

GWC 2018

Proceedings of the 9th Global Wordnet Conference

**Francis Bond, Takayuki Kuribayashi,
Christiane Fellbaum, Piek Vossen (Eds.)**

8–12 January, 2018
Nanyang Technological University (NTU)
Singapore



**Global
WordNet
Association**

©2018 Global Wordnet Association

ISBN 978-981-11-7087-4

Foreword

The Ninth Global Wordnet Conference was held at Nanyang Technological University, Singapore from 9–12th January 2018.

The program combined the main conference with a special day on wordnets and word-embeddings and finished with a half day workshop on technology enhanced learning (TEL). There were 4 invited talks, 41 full papers, 15 posters and 4 invited talks on TEL. Including the papers on embeddings, there were 15 rejections: the acceptance rate for full papers was 58% a sign of the consistently high quality of papers submitted to the conference. Copyrights for the papers reside with the original authors.

The invited papers were *One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction* by Ng Hwee Tou (National University of Singapore), *How are you two related? Corpus-based Learning of Lexical Semantic Relations* by Vered Shwartz (Bar-Ilan University), *Inducing Interpretable Word Senses for WSD and Enrichment of Lexical Resources* by Alexander Panchenko (University of Hamburg) and *Using a Grammar Implementation to Teach Writing Skills* by Dan Flickinger (Stanford). As well as many papers on distributional semantics, there were some on extending the coverage of existing wordnets, linking wordnets to new resources (especially in the medical domain), using wordnets for teaching and many other topics. There were papers from 24 different countries with every continent except Antarctica represented.

The conference and workshops were partially supported by the NTU Centre for Liberal Arts and Social Sciences (CLASS) and the Singapore MOE TRF *Grant Syntactic Well-Formedness Diagnosis and Error-Based Coaching in Computer Assisted Language Learning using Machine Translation Technology*. Support for students came from the Global Wordnet Association. We would like to thank the programme committee for their thoughtful and timely reviews.

The conference homepage is <http://compling.hss.ntu.edu.sg/events/2018-gwc/>.

Francis Bond, Nanyang Technological University

Takayuki Kuribayashi, Nanyang Technological University

Christiane Fellbaum, Princeton University

Piek Vossen, VU University Amsterdam

January 2018

Organizers

Local Chairs Francis Bond and Takayuki Kuribayashi

Program Chairs Christiane Fellbaum, Francis Bond and Piek Vossen

Workshop Chairs Christiane Fellbaum, Francis Bond and Erhard Hinrichs

Program Committee

Adam Pease	Articulate Software, San Jose, USA
Ales Horak	Masaryk University
Alexandre Rademaker	IBM Research Brazil and EMAP/FGV
Antoni Oliver	Universitat Oberta de Catalunya
Bolette Pedersen	University of Copenhagen
Corina Forăscu	Alexandru Ioan Cuza University of Iași
Dan Cristea	Alexandru Ioan Cuza University of Iași
Darja Fiser	University of Ljubljana
Diptesh Kanojia	IIT Bombay
Emiel van Miltenburg	Vrije Universiteit Amsterdam
Eneko Agirre	University of the Basque Country
Erhard Hinrichs	University of Tübingen
Ewa Rudnicka	Wroclaw University of Technology
Gerard de Melo	Rutgers University
German Rigau	University of the Basque Country (EPV/EHU)
Haldur Oim	University of Tartu
Heili Orav	University of Tartu
Horacio Rodriguez	Universitat Politècnica de Catalunya
Isahara Hitoshi	Toyohashi University of Technology
Janos Csirik	University of Szeged
John P. Mccrae	National University of Ireland, Galway
Kadri Vider	University of Tartu
Kevin Scannell	Saint Louis University
Kyoko Kanzaki	Toyohashi University of Technology
Maciej Piasecki	Wrocław University of Technology
Magnini Bernardo	FBK-IRST
Marten Postma	VU University Amsterdam
Mikhail Khodak	Princeton University
Rada Mihalcea	University of North Texas / Oxford University
Roxane Segers	VU University Amsterdam
Shan Wang	The Education University of Hong Kong
Shu-Kai Hsieh	National Taiwan Normal University

Sonja Bosch
Ted Pedersen
Timothy Baldwin
Tomaž Erjavec
Verginica Mititelu

Department of African Languages, University of South Africa
University of Minnesota, Duluth
The University of Melbourne
Dept. of Knowledge Technologies, Jožef Stefan Institute
Romanian Academy Research Institute for Artificial Intelligence

Invited Speakers

- Ng Hwee Tou, National University of Singapore
- Vered Shwartz, Bar-Ilan University
- Alexander Panchenko, University of Hamburg
- Dan Flickinger, Stanford

Invited Talks

Ng Hwee Tou: One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction

Supervised word sense disambiguation (WSD) systems have achieved the best performance when evaluated on standard benchmark datasets. However, the lack of large amounts of sense-tagged data poses a major hurdle to scaling up supervised WSD systems to disambiguate all words of English. In this talk, I will present a semi-automatic approach to extract and annotate a large sense-tagged corpus. This one-million-word sense-tagged corpus has been publicly released since 2015 and has been used by other researchers working on automated WSD. When trained on this one-million-word sense-tagged corpus, the open source IMS (It Makes Sense) WSD system created in my research group achieves good performance on standard WSD tasks and another word sense induction task.

Vered Shwartz: How are you two related? Corpus-based Learning of Lexical Semantic Relations

Recognizing lexical semantic relations between words is an essential component in semantic applications such as question answering and recognizing textual entailment. In order to overcome lexical variability, such systems traditionally relied heavily on lexical resources such as WordNet.

In the main part of the talk I will discuss our work on automatic detection of lexical semantic relations from free text. This task stems from the limited coverage of lexical resources, both in terms of missing lexical items (proper names, new words) and missing relations between existing items. Typical approaches to address this task are either distributional, i.e. based on the word embeddings of the two target words, or path-based (pattern-based) approach, based on the words co-occurrences in the corpus. I will present our integrated path-based and distributional method for recognizing lexical semantic relations, which is currently the state-of-the-art in this task.

In the second part, I will raise some questions about the interplay of WordNet and word embeddings: is external lexical knowledge obsolete in the deep learning era? And if it isn't, then how can lexical knowledge from WordNet and other resources be incorporated into neural models for semantic applications?

Alexander Panchenko: Inducing Interpretable Word Senses for WSD and Enrichment of Lexical Resources

In this talk, we will discuss induction of sparse and dense word sense representations using graph-based approaches and distributional models. Induced senses are represented by a vector, but also a set of hypernyms, images, and usage examples, derived in an unsupervised and knowledge-free manner, which ensure interpretability of the discovered senses by humans. We showcase the usage of the induced representations for the tasks of word sense disambiguation and enrichment of lexical resources, such as WordNet.

Dan Flickinger: Using a Grammar Implementation to Teach Writing Skills

This paper presents an approach to grammar checking, using a large-scale HPSG grammar of English. The system has been used in a Language Arts & Writing course for McGraw-Hill Education in U.S. classrooms for the past ten years. It has helped over 50,000 primary school students, mostly native English speakers. We have given feedback on over 10 million sentences. The feedback is generated using mal-rules that identify errors with high precision. We are currently looking at extending the system to non-native speakers' English.

Table of Contents

BanglaNet: Towards a WordNet for Bengali Language	1
<i>K.M. Tahsin Hassan Rahit, Khandaker Tabin Hasan, Md. Al- Amin and Zahiduddin Ahmed</i>	
WME 3.0: An Enhanced and Validated Lexicon of Medical Concepts	10
<i>Anupam Mondal, Dipankar Das, Erik Cambria and Sivaji Bandyopadhyay</i>	
Using Context to Improve the Spanish WordNet Translation	17
<i>Alfonso Methol, Guillermo López, Juan Álvarez, Luis Chiruzzo and Dina Wonsever</i>	
Towards Cross-checking WordNet and SUMO Using Meronymy	25
<i>Javier Álvez and German Rigau</i>	
Comparing Two Thesaurus Representations for Russian	34
<i>Natalia Loukachevitch, German Lashevich and Boris Dobrov</i>	
Towards Mapping Thesauri onto plWordNet	44
<i>Marek Maziarz and Maciej Piasecki</i>	
Investigating English Affixes and their Productivity with Princeton WordNet	53
<i>Verginica Mititelu</i>	
Mapping WordNet Instances to Wikipedia	61
<i>John P. McCrae</i>	
Mapping WordNet Concepts with CPA Ontology	69
<i>Svetla Koeva, Cvetana Dimitrova, Valentina Stefanova and Dimitar Hristov</i>	
Improving Wordnets for Under-Resourced Languages Using Machine Translation	77
<i>Bharathi Raja Chakravarthi, Mihael Arcan and John P. McCrae</i>	
Semantic Feature Structure Extraction From Documents Based on Extended Lexical Chains	87
<i>Terry Ruas and William Grosky</i>	
Toward a Semantic Concordancer	97
<i>Adam Pease and Andrew Cheung</i>	
Using OpenWordnet-PT for Question Answering on Legal Domain	105
<i>Pedro Delfino, Bruno Cuconato, Guilherme Paulino Passos, Gerson Zaverucha and Alexandre Rademaker</i>	
Implementation of the Verb Model in plWordNet 4.0	113
<i>Agnieszka Dziob and Maciej Piasecki</i>	
Public Apologies in India - Semantics, Sentiment and Emotion	123
<i>Sangeeta Shukla and Rajita Shukla</i>	
Derivational Relations in Arabic WordNet	136
<i>Mohamed Ali Batita and Mounir Zrigui</i>	
Extending Wordnet to Geological Times	145
<i>Henrique Muniz, Fabricio Chalub, Alexandre Rademaker and Valeria De Paiva</i>	
Towards Emotive Annotation in plWordNet 4.0	153
<i>Monika Zaśko-Zielińska and Maciej Piasecki</i>	
The Company They Keep: Extracting Japanese Neologisms Using Language Patterns	163

<i>James Breen, Timothy Baldwin and Francis Bond</i>	
Lexical-semantic resources: yet powerful resources for automatic personality classification	172
<i>Xuan-Son Vu, Lucie Flekova, Lili Jiang and Iryna Gurevych</i>	
Towards a principled approach to sense clustering –a case study of wordnet and dictionary senses in Danish	182
<i>Bolette Pedersen, Manex Agirrezabal, Sanni Nimb, Ida Olsen and Sussi Olsen</i>	
WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation . .	190
<i>Tomasz Naskręt, Agnieszka Dziob, Maciej Piasecki, Chakaveh Saedi and António Branco</i>	
Translation Equivalence and Synonymy: Preserving the Synsets in Cross-lingual Wordnets	200
<i>Oi Yee Kwong</i>	
Lexical Perspective on Wordnet to Wordnet Mapping	209
<i>Ewa Rudnicka, Francis Bond, Łukasz Grabowski, Maciej Piasecki and Tadeusz Piotrowski</i>	
ReferenceNet: a semantic-pragmatic network for capturing reference relations.	219
<i>Piek Vossen, Filip Ilievski and Marten Postrma</i>	
Wordnet-based Evaluation of Large Distributional Models for Polish	229
<i>Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski and Paweł Kędzia</i>	
Distant Supervision for Relation Extraction with Multi-sense Word Embedding	239
<i>Sangha Nam, Kijong Han, Eun-Kyung Kim and Key-Sun Choi</i>	
Cross-Lingual and Supervised Learning Approach for Indonesian Word Sense Disambiguation Task	245
<i>Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung and Mirna Adriani</i>	
Recognition of Hyponymy and Meronymy Relations in Word Embeddings for Polish	251
<i>Gabriela Czachor, Maciej Piasecki and Arkadiusz Janz</i>	
Simple Embedding-Based Word Sense Disambiguation	259
<i>Dieke Oele and Gertjan Van Noord</i>	
Semi-automatic WordNet Linking using Word Embeddings	266
<i>Kevin Patel, Diptesh Kanojia and Pushpak Bhattacharyya</i>	
Multilingual Wordnet sense Ranking using nearest context	272
<i>E Umamaheswari Vasanthakumar and Francis Bond</i>	
Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton WordNet .	284
<i>Kiril Simov, Alexander Popov, Iliana Simova and Petya Osenova</i>	
An Iterative Approach for Unsupervised Most Frequent Sense Detection using WordNet and Word Embeddings	293
<i>Kevin Patel and Pushpak Bhattacharyya</i>	
Automatic Identification of Basic-Level Categories	298
<i>Chad Mills, Francis Bond and Gina-Anne Levow</i>	
African Wordnet: facilitating language learning in African languages	306
<i>Sonja Bosch and Marissa Griesel</i>	
Hindi Wordnet for Language Teaching: Experiences and Lessons Learnt	314
<i>Hanumant Redkar, Rajita Shukla, Sandhya Singh, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia,</i>	

<i>Preethi Jyothi, Malhar Kulkarni and Pushpak Bhattacharyya</i>	
An Experiment: Using Google Translate and Semantic Mirrors to Create Synsets with Many Lexical Units	324
<i>Ahti Lohk, Mati Tombak and Kadri Vare</i>	
Context-sensitive Sentiment Propagation in WordNet	329
<i>Jan Kocoń, Arkadiusz Janz and Maciej Piasecki</i>	
ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities	335
<i>Bolette Pedersen, John McCrae, Carole Tiberius and Simon Krek</i>	
Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic Analysis in the Domain of Theology	341
<i>Laura Slaughter, Wenjie Wang, Luis Morgado Da Costa and Francis Bond</i>	
Estonian Wordnet: Current State and Future Prospects	347
<i>Heili Orav, Kadri Vare and Sirli Zupping</i>	
Further expansion of the Croatian WordNet	352
<i>Krešimir Šojat, Matea Filko and Antoni Oliver</i>	
Linking WordNet to 3D Shapes	358
<i>Angel X Chang, Rishi Mago, Pranav Krishna, Manolis Savva and Christiane Fellbaum</i>	
Multisłownik: Linking plWordNet-based Lexical Data for Lexicography and Educational Purposes	364
<i>Maciej Ogrodniczuk, Joanna Bilińska, Zbigniew Bronk and Witold Kieraś</i>	
Putting Figures on Influences on Moroccan Darija from Arabic, French and Spanish using the WordNet	372
<i>Khalil Mrini and Francis Bond</i>	
pyiwn: A Python based API to access Indian Language WordNets	378
<i>Ritesh Panjwani, Diptesh Kanojia and Pushpak Bhattacharyya</i>	
Sinitic Wordnet: Laying the Groundwork with Chinese Varieties Written in Traditional Characters	384
<i>Chih-Yao Lee and Shu-Kai Hsieh</i>	
Synthesizing Audio for Hindi WordNet	388
<i>Diptesh Kanojia, Preethi Jyothi and Pushpak Bhattacharyya</i>	
Toward Constructing the National Cancer Institute Thesaurus Derived WordNet (ncitWN)	394
<i>Amanda Hicks, Selja Seppälä and Francis Bond</i>	
Towards a Crowd-Sourced WordNet for Colloquial English	401
<i>John P. McCrae, Ian Wood and Amanda Hicks</i>	
WordNet Troponymy and Extraction of "Manner-Result" Relations	407
<i>Aliaksandr Huminski and Hao Zhang</i>	
SardaNet: a Linguistic Resource for Sardinian Language	412
<i>Manuela Angioni, Franco Tuveri, Maurizio Viridis, Laura Lucia Lai and Micol Elisa Maltesi</i>	
Extraction of Verbal Synsets and Relations for FarsNet	420
<i>Fatemeh Khalghani and Mehrnoush Shamsfard</i>	

BanglaNet: Towards a WordNet for Bengali Language

K.M. Tahsin Hassan Rahit

American International
University -Bangladesh
tahsin.rahit@gmail.com
Md. Al- Amin
American International
University -Bangladesh
alamin@aiub.edu

Khandaker Tabin Hasan

American International
University -Bangladesh
tabin@aiub.edu
Zahiduddin Ahmed
American International
University -Bangladesh
zahid@aiub.edu

Abstract

Despite being a popular language in the world, the Bengali language lacks in having a good wordnet. This restricts us to do NLP related research work in Bengali. Most of the today's wordnets are developed by following *expand approach*. One of the key challenges of this approach is the cross-lingual word-sense disambiguation. In our research work, we make semantic relation between Bengali wordnet and Princeton WordNet based on well-established research work in other languages. The algorithm will derive relations between concepts as well. One of our key objectives is to provide a panel for lexicographers so that they can validate and contribute to the wordnet.

1 Introduction

The Princeton WordNet (PWN) (Miller, 1995; Fellbaum, 1998) is one of the most semantically rich English lexical database which is widely used as a resource in many research and development. It is not only an important resource for NLP applications in each language, but also for inter-linking WordNets of different languages to develop multilingual applications to overcome the language barrier. In the

present, there are roughly 6,500 languages ¹. Among those, Bengali is the 7th most popular language ² in the world. Yet, there is a lack of work for Bengali wordnet. Global WordNet Association (GWA) has enlisted almost all wordnets in several levels depending on availability and how rich it is. At first level, there are 34 Open Multi-lingual WordNet ³ that are merged into Global WordNet Grid. But in spite of being a popular language, Bengali is not one of them. GWA also enlist other available wordnets. Among those 80 wordnets, there are two Bengali wordnets which are developed in India.

In this research work, a baseline for BanglaNet has been developed which is a wordnet for the Bengali language. To link the wordnet with Princeton WordNet, semi-automatic cross-lingual sense mapping approach is used. We align the Princeton WordNet synset into a bilingual dictionary through the English equivalent and its part-of-speech (POS). Manual translation and link-up can also be employed after the alignment. This paper covers previous works for other wordnets including previous

¹ How many spoken languages are there in the world, <http://www.infoplease.com/askeds/many-spoken-languages.html> (Accessed 2016-10-22)

²Most widely spoken languages in the world, <http://www.infoplease.com/ipa/A0775272.html> (Accessed 2016-10-22)

³Open Multilingual WordNet, <http://compling.hss.ntu.edu.sg/omw/> (Accessed 23-10-2016)

attempts of developing Bengali WordNet, describe initiative taken for BanglaNet and our design and execution process in depth. Lastly, analysis of resultant lexical database has been presented. We aim to include BanglaNet into GlobalWordNet in future. Intending to doing so, relation with Princeton WordNet is maintained as much as possible as per the convention. Additionally, a web-based collaborative tool, called *Oikotan* which is BanglaNet Lexicography Development Panel (LDP) has been developed for revising the result of synset assignment and provide a framework to create BanglaNet via the linkage with synsets.

2 Background Study

2.1 WordNet Development Techniques

To this date, there are two ways develop wordnet for a particular language.

Merge Approach is used to build the word net from scratch. The Princeton WordNet is built in this approach. The taxonomies of the language, synsets, relations among synsets are developed first. Experienced work power, lexicographer and time are needed to develop for this approach (Taghizadeh and Faili, 2016). Mapping resultant wordnet with the Princeton WordNet is also required extensive work and cross-language expert.

Expand Approach is used to map or translate local words directly to the Princeton WordNet's synsets by using the existing bilingual dictionaries. Most of the WordNet available currently is developed by following this approach. This process can be made easy by semi-automatically doing many tasks and then refactoring it for further proofing.

2.2 Related Works

2.2.1 International Languages

The first attempt for developing WordNet in another language other than English started

in 1996. EuroWordNet (Vossen, 2002) began as an EU project, with the goal of developing wordnets for Dutch, Spanish and Italian and linking these wordnets to the English WordNet in a multilingual database. Later in 1997, it was extended and German, French, Czech and Estonian included. Balkan WordNet (Tufis et al., 2004) - which was developed in the BalkaNet project was developed with an aim to develop a multilingual semantic network for Balkan languages such as reek, Turkish, Romanian, Bulgarian, Czech and Serbian. In developing BalkaNet semantic relations are classified in the independent WordNets according to a shared ontology. BalkaNet was integrated along with EuroWordNet through a WordNet Management System. Relations among synsets have been built mostly automatically (Pala and Smrz, 2004) and these relations are developed based on Princeton WordNet. However, to achieve high accuracy rate developer needs to pay special attention to the problem of the translation equivalents.

There are open challenges in NLP research to automate development of semantic resources constitutes. In WOLF (Wordnet Libre du Français, Free French Wordnet) (Apidianaki and Sagot, 2012) development, multiple NLP algorithms including cross-lingual word sense disambiguation is used. WOLF is free wordnet for the French language. In Asian region, Japanese WordNet (Isahara et al., 2008) was developed using *expand approach*. Korean WordNet (Lee et al., 2002) was developed using extracting semantic hierarchy by utilizing a monolingual MRD and an existing thesaurus in *expand approach*. Thai WordNet was (Sathapornrungskij and Pluem-pitiwiriawej, 2005) also developed by following this same approach. Another large work in Asian region includes IndoWordNet (Prabhu et al., 2012) developed in India to incorporate language used in Indian sub-continent. In-

doWordNet was also developed using existing WordNets.

Word-Sense Disambiguation (WSD) technique played a major role in most of the wordnet development. Lefever, Els and Hoste, Veronique have presented review on cross-lingual disambiguation (Lefever and Hoste, 2010) (Lefever and Hoste, 2013). They found out that languages where the ratio of word against sense is low, it becomes hard to extract translation for that language since the number of translation for a particular word in another language becomes greater. Hence, a particular word contains multiple translations in counter language.

French encountered the similar problem like us. It had no corpus with predicate-argument annotations which help to express semantic relation build-up. Van der Plas et al. researched on predicate labeling in French (van der Plas and Apidianaki, 2014) to overcome this issue using Word Sense Disambiguation.

There are two terms in cross-lingual WSD. One is *best* match and another one is *Out-of-five*. In *best* mode, the word or sense with the best probability score tagged with its counter word or sense. In case of, *Out-of-five* approach, if multiple senses or word belongs to candidate conceptualization, best five probability candidates are considered for further analysis. Further analysis can be done manually or automatically. It can be semi-automatic as well.

WSD process performance can be improved by using the Direct Semantic Transfer (DST) technique developed by Van der Plas et al. (Van der Plas et al., 2011). It tells us that the senses which can be directly transferred to another language if and only if both share same semantic property.

Surtani et al. developed a system where it can predict the paraphrases based on corpus (Surtani et al., 2013). In their system, they have a semantic relation prediction model.

Recently, BabelNet⁴ (Navigli and Ponzetto, 2012a) has become a good example of multilingual language resource. BabelNet simplified WSD process by incorporating coding API (Navigli and Ponzetto, 2012b). Primarily, it uses open-source resources such as Wikipedia. However, BabelNet does not create any WordNet for a particular language. It is a huge standalone network of multilingual resources which utilizes Princeton WordNet along with other resources to make relations.

2.2.2 Bengali

Between two of Bengali wordnets listed in GWA, one is developed by Indian Institute of Statistics under Indradhanush Project⁵. It has an online browser which does not provide the semantic relation between synsets and only provides different concept available for a word. Another Bengali wordnet is developed as part of IndoWordNet by Center for Indian Language Technology (CILT) and Indian Institute of Technology (IIT-Bombay) (Prabhu et al., 2012). A notable point in this WordNet is - it is built by following the *expand approach*. It does have the semantic relation between synset to some extent. This is the most mature and contextually rich Bengali WordNet to this date. Both WordNets are browsable and closed source. These are neither publicly available for development, use or extend nor it provides any API for general use.

There was an effort for developing Bengali WordNet in BRAC University's Center for Research on Bengali Language Processing. In their development process they followed *merge approach* (Faruqe and Khan, 2010).

⁴BabelNet can be found on <http://babelnet.org> (Accessed 2016-12-07)

⁵Indradhanush Project, <http://indradhanush.unigoa.ac.in> (Accessed 2016-10-22.)

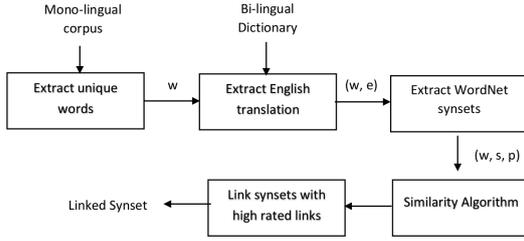


Figure 1: Proposed method for BanglaNet

3 Architecture

It has been discussed above that *expand approach* is followed to construct the BanglaNet by translating the synsets in the Princeton WordNet to the Bengali language. Both automatic and manual methods are applied in the process. Ambiguity is one of the concerns for automatic concept mapping. This cross-lingual ambiguity can come in different form. For instance - one-to-one, one-to-many, many-to-one, many-to-many. In this research work, uni-directional ambiguity in one-to-one and one-to-many has been addressed.

Based on our research on other languages’ WordNet and past works in Bengali WordNet, this paper proposes to follow methodology described in Fig 1 for BanglaNet development.

- i) Extract monosemous literals w from Bengali lexicon.
- ii) Translate each Bengali literal to English literals e using bilingual dictionary.
- iii) For each English literals, extract concept(s) available in Princeton WordNet p .
- iv) Run similarity score calculation algorithm using the e and p we found for two different Bengali sense. We take different synset available for sense w and compare their English counterpart.
- v) Based on similarity score, map Bengali concept with pwn concept.

- vi) Lexicographer validation for resultant mapping.

3.1 Similarity Matrices

In step iv, similarity algorithm is used. Similarity algorithm calculates similarity in a sense between two words in Princeton WordNet. Similarity can be calculated in several ways. There are well-established algorithms (Pedersen et al., 2004; Meng et al., 2013) to calculate similarity score. Few of those algorithms are -

- i) Path Similarity (Meng et al., 2013) calculates the score in a range of 0 to 1 based on the shortest path that connects the senses in “is-a” (hypernym/hyponym) relation.
- ii) Leacock-Chodorow Similarity (Bruce and Wiebe, 1994) scores based on the shortest path that connects the senses (identical to Path Similarity) and the maximum depth of the taxonomy in which the senses occur.
- iii) Wu-Palmer Similarity (Wu and Palmer, 1994) uses depth of the two senses in the taxonomy considering their most specific ancestor node are used to calculate the score.

There are other algorithms like Resnik Similarity (Resnik, 1995), Jiang-Conrath Similarity (Jiang and Conrath, 1997), Lin Similarity (Lin, 1998). To calculate the similarity between two concepts, we use Wu & Palmer’s similarity algorithm as it takes the hierarchical position of concepts C_1 and C_2 in the taxonomy tree relatively to the position of the most specific common concept $Iso(c_1, c_2)$ into account. It assumes that the similarity between two concepts is the function of path length and depth in path-based measures (Wu and Palmer, 1994).

$$sim_{WP}(C_1, C_2) = \frac{2 * \text{depth}(Iso(c_1, c_2))}{len(c_1, c_2) + 2 * \text{depth}(Iso(c_1, c_2))} \quad (1)$$

4 BanglaNet Development

The primary task for WordNet development using *expand approach* is to generate base lexicons and concepts. Full system including the database of Princeton WordNet is downloadable from its official website. It is possible only to get the database files without the system as well. Lexical database files can be downloaded separately as well. For base concepts, a dataset which is available on GitHub⁶ has been used. It provides conceptual gloss in Bengali for words along with its synonymy. This dataset made our work more focused on cross-lingual mapping rather than local synset construction. This research work is focused more on making relation with PWN concept rather than producing concepts. After analyzing the list of concept retrieved from the dataset, at first synsets for each concept is generated. A concept can be represented using multiple words; it ensures that we have synonyms for every concept.

Moreover, There is a POS tag available for each concept representing the word.

4.1 Word to Word Translation

Currently, a list of concepts with its gloss and synset is available. Now, English translation for each word needed to be determined. A word in one language can be represented by multiple words in another language. This is true for concept also. But for now, English translation for the enlisted words is needed. Nevertheless, for a Bengali word, there can be multiple English meaning. For example: "বল" means 'Ball' in English. It means 'Force' as well. A bilingual dictionary is needed to collect these translations. In this step, candidate translations from Bengali to English bilingual dictionary is stored. The reason behind collecting English translation using a dictionary is to

⁶Bengali Synsets Data available on GitHub, Soumen-ganguly. <https://github.com/soumenganguly/Bangla-Wordnet/> (Accessed 2016-10-22)

get the proper concept from WordNet. This is achieved through the WordNet concept selection algorithm which is explained in later part of this paper. For now, let's see how dictionary translations are processed.

At first, every possible English translation for each of the words in the lexicon is needed. This translation is achieved by iterating through each Bengali word in our lexicon. Bi-lingual (Bengali to English) dictionaries are used to get translations of each of the words. This translation can be from multiple parts of speech. POS for this translation is considered as well so that it can be used to properly identify correct translation in later steps. However, not all words have its counter English words. These words can be a concept which is only available in Bengali concept only. These words can also be a proper noun. For instance, the name of the places, location, river or person, scientific terms. Although, it is also possible to collect this information in run-time, to reduce time latency and run-time processing, translations along with the POS are temporarily stored.

4.2 Linking with Princeton WordNet using Probabilistic Model

It is mentioned earlier that, automated and semi-automated WordNet mostly depends on well-crafted algorithms of Natural Language Processing (NLP) and data processing. These statistical and probabilistic heuristic algorithms are good enough to create the relation between words, sense. It is obvious that the results are not always 100% accurate. Hence, lexical post-verification steps then come in place to fine tune the results.

After having the candidate translation, now it is possible to calculate the score of the probable concept from Princeton WordNet for a BanglaNet concept. Let's assume, S_c is the

synset for a Bengali concept c . We have a set of candidate translation CT_w for a particular Bengali word w . w belongs to the concept c . POS tag associated with w is a .

$$S_c = \{s \mid s \in \text{Bengali word}\} \quad (2)$$

Now, translation for each Bengali word s_i in S_c is taken:

$$ST_{s_i} = \{st_i \mid s_i \in S_c, st_i \in CT_{s_i}\} \quad (3)$$

Combining ST_{s_i} for all S_c .

$$ST_c = \{st \mid \forall st \in \bigcup_{i=0}^n ST_{s_i} \rightarrow s_i \in S_c\} \quad (4)$$

According to set theory, ST_c will contain all unique English translations for the words in Synset S_c . Synset from Princeton WordNet for each words in the set CT_w and ST_c is retrieved. POS tag for the synsets should match with a . Assuming, u as an English word -

$$syn_{u,a} = \{x \mid x \in \text{PWN Synset for } u \text{ and } x \in a\} \quad (5)$$

$$P_1 = \{x \mid \forall x \in \bigcup_{u=CT_w} syn_{u,a}\} \quad (6)$$

$$P_2 = \{x \mid \forall x \in \bigcup_{u=ST_c} syn_{u,a}\} \quad (7)$$

We take cross product of elements of P_1 against each elements of P_2 .

$$P = \{(m,n) \mid m \in CT_w \text{ and } n \in ST_c\} \quad (8)$$

After having the cross product, a similarity algorithm on each tuple is run. To calculate similarity score, equation (1) on each tuple is used. Sorting the synset P_1 according to the summation of each synset's score which is probability score for the synset, the tuple with maximum similarity score is chosen. Algorithm for this task is transcribed in Algorithm 4.1 Now, the probability score for all probable synset in Princeton WordNet for the Bengali concept is c . Bengali synset is linked with Princeton WordNet synset using

Algorithm 4.1: Algorithm for calculating probability score

```

1 Function CalculateProbabilityScore ( $P$ )
   Input:  $P$ 
   Output: Sprted scores of  $P$  based on
           probability score
2  $scores[] := \emptyset;$ 
3 foreach ( $m,n$ )  $\in P$  do
4   if  $scores[m] \neq \emptyset$  then
5      $scores[m] \leftarrow$ 
6        $scores[m] + sim_{wp}(m,n);$ 
7   else
8      $scores[m] \leftarrow sim_{wp}(m,n);$ 
9   end
10 return  $sort(scores);$ 

```

algorithm 4.2. To link Bengali concept with Princeton WordNet, multiple procedures have used to ensure correctness as much as possible. First of all, Princeton WordNet concept is assigned to those concepts in BanglaNet which have only one possible item in P_1 . Secondly, if and only if there is only one concept available for the word w , in that case, the concept from Princeton WordNet which scored high probability in probability calculation algorithm would be chosen. A point to be noted is, if any of the synonyms (word) in synset of a concept has only one concept tagged to it, it can be linked using this method. By using this first pass on all over the concepts, Princeton WordNet concepts is assigned.

5 Results and Analysis

In the initial dataset, there were 27239 unique concepts. These concepts are represented using 40158 unique words tagged with different parts of speech. Table 1 shows statistics of our initial data. In total, almost 65% of the whole concepts are tagged with Noun parts of speech.

English translation for 13029 words has

Algorithm 4.2: Algorithm for linking concept- first pass

```

1 Function LinkSynset (w)
   Input: w
2   concept count := number of concepts
   for the word w;
3   P := Generate synset cross product ;
4   sorted_scores[] :=
   CalculateProbabilityScore(P);
5   if length of sorted_scores = 1 or
   concept_count = 1 then
6     C := concepts for the word w;
7     foreach c ∈ C do
8       c.pwn ←
       sorted_scores.top().key();
9     end
10  end

```

		Noun	Adj	Verb	Adv	Total
Initial	synsets	18311	5713	2777	438	27239
	words	28311	8136	2923	788	40158
Linked	synsets	3174	1352	73	66	4665
	words	7477	2971	130	170	10748

Table 1: Status of linked Synset and Words from initial dataset

been retrieved. After applying concept linking, 4665 concepts are linked with Princeton WordNet. In total, 10748 words are linked with Princeton WordNet.

To link this 4665 concepts with Princeton WordNet, 3729 Princeton WordNet concepts are used. That means, there are cross multiple concepts within two WordNet.

Cross-lingual word-sense disambiguation can be shown using another example. For the word "ফলকপি" there are two concept available in Bengali. In English it has two concepts too.

cauliflower.n.01 a plant having a large edible head of crowded white flower buds

cauliflower.n.02 compact head of undevel-

oped white flowers

The algorithm predicted both English concepts for the two concepts available. For ফলকপি .n.01 probability score for English concepts are 4.4419589754 and 4.4419589754 respectively. On the other hand, ফলকপি .n.02 score is 6.84959684439 and 6.20774295822. It is observed that for both cases these scores are too close to prioritize probability.

Although the algorithm used in BanglaNet is directed from Bengali to English synset matching, this development can also be implied from another way around. In that case, Bengali word which represents a particular concept in Princeton WordNet can be used to verify and add more confidence to concept linking. As a result, more link up can be achieved.

Our initial synset contains gloss. But our approach does not take gloss into consideration. As a consequence, BanglaNet can be expanded using the same approach in future even if gloss for a synset is not available.

5.1 Future Works

There is a big opportunity to work on BanglaNet expansion and development. In this algorithm, the gloss is not taken into consideration. The accuracy of the algorithm can be noticeably improved by incorporating the gloss. However, a bilingual corpus will be required to achieve this. It has been found out that there is a lack of good corpus for Bengali. Good corpus is one of the key components of Natural Language Processing. However, our literature review discussed BabelNet. It's data sources and approach can be useful to map concepts. In this research work, first pass or first level linking is done. In the second pass, new algorithm needed to connect concepts which have multiple synsets in either end (BanglaNet or Princeton WordNet). We propose to use, Variable Neighborhood Search (VNS) ("Hansen and Mladenović, Nenad and MorenoA Pérez,

José A.”, 2010).

6 Conclusion

Developing wordnet is an immense task. It is our distinct pleasure that in this research work, a basic layer of the system has been laid down for Bengali wordnet from where further development can be made. Suggestion generation task for validation can be achievable through the result of this research work. Our result analysis shows that around 5000 words from initially collected data are automatically linked up with Princeton WordNet. Although there is a long way to go in the development of Bengali wordnet, this research work is starting stage for further development.

Acknowledgments

This research is funded by ICT Division - Government of the People’s Republic of Bangladesh under its fellowship program.

References

- [Apidianaki and Sagot2012] Marianna Apidianaki and Benoît Sagot. 2012. Applying cross-lingual wsd to wordnet development. In *LREC 2012-Eighth International Conference on Language Resources and Evaluation*.
- [Bruce and Wiebe1994] Rebecca Bruce and Janyce Wiebe. 1994. A new approach to word sense disambiguation. In *Proceedings of the Workshop on Human Language Technology, HLT 94*, pages 244–249.
- [Faruqe and Khan2010] Farhana Faruqe and Munit Khan. 2010. Bwn-a software platform for developing bengali wordnet. In *Innovations and Advances in Computer Sciences and Engineering*, pages 337–342. Springer.
- [Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.
- [Hansen and Mladenović, Nenad and MorenoA Pérez, José A.”2010] Pierre Hansen and Mladenović, Nenad and MorenoA Pérez, José A.”. 2010”. ”variable neighbourhood search: methods and applications”. *Annals of Operations Research*, ”175”(”1”):”367–407”.
- [Isahara et al.2008] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the japanese wordnet. In *LREC*.
- [Jiang and Conrath1997] Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008.
- [Lee et al.2002] Juho Lee, Koanghi Un, Hee-Sook Bae, and Key-Sun Choi. 2002. A korean noun semantic hierarchy (wordnet) construction. In *GLOBAL WORDNET CONFERENCE*.
- [Lefever and Hoste2010] Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 15–20. Association for Computational Linguistics.
- [Lefever and Hoste2013] Els Lefever and Veronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166.
- [Lin1998] Dekang Lin. 1998. An information-theoretic definition of similarity. *Proceedings of ICML*, pages 296–304.
- [Meng et al.2013] Lingling Meng, Runqing Huang, and Junzhong Gu. 2013. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12.
- [Miller1995] George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [Navigli and Ponzetto2012a] Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

- [Navigli and Ponzetto2012b] Roberto Navigli and Simone Paolo Ponzetto. 2012b. Multilingual wsd with just a few lines of code: the babelnet api. In *Proceedings of the ACL 2012 System Demonstrations*, pages 67–72. Association for Computational Linguistics.
- [Pala and Smrz2004] Karel Pala and Pavel Smrz. 2004. Building czech wordnet. *Romanian Journal of Information Science and Technology*, 7(2-3):79–88.
- [Pedersen et al.2004] T Pedersen, S Patwardhan, and J Michelizzi. 2004. Wordnet similarity - measuring the relatedness of concepts. *Proceedings - Nineteenth National Conference on Artificial Intelligence (AAAI-2004): Sixteenth Innovative Applications of Artificial Intelligence Conference (IAAI-2004)*, pages 1024–1025.
- [Prabhu et al.2012] Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, NR Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2012. An efficient database design for indowordnet development using hybrid approach. *3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)*.
- [Resnik1995] Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95*, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Sathapornrungskij and Pluempitiwiriyaewj2005] Patanakul Sathapornrungskij and Charnyote Pluempitiwiriyaewj. 2005. Construction of thai wordnet lexical database from machine readable dictionaries. *Proc. 10th Machine Translation Summit, Phuket, Thailand*.
- [Surtani et al.2013] Nitesh Surtani, Arpita Batra, Urmi Ghosh, and Soma Paul. 2013. Iiith: A corpus-driven co-occurrence based probabilistic model for noun compound paraphrasing. *Atlanta, Georgia, USA*, page 153.
- [Taghizadeh and Faili2016] Nasrin Taghizadeh and Hesham Faili. 2016. Automatic wordnet development for low-resource languages using cross-lingual wsd. *Journal of Artificial Intelligence Research*, 56:61–87.
- [Tufis et al.2004] Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.
- [van der Plas and Apidianaki2014] Lonneke van der Plas and Marianna Apidianaki. 2014. Cross-lingual word sense disambiguation for predicate labelling of french. In *Proceedings of the 21st TALN (Traitement Automatique des Langues Naturelles) conference*, pages 46–55.
- [Van der Plas et al.2011] Lonneke Van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.
- [Vossen2002] Piek Vossen. 2002. Wordnet, eurowordnet and global wordnet. *Revue française de linguistique appliquée*, 7(1):27–38.
- [Wu and Palmer1994] Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94*, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

WME 3.0: An Enhanced and Validated Lexicon of Medical Concepts

Anupam Mondal¹ Dipankar Das¹ Erik Cambria² Sivaji Bandyopadhyay¹

¹Department of Computer Science and Engineering Jadavpur University, Kolkata, India ²School of Computer Science and Engineering Nanyang Technological University, Singapore

¹anupam@sentic.net, ¹dipankar.dipnil2005@gmail.com

²cambria@ntu.edu.sg, ¹sivaji_cse_ju@yahoo.com

Abstract

Information extraction in the medical domain is laborious and time-consuming due to the insufficient number of domain-specific lexicons and lack of involvement of domain experts such as doctors and medical practitioners. Thus, in the present work, we are motivated to design a new lexicon, WME 3.0 (WordNet of Medical Events), which contains over 10,000 medical concepts along with their part of speech, gloss (descriptive explanations), polarity score, sentiment, similar sentiment words, category, affinity score and gravity score features. In addition, the manual annotators help to validate the overall as well as individual category level of medical concepts of WME 3.0 using Cohen's Kappa agreement metric. The agreement score indicates almost correct identification of medical concepts and their assigned features in WME 3.0.

1 Introduction

In the clinical domain, the representation of a lexical resource is treated as a crucial and contributory task because of handling several challenges. The challenges are the identification of medical concepts, their categories and relations, disambiguation of polarities, recognition of semantics whereas the scarcity of structured clinical texts doubles the challenges. In the last few years, several researchers were involved in developing various domain-specific lexicon such as Medical WordNet and UMLS (Unified Medical Language System) to cope up with such challenges. These lexicons help to bridge the gap between medical experts such as doctors or medical practitioners and non-experts such as patients (Cambria et al., 2010a; Cambria et al., 2010b).

However, medical text is in general unstructured since doctors do not like to fill forms and prefer free-form notes of their observations. Hence, a lexical design is difficult due to lack of any prior knowledge of medical terms and contexts. Therefore, we are motivated to enhance a medical lexicon namely WordNet of Medical Events (WME 2.0) which helps to identify medical concepts and their features. In order to enrich this lexicon, we have employed various well-known resources like conventional WordNet, SentiWordNet (Esuli and Sebastiani, 2006), SenticNet (Cambria et al., 2016), Bing Liu (Liu, 2012), and Taboada's Adjective list (Taboada et al., 2011) and a preprocessed English medical dictionary¹ on top of WME 1.0 and WME 2.0 lexicons (Mondal et al., 2015; Mondal et al., 2016). WME 1.0 contains 6415 number of medical concepts and their glosses, POS, polarity scores, and sentiment. Thereafter, Mondal et. al., (2016) enhanced WME 1.0 by adding few more features as affinity score, gravity score, and SSW to the medical concepts and presented as WME 2.0. The affinity and gravity scores present the hidden link between the pair of medical concepts and the concept with the various source of glosses respectively. SSW of a medical concept refers the similar sentiment words (SSW) which follow the common sentiment property.

In the current research, we have focused on enriching WME 2.0 with more number of medical concepts and including an additional feature i.e medical category. In order to develop such updated version of WME namely WME 3.0, we have taken the help of WME 1.0 and WME 2.0. We have also noticed that the previous versions of WMEs are unable to extract knowledge-based information such as the category of the medical concepts and its coverage is also lower.

¹[http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+\(Malestrom\).pdf](http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+(Malestrom).pdf)

Therefore, we have enhanced the number of medical concepts as well as add category feature on top of WME 2.0. The current version, WME 3.0 contains 10,186 number of medical concepts and their category, POS, gloss, sentiment, polarity score, SSW, affinity and gravity scores. For example, WME 3.0 lexicon presents the properties of a medical concept say *amnesia* as of category (*disease*), POS (*noun*), gloss (*loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.*), sentiment (*negative*), polarity score (*-0.375*), SSW (*memory_loss, blackout, fugue, stupor*), affinity score (*0.429*) and gravity score (*0.170*).

Moreover, to enhance and validate lexicon with the newly added medical concepts and categories, we have summarized our contributions as follows.

(a) *Enriching the number of medical concepts in the existing lexicon, WME 2.0:* In order to meet up this issue, we have employed a preprocessed English medical dictionary² and various well-defined lexicons such as SentiWordNet, SenticNet, and MedicineNet etc. They helped to enhance the number of medical concepts of the proposed lexicon.

(b) *Overall validation of the current lexicon:* To resolve the issue, we have taken the help of two manual annotators as medical practitioners. The annotators provided agreement scores that are processed using Cohen's Kappa and obtained a κ score which assists in validating the overall lexicon as well as the individual features of WME 3.0 (Viera et al., 2005).

(c) *Evaluate various individual feature of the medical concepts:* In order to extract the subjective and knowledge-based features, we have applied our evaluation scripts on the mentioned resources. The scripts assist in identifying the affinity and gravity scores as feature values for the concepts. Also, the resources are used to assign the SSW as semantics and glosses for the concepts. On the other hand, a supervised classifier helps to add the category feature in the proposed lexicon.

The remainder of the paper is organized as follows: Section 2 presents the related works for building a medical lexicon; Section 3 and Section 4 describe the previous versions of WMEs like WME 1.0 and WME 2.0 and the development

steps of WME 3.0; Section 5 discusses the validation process of the proposed lexicon; finally, Section 6 illustrates the concluding remarks and future scopes of the research.

2 Background

Biomedical information extraction is treated as one of the challenging research tasks as it deals with available medical corpora that are either unstructured or semi-structured. Hence, a domain-specific lexicon becomes an essential component to convert a structured corpus from the unstructured corpus (Borthwick et al., 1998). Also, it helps in extracting the subjective and conceptual information related to medical concepts from the corpus. Various researchers have tried to build various ontologies and lexicons such as UMLS, SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), MWN (Medical WordNet), SentiHealth, and WordNet of Medical Events (WME 1.0 and WME 2.0) etc. in the domain of healthcare (Miller and Fellbaum, 1998; Smith and Fellbaum, 2004; Asghar et al., 2016; Asghar et al., 2014). UMLS helps to enhance the access to biomedical literature by facilitating the development of computer systems that understand biomedical language (Bodenreider, 2004). SNOMED-CT is a standardized, multilingual vocabulary that contains clinical terminologies and assists in exchanging the electronic healthcare information among physicians (Donnelly, 2006).

Furthermore, Fellbaum and Smith (2004) proposed Medical WordNet (MWN) with two sub-networks e.g., Medical FactNet (MFN) and Medical BeliefNet (MBN) for justifying the consumer health. The MWN follows the formal architecture of the Princeton WordNet (Fellbaum, 1998). On the other hand, MFN aids in extracting and understanding the generic medical information for non-expert groups whereas MBN identifies the fraction of the beliefs about the medical phenomena (Smith and Fellbaum, 2004). Their primary motivation was to develop a network for medical information retrieval system with visualization effect. SentiHealth lexicon was developed to identify the sentiment for the medical concepts (Asghar et al., 2016; Asghar et al., 2014). WME 1.0 and WME 2.0 lexicons were designed to extract the medical concepts and their related linguistic and sentiment features from the corpus (Mondal et al., 2015; Mondal et al., 2016).

²[http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+\(Malestrom\).pdf](http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+(Malestrom).pdf)

These mentioned ontologies and lexicons assist in identifying the medical concepts and their sentiments from the corpus but unable to provide the complete knowledge-based information of the concepts. Hence, in the current work, we are motivated to design a full-fledged lexicon in healthcare which provides the linguistic, sentiment, and knowledge-based features together for the medical concepts.

3 Attempts for WordNet of Medical Events

In healthcare, a domain-specific lexicon is required for identifying the conceptual and knowledge-based information such as category, gloss, semantics, and sentiment of the medical concepts from the clinical corpora (Cambria, 2016). We have borrowed the knowledge from a domain-specific lexicon namely WordNet of Medical Events (WME) with its two different versions such as WME 1.0 and WME 2.0. These versions are distinguished according to the versatility and variety of medical concepts and their features.

3.1 WME 1.0

WME 1.0 contains 6415 numbers of medical concepts and their linguistic features such as gloss, parts of speech (POS), sentiment and polarity score (Mondal et al., 2015). The gloss and POS represent the descriptive definition and linguistic nature of the medical concepts whereas the sentiment and polarity score refer the classes as positive, negative, and neutral and their corresponding strength (+1) and weakness (-1). The resource was prepared by employing the trial and training datasets of SemEval-2015 Task-6³ which initially contains only 2479 medical concepts. Thereafter, the extracted concepts were updated using WordNet and a preprocessed English medical dictionary as mentioned earlier for enriching the number of concepts and identifying gloss and POS of them. However, sentiment and polarity scores were added afterwards using sentiment lexicons such as SentiWordNet⁴, SenticNet⁵, Bing Liu's subjective list⁶, and Taboada's adjective list⁷ (Cambria et al., 2016; Taboada et al., 2011; Esuli and Sebastiani, 2006).

³<http://alt.qcri.org/semeval2015/task6/>

⁴<http://sentiwordnet.isti.cnr.it/>

⁵<http://sentic.net/downloads/>

⁶<https://www.cs.uic.edu/>

⁷<http://neuro.imm.dtu.dk/wiki/>

For example, the medical concept *abnormality* appears with the following gloss, POS as noun, negative sentiment and polarity score of -0.25 in WME 1.0.

3.2 WME 2.0

The next version of WME, i.e., WME 2.0, extracts more semantic features of medical concepts (Mondal et al., 2016) and added with the existing features of WME 1.0. While updated WME 2.0 with affinity score, gravity score, and SSW, the number of concepts in WME 2.0 remains same, but the features of each concept are included (Mondal et al., 2016).

Affinity score indicates the strength of a medical concept and its corresponding SSWs by assigning a probability score. SSW of a medical concept presents the SSW shared through their common sentiment property. The affinity score '0' indicates no relation whereas '1' suggests a strong relationship between a pair of concepts. On the other hand, gravity score helps to extract the sentiment relevance between a concept and its glosses. It ranges from -1 to 1 including 0 while '-1' suggests no relation, '0' describes neutral situations of either concept or gloss without sentiment, and '1' indicates strong relations either positive or negative. It is used to prove the knowledge-based relevance between a concept and its gloss. In order to extract the features, the authors used WordNet, SentiWordNet, SenticNet, and a preprocessed English medical dictionary. Figure 1 shows the presentation of WME 2.0 lexicon for a medical concept *abnormality*.

In the present research, we have enriched the number of medical concepts and category feature with WME 2.0 lexicon and presented the enhanced version WME 3.0. The following section discusses the steps of WME 3.0 building.

4 Development of WME 3.0

A large number of daily produced medical corpora and their adaptable natures introduce the difficulty to build a full-fledged medical lexicon in healthcare domain. In order to resolve the issue, we have proposed a new version of WordNet of Medical Events namely WME 3.0. It is observed that WME 3.0 helps to extract more medical concepts and features from the unstructured corpus with respect to the previous version of WME, i.e., WME 2.0.

```

<Concept>
<Title>abnormality</Title>
<Properties>
  <Affinity_score>0.692</Affinity_score>
  <Gloss>An abnormal physical condition resulting from
  defective genes or developmental deficiencies.</Gloss>
  <Gravity_score>0.125</Gravity_score>
  <Polarity_score>-0.25</Polarity_score>
  <POS>Noun</POS>
  <Sentiment>Negative</Sentiment>
  <SSW>Anomaly,Peculiarity,Extraordinariness</SSW>
</Properties>
</Concept>

```

Figure 1: An example of assigned features of a medical concept *abnormality* under WME 2.0 lexicon.

Another 3771 number of medical concepts and an additional category feature were newly added into WME 3.0. Finally, WME 3.0 contains 10,186 medical concepts and their POS, categories, affinity scores, gravity scores, polarity scores, sentiments and SSW. To identify the additional medical concepts, we have employed the conventional WordNet⁸ and MedicineNet⁹ resource. Thereafter, we have written a script to extract new medical concepts, which are semantically (like common POS as well as sentiment) related with medical concepts of WME 2.0. Besides, SentiWordNet, SenticNet, Bing Liu subjective list, Taboada’s adjective list, and previously mentioned preprocessed medical dictionary help to assign all features except category to 3771 medical concepts which were added.

Thereafter, we newly considered four different types of categories namely diseases, drugs, symptoms, and human_anatomy for this research after examining the nature of medical concepts. In WME 3.0, all concepts are tagged with either the above-mentioned four categories or MMT category. MMT represents the miscellaneous medical terms which refer to the uncategorized and unrecognized medical concepts. In order to assign the category to the medical concepts, we have applied a well-known machine learning classifier, Naïve Bayes on top of WME 3.0 driven features. The classifier learns through the manually annotated 2000 medical concepts and their categories. Thereafter, rest of 8186 medical concepts of WME 3.0 were processed by the classifier by predicting the category (Mondal et al., 2017a).

⁸<https://wordnet.princeton.edu/>

⁹<http://www.medicinenet.com/script/main/hp.asp>

For example, the medical concept *ranitidine* represents the category, *drug* in WME 3.0 lexicon. Table 1 illustrates a comparative analysis and progress reports on WME 1.0, WME 2.0, and WME 3.0 with respect to the coverage of medical concepts, n-gram counts, and other different features such as POS, sentiment, polarity score, affinity score, gravity score, and category.

We have also noticed that the proposed WME 3.0 primarily contains POS as a noun, sentiment as negative, category as disease and drug, and n-gram feature as uni-grams and bi-grams. The observations could help to understand the characteristic of the lexicon and assist in designing various applications viz. medical annotation and concept network systems etc. The lexicon is very much demanding to identify four different types of categories and each medical concepts related gloss from a medical corpus, which presents the difference between WME 3.0 and already established very large scale semantic networks, such as UMLS. Also, the lexicon-driven medical concepts and their features also assist in emulating human thought as a recommendation of medical advice, serving a potential foundation of a higher-order cognitive model under natural language processing (Cambria and Hussain, 2015; Cambria et al., 2011). Finally, the evaluation process of WME 3.0 as overall and its individual feature levels are discussed in the following section.

5 Evaluation

In order to validate our proposed WME 3.0 lexicon, we have conducted the following result analysis. The result shows the agreement between two manual annotators to explain the acceptance

Features		WME 1.0	WME 2.0	WME 3.0
No. of Concepts		6415	6415	10186
n-grams	Uni-gram	2956	2956	3722
	Bi-gram	2837	2837	3866
	Tri-gram	622	622	1762
POS	Noun	4248	4248	7677
	Verb	2056	2056	2352
	Adjective	111	111	157
Sentiment and Polarity_score	Positive (≥ 1)	2800	2800	3227
	Negative (< 1)	3615	3615	6959
Affinity_score	0 to 0.5	-	4325	7177
	0.5 to 1	-	2090	3009
Gravity_score	less than zero	-	2320	3783
	equal to zero	-	732	1961
	grater than zero	-	3363	4442
Category	Disease	-	-	3243
	Drug	-	-	3390
	Symptom	-	-	1409
	Human_Anatomy	-	-	227
	MMT	-	-	1917

Table 1: [Color online] A comparative statistics for various features of medical concepts present in WME 1.0 (Blue), WME 2.0 (Green), and WME 3.0 (Yellow).

of overall lexicon as well as its individual features. The agreement has been calculated using Cohen’s Kappa coefficient score κ which is defined in Equation 1 (Viera et al., 2005).

$$\kappa = \frac{Pr_a - Pr_e}{1 - Pr_e}, \quad (1)$$

where Pr_a is the observed proportion of full agreement between two annotators. In addition, Pr_e is the proportion expected by a chance which indicates a kind of random agreement between the annotators.

5.1 Overall Validation of WME 3.0

WME 3.0 has been validated by two manual annotators, where the annotators are medical practitioners. The annotators have verified both medical concepts and their category, POS, gloss, affinity score, gravity score, polarity score, SSW, and sentiment features and presented as a number of yes (agreed) and number of no (disagreed) values. Table 2 indicates the values provided by both of the annotators in terms of agreement-based scores. The scores produced 0.79 κ score using equation 1. The κ score shows significantly approved result for WME 3.0 lexicon.

No. of Concepts: 10186		Annotator-1	
		Yes	No
Annotator-2	Yes	8629	189
	No	285	1083

Table 2: An agreement analysis between two annotators to validate medical concepts and their all features under WME 3.0.

5.2 Individual Feature based Validation of WME 3.0

On the other hand, the same annotators also assist in validating the individual feature of WME 3.0 with respect to the medical concepts. Hence, we have split the proposed lexicon into five parts where each of the parts contains the medical concepts and its corresponding primary features viz. category, POS, gloss, SSW, and sentiment individually. We have not considered rest of the three features namely affinity, gravity, and polarity scores of WME 3.0 because these features were derived from the above-mentioned five primary features. Thereafter, the annotators help to validate the five parts by counting the number of yes (agreed) and no (disagreed) individually. The provided agreement counts are processed with Equation 1 and get 0.89, 0.91, 0.88, 0.82, and 0.92 κ scores for category, POS, gloss, SSW, and sentiment, respectively.

The κ scores prove the usefulness and quality of individual features of the medical concepts for WME 3.0. Table 3 shows the agreement statistics between two annotators for validating the features of WME 3.0 lexicon.

No. of Concepts: 10186			Annotator-1		κ score
			Yes	No	
Annotator-2	Category	Yes	8778	93	0.89
		No	161	1154	
	POS	Yes	9229	52	0.91
		No	92	813	
	Gloss	Yes	8805	97	0.88
		No	172	1112	
	SSW	Yes	8767	137	0.82
		No	256	1026	
	Sentiment	Yes	8727	67	0.92
		No	124	1268	

Table 3: An agreement analysis between two annotators to validate category, POS, Gloss, SSW, and Sentiment features of medical concepts of WME 3.0.

We have analyzed the agreement scores for the features of WME 3.0. It is found that all the features of medical concepts are quite correctly labeled in the lexicon as presented in Table 3. We have also observed that the disagreement has been occurred due to the conceptual mismatch between two annotators or place of the usage of a few medical concepts for each of the features.

For example, the medical concept *blood clot* is tagged with either *symptom* or *disease* category. In case of POS, the medical concept *abnormality* is either labeled as an *adjective* or a *noun* whereas *menstrual cycle* refers *positive* or *negative* sentiment. Such types of disagreements are treated as very difficult task for the contextual behavior of medical corpora.

Besides, we have studied each type of the categories such as disease, symptom, and drug etc. to justify their presence in WME 3.0 lexicon. The annotators again help to validate each of the assigned categories using agreement analysis as shown in Table 4. The supplied agreement counts have been applied on Equation 1 and we found 0.89, 0.87, 0.88, 0.90, and 0.91 κ scores for disease, symptom, drug, human_anatomy, and MMT categories, respectively.

Finally, we can conclude that, WME 3.0 lexicon assists in increasing the coverage of the medical concepts as well as features and may be pre-

No. of Concepts			Annotator-1		κ score
			Yes	No	
Annotator-2	Disease (3243)	Yes	2794	31	0.89
		No	51	367	
	Symptom (1409)	Yes	1214	14	0.87
		No	26	155	
	Drug (3390)	Yes	2922	34	0.88
		No	53	381	
Human_anatomy (227)	Yes	196	2	0.90	
	No	3	26		
MMT (1917)	Yes	1652	12	0.91	
	No	28	225		

Table 4: An agreement analysis between two annotators to validate individual categories of WME 3.0.

sented as a full-fledged lexicon in the healthcare domain. Also, the lexicon can take a crucial role to design various applications such as medical annotation, concept network, and relationship identification system in healthcare (Mondal et al., 2017b).

6 Conclusion and Future Work

The present task has been motivated to enrich a medical lexicon with additional medical concepts and a feature called category in WME 3.0. In order to prepare the current version, we have employed previous two versions of WME viz. WME 1.0 and WME 2.0 along with various well-defined lexicons and a machine learning classifier. WME 3.0 contains 10,186 medical concepts and eight different types of useful features such as category and gloss etc.

In addition, we have also validated WME 3.0 from two different aspects, namely overall evaluation and usefulness of individual feature with the help of two manual annotators. The annotators provided agreement scores that were processed using Cohen’s kappa agreement analysis. Finally, the κ scores showed the importance of WME 3.0 in healthcare. In future, we will attempt to enhance WME 3.0 with more number of medical concepts as well as syntactic and semantic features for improving the coverage and quality.

References

- Muhammad Z Asghar, Aurangzeb Khan, Fazal M Kundi, Maria Qasim, Furqan Khan, Rahman Ullah, and Irfan U Nawaz. 2014. Medical opinion lexicon: an incremental model for mining health reviews. *International Journal of Academic Research*, 6(1):295–302.

- Muhammad Zubair Asghar, Shakeel Ahmad, Maria Qasim, Syeda Rabail Zahra, and Fazal Masud Kundi. 2016. SentiHealth: creating health-related sentiment lexicon using hybrid approach. *Springer-Plus*, 5(1):1139.
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the Sixth Workshop on Very Large Corpora*, volume 182.
- Erik Cambria and Amir Hussain. 2015. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer, Cham, Switzerland.
- Erik Cambria, Amir Hussain, Tariq Durrani, Catherine Havasi, Chris Eckl, and James Munro. 2010a. Sentic computing for patient centered applications. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1279–1282. IEEE.
- Erik Cambria, Amir Hussain, Catherine Havasi, Chris Eckl, and James Munro. 2010b. Towards crowd validation of the UK national health service. In *WebSci*, Raleigh.
- Erik Cambria, Thomas Mazzocco, Amir Hussain, and Chris Eckl. 2011. Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. In D Liu, H Zhang, M Polycarpou, C Alippi, and H He, editors, *Advances in Neural Networks*, volume 6677 of *Lecture Notes in Computer Science*, pages 601–610, Berlin. Springer-Verlag.
- Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. 2016. SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. In *COLING*, pages 2666–2677.
- Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.
- Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.
- Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- George Miller and Christiane Fellbaum. 1998. WordNet: An electronic lexical database.
- Anupam Mondal, Iti Chaturvedi, Dipankar Das, Rajiv Bajpai, and Sivaji Bandyopadhyay. 2015. Lexical Resource for Medical Events: A Polarity Based Approach. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1302–1309. IEEE.
- Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay. 2016. WME: Sense, Polarity and Affinity based Concept Resource for Medical Events. *Proceedings of the Eighth Global WordNet Conference*, pages 242–246.
- Anupam Mondal, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2017a. Auto-categorization of medical concepts and contexts. In *IEEE Symposium Series on Computational Intelligence (SSCI 2017)*, Honolulu, Hawaii, USA.
- Anupam Mondal, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2017b. MediConceptNet: An Affinity Score Based Medical Concept Network. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, pages 335–340.
- Barry Smith and Christiane Fellbaum. 2004. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, page 371. Association for Computational Linguistics.
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding inter-observer agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

Using Context to Improve the Spanish WordNet Translation

Alfonso Methol, Guillermo López, Juan Miguel Álvarez, Luis Chiruzzo, Dina Wonsever
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

Abstract

We present some strategies for improving the Spanish version of WordNet, part of the MCR, selecting new lemmas for the Spanish synsets by translating the lemmas of the corresponding English synsets. We used four simple selectors that resulted in a considerable improvement of the Spanish WordNet coverage, but with relatively lower precision, then we defined two context based selectors that improved the precision of the translations.

1 Introduction

This paper presents an approach at the expansion of the lexical database WordNet in Spanish using an automatic translation processes. We implemented some previously proposed strategies for improving the coverage of the lexical database in Spanish, then we analyzed the results that these strategies produced and finally we designed new strategies in order to improve the quality of the translated lemmas.

The rest of the paper is structured as follows: section 2 presents the lexical database we aim to improve and describes related work in the area, section 3 describes the translation sources we used and how they were preprocessed, section 4 details the different strategies implemented for translation, section 5 shows the results obtained by the strategies and their evaluation, finally section 6 shows our conclusions and some future research directions.

2 Background

The Multilingual Central Repository, MCR (González-Agirre et al., 2012), is a multilingual lexical database that contains linked WordNet versions for English and five languages spoken in the

Iberian peninsula: Spanish, Catalan, Basque, Galician and Portuguese. The same Princeton WordNet synsets structure is used for all languages. The central component of this lexical database is the Inter-Lingual-Index (ILI), which allows the mapping of concepts of different languages through the use of identifiers. The identifiers are composed of four values: language, version of MCR, synset offset and part of speech.

Synsets in different languages that have the same meaning share the offset, version and part of speech, varying the language. For example, “house” (eng-30-03544360-n) corresponds to “casa” (spa-30-03544360-n) and both synsets are related through the ILI code “ili-30-03544360-n”.

The first attempts at building a Spanish version of WordNet are described in (Atserias et al., 1997), using bilingual English-Spanish dictionaries and a large monolingual Spanish dictionary. A different approach is proposed in (Oliver and Climent, 2011) for Spanish and Catalan, using machine translation systems to translate the semantically annotated SemCor (Miller et al., 1993) corpus and select the translations for variants based on the relative frequencies of words in the corpus with the following strategies:

- Algorithm A: Order the English synsets by frequency in the original corpus. Starting with the most frequent synset, build a subset of the automatically translated corpus with the sentences that contain a member of the synset. Choose the most frequent lemma from the translated corpus that has the same POS as the original synset. This process is repeated for each synset in order of frequency.
- Algorithm B: The same as algorithm A, but choose a lemma only if its frequency is at least twice the frequency of the next lemma. This process has considerably better precision than the previous one.

In (Pradet et al., 2014) the authors present a method for improving the French version of WordNet. They compile a collection of possible translations for the variants from several bilingual sources and design strategies for selecting the appropriate translation, these strategies are called “selectors”. A selector is a heuristic strategy that takes a synset and a set of candidate lemmas in the target language, and returns the most appropriate lemma that should be associated to the synset. A similar approach was followed by (Herrera et al., 2016) for the expansion of Spanish WordNet, defining five selectors and obtaining good results for a subset of synsets from Princeton WordNet (92% accuracy for simple selectors and 74% accuracy for the distributional selector). The selectors were only applied on a subset of the synsets due to the long execution times, also some problematic synsets (such as multiword expressions) were not considered, which might explain in part the high accuracy of the simple selectors.

The authors of (Oliver, 2016) also use a dictionary based approach, combining several linguistic resources in a variety of languages for improving the WordNet translation in each of those languages.

3 Translation sources

Translation sources are key elements in the process of building WordNet in Spanish. They provide, for the English lemmas, the lemmas in Spanish that will be used by the selectors as translation candidates.

Two types of sources were used: dictionaries and statistical machine translators. The dictionaries are made up of tuples [*English word*, *Spanish word*, *POS*]. They are generated manually so they are very reliable, but with a limited volume of translations. The machine translators used are statistical systems that allow to translate words and also complete sentences taking the context into consideration, a property that will be exploited by some of the selectors.

3.1 Dictionaries

- Apertium: It is a rule based machine translation system (Forcada et al., 2011) developed with the joint financing of the Spanish government and the Generalitat de Catalunya at the University of Alicante. The software as well as the linguistic data is free and

it is released under the terms of the GNU GPL license. A dictionary was created from the “.dix” file of Apertium corresponding to the translations from English into Spanish. The version used has 26,643 translations, and covers 42,996 WordNet lemmas, which accounts for 20.67% of it.

- Wiktionary¹: It is a project of the Wikimedia foundation that aims to create a free multilingual dictionary, based on the massive collaboration of volunteers through the wiki technology for the elaboration of its content. It is currently available in more than 170 languages and has more than 15 million entries. Because of the considerable volume of its data and its well defined structure, it is particularly useful for our processing. The version used contains 40,166 possible translations into Spanish for 47,982 lemmas, covering the 23.06% of WordNet lemmas.
- Eurovoc: Published by the Publications Office of the European Union, it is a multidisciplinary thesaurus focused on the terminology used in the different areas of activity of the European Union (Maciá, 1995), and it covers the 23 official languages of the region. Due to the scope of the thesaurus, this translation source has few general terms, which considerably restricts its broad applicability in this project, but it contains specific data that can be very useful for translation of diplomatic documents. Out of 6945 lemmas contained in EuroVoc, 2032 appear in WordNet, which represents 1.38% of the lemmas.

3.2 Machine translators

- Google Translate: It is a statistical machine translation system capable of translating texts, speech, images, websites among more than 100 languages. Provides a free access web tool² as well as a service included in Google Cloud Platform.
- Microsoft Translator: It is a statistical machine translation web service³ provided by Microsoft, which can be used through an API that provides translation of text, voice and text to speech.

¹<https://www.wiktionary.org/>

²<https://translate.google.com>

³<https://www.bing.com/translator>

- Yandex: They offer a statistical machine translator⁴ for many pairs of languages, including Spanish and English. The translator uses a combination of dictionaries of words and expressions with probabilistic information and also linguistic rules. It can be queried using a web API.

3.3 Cleaning sources

To solve some of the limitations and reduce the costs of access to the selected translation sources, a single format was defined and stored in the same database. For each translation source a table was created with the following columns:

- English word
- Spanish translation
- Part of Speech

The dictionaries did not need any extra processing and only these fields are stored. The tables corresponding to machine translation systems have another field:

- Snapshot date

We decided to take a snapshot of the translation of all WordNet lemmas by each of the machine translation systems at a specific time. This was motivated by the different limitations in the use of online APIs and their response times. Using the snapshot approach, we can use the machine translation systems as if they were just another dictionary. Although we might not have completely up to date information in each run, we consider the translations we use should not vary much in time and the execution time is greatly improved respect to the online execution of the APIs. The snapshot date is stored, so we can later on take a new snapshot, compare the differences and adjust the methods accordingly.

None of the three machine translation used systems return the POS along with the translation, so we used FreeLing (Padró and Stanilovsky, 2012) for POS tagging. We detected many translation errors, where a different POS was returned because of the lack of context, so we did some improvements to the translation heuristic, such as adding the prefix “to” to verbs in English in order to force the translator to consider them as verbs. We also used FreeLing dependency parser to assign the POS in multiword expressions.

⁴<https://translate.yandex.com/>

3.4 Coverage

We analyzed the coverage of MCR over a corpus of 850 million words of news text in Spanish (Bonanata and Stecanella, 2013)

The coverage before our process is shown in the following table:

POS	Lemmas in corpus	Lemmas in MCR
Adj	42,604	5,592 (13.12%)
Adv	10,676	523 (4.90%)
Noun	104,811	11,523 (10.99%)
Verb	37,522	8,821 (23.51%)
All	195,613	26,459 (13.53%)

Table 1: MCR Coverage over news corpus

We can observe a low coverage of the corpus MCR. This is due in part to the number of lemmas available in Spanish.

4 Translation process

We first implemented some of the already defined selectors and applied them to the whole collection of synsets. As these selectors resulted in poor precision, we created new selectors that exploit contextual information in order to improve the precision of the translation.

4.1 Simple selectors

Following the strategies of (Pradet et al., 2014) and (Herrera et al., 2016), we reimplemented some of the selectors that have been previously executed for only a fraction of the English synsets and applied them to all the synsets.

- Monosemy

This strategy works with English lemmas which appear in a single synset regardless of their part of speech. The assumption behind this is that this uniqueness condition implies the meaning of the lemma is unambiguous. The translations of all the sources for each compliant lemma are assigned to the corresponding synset.

For example: Consider the English lemma “advisable” which only appears in the English synset “eng-30-00067038-a”. The selector then assigns all of the lemmas translations to the corresponding Spanish synset “spa-30-00067038-a”, in this case: “aconsejable”, “recomendable” and “conveniente”.

- Single Translation

This selector takes into account only those lemmas that have a unique translation into Spanish. This translation is added in all the synsets in Spanish corresponding to the synsets in English that contain this lemma.

For example: Consider the lemma “flavor”, which occurs in the synsets “eng-30-14526182-n”, “eng-30-05715864-n” and “eng-30-05844282-n”. There is a unique translation for this lemma that is “sabor”. This translation is selected for the corresponding synsets in Spanish. However, for the lemma “play” occurring in the synsets “eng-30-01072949-v”, “eng-30-02370650-v” and “eng-30-01725051-v” (among other 35 synsets in total), our translation sources give four possible lemmas: “jugar”, “reproducir”, “tocar” and “interpretar”. Because of this the selector discards these translations.

- Factorization

Unlike previous selectors, this one runs at synset level. For each lemma of a synset it obtains all its translations and generates a translation set. Once the sets of translations of each lemma of the synset are obtained, the selector keeps those translations common to all sets.

For example: The synset “eng-30-00011516-r” contains the lemmas “poorly”, “badly” and “ill”, where their translations are:

- poorly: mal, pobremente.
- badly: mal, malamente.
- ill: mal, enfermo.

In this example, the only translation common to the three lemmas that is selected for the corresponding synset in Spanish is “mal”, the remaining translations (“pobremente”, “malamente” and “enfermo”) are discarded.

- Derived Adverb

This selector is executed for the adverb synsets and is the only one that uses a semantic relation of those defined in MCR, the `is_derived_from` relation. From an adverb synset, look up with which adjective

synsets it is related. For each adjective obtain the translations, and use morphological derivation rules to convert them into possible adverbs.

The morphological rules applied are as follows:

- If the adjective ends with the letter “o”, it is replaced by the sequence “amente”, for example, for “rápido” the result is “rápidamente”.
- If the adjective ends with the letter “r” or “n”, the sequence “amente” is attached, for example, for “alentador” the result is “alentadoramente”.
- If the adjective does not fit into the above categories, only the sequence “mente” is attached, for example, for “vil” the result is “vilmente”.

As these rules are heuristics, not all results obtained after the process are valid adverbs. For example, when applying the rules to the adjective “rojo” we get the adverb “rojamente”, which does not exist as a valid word in the Spanish language. To solve this problem the adverbs generated were validated against a list of adverbs that occur in a corpus (Bonanata and Stecanella, 2013).

For example: The synset “eng-30-00010466-r” has the lemmas “fully”, “full” and “to_the_full” and is related to the adjective synset “eng-30-00522885-a” containing the lemmas “total” and “full”. The translations for the adjectives are: “pleno”, “repleto”, “lleno”, “completo” and “total”. Applying the morphological rules we get: “plenamente”, “repletamente”, “llenamente”, “completamente” and “totalmente”. These are checked using the corpus and added to the corresponding synset.

4.2 Selectors based on contextual information

After analyzing the performance of the original selectors, which will be shown in section 5, we realized that many of the errors happened because these selectors do not take in consideration the context the words could be used in. We defined two new selectors that try to use the context provided by the examples of the synsets to improve the quality of the candidate translations.

We translate all the examples contained in WordNet using Google Translate and generated a parallel corpus associated with synsets. 27.71% of the MCR English synsets have examples, adding up to 41,305 candidates, which gives us an upper bound to the number of synsets we might translate with these strategies.

- Filtering selector

This selector works by analyzing which of the generated translations of the present lemma in an example in English, are in the translation of the example to Spanish. The check of occurrences of both the lemmas in the example in English, and their translations in the translated example is done in two stages. It is called filtering because it leverages the information from the dictionaries, trying to filter which of the candidates are present in the example and its translation. The procedure is as follows: First check if lemma and translation occur in the example and the translated example. If this does not happen, apply FreeLing to the text to obtain the lemma and POS of each word. Then iterate them by re-checking the occurrences. This second stage tries to detect words or translations that occur in the examples in a conjugated form, as is the case for many verbs. Otherwise we would be losing many valid translations. This is done as a second step because using FreeLing to get the lemmas is an expensive process.

For example: When we apply the selector to the example “his last words” associated with the synset “eng-30-00004296-a”, it detects that the only lemma of the synset (“last”) occurs directly in the example. Once this is detected, the translations are obtained. In this case, the lemma “último” is the only translation candidate.

The translated example is “sus últimas palabras”. The candidate lemma is not present so FreeLing is used to obtain the lemmas and POS of the translated example, getting the following information: “[(su, D), (último, A), (palabra, N)]”. Since we are dealing with an adjective synset, we compare to the adjectives returned by FreeLing and we get a match with the lemma “último”, which is selected as the translation to the corresponding

synset (“spa-30-00004296-a”) in Spanish.

- Structure based selector

This selector focuses on the use of translated examples as a parallel corpus where it is possible to align the different parts of the sentences in both languages. We use the path from the root to the word in a dependency parse tree, and try to match the corresponding path in the tree of the translated example. In this way, we use the internal structure of the sentences and the relative positions of the words, such as their location within a subject or a predicate.

We begin by obtaining the dependency parses of the example and its translation using FreeLing. This construction allows the analysis of the different components of sentences and their relationships. Using the dependency structures, we identify the lemma to be translated from the sentence and its syntactic (subject or predicate) location, and take note of the labels belonging to the shortest path from the root of the tree. The same path is followed in the translated example, taking in consideration the differences in label names for both languages, and we return a lemma if it is in the appropriate position in the tree and has the expected POS.

Example: We want to translate the lemma “bond” for the English synset “eng-30-13792183-n” using the example: “their friendship constitutes a powerful bond between them”. The dependency tree for this sentence is shown in figure 1.

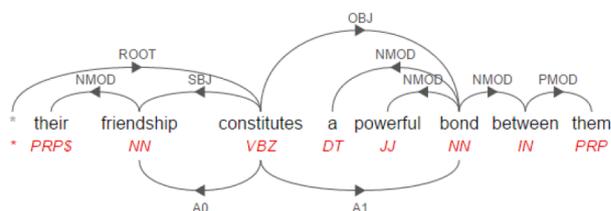


Figure 1: Dependency parsing of “their friendship constitutes a powerful bond between them”

The corresponding translation for this example in Spanish is “su amistad constituye un poderoso vínculo entre ellos”, whose dependency tree is shown in figure 2. In this tree we find the lemma “vínculo” in the correspond-

Selector	Generated	MCR	Intersection	Overlap	New
Monosemy	183386	146501	47632	32.51%	74.03%
Single Transl.	81058	146501	38505	26.28%	52.50%
Factorization	111919	146501	34400	23.48%	69.26%
Derived Adv.	5161	3583	1907	53.22%	63.05%
All Simple	256852	146501	72674	50.39%	71.71%
Filtering	22401	146501	12680	8.66%	43.40%
Structure	12168	146501	6857	4.68%	43.65%
All Context	25223	146501	13291	9.07%	47.31%
All	264105	146501	75416	51.48%	71.44%

Table 2: Number of generated lemmas, overlap with MCR lemmas and generated lemmas that are new by selector.

ing position, so this lemma gets selected for the Spanish synset “spa-30-13792183-n”.

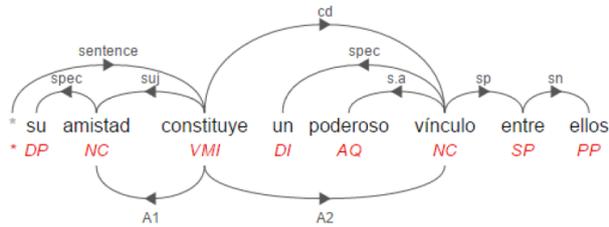


Figure 2: Dependency parsing of “su amistad constituye un poderoso vínculo entre ellos”

In this case the lemma and its translation was easy to locate both in the original sentence and the translation: it is a single name located in the direct object of the sentence in both cases, so it quickly follows that the translation of “bond” is “vínculo”.

However, this is not always the case. Among the most common errors in the execution of this selector are situations in which the root of the example in English changes considerably when translated. This is because in many cases the English and Spanish parsers use different criteria. That is the case of the sentence: “Can you read Greek?” (figure 3), whose translation is “¿Puede usted leer griego?” (figure 4). The lemma that we want to translate is “read”, and is located in the sentence predicate in the original version, but becomes the root of the tree in the translated version. Even though both sentences have similar structure in English and Spanish, the parsing process treats them differently.

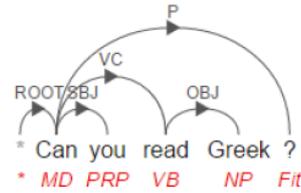


Figure 3: Dependency parsing of “Can you read greek?”

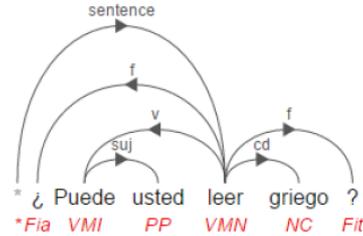


Figure 4: Dependency parsing of “¿Puede usted leer griego?”

5 Evaluation

Evaluation was one of the hardest tasks due to the complexity of the evaluation of some semantic notions, as well as the volume of data involved. Because of this, we decided to use two methods of evaluation: evaluation by overlap and evaluation by sampling.

5.1 Overlap evaluation

The overlap evaluation consists in comparing the translations generated with those already found in Spanish MCR. This could be seen as a kind of recall, giving an idea of how good our heuristics are at capturing the information we already knew. The overlap by phase and selector is shown in table 2.

Notice that the lemmas translated using the con-

text based selectors are fewer than the ones generated with the simple translators. This was an expected result, because these selectors use the synset examples. Not all synsets have examples, and even the ones that contain examples do not necessarily have them for every lemma. This coverage could be greatly improved using more data.

5.2 Sample evaluation

Due to the large volume of translations generated we could not evaluate the correctness of each one of the terms. For this reason we carried out a sampling evaluation consisting of taking a random sample of 3,000 synsets and evaluating them manually. For the initial phase, 750 synsets by POS were selected, in the contextual information phase, 1500 were selected per selector (375 by POS). We built a special tool that aids in the process of evaluating the correctness of the sampled translations. The result of this method of evaluation is an estimation of the precision for each selector and each phase. The precision is shown in table 3.

Selector	Sampled	Correct
Monosemy	3,603	2,367 (65.70%)
Single Transl.	2,471	1,927 (73.65%)
Factorization	3,193	2,057 (64.42%)
Derived Adv.	1,164	852 (73.20%)
All simple	10,431	7,203 (69.05%)
Filtering	1,695	1,424 (83.96%)
Structure	1,674	1,361 (81.30%)
All contextual	3,369	2,785 (82.67%)

Table 3: Precision by selector, showing the number of tested lemmas and the number of correct ones for each selector.

Table 4 shows the precision achieved for each POS, separated in the two phases: simple selectors and selectors with contextual information.

POS	Simple Sel.	Contextual Sel.
Adj	74.89%	87.34%
Adv	73.65%	88.42%
Noun	57.51%	80.24%
Verb	52.47%	74.12%

Table 4: Precision by POS, showing the overall precision for simple selectors and selectors with contextual information.

As we can see, the precision for the initial selectors was lower than the one reported in (Herrera et

al., 2016). There are several causes for this, first of all we transformed the whole collection of synsets and took a larger evaluation sample, even considering multiword expressions and their translations. In one of the cases the precision only for simple lemmas got 81%, while for multiword expressions it dropped to 66%. Also, on occasions the machine translation systems returned results that contained an unnecessary determinant (e.g. translating “immigration” as “*la inmigración*”). However, at many times the error was caused by selecting a translation that would be unfit for the context, for example it translated “ring” from synset “eng-30-07391863-n” (“the sound of a bell ringing”) as “*anillo*”, which is an appropriate translation for the other sense in synset “eng-30-04092609-n” (“jewelry consisting of a circlet of precious metal...”). The low precision of these methods motivated the contextual information approach, which obtained fewer translations but with better precision for all parts of speech.

5.3 Impact over MCR

The contribution to Spanish MCR is shown in Table 5.

POS	Spanish MCR Lemmas	New Lemmas	Increase
Adjectives	6,967	19,140	274.72%
Adverb	1,051	8,689	826.74%
Noun	39,142	183,880	469.78%
Verb	10,829	21,355	197.20%

Table 5: Contribution to Spanish MCR

Reanalyzing the coverage of MCR over the news text based corpus (Bonanata and Stecanella, 2013) including the newly generated lemmas we obtained the new coverage shown in table 6.

6 Conclusions

We implemented four simple selectors and two contextual based selectors for the translation of English WordNet synsets to Spanish, in order to expand the Spanish version of WordNet present in MCR. Using the simple selectors, we obtained 182,051 nouns, 19,683 verbs, 17,384 adjectives and 8,436 adverbs with 69.05% precision. The precision of these selectors was lower than the one reported in previous works, probably because in our case we evaluated the whole collection

POS	Lemmas in corpus	Lemmas in MCR	MCR + new lemmas
Adj	42,604	5,592 (13.12%)	18,063 (42,40%)
Adv	10,676	523 (4.90%)	7,105 (66,55%)
Noun	104,811	11,523 (10.99%)	35,535 (33,90%)
Verb	37,522	8,821 (23.51%)	22,427 (59,77%)
All	195,613	26,459 (13.53%)	83,130 (42,50%)

Table 6: Coverage of MCR with new lemmas.

of synsets, even processing multiword lemmas. In order to improve this precision, we designed and implemented two new selectors that use the contextual information, whose execution obtained 5,339 nouns, 4,441 verbs, 6,444 adjectives and 1,747 adverbs with 82.67% precision. The context based selectors yield much fewer results because they depend on the existence of examples in the corresponding WordNet synsets.

During the course of the project we detected several directions that could be explored in the future. First of all, we would need to analyze the cases in which the simple selectors did not give any results. This could mean expanding the set of translation sources in order to cover all the vocabulary of the original WordNet, as this coverage is the upper bound to what we might be able to translate.

For the contextual information selectors, we could obtain a larger parallel corpus of examples. One possibility is using the SemCor corpus that has been used in other projects, another possibility would be performing word sense disambiguation over a large parallel corpus, taking into account that this process would probably not select the correct synset every time. The structure selector is particularly interesting to analyze and extend, because this selector applies syntactic notions and heuristic rules that could be expanded and improved in order to add coverage and accuracy.

It would also be interesting to design new selectors based on the notions of distributed semantics, such as the use of word embeddings. The relations contained in WordNet could be used to guide the selection of new lemmas given the word embed-

dings property that words close in the vector space tend to have similar or related meanings.

References

- Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338.
- Jairo Bonanata and Rodrigo Stecanella. 2013. Extracción de opiniones de prensa. Proyecto de grado, Ingeniería en Computación, Facultad de ingeniería, Universidad de la República, Uruguay.
- Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.
- Aitor González-Agirre, Egoitz Laparra, German Rigau, and Basque Country Donostia. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118.
- Matías Herrera, Javier González, Luis Chiruzzo, and Dina Wonsever. 2016. Some strategies for the improvement of a spanish wordnet. In *Proceedings of the Global WordNet Conference*.
- Mateo Maciá. 1995. El tesoro eurovoc. *Revista General de Información y Documentación*, 5(2):265–284.
- George A Miller, Claudia Leacock, Randee Teng, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.
- Antoni Oliver and Salvador Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. Sociedad Española para el Procesamiento del Lenguaje Natural.
- Antoni Oliver. 2016. Extending the wn-toolkit: dealing with polysemous words in the dictionary-based strategy. In *Proceedings of the Global WordNet Conference*.
- Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.
- Quentin Pradet, Gaël de Chalendar, and Jeanne Bague-nier Desormeaux. 2014. Wonef, an improved, expanded and evaluated automatic french translation of wordnet. *Volume editors*, page 32.

Towards Cross-checking WordNet and SUMO Using Meronymy

Javier Álvez

LoRea Group

University of the Basque Country
UPV/EHU

javier.alvez@ehu.eus

German Rigau

IXA Group

University of the Basque Country
UPV/EHU

german.rigau@ehu.eus

Abstract

We describe the practical application of a black-box testing methodology for the validation of the knowledge encoded in WordNet, SUMO and their mapping by using automated theorem provers. In this paper, we concentrate on the part-whole information provided by WordNet and create a large set of tests on the basis of few question patterns. From our preliminary evaluation results, we report on some of the detected inconsistencies.

1 Introduction

Despite being created manually, knowledge resources such as WordNet (Fellbaum, 1998) and SUMO (Niles and Pease, 2003) are not free of errors and inconsistencies. Unfortunately, improving, revising, and correcting such large knowledge bases is a never ending task that have been mainly carried out also manually. A few automatic approaches have been also applied focusing on checking certain structural properties on WordNet (e.g. (Daudé et al., 2003), (Richens, 2008)) or using automated theorem provers on SUMO (e.g. (Horrocks and Voronkov, 2006), (Álvez et al., 2012)). Just a few more have studied automatic ways to validate the knowledge content encoded in these resources by cross-checking them. For instance, Álvez et al. (2008) exploit the EuroWordNet Top Ontology (Rodríguez et al., 1998) and its mapping to WordNet for detecting many ontological conflicts and inconsistencies in the WordNet nominal hierarchy.

In Álvez et al. (2017), we propose a method for the automatic creation of *competency questions* (CQs) (Grüninger and Fox, 1995), which enable to evaluate the competency of SUMO-based ontologies. Our proposal is based on several predefined question patterns (QPs) that are instantiated using

information from WordNet (Fellbaum, 1998) and its mapping into SUMO (Niles and Pease, 2003). In addition, we also describe an application of automated theorem provers (ATPs) for the automatic evaluation of the proposed CQs.

The main contribution of this paper is to demonstrate the practical capabilities of the method introduced in Álvez et al. (2017) for the detection of semantic agreements and inconsistencies between WordNet and SUMO thanks to their mapping. For this purpose, we propose a new set of CQs that is obtained on the basis of the part-whole data of WordNet. In our ongoing experimentations using the ATPs Vampire (Kovács and Voronkov, 2013) and E (Schulz, 2002), we have automatically detected some knowledge discrepancies and disagreements that were hidden in both WordNet, SUMO and their mapping.

Outline of the paper. In the following three sections, we introduce WordNet, SUMO, and their mapping. Then, we describe our formal interpretation of the mapping information in Section 5 and the proposed question patterns for the creation of competency questions in Section 6. Next, we discuss our preliminary evaluation results in Section 7. Finally, we report on the ongoing work in Section 8 and provide some conclusions in Section 9.

2 Meronymy Relations in WordNet

In WordNet, meronymy —the part-whole relation— holds between synsets like *backrest*_n¹ and *seat*_n¹ (i.e. parts) and *chair*_n¹ (i.e. whole). Parts are inherited from their superordinates: if a chair has a seat, then an armchair has a seat as well. But parts are not inherited “upward” as they may be characteristic only of specific kinds of things rather than the class as a whole: chairs and kinds of chairs have a seat, but not all kinds of furnitures have a seat.

There exist 3 main meronymy relations in WordNet v3.0 that relate noun synsets: *part*, the

general meronymy relation; *member*, which relates particulars and groups; *substance*, which relates physical matters and things. In total, there are 22,187 (ordered) synset pairs: 9,097 pairs using *part*, 12,293 pairs using *member* and 797 pairs using *substance*. For example, the synsets *committee*_n¹ and *committee_member*_n¹ are related by *member*, while *grape*_n¹ and *wine*_n¹ are related by *substance*.

3 SUMO and Adimen-SUMO

SUMO¹ (Niles and Pease, 2001) has its origins in the nineties, when a group of engineers from the IEEE Standard Upper Ontology Working Group pushed for a formal ontology standard. Their goal was to develop a standard upper ontology to promote data interoperability, information search and retrieval, automated inference and natural language processing.

Currently, SUMO consists of about 20,000 terms and about 70,000 axioms organized in several levels. In the the upper two levels —*Top* and *Middle* levels— one can find the concepts, relations and axioms that are meta, generic or abstract. From now on, we refer to the upper two levels of SUMO as its *core*. On the basis of these two levels, concepts that are specific to particular domains are in the so-called *domain* ontologies. Adimen-SUMO (Álvarez et al., 2012) is obtained by means of a suitable transformation of the knowledge in the core of SUMO into FOL, which enables its use by FOL ATPs such as Vampire (Kovács and Voronkov, 2013) and E (Schulz, 2002). Adimen-SUMO inherits all the axioms in the core of SUMO that can be expressed in FOL (around an 88% of the axioms).

The knowledge in SUMO is organized around the notions of *individuals* and *classes* —the main SUMO concepts. These concepts are respectively defined in Adimen-SUMO by means of the meta-predicates *\$instance* and *\$subclass*. SUMO individuals and classes are not disjoint, since every SUMO class is defined to be instance of *Class* and, thus, every SUMO class is also a SUMO individual. Additionally, SUMO also differentiates *relations* and *attributes*. In particular, SUMO distinguishes between *individual* relation and attributes —that is, instances of the SUMO classes *Relation* and *Attribute* respectively— and *classes* of relations and attributes —that is, subclasses of the

SUMO classes *Relation* and *Attribute* respectively.

SUMO provides specific predicates for dealing with relations and attributes. Among others, we currently use the next ones in Adimen-SUMO:

- *subrelation*, which relates two individual SUMO relations (that is, two instances of the SUMO class *Relation*).
- *subAttribute*, which relates two individual SUMO attributes (that is, two instances of the SUMO class *Attribute*).
- *holds^k*, which relates an individual SUMO relation (that is, an instance of the SUMO class *Relation*) with a *k*-tuple of SUMO concepts, where *k* ranges from 2 to 5.
- *attribute*, which relates a SUMO individual² with an individual SUMO attribute (that is, an instance of the SUMO class *Attribute*).

For simplicity, from now on we denote the nature of SUMO concepts by adding as subscript the symbols *o* (SUMO individuals that are neither classes nor relations nor attributes), *c* (SUMO classes that are neither classes of relations nor classes of attributes), *r* (individual SUMO relations), *a* (individual SUMO attributes), *R* (classes of SUMO relations) and *A* (classes of SUMO attributes). For example: *Cell_c*, *member_r* and *Larval_a*.

4 The Mapping Between WordNet and SUMO

WordNet is linked with SUMO by means of the mapping described in Niles and Pease (2003). This mapping connects synsets of WordNet to terms of SUMO using three relations: *equivalence*, *subsumption* and *instance*.³ *equivalence* denotes that the related WordNet synset and SUMO concept are equivalent in meaning, whereas *subsumption* and *instance* indicate that the WordNet synset is subsumed by the SUMO concept or is an instance of the SUMO concept respectively. Additionally, the mapping also uses the complementaries of *equivalence* and *instance*. We de-

²The individual in the first argument of *attribute* is restricted to be instance of *Object* by the *domain* axioms provided by SUMO.

³Note that *instance* denotes the relation that is used in the mapping between WordNet and SUMO (for example, in *Integer*®), while *\$instance* denotes the meta-predicate that is used in the axiomatization of SUMO.

¹<http://www.ontologyportal.org>

SUMO Concept Type	Mapping Relation					Total
	=	+	@	^		
Individuals	132 (0)	171 (0)	15 (0)	0 (0)		318 (0)
Classes	1,564 (0)	57,018 (546)	8,991 (337)	30 (0)		67,520 (883)
Relations	77 (0)	538 (0)	0 (0)	0 (0)		615 (0)
Attributes	340 (0)	12,762 (250)	570 (0)	0 (0)		13,662 (250)

Table 1: The mapping between WordNet and the core of SUMO

note mapping relations by concatenating the symbols ‘=’ (*equivalence*), ‘+’ (*subsumption*), ‘@’ (*instance*), ‘≐’ (complementary of *equivalence*) and ‘+̂’ (complementary of *subsumption*) to the corresponding SUMO concept. For example, the synsets $horse_n^1$ and $education_n^4$ are connected to $Horse_c=$ and $EducationalProcess_c+$ respectively.

From the 82,115 noun synsets defined in WordNet v3.0, 73,472 noun synsets are directly connected to concepts that are defined in the core of SUMO—and, thus, in Adimen-SUMO—, while only 7,578 synsets are linked to SUMO concepts defined in domain ontologies. As described in Álvez et al. (2017), those synsets linked to concepts defined in domain ontologies are connected to concepts from the core of SUMO by means of the SUMO structural relations $\$subclass$, $subrelation_r$ and $subAttribute_r$. For example, the synset $frying_n^1$ is connected to $Frying_c=$, which does not belong in the core of SUMO: $Frying_c$ is defined in the domain ontology *Food* to be subclass of the SUMO core concept $Cooking_c$. Thus, by means of $\$subclass$, we can connect $frying_n^1$ to $Cooking_c+$ in order to obtain a whole mapping between WordNet and the core of SUMO.

It is worth to remark that some noun synsets are connected to several SUMO concepts. Concretely, 1,043 synsets.

In Table 1, we provide some figures about the mapping between WordNet and the core of SUMO. More specifically, we provide the amount of noun synsets that are respectively connected to SUMO individuals, classes, relations and attributes by mapping relation. In addition, we also provide the number of multiple connections—or *multiple mappings*— between brackets. It is easy that there is no multiple mapping involving individuals and relations. Furthermore, most of the synsets are connected to SUMO classes and attributes (in total, 81,182 synsets), while only 933 synsets are connected to SUMO individuals and relations.

5 Formal Interpretations of the Mapping Between WordNet and SUMO

The automatic validation of WordNet and SUMO on the basis of CQs and ATPs requires to translate all the information into a formal language. By means of Adimen-SUMO (Álvez et al., 2012), the core information of SUMO is already written in FOL. However, WordNet and its mapping to SUMO are not formally characterized. Therefore, we next describe and compare two possible formal interpretations of the mapping between WordNet and SUMO.

The first possible interpretation is just to literally follow the definition of the mapping relations provided in Niles and Pease (2003). That is:

- *equivalence* is synonymy.
- *subsumption* indicates that the SUMO concept is a hypernym of the associated synset.
- *instance* designates the synset as an individual of the connected SUMO concept.

However, the above literal interpretation of the mapping suffers from several problems. On one hand, *subsumption* and *instance* lack an obvious interpretation when referred to SUMO individuals:⁴ it is non-sense to assert that an individual has hyponyms or individuals and, in addition, there is only one SUMO predicate for dealing with relations (i.e. $subrelation_r$) and attributes ($subAttribute_r$) respectively. On the other hand, the literal interpretation of the mapping may yield to inconsistent statements when applied to synsets that are connected to several SUMO concepts. For example, $male_horse_n^1$ is connected to both $Male_a+$ and $Horse_c+$. Thus, $male_horse_n^1$ would be interpreted of hyponym of both $Male_a$ and $Horse_c$. For this purpose, we would use the

⁴Note that most of the SUMO relations and attributes are individuals.

SUMO predicates $subAttribute_r$ and $\$subclass$ respectively. However, these two predicates are defined to relate incompatible SUMO concepts: $Attribute_c$ and $Class_c$ are disjoint classes.⁵

The second possibility is to interpret all the mapping relations exclusively in terms of SUMO individuals. Under this interpretation, we consider synsets to be related to sets of SUMO individuals that are characterized by a) the particular SUMO concept to which the synset is connected and b) the mapping relation that is used in the linking. The set of SUMO individuals that are potentially related to a given synset can be represented using SUMO statements. For the construction of those statements, we associate a different variable to each synset and choose the most suitable SUMO predicate depending of the nature of the SUMO concept to which the synset is connected: $equal$ for SUMO individuals, $\$instance$ for SUMO classes and $attribute_r$ for SUMO individual attributes.⁶ The interested reader is referred to Álvarez et al. (2017) for further details. For example, the synsets $malacosoma_americana_n^1$ and $genus_malacosoma_n^1$ are connected to $Insect_c+$ and $Larval_a+$ respectively. By associating the variables $?X$ and $?Y$ to each synset, we generate the following Adimen-SUMO statements:

($\$instance$?X Insect) (1)
(attribute ?Y Larval) (2)

On the basis of the above Adimen-SUMO statements that restrict the set of potential SUMO individuals related to a synset, the second interpretation of the mapping information is completed according to the mapping relation that links the synset and the SUMO concept:

- If the synset is connected using *equivalence* (resp. the negation of *equivalence*), then we can assume that the synset is related to all (resp. is not related to any of) the potential SUMO individuals that satisfy the Adimen-SUMO statement proposed above. For this purpose, the variable associated to the given synset is considered to be universally quantified.

⁵It is worth to recall that $subAttribute_r$ relates SUMO individual attributes, which are instance of $Attribute_c$, while $\$subclass$ relates SUMO classes, which are instance of $Class_c$.

⁶The linkings to SUMO relations are discarded.

- Otherwise —the synset is connected using *subsumption* (resp. the negation of *subsumption*) or *instance*—, we can only assume that the synset is related to (resp. is not related to) some of the potential SUMO individuals the Adimen-SUMO statement proposed above. This means that the variable associated to the given synset is considered to be existentially quantified.

This second interpretation of the mapping information takes advantage from the fact that most of the SUMO knowledge is based on the notion of *individuals* and that only a few of SUMO predicates provide information at the level of *classes*. From this point of view, this interpretation enables a more precise use of the knowledge of SUMO. In addition, the problem with synsets connected to several SUMO concepts is overcome. Going back to the example about $male_horse_n^1$, its mapping to $Male_a+$ and $Horse_c+$ can be translated as

(and (3)
(attribute ?S Male)
(\$instance ?S Horse))

where its associated variable $?S$ stands for all the SUMO individuals that are related to $male_horse_n^1$.

6 Competency Questions Based on Meronymy

In this section, we describe the set of CQs that is created on the basis of the part-whole data provided by WordNet.

For this purpose, we consider the second interpretation of the mapping information introduced in Section 5. Since that interpretation does not distinguish between *subsumption* and *instance*, we only consider two linking options for WordNet synsets: synsets connected by *equivalence* (or its negation) and synsets connected by (the negation of) *subsumption* or *instance*. Therefore, there are just 4 possible combinations of mapping relations in the 12,293 ordered synset pairs provided by WordNet and we propose a different question pattern for each of them.

Given an ordered synset pair, the corresponding question pattern is instantiated according to a) the WordNet meronymy relation and b) the SUMO concepts to which synsets are connected.

With respect to WordNet meronymy relations, we have inspected SUMO in order to find the relations that are synonym or semantically similar to

```

(exists (?X, ?Y)
  (and
    < s_part, ?X >
    < s_whole, ?Y >
    (< SUMO predicate > ?X ?Y)))

```

Figure 1: First question pattern for $\langle s_part, s_whole \rangle$ meronymy pairs

them. In SUMO, the main meronymy relation is $part_r$, and we can find 30 different subrelations of $part_r$ in its core. Among them, we have selected the SUMO predicates $part_r$, $member_r$, $piece_r$ as counterpart of the WordNet relations $part$, $member$ and $substance$ respectively. As for every SUMO relation, SUMO provides *domain* axioms that restrict the set of SUMO individuals that can be related by the above predicates as follows:

- $part_r$ relates pairs of $Object_c$ individuals.
- $member_r$ relates $SelfConnectedObject_c$ individuals (first argument) to $Collection_c$ individuals (second argument).
- $piece_r$ relates pairs of $Substance_c$ individuals.

Additionally, SUMO also defines several incompatibilities between SUMO individuals. Among others, individuals of $CorpuscularObject_c$ are not compatible with neither $Collection_c$ nor $Substance_c$ because $CorpuscularObject_c$ and $Collection_c$ (also $Substance_c$) are defined as disjoint classes.

On the basis of individual SUMO incompatibilities, we can already detect some errors. For example, the synsets $grape_n^1$ and $wine_n^1$ are related by $substance$ (as introduced in Section 2) and respectively connected $FruitOrVegetable_c+$ and $Wine_c=$. In SUMO, $FruitOrVegetable_c$ is defined to be subclass of $CorpuscularObject_c$. Consequently, $FruitOrVegetable_c$ is incompatible with $Substance_c$, which prevents the use of $piece_r$ for relating synsets pairs with individuals of $FruitOrVegetable_c$ in the first place. The source of this error is discussed in Section 7.

After choosing the most suitable SUMO predicate for a given synset pair, the instantiation of the corresponding question pattern is finished according to the SUMO concepts to which synsets are

```

(forall (?X)
  (=>
    < s_part, ?X >
    (exists (?Y)
      (and
        < s_whole, ?Y >
        (< SUMO predicate > ?X ?Y))))))

```

Figure 2: Second question pattern for $\langle s_part, s_whole \rangle$ meronymy pairs

connected. More specifically, we apply the second interpretation of the mapping information in order to obtain a Adimen-SUMO statement for each synset. The resulting Adimen-SUMO statements are directly used for the instantiation of question patterns.

In the next subsections, we describe the proposed question patterns.

6.1 First Question Pattern

The first question pattern is designed for its application to meronymy pairs where both synsets are connected using (the negation of) $subsumption$ or $instance$.

In Figure 1, we describe the combination of the selected SUMO predicate and the statements that are obtained by following the second interpretation of the mapping information introduced in Section 5. In that combination, the variables associated to both synsets are considered to be existentially quantified.

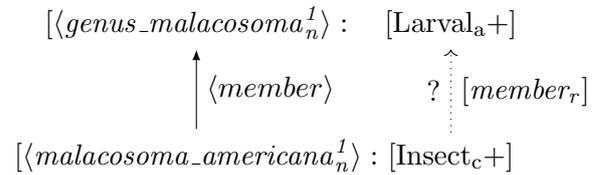


Figure 3: $malacosoma_americana_n^1$ and $genus_malacosoma_n^1$.

Next, we illustrate the instantiation of the resulting question pattern by considering again the synsets $malacosoma_americana_n^1$ and $genus_malacosoma_n^1$, which are related by $member$ and connected to $Insect_c+$ and $Larval_a+$ respectively as described in Figure 3. The combi-

nation of the SUMO statements (1,2) that result from their mapping information with the SUMO predicate $member_r$ yields the following CQ:

```
(exists (?X, ?Y)                                     (4)
  (and
    ($instance ?X Insect)
    (attribute ?Y Larval)
    (member ?X ?Y)))
```

6.2 Second Question Pattern

The second question pattern is designed for meronymy synset pairs $\langle s_part, s_whole \rangle$ where s_part is connected by (the negation of) *equivalence* and s_whole is connected by (the negation of) *subsumption* or *instance*.

In this case, the variable associated to s_whole is considered to be universally quantified, while the variable associated to s_part is considered to be existentially quantified. The resulting question pattern is described in Figure 2.

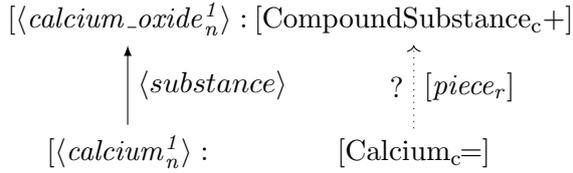


Figure 4: $calcium_n^1$ and $calcium_oxide_n^1$.

In order to illustrate the instantiation of this second question pattern, we consider the synset pair $substance(calcium_n^1, calcium_oxide_n^1)$, where the involved synsets are respectively connected to $Calcium_{c=}$ and $CompoundSubstance_{c+}$ as described in Figure 4. On the basis of the above mapping information, the selected SUMO predicate is $piece_r$ and we obtain the following CQ:

```
(forall (?X)                                         (5)
  (=>
    ($instance ?X Calcium)
    (exists (?Y)
      (and
        ($instance ?Y CompoundSubstance)
        (piece ?X ?Y))))))
```

6.3 Third Question Pattern

The third question pattern is the dual of the second one because it is designed for meronymy synset pairs $\langle s_part, s_whole \rangle$ where s_part is connected by (the negation of) *subsumption* or *instance*, and s_whole is connected by (the negation of) *equivalence*.

Consequently, the variables associated to s_whole and s_part are considered to be universally and existentially quantified respectively.

This third question pattern is applied to synset pairs like $member(committee_n^1, committee_member_n^1)$, where synsets are respectively connected to $Human_{c+}$ and $Commission_{c=}$. By using the SUMO predicate $member_r$, the resulting CQ is:

```
(forall (?Y)                                         (6)
  (=>
    ($instance ?Y Commission)
    (exists (?X)
      (and
        ($instance ?X Human)
        (member ?X ?Y))))))
```

6.4 Fourth Question Pattern

The last question pattern is designed for its application to meronymy pairs where both synsets are connected using (the negation of) *equivalence*.

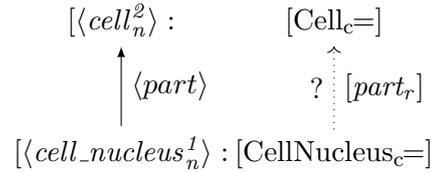


Figure 5: $cell_n^2$ and $cell_nucleus_n^1$.

In this case, the question pattern is obtained by the conjunction of the second and the third question patterns. In order to illustrate its application, we consider the synset pair $part(cell_n^2, cell_nucleus_n^1)$, where synsets are respectively connected to $Cell_{c=}$ and $CellNucleus_{c=}$ as described in Figure 5. The resulting CQ is:

```
(and                                                 (7)
  (forall (?X)
    (=>
      ($instance ?X CellNucleus)
      (exists (?Y)
        (and
          ($instance ?Y Cell)
          (part ?X ?Y))))))
  (forall (?Y)
    (=>
      ($instance ?Y Cell)
      (exists (?X)
        (and
          ($instance ?X CellNucleus)
          (part ?X ?Y))))))
```

7 Discussion

In this section, we discuss the results obtained from our ongoing validation of WordNet and SUMO by applying the evaluation framework proposed in Álvez et al. (2017).

In Table 2, we report on some figures about the instantiation of the 4 question patterns introduced in the above section using the 22,187 meronymy pairs provided by WordNet. The information is organized in 11 columns as follows: according to the different WordNet meronymy relations (first column), we first provide the total amount of synset pairs (second column) and the number of synset-pairs which do not satisfy SUMO domain restrictions (third column); in the remaining 8 columns, we respectively provide the amount of synset pairs (Pairs columns) that have been applied to each question pattern and the number of resulting competency questions (CQs columns). To sum up, we have obtained 2,137 different CQs—1,418 + 447 + 197 + 75 CQs— from 7,674 synset pairs, while 14,513 pairs have not been used due to SUMO incompatibilities. Most of those synset pairs (11,920) are related by *member*, which relates *SelfConnectedObject_c* individuals (first argument) to *Collection_c* individuals (second argument).⁷ By a manual inspection, we discover that the source of the problem in more than 8,000 pairs is the same: pairs where both synsets are connected to the same concept although the first synset denotes an individual organism and the second one the species, genus or family to which the organism belongs. For example, *bear_n¹* and *Ursidae_n¹* are both connected to *Mammal_c⁺*, which is subclass of *SelfConnectedObject_c*. In those cases, we decide that the mapping is not consistent because it does not correctly characterize the knowledge of WordNet in terms of SUMO: *Ursidae_n¹* does not refer to any particular mammal, but to a group of mammals.

Another divergence between the knowledge of WordNet and SUMO that can be detected by means of SUMO incompatibilities is given by the pair *substance(grape_n¹, wine_n¹)*, as described in Section 6. In this case, the WordNet pair is not complete, since *grape-juice_n¹* is neither related to *grape_n¹* nor *wine_n¹*.

Regarding our preliminary experimental results using ATPs, we have already checked that the pro-

posed CQs enable to validate some pieces of the information of WordNet, SUMO and their mapping, and also to detect some conflicts. For example, the following CQ

$$\begin{aligned}
 &(\text{forall } (?Y) && (8) \\
 &(\Rightarrow \\
 &(\text{attribute } ?Y \text{ PoliceOfficer}) \\
 &(\text{exists } (?X) \\
 &(\text{and} \\
 &(\$instance ?X \text{ PoliceOrganization}) \\
 &(\text{member } ?X ?Y))))
 \end{aligned}$$

is obtained from the synset pair *member(police_officer_n¹, police_force_n¹)* by applying the third question pattern, since *police_officer_n¹* is connected to *PoliceOfficer_a=* and *police_force_n¹* is connected to *PoliceOrganization_c⁺*. ATPs are able to prove conjecture (8), consequently both the WordNet meronymy pair, the mapping of the related synsets and the involved SUMO information are validated. On the contrary, ATPs do not find any proof for conjecture (6) or its negation. This fact leads us to discover that SUMO lacks from information conveniently relating the concepts of *Human_c* and *Commission_c* by *member_r*.

In the rest of this section, we proceed to illustrate three different kinds of discrepancies or disagreements that can be detected by the application of ATPs to the proposed CQs as described in Álvez et al. (2017).

In the first place, the use of ATPs enables to detect additional inconsistencies in the mapping between WordNet and SUMO. For example, ATPs are able to prove the negation of conjecture (4), which reveals the existence of a problem with the synsets *malacosoma_american_n¹* and *genus_malacosoma_n¹*. More specifically, the mapping of *genus_malacosoma_n¹* to *Larval_a⁺* is not suitable.

Secondly, our proposal enables to detect conflicts which are due to the knowledge represented in SUMO. For example, the negation of conjecture (5) is proven by ATPs. By inspecting the proof, we discover that the problem is related to the following SUMO axiom (described in Adimen-SUMO syntax):

$$\begin{aligned}
 &(\Rightarrow && (9) \\
 &(\text{piece } ?SUBSTANCE1 ?SUBSTANCE2) \\
 &(\text{forall } (?CLASS) \\
 &(\Rightarrow \\
 &(\$instance ?SUBSTANCE1 ?CLASS) \\
 &(\$instance ?SUBSTANCE2 ?CLASS))))
 \end{aligned}$$

⁷It is worth to recall that *SelfConnectedObject_c* and *Collection_c* are disjoint classes.

Meronymy relations	Pairs		1 st QP		2 nd QP		3 rd QP		4 th QP	
	Total	Error	Pairs	CQs	Pairs	CQs	Pairs	CQs	Pairs	CQs
<i>part</i>	9,097	2,221	5,974	1,252	725	430	116	104	61	59
<i>member</i>	12,293	11,920	348	78	14	14	10	7	1	1
<i>substance</i>	797	372	248	83	152	89	10	10	15	15
Total	22,187	14,513	6,570	1,418	745	447	282	197	77	75

Table 2: Instantiation of question patterns

In particular, $Calcium_c$ is subclass of $ElementalSubstance_c$, which is disjoint with $CompoundSubstance_c$. Therefore, no individual of $CompoundSubstance_c$ can inherit the property of being instance of $Calcium_c$.

Finally, we can also detect inconsistencies which are related to WordNet meronymy pairs. For example, ATPs are able to prove the negation of conjecture (7), thus revealing a problem related to the pair $part(cell_n^2, cell_nucleus_n^1)$. More specifically, that pair is incompatible with the fact that some cells lack a nucleus, as stated by the following SUMO axiom (described in Adimen-SUMO syntax):

$$\begin{aligned}
 (= > & \hspace{15em} (10) \\
 (\$instance ?C RedBloodCell) & \\
 (not & \\
 (exists (?N) & \\
 (and & \\
 (\$instance ?N CellNucleus) & \\
 (part ?N ?C)))) &
 \end{aligned}$$

Consequently, the synset pair $part(cell_n^2, cell_nucleus_n^1)$ is not consistent.

8 Ongoing Work

Currently, we are finishing our experimental evaluation of WordNet, SUMO and their mapping by applying the methodology proposed in Álvarez et al. (2017). For this purpose, we are using the ATPs Vampire (Kovács and Voronkov, 2013) and E (Schulz, 2002) for checking whether the conjectures resulting from the set of CQs proposed in this paper are entailed or not by Adimen-SUMO. All the resources—the ontology, the set of CQs and conjectures, and the resulting execution reports—will be available at <http://adimen.si.ehu.es/web/AdimenSUMO>.

By analysing our preliminary experimentation results, we can conclude that our proposal enables a sophisticated cross-checking of the information

in WordNet, SUMO and their mapping. In particular, by means of practical examples, we have illustrated that the proposed system enables (a) the validation of some pieces of information and (b) the detection of missing information and inconsistencies. Further, our preliminary experimental results also demonstrate the suitability of the involved resources for its application to practical tasks related to natural language processing.

9 Concluding Remarks

In this work, we enlarge the set of CQs proposed in Álvarez et al. (2017) by means of part-whole data of WordNet, which illustrates the fact that our proposal can be generally applied to any data extracted from WordNet. Nowadays, our complete set of CQs includes around 3,000 CQs obtained from antonymy and around 2,000 CQs obtained from *Morphosemantic Links* database of WordNet. In the last case, we exclusively concentrate on the relations *event*, *agent*, *instrument* and *result*. In the next future, we plan to extend our benchmark by considering additional WordNet relations.

Acknowledgments

This work has been partially funded by the Spanish Projects TUNER (TIN2015-65308-C5-1-R) and COMMAS (TIN2013-46181-C2-2-R), the Basque Project LoRea (GIU15/30), the UPV/EHU project OEBU (EHUA16/33) and grant BAILab (UFI11/45).

References

- J. Álvarez, J. Atserias, J. Carrera, S. Climent, E. Larraza, A. Oliver, and G. Rigau. 2008. Complete and consistent annotation of WordNet using the Top Concept Ontology. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proc. of the 6th Int. Conf. on Language Resources and Evaluation (LREC 2008)*, pages 1529–

1534. European Language Resources Association (ELRA), may.
- J. Álvarez, P. Lucio, and G. Rigau. 2012. Adimen-SUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semantic Web Inf. Syst.*, 8(4):80–116.
- J. Álvarez, P. Lucio, and G. Rigau. 2017. Black-box testing of first-order logic ontologies using WordNet. *CoRR*, abs/1705.10217.
- J. Daudé, L. Padró, and G. Rigau. 2003. Making WordNet mapping robust. *Procesamiento del lenguaje natural*, 31.
- C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- M. Grüninger and M. S. Fox. 1995. Methodology for the design and evaluation of ontologies. In *Proc. of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995)*.
- I. Horrocks and A. Voronkov. 2006. Reasoning support for expressive ontology languages using a theorem prover. In J. Dix et al., editor, *Foundations of Information and Knowledge Systems*, LNCS 3861, pages 201–218. Springer.
- L. Kovács and A. Voronkov. 2013. First-order theorem proving and Vampire. In N. Sharygina and H. Veith, editors, *Computer Aided Verification*, LNCS 8044, pages 1–35. Springer.
- I. Niles and A. Pease. 2001. Towards a standard upper ontology. In Guarino N. et al., editor, *Proc. of the 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM.
- I. Niles and A. Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In H. R. Arabnia, editor, *Proc. of the IEEE Int. Conf. on Inf. and Knowledge Engin. (IKE 2003)*, volume 2, pages 412–416. CSREA Press.
- T. Richens. 2008. Anomalies in the WordNet verb hierarchy. In *Proc. of the 22nd Int. Conf. on Computational Linguistics-Volume 1*, pages 729–736. Association for Computational Linguistics.
- H. Rodríguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, and A. Roventini. 1998. The top-down strategy for building EuroWordNet: Vocabulary coverage, base concepts and top ontology. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 45–80. Springer.
- S. Schulz. 2002. E - A brainiac theorem prover. *AI Communications*, 15(2-3):111–126.

Comparing Two Thesaurus Representations for Russian

Natalia Loukachevitch
Lomonosov Moscow State
University, Moscow, Russia
Tatarstan Academy of Sciences,
Kazan, Russia
louk_nat@mail.ru

German Lashevich
Kazan Federal University
Kazan, Russia
design.ber@gmail.com

Boris Dobrov
Lomonosov Moscow State
University, Moscow, Russia
dobrov_bv@mail.ru

Abstract

In the paper we presented a new Russian wordnet, RuWordNet, which was semi-automatically obtained by transformation of the existing Russian thesaurus RuThes. At the first step, the basic structure of wordnets was reproduced: synsets' hierarchy for each part of speech and the basic set of relations between synsets (hyponym-hypernym, part-whole, antonyms). At the second stage, we added causation, entailment and domain relations between synsets. Also derivation relations were established for single words and the component structure for phrases included in RuWordNet. The described procedure of transformation highlights the specific features of each type of thesaurus representations.

1 Introduction

WordNet thesaurus is one of the popular language resources for natural language processing (Fellbaum, 1998). The projects for creating WordNet-like resources have been initiated for many languages in the world (Vossen, 1998; Bond and Paik, 2012). Other thesaurus models are rarely discussed, created and used in NLP.

In several works, S.Szapkowicz and co-authors (Jarmasz and Szpakowicz, 2004; Aman and Szpakowicz, 2008; Kennedy and Szpakowicz, 2008) evaluated two versions of Roget's thesaurus in several applications. Borin and colleagues (Borin and Forsberg, 2009; Borin et al. 2013) compared the structure of the Swedish thesaurus Saldo with the WordNet structure. In (Borin et al., 2014) automatic generation of Swedish Roget's thesaurus and its comparing

with the existing Roget-style thesaurus for Swedish is discussed.

For the Russian language, RuThes thesaurus has been created more than fifteen years ago (Loukachevitch and Dobrov, 2002). It was utilized in various information-retrieval and NLP applications (Loukachevitch and Dobrov, 2014). RuThes was successfully evaluated in text summarization (Mani et al., 2002), text clustering (Dobrov and Pavlov, 2010), text categorization (Loukachevitch and Dobrov, 2015), detecting Russian paraphrases (Loukachevitch et al., 2017), etc.

Using the RuThes model for the concept representation, several domain-specific thesauri have been created for NLP and domain-specific information-retrieval applications including Sociopolitical thesaurus (Loukachevitch and Dobrov, 2015), Ontology on Natural Sciences and Technology (Dobrov and Loukachevitch, 2006), Banking thesaurus (Nokel and Loukachevitch, 2016) and others. Currently, RuThes concepts provide a basis for creating the Tatar Socio-Political Thesaurus (Galieva et al., 2017).

In 2013, RuThes was partially published for non-commercial use (Loukachevitch et al., 2014). But people would like to have a large Russian wordnet. Therefore, we have initiated a transforming procedure from the published version of RuThes (RuThes-lite) to the largest Russian WordNet (RuWordNet¹), which we describe in this paper. This transformation allows us to show similarities and differences between two resources in a detailed way. RuWordNet currently includes 115 thousand unique words and phrases.

¹ <http://ruwordnet.ru/en/>

The structure of this paper is as follows. In Section 2, we describe related work. Section 3 presents the structure of RuThes thesaurus, including the set of relations and principles of work with multiword expressions. Section 4 describes the main stages for creating the basic structure of RuWordNet. Section 5 is devoted to enrichment of the basic RuWordNet relations.

2 Related work

Creating large lexical resources like WordNet from scratch is a complex task, which requires effort for many years (Azarowa, 2008). To speed up the development of a wordnet for own language, the first version of such a resource can be created by automatically translating Princeton WordNet into the target language (Vossen, 1998; Gelfenbein et al., 2003; Sukhonogov et al. 2005), but then considerable effort is required to proof-read and correct the obtained translation.

As an intermediate approach, researchers propose a two-stage creation of a wordnet for a new language: first translating and transferring the relations of the top concepts of Princeton WordNet (the so-called core WordNet), and then manually replenishing hierarchies based on dictionaries and text corpora. This approach was used in the creation of such resources as DanNet (Pedersen, 2010) and EuroWordNet (Vossen, 1998).

After analyzing the existing approaches to the development of wordnets, the creators of the Finnish wordnet (FiWN) decided to translate Princeton WordNet manually, using the work of professional translators. As a result, the Finnish wordnet was created on the basis of translation of more than 200 thousand word senses of Princeton WordNet words within 100 days (Lindén and Niemi, 2014).

In work (Braslavsky et al., 2012), it was proposed to develop a new Russian wordnet (YARN) using the Russian Wiktionary and crowdsourcing. The authors planned to attract a large number of students and interested people to create a new resource.

There are at least four known projects for creating a wordnet for the Russian language. In RussNet (Azarova et al., 2004), the authors planned to create the Russian wordnet from scratch, guided by the principles of Princeton WordNet. In two different projects described in (Gelfenbein et al., 2003; Sukhonogov et al. 2005), attempts were made to automatically translate WordNet into Russian, with all the orig-

inal thesaurus structure preserved. The results of (Gelfenbein et al., 2003) are published, but the analysis of the thesaurus generated in this way shows that it requires considerable editing or the use of better algorithms.

The last project YARN (Yet Another Russian wordNet) was initiated in 2012 and initially was created on the basis of crowdsourcing, i.e. participation in the work of filling the thesaurus by a large number of participants. Currently, YARN contains a significant number of synsets with a small number of relationships between them. The published version² of the YARN thesaurus contains too many similar or partially similar synsets.

In (Azarova et al., 2016), the authors describe the project on the integration of the thesaurus RussNet (Azarowa., 2008) and the thesaurus YARN (Braslavsky et al., 2012) into a single linguistic resource, where the expert approach and the crowdsourcing will be combined.

In (Khodak et al., 2017), a new approach to automatic wordnet construction is presented and tested on a specially prepared Russian dataset comprising senses of 600 words (200 nouns, 200 verbs, and 200 adjectives). The approach is based on translation of English synsets, and a number of techniques of clustering and assessing the obtained translation. For Russian, the authors report 60% F-measure on the above-mentioned tests. However, the analysis of the dataset showed that the presented Russian words have much more senses than it is usually presented in Russian dictionaries. For example, word *опасность* (*danger*) is usually described as having 2 senses. But in the dataset it has 6 senses. Word *оборудование* (*equipment*) is usually described with 2 senses, but in the dataset it has 8 senses. It looks that the expert labeling of Russian senses for the dataset was somehow biased to English and its representation in Princeton WordNet.

3 RuThes Structure and Relations

RuThes (Loukachevitch and Dobrov, 2014; Loukachevitch et al., 2014) and WordNet are both thesauri, i.e. lexical resources in that words similar in meaning are gathered into synsets (WordNet) or concepts (RuThes), between which relations are established. When applying the two thesauri to text processing, similar steps should be carried out, including a comparison of the text

² <https://russianword.net/>

with the thesaurus, and the use of the described relations if necessary. There are also significant differences between the thesauri.

Firstly, in RuThes there is no division into lexical networks by parts of speech. Any part of speech can be associated with the same RuThes concept, if they mean the same (so-called part-of-speech synonyms). Each thesaurus concept has a unique name.

To provide morpho-syntactic information for a word, each RuThes entry has parts of speech labels. The morpho-syntactic representation of a multiword expression contains the syntactical type of the whole group, the head word, parts of speech and lemmatized forms for each component word.

Therefore, secondly, when establishing relations in RuThes, it is often impossible to apply synonym tests based on the interchangeability of words in different contexts (Miller, 1998). Instead, tests are used to detect the denotative similarity of word meanings, for example, "if the entity X in different situations can be called W_1 , can it always be called W_2 ", and vice versa.

Thus, because of the above-mentioned differences (denotative tests, unique names of concepts), RuThes is closer to ontologies on an imaginary scale from lexical resources to formal ontologies than WordNet-like thesauri (Loukachevitch and Dobrov, 2014).

3.1 Relations in RuThes.

Different models of the knowledge description presuppose different sets of relations.

In RuThes, the relations are established only between concepts. The main class-subclass relation roughly corresponds to the relation of hyponym-hypernym in WordNet (Miller, 1998).

Also, RuThes has the part-whole relationship, but unlike WordNet, it is only established when the part always (or at least in the vast majority of cases) refers to the specified whole, i.e. cannot belong to a number of alternative wholes. This makes it possible to use the transitivity of the part-whole relations with greater reliability (Loukachevitch, Dobrov, 2014). There are some techniques allowing representation of part-whole relations in other cases.

When the above-mentioned conditions for establishing the part-whole relationship are imposed, a fairly broad interpretation of the part-whole relationship is adopted in RuThes:

- between physical objects (*storey – building*);

- between regions (*Europe – Eurasia*);
- between substances;
- between sets (*battalion – company*);
- between parts of the text (*strophe – poem*);
- between processes (*production cycle – industrial manufacturing*).

Also, the part-whole relations are established for connections between entities, one of which is internal, dependent on another (Guarino, 2009) such as: characteristics of an entity (*displacement – ship*); role in the process (*investor – investment*); participant in the field of activity is the sphere of activity (*industrial plant – industry*).

In addition, one of the main relations in RuThes is the relation of ontological dependence, which shows the dependence of the existence of one concept on another. An example of such an attitude is the relationship between the concepts *Tree – Forest*, where *Forest* is a dependent concept requiring the existence of the *Tree* concept.

The relation of the ontological dependence is denoted as directed association $asc_1 – asc_2$. In fact, this directed association represents a more formalized form of the association relations in traditional information-retrieval thesauri (Z39.19, 2005). Symmetric associations are also possible in only restricted number of cases.

Thus, the structure and the set of relations in the thesaurus RuThes are significantly different from the structure and relations of WordNet. It is also important to stress the differences in the properties of the relationships in the thesauri WordNet and RuThes. In WordNet, basically, only the transitivity of hyponym-hypernym relations is used. In RuThes, in addition to the transitivity of the class-subclass relationship, the following relations are also postulated:

- transitivity of the part-whole relations:

$$whole(c_1, c_2) \wedge whole(c_2, c_3) \rightarrow whole(c_1, c_3);$$
- inheritance of the whole relationship to subclasses:

$$class(c_1, c_2) \wedge whole(c_2, c_3) \rightarrow whole(c_1, c_3);$$
- inheritance of dependence association relations and symmetric association relations on types and parts:

$$\text{class}(c_1, c_2) \wedge \text{asc}_1(c_2, c_3) \rightarrow \text{asc}_1(c_1, c_3);$$
$$\text{class}(c_1, c_2) \wedge \text{asc}(c_2, c_3) \rightarrow \text{asc}(c_1, c_3);$$
$$\text{whole}(c_1, c_2) \wedge \text{asc}_1(c_2, c_3) \\ \rightarrow \text{asc}_1(c_1, c_3);$$
$$\text{whole}(c_1, c_2) \wedge \text{asc}(c_2, c_3) \rightarrow \text{asc}(c_1, c_3)$$

Considering all possible relation paths existing between two thesaurus concepts C_1 and C_2 , it was supposed that those paths that can be reduced to a single relation with the application of the above-mentioned rules of transitivity and inheritance indicate semantic relatedness between concepts C_1 and C_2 , so called semantic paths. Word and phrases presented as thesaurus entries assigned to the concepts C_1 and C_2 are also considered semantically related even if the length of the path is quite large (five and more relations). Such defined semantic similarity between words and phrases included in RuThes is used for query expansion in information retrieval, thematic text representation (Loukachevitch and Alekseev, 2014), representation of categories in knowledge-based text categorization (Loukachevitch and Dobrov, 2015), and automatic word sense disambiguation.

The properties of the RuThes relations and defined paths were used to infer some types of relationships for RuWordNet.

3.2 Multiword Expressions in RuThes

Another issue, which is important in transformation of data from RuThes to RuWordNet, is the representation of multiword expressions (Loukachevitch and Lashevich, 2016).

The distinctive feature of RuThes is that it contains many multiword expressions. Experts are recommended to introduce new multiword expressions into RuThes if they can substantiate their decision with the necessity to represent the expression in the thesaurus. The expert should show that adding the expression to the thesaurus gives useful information that does not follow from the component structure of this expression. Such information is usually expressed in form of additional thesaurus relations (or their deliberate exclusion), which enriches the thesaurus knowledge.

In fact, we shift the often discussed question on compositionality vs. non-compositionality of a multiword expression to the more visible question of adding information to a thesaurus. The employed principles of introducing multiword expressions into RuThes can be subdivided as follows:

- absence of meaningful relations between an expression and senses of component words (idioms),
- synonym to own component word or its derivative (multisynonyms),
- additional relationships to other single words and multiword expressions.

In RuThes, multiword expressions that are synonymous its own component or its derivative are specially collected. The examples of such expressions include *политическая партия* (*political party*) – *партия* (*party*), the phrase is quite frequent in Russian as well as its translation in English. Another example is *компьютерная программа* (*computer program*) – *программа* (*program*). The example of a multisynonym to the component derivative is: *участвовать* (*participate*) – *принимать участие* (*take participation*).

In creating RuThes, the introduction of such multiword synonyms was especially encouraged, because the important feature of these expressions is that their components can be ambiguous, but the whole expression is often unambiguous. Thus, if the expression is known and described in a thesaurus there are no problems with disambiguation of its components and with the semantic interpretation of the whole expression. In fact, these expressions can improve the recognition of their own concepts.

In addition, the inclusion of such expressions in a synset often clarifies the sense of the synset. It is clear that introduction of these expressions does not require additional concepts.

Such multisynonyms are very common in the Russian language. Currently, the published version of RuThes – RuThes 2.0 (Loukachevitch et al., 2014) contains more than 13 thousand multiword synonyms.

Numerous examples of multisynonyms can be found also in English and can be met in WordNet. For example, *plant* – *industrial plant*, *platform* – *political platform*, *park* – *car park* – *parking lot*. But in RuThes, multisynonyms were specially searched and added.

RuThes also includes multiword expressions with so called *relational idiosyncrasy*, that is multiword expressions that look like compositional ones but they have specificity in relations with other single words and/or expressions, which usually means that these expressions denote some important concepts, entities or situations (Loukachevitch and Gerasimova, 2017).

For example, such phrase as *дорожное движение* (*road traffic*) seems to be compositional one, but it has hyponyms: *левостороннее движение* (*left-hand traffic*) and *правостороннее движение* (*right-hand traffic*): the existence of such hyponyms cannot be inferred from its component words.

Currently, all multiword expressions (54 thousand of 115 thousand entries) of RuThes-lite were transferred to RuWordNet. In such a way, it is possible to say that RuWordNet contains the maximal share of phrases in synsets among other WordNet-like resources. It means that the representation of phrases in RuWordNet requires special attention.

4 Creating Basic Structure of RuWordNet

In our opinion, one of the most distinctive features of WordNet-like resources is their division into synset nets according to parts of speech. Therefore, all text entries of RuThes-lite 2.0 were subdivided into three parts of speech: nouns (single nouns, noun groups, or preposition groups), verbs (single verbs and verb groups), adjectives (single adjectives and adjective groups). We have obtained 29,297 noun synsets, 12,865 adjective synsets, and 7,636 verb synsets (Table 1).

This subdivision was based on the morpho-syntactic representation of RuThes-lite 2.0 text entries, which was fulfilled semi-automatically. Therefore, a small number of mistakes because of particle treatment (verbs or adjectives) or nominalized adjectives can appear. For example, Russian phrase *любитель подраться* (=драчун) (*brawler, scrapper*) was treated in this procedure as a verb group and was assigned to the verb synsets. Currently all found mistakes are corrected.

Part of speech	Number of synsets	Number of unique entries	Number of senses
Noun	29,296	68,695	77,153
Verb	7,634	26,356	35,067
Adj.	12,864	15,191	18,195

Table 1. Quantitative characteristics of synsets and entries in RuWordNet

The divided synsets were linked to each other with the relation of part-of-speech synonymy.

The hyponym-hypernym relations were established between synsets of the same part of speech. These relations include direct hyponym-

hypernym relations from RuThes-lite 2.0. In addition, the transitivity property of hyponym-hypernym relations was employed in cases when a specific synset did not contain a specific part of speech but its parent and child had text entries of this part of speech. In such cases, the hypernymy-hyponymy relation was established between the child and the parent of this synset.

Similar to the current version of Princeton WordNet, in RuWordNet class-instance relations are also established. By now, they had been generated semi-automatically for geographical objects.

The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources. Now RuWordNet contains 3.5 thousand part-whole relations. The part-whole relations include the following subtypes:

- functional parts (*nostrils – nose*),
- ingredients (*additives – substance*),
- geographic parts (*Seville – Andalusia*),
- members (*monk – monastery*),
- dwellers (*Moscow citizen – Moscow*),
- temporal parts (*gambit – chess party*)
- inclusion of processes, activities (*industrial production – industrial cycle*)

Adjectives in RuWordNet similarly to German or Polish wordnets (Gross and Miller, 1990; Maziarz et al., 2012; Kunze and Lemnitzer, 2010) are connected with hyponym-hypernym relations. For example, word *цветовой* (*colored*) is linked to such hyponyms as *красный* (*red*), *синий* (*blue*), *зеленый* (*green*), etc.

Part of speech	Hyper-nyms	Inst-ance	Holo-nyms	POS-syn.	Ant-o-nyms
Noun	39,155	1863	10,010	18,179	454
Verb	10,304	0	0	7,143	20
Adj.	16,423	0	0	13,794	456

Table 2. Quantitative characteristics of basic relations in RuWordNet

Adjectives often have POS-synonymy links to nouns, but also can have POS-synonyms to verb synsets. For example, word *строительный* (*building* as an adjective) has two POS-synonymy relations: to the noun synset {*стройка, постройка, возведение*,

сооружение..} (*building* as a noun) and to the verb synset {*строить, построить, возводить ...*} (*to build*).

Antonymy relations are conceptual relations in RuWordNet, that means they link synsets, not single lexemes. They are introduced for all parts of speech, mainly for synsets denoting properties and states, for example:

- noun synset {*легкость, с легкостью, без труда, без затруднений*} (*easiness*) is antonymous to synset {*тяжесть, трудность*} (*difficulty*),
- adjective synset {*легкий, легкий для выполнения, легкий для осуществления, нетрудный*} (*easy*) is antonymous to synset {*тяжелый, трудный, тяжелый, трудный для выполнения, нелегкий ...*} (*difficult*),
- verb synset {*не соответствовать действительности*} (*to be contrary to the fact*) is antonymous to synset {*соответствовать истине, соответствовать действительности*} (*to be in accordance with the truth*).

The current numbers of basic relations described in RuWordNet are presented in Table 2.

5 Enrichment of Basic Relations of RuWordNet

Basic relations in the RuWordNet thesaurus were supplemented by several types of relations, including the relations of causation and entailment, the domain relation, the relations of word derivation and the relations between phrases and their components.

5.1 Causation and entailment

The relationships of entailment and causation were treated in the same way as in WordNet. The WordNet entailment relation is a relation between two verbs V_1 and V_2 that holds when the sentence "*Someone V₁*" logically entails "*Someone V₂*" and there is the temporal inclusion of event V_1 into V_2 or vice versa (Fellbaum, 1998). The causation relation can be also considered as a subtype of a general logical entailment relation but there is not temporal inclusion between corresponding situations (Fellbaum, 1998).

To automate the introduction of the relations of causation and entailment into RuWordNet, the RuThes directed associations between concepts containing verbs were extracted. This relation means in this case that the emergence of one sit-

uation (process, action) somehow requires the emergence of another situation (process, action).

The prepared lists of relations between verbs were checked out by linguists, resulting in the following relations:

- 97 relations of antonymy, denoting the opposite of what was before, for example, *откупорить (uncork) – закупорить (cork)*,
- 610 relations of causation, for example, *сажать (sit) - сесть (sit down)*. This relation in RuWordNet often connects the synsets corresponding to the reflexive forms of the verbs, for example, the synset *купать, выкупать, докупать, искупать (give a bath)* is the cause of *купаться, выкупаться, искупаться, покупаться (to bathe, cleanse own body)*.
- 943 entailment relationships, for example, the synset *сниться (to dream)* is related by the entailment relation with synset *спать, поспать, почитать (to sleep)* because if someone dreams something, then this someone is sleeping.

5.2 Domain relations

Since relations in such thesauri as WordNet are mostly generic (hyponym-hypernym), there exists a so-called "tennis problem" (Miller, 1998), which is that synsets from the same domain (for example, related to tennis: *tennis player, racket, court*) are very far from each other in the WordNet hierarchy.

To solve this problem in part, a hierarchical system of domains (domains)³ has been proposed, and WordNet synsets were semi-automatically assigned to one or more domains (Magnini, Pianta, 2000; Bentivogli et al., 2004). This domain system is now partially transferred to RuWordNet.

The mechanism of introducing domains for the RuWordNet synsets was as follows. The existing domain system for Princeton WordNet was taken. First, the domain list was refined: the subject areas that were not presented in the RuWordNet thesaurus were removed (i.e. *Heraldry*), and several new domains were added. For example, domain labels corresponding to world religions and some confessions were introduced. Currently, RuWordNet has 156 domains.

The domains labels can be considered as a list of categories for a knowledge-based categoriza-

³ <http://wndomains.fbk.eu/>

tion system. RuThes has a special interface for linking categories with thesaurus concepts and hierarchies.

Each domain was linked to one or more "supporting" concepts of the RuThes thesaurus. Using the RuThes relation properties, the list of supporting concepts was expanded by lower-level concepts (subclasses, parts, associations). This can be done, because in RuThes the relation to the sphere of activity is one of the types of the part-whole relationship, and therefore it is explicitly indicated in the thesaurus.

The generated list of concepts for each domain was looked through and cleaned by experts. Also, for each domain, a noun synonym of RuWordNet was assigned as the domain title.

As a result, a chain of relations has been created:

- (1) RuWordNet synsets,
- (2) Initial concepts of the RuThes thesaurus for these synsets,
- (3) Domain labels presented as categories over RuThes concepts,
- (4) RuWordNet synsets, assigned as a label to each subject domain.

Such a chain makes it possible to introduce direct domain relations between RuWordNet synsets: (1) -> (4).

For example, domain "Art" is described as RuThes concept *Art* with full expansion, which adds to the Art domain all hyponyms, parts, dependent concepts obtained by logical inference using the properties of transitivity and inheritance (Section 3.1). As a result, "Art" concepts comprise more than 700 RuThes concepts, including *Jazzman*, *Piece of painting*, *Harp*, etc. Then RuWordNet synsets originated from these RuThes concepts were also assigned to the Art domain.

5.3 Derivational relations

For RuWordNet, the derivational relations were also introduced (Leseva et al., 2015; Pala and Hlaváčková, 2007, Piasecki, et al, 2012). These relations are lexical, that is established between lexical entries. At the moment, these relations are established for those words that have the same beginning of the word (without prefixes).

The derivation relations were established between words if two conditions were fulfilled:

- the words have the same beginnings,
- these words refer to concepts that either have a direct relationship in the RuThes thesaurus or the relationship can be de-

rived from the properties of transitivity and inheritance established in RuThes.

For example, for the word *аренда* (*lease*), the following words with the same root are indicated: *арендатор* (*lessee*), *арендаторский* (*lessee* as an adjective), *арендователь* (*lessee*), *арендаторша* (*lessee-woman*), *арендный* (*lease* as an adjective), *арендование* (*leasing*), *арендовать* (*to lease*), *арендодатель* (*leaseholder*). Such relations allow us to present semantic relations between words for which there is no other suitable relationships in RuWordNet.

5.4 Relations between phrases and its components

According to the accepted rules for the RuThes thesaurus, experts try to find all possible words and phrases that can express a specific concept (Loukachevitch and Lashevich, 2016). In addition, as described in subsection 3.2, a new concept can be introduced if a phrase carries information that does not follow from the meanings of the word-components of this phrase. For example, RuThes contains the concept *Increase of prices*, which have an important relation to the concept of *Inflation*. Text entries of the concept in RuThes comprise a variety of phrases as: *price growth*, *increase prices*, *price increases*, etc.

This decision in RuThes is supported with the existing system of relations. For example, we can easily describe relations between concepts *Price*, *Increase of prices* and *Inflation* using directed associations.

Type of relation between word and phrase	Number of relations
Phrase and its component are in the same synset (<i>political party – party</i>)	13,367
Pos-synonym relations (<i>participate – take participation</i>)	6,285
Other relations from RuWordNet	16,279
Direct RuThes relations, not included in RuWordNet	15,677
Relations inferred using the RuThes relations properties	12,513

Table 3. Quantitative characteristics of the relationships between phrases and their components in RuWordNet

All these solutions lead to a large number of multiword expressions in RuThes. When RuWordNet has been generated, the phrases were also transferred to it from RuThes. However, the RuWordNet relationship system is different, and for a large number of compositional phrases, the relationship between the phrase and its component words can be lost, which can negatively affect the use of the RuWordNet thesaurus in natural language processing. Therefore, in RuWordNet additional types of relations have been introduced: for the phrase (*has_component*) and for individual words that are phrase components (*component_for*).

These relations were obtained automatically on the basis of direct relations in the thesaurus RuThes, and also on the basis of a logical inference on the relation properties (Section 3.1). Table 3 shows the quantitative results for the established relations between phrases and its components in RuWordNet.

Conclusion

In the paper, we presented a new Russian wordnet, RuWordNet, which was obtained by semi-automatic transformation of the existing Russian thesaurus RuThes. At the first step, the basic structure of wordnets was reproduced: synsets' hierarchies for each part of speech and the basic set of relations between synsets (hyponym-hypernym, part-whole, antonyms).

At the second stage, we added causation, entailment and domain relations between synsets. Also, derivation relations were described for single words and component structure for phrases included in RuWordNet.

It can be seen that RuThes relations are unusual for wordnet-like resources but they give the possibility:

- to introduce a multiword expression into the thesaurus if it gives new information,
- infer domain labels because in RuThes the domain relation is a subtype of the part-whole relation,
- infer derivation relations between lexical entries using the RuThes relation properties.

Acknowledgments

This work is partially supported by Russian Scientific Foundation, according to the research project No. 16-18-020.

References

- Saima Aman and Stan Szpakowicz. 2008. Using Roget's Thesaurus for Fine-grained Emotion Recognition, *Proceedings of IJCNL-2008*: 312-318.
- Irina Azarowa. 2008. RussNet as a Computer Lexicon for Russian, *Proceedings of the Intelligent Information systems IIS-2008*: 341-350.
- Irina Azarova, Pavel Braslavski, Viktor Zakharov, Yuri Kiselev, Dmitrii Ustalov and Maria Khohlova. 2016. Integration of thesauri RussNet и YARN. Proceedings of "Internet and Modern Society" conference IMS-2016 (in Russian).
- Valentina Balkova, Andrey Suhonogov, and Sergey Yablonsky. 2008. Some Issues in the Construction of a Russian WordNet Grid. In *Proceedings of the Forth International WordNet Conference*, Szeged, Hungary:44-55.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, B., and Emanueke Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*: 101-108.
- Francis Bond, and Paik Kyonghee. 2012. A survey of wordnets and their licenses. *Proceedings of Global Wornet Conference GWC-2012*: 64-71.
- Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources*, Odense.
- Lars Borin, Markus Forsberg and Lennart Lonngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191-1211.
- Lars Borin, Jens Allwood, and Gerard de Melo. 2014 .Bring vs. MTRoget: Evaluating automatic thesaurus translation. *Proceedings of LREC-2014*, Reykjavik, ELRA: 2115-2121.
- Pavel Braslavski, Dmitrii Ustalov and Mikhail Mukhin. 2014, A Spinning Wheel for Yarn: User Interface for a Crowdsourced Thesaurus, *Proceedings of EACL-2014*, Gothenberg, Sweden: 101-104.
- Boris Dobrov and Natalia Loukachevitch. 2006. In Development of Linguistic Ontology on Natural Sciences and Technology. In *Proceedings of LREC-2006*.
- Boris Dobrov and Andrey Pavlov. 2010. A Basic line for news clusterization methods evaluation. *Proceedings of RCDL-2010*: 287-295.
- Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

- Alfiya Galieva, Olga Nevzorova and Dilyara Yakubova . 2017. Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project. In *Proceedings of Recent Advances in Natural Language Processing Conference (RANLP-2017)*: 245-252
- Ilia Gelfenbeyn, Artem Goncharuk, Vlad Lehelt, Anton Lipatov, and Viktor Shilo. 2003. Automatic translation of WordNet semantic network to Russian language. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003*.
- Derek Gross and Katherine Miller. 1990. Adjectives in WordNet, *International Journal of Lexicography*, 3(4):.265-277.
- Nicola Guarino. 2009. The Ontological Level: Revisiting 30 Years of Knowledge Representation. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*. Lecture Notes in Computer Science, 5600, Berlin and Heidelberg, Germany: Springer-Verlag: 52–67.
- Mario Jarmasz and Stan Szpakowicz. 2004. Roget’s thesaurus and semantic similarity, *Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 2004*: 111-120.
- Alistair Kennedy, and Stan Szpakowicz. 2008. Evaluating Roget’s Thesauri, *Proceedings of ACL-2008*: 416-424.
- Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated WordNet Construction Using Word Embeddings. *Proceedings of SENSE 2017*: 12-23.
- Claudia Kunze and Lothar Lemnitzer. 2010. Lexical-Semantic and Conceptual relations in GermaNet. In *Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives*, 28:163-183.
- Svetlozara Leseva, Maria Todorova, Tsvetlana Dimitrova, Borislav Rizov, Ivelina Stoyanova, Svetla Koeva. 2015. Automatic classification of wordnet morphosemantic relations. In *The 5th Workshop on Balto-Slavic Natural Language Processing*: 59-64.
- Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation, *Language resources and evaluation*, 48.2: 191-201.
- Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Language RuThes. *Proceedings of workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation. LREC-2002*: 65-70.
- Natalia Loukachevitch and Boris Dobrov. 2014. RuThes Linguistic Ontology vs. Russian Wordnets. *Proceedings of Global WordNet Conference GWC-2014, Tartu*: 154-162.
- Natalia Loukachevitch, Boris Dobrov and Ilia Chetviorkin. 2014. RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014*, 340-350.
- Natalia Loukachevitch and Aleksey Alekseev. 2014. Summarizing News Clusters on the Basis of Thematic Chains. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.
- Natalia Loukachevitch and Boris Dobrov. 2015. The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 21.2: 237-262.
- Natalia Loukachevitch and German Lashevich. 2016. Multiword expressions in Russian thesauri RuThes and RuWordNet. In *Artificial Intelligence and Natural Language Conference (AINL-2016)*, IEEE: 1-6.
- Natalia Loukachevitch, Alexander Shevelev and Valeria Mozharova. 2017. Testing Features and Methods in Russian Paraphrasing Task. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2017, V.1*:. 135-146.
- Natalia Loukachevitch and Anastasia Gerasimova. 2017. Human Associations Help to Detect Conventionalized Multiword Expressions. *arXiv preprint arXiv:1709.03925*.
- Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of Language Resources and Evaluation Conference LREC-2000*: 1413-1418.
- Inderjeet Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8 (1), 43-68.
- Marek Maziarz, Stanoslaw Szpakowicz and Maciej Piasecki. 2012. Semantic relations among adjectives in Polish WordNet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies*, 12:.,149-179.
- George Miller. 1998. Nouns in WordNet. In *WordNet – An Electronic Lexical Database*, edited by Christiane Fellbaum, Cambridge, MA: The MIT Press: 23–47.
- George Miller and Florentina Hristea. 2006. WordNet Nouns: Classes and Instances. *Journal of Computational linguistics*, 32(1):1-3.

- Michael Nokel and Natalia Loukachevitch. 2016. Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of 12th Workshop on Multiword Expressions, ACL 2016*: 44-49.
- Karel Pala and Dana Hlaváčková. 2007. Derivational relations in czech wordnet. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. Association for Computational Linguistics.
- Bolette Pedersen, Sanni Nimb, Jorg Asmussen, Nicolai Sørensen, Lars Trap-Jensen L and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary, *Language resources and evaluation*, 43(3): 269-299.
- Maciej Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012. Automated generation of derivative relations in the Wordnet expansion perspective. In *Proceedings of the 6th Global Wordnet Conference GWC 2012*: 273–280.
- Piek Vossen. 1998. Introduction to EuroWordNet. In *EuroWordNet: A multilingual database with lexical semantic networks*, Springer Netherlands: 1-17.
- Z39.19. 2005. *Guidelines for the Construction, Format and Management of Monolingual Thesauri*. NISO.

Towards Mapping Thesauri onto plWordNet

Marek Maziarz, Maciej Piasecki

G4.19 Research Group, Department of Computational Intelligence

Wrocław University of Technology, Wrocław, Poland

marek.maziarz|maciej.piasecki@pwr.edu.pl

Abstract

plWordNet, the wordnet of Polish, has become a very comprehensive description of the Polish lexical system. This paper presents a plan of its semi-automated integration with thesauri, terminological databases and ontologies, as a further necessary step in its development. This will improve linking of plWordNet into Linked Open Data, and facilitate applications in, e.g., WSD, keyword extraction or automated metadata generation. We present an overview of resources relevant to Polish and a plan for their linking to plWordNet.

1 Introduction

After more than 12 years of continuous development plWordNet – the wordnet of Polish – with the version 3.0 emo (Maziarz et al., 2016) has become a very comprehensive description of the Polish lexical system including: 197,721 synsets, 179,125 lemmas and 260,214 Lexical Units (henceforth LUs¹) described by about 650,000 relation links. It provides also a very good coverage of large corpora of Polish, cf (Maziarz et al., 2016). This is much more than it could have been expected at the beginning, especially if we take into account that plWordNet has been constructed from scratch on the basis of the corpus-based wordnet development method (Maziarz et al., 2013). Moreover, plWordNet has been also manually mapped onto Princeton WordNet on the synset level to a very large extent (>200K mapping relation instances) and onto Wikipedia on the

LU (sense) level (55K mapping relations). Selected statistics are presented in Tab. 1. It includes also emotive annotation for more than 31,000 LUs (Zaśko-Zielińska et al., 2015).

mapping	relation type	instances
plWN-WordNet	<i>I-synonymy</i>	44K
plWN-WordNet	<i>I-near-synonymy</i>	7K
plWN-WordNet	<i>I-hyponymy</i>	125K
plWN-Wikipedia	<i>exactMatch</i>	55K

Table 1: Mappings from plWordNet to Princeton WordNet and to *Wikipedia*.

The question is whether it is the final stage of the development of a wordnet of Polish, or more generally, an example of the final stage of a wordnet in general? The immediate answer is no. A complete wordnet is a moving target that evolves along two dimensions: increasing understanding of the effective use of a wordnet as a tool in describing the lexical system of the natural language, and growing expectations of the wordnet applications developers. In this paper we are going to focus on the latter. Wordnets have to compete with statistical models that are relatively easy to extract from very large corpora. However a wordnet is (or must be) a trustworthy language resource of high quality, providing description of the lexical meanings and the lexical system. Its advantage is in description of infrequent lemmas and LUs that is beyond the scope of Distributional Semantics methods (including word embeddings). Next, an appropriate, high quality means of linking a wordnet with knowledge resources must be provided to facilitate its applications in WSD, keyword and semantic meta-data extraction from text, semantic text classification etc. Our goal is to design a linking mechanism between plWordNet and a rich

¹ A lexical unit is defined here technically as a triple: (Part of Speech, lemma, sense id.)

cloud of heterogeneous terminological and ontological resources, as well as Linked Open Data (LOD), and next to develop an efficient method for building this mechanism in a semi-automated way. In this paper, we focus on linking with terminological resources as a natural extension to the wordnet.

2 Terminology, Terms and Lexical Units

2.1 Ontologies, thesauri, wordnets

The word *ontology* means many things. Most prominent semantic distinction is between ‘metaphysics’ vs ‘a specific kind of computer science object’, however, there is a huge debate on how to define the word in the latter sense:

“Ontology has become, at least for a time, a prevalent buzzword in computer science. An unfortunate side-effect is that the term has become less meaningful, being used to describe everything from what used to be identified as taxonomies or semantic networks all the way to formal theories in logic.” (Pease, 2011).

According to (Roussey et al., 2011) several types of ontologies can be distinguished in relation to their components and structure, including:

Formal ontologies focus mainly on instances (individuals), concepts and their logical definitions (a.k.a. *axioms*) combine logic operators and quantifiers with relations between concepts, and thus enable reasoning.

Software implementation driven ontologies “provide conceptual schemata whose main focus is normally on data storage and data manipulation, and are used for software development activities, with the goal of guaranteeing data consistency” (*ibidem*).

Linguistic ontologies² focus mainly on labels and relations between them:

- *glossaries* - are simple, subject oriented lists of terms and their meanings;
- *dictionaries* - expand term lists with sense/concept textual definitions, often beyond one given subject domain;

² *Lexical* ontologies lack formalization which is characteristic property of formal ontologies, but the former might be comparable to the latter in taxonomic parts (like biology vocabulary), cf. (Hirst, 2009).

- *taxonomies* arrange vocabulary (terms) by hierarchical relations (hypo-/hypernymy, type-/instance, broader/narrower, see (Mitkov and Matsumoto, 2004)),
- *thesauri* are based on a more complex relation system: apart from sub-/superordinate relations also other lexico-semantic links are involved, cf (Currás, 2010),
- *lexical databases* - like WordNet - use a couple dozen lexico-semantic relations between (sets of) senses (concepts), mixing them with textual definitions and other properties (register labels, frequency information, semantic domains, valence frames etc.).

Information ontologies – used by humans in project development processes – aim at capturing relations between concept instances in diagrams in order to clarify the ideas of collaborators.

We adopt here the term *formal ontology* in the meaning: “a formal, explicit specification of a shared conceptualization” (Studer et al., 1998).³ The term *lexical resource* will be used instead of *ontology* (in its broader sense) for all types of computer science objects comprising concepts, their instances, properties, labels and relations between them in various configurations.

Several phenomena arise in vocabulary formalisation. Mapping between concepts and their lexicalisations is not one to one. Existence of near-synonymy and sense vagueness cause that there is no clear cut between many semantically related word senses, and they often overlap. Only subtle differences constitute the distinctions (Fellbaum, 2011). This is captured by a concept of *near-synonymy*, a relation that links word senses close in meaning, being equivalents (interchangeable) in some, but not in all contexts.

In fact, also mapping from words to concepts is not straightforward due to polysemy. Especially many frequent words possess two or more meanings, which is an unusual situation in a formal ontology.

³ The word “conceptualization” means here ‘an abstract, simplified view of the world that we wish to represent for some purpose’ (Guarino et al., 2009). This knowledge ought to be shared by a group of people / a community (e.g., specialists in a given field), and the specification should be so intuitive that most stakeholders could agree with it. (Vrandečić, 2009). Moreover, an ontology should be formally specified and formal logic (usually first order logic or Description Logic) should be used for description purposes to avoid any ambiguities (Prévoit et al., 2010).

Structural (lexical) gaps are also problematic: the mental lexicon does not lexicalise all concepts people have in mind, so there appear gaps in lexical taxonomies (Vossen, 2004).

Natural language is not a *formal* language and the formalization of a vocabulary, even the formalization of relational dictionary, is not an easy task. Consider group / mass nouns *armament* – *weaponry* and try to ascribe them a relation type. Would it be meronymy or hyponymy?

Lexicon is *not* a formal ontology, nevertheless

“a formal ontology without natural language labels attached to classes or properties is almost useless, because without this kind of grounding it is very difficult, if not impossible, for humans to map an ontology to their own conceptualization, i.e. the ontology lacks human-interpretability.” (Völker et al., 2007), after (Hirst, 2009)

2.2 Terms and lexical units

Dictionaries, thesauri, wordnets and formal ontologies in a way deal with vocabulary. A formal ontology uses words as labels that help people to find out the meanings of ontology concepts. A dictionary concentrates on words – describes words, their meaning, grammatical properties and usage. Thesauri and wordnets interlink words and their senses into a lexical net, encoding their description by lexico-semantic relations.

Apart from words, all these resources tend to house some multi-word expressions (MWEs), either fixed (lexicalised) or free. The distinction between what is a part of a vocabulary (what is a multi-word LU) and what is a free syntactic word combination (a collocation)⁴, although not entirely clear, is valid for dictionaries, terminological thesauri, and some wordnets (plWordNet, Germanet). However, in formal ontologies, many domain thesauri and WordNet, words, fixed and free phrases are mixed up. For instance, in the thesaurus of European Union *Eurovoc* we may find free word combinations: *regions and regional policy* or *water management in agriculture*. Similarly in *MeSH* we spot MWEs *Chemicals and Drugs* and *Virus Diseases* (plural). In WordNet we notice word combinations *wheeled vehicle* and

⁴ We call semantically or syntactically fixed MWEs *multi-word lexical units* (MWLU, cf. (Zgusta, 1967)). According to some linguists semantic or syntactic fixedness of MWEs is merely a symptom of being a part of one’s mental lexicon, see (Svensén, 2009; Müller, 2015; Sprenger, 2003).

*horse-drawn vehicle*⁵. Many entries occurring in these lexical resources are domain specific. This leads us to the problem of demarcation between terminology and ordinary phrases and words. The distinction lies in the specialist nature of terminology and the natural provenance of ordinary vocabulary. Terminology is known mostly to specialists, while ordinary language is spoken by all of us.⁶

In ISO 1087-1 *term* is a “verbal designation of a general concept in a specific subject field”. (Wright and Budin, 2001, p. 325) defines *terminology* as “the (structured) set of concepts and their representations in a specific subject field”. These two exemplar definitions suggest that concepts dominate over their lexical manifestations within terminology. Conceptual structure of a theory may enforce morphological shape of words (like in chemistry nomenclature) or can influence formation of MWEs (e.g. in biological taxonomy).

Despite the dissimilar provenance of ordinary and specialist vocabulary, they do not differ with regard to their relation to meaning:

“[T]he relationship between concept and terms is *formally equivalent* to the relationship between meaning and words.” (...) “The traditional theory of terminology [claims] that the concept is the meaning of the term”. (Kageura, 2002)

Terms consist of phonemes, they have their morphemes, inflect like ordinary words or are composed of words like ordinary compositions and have inflection like ordinary phrases. Like ordinary lexemes they do have their meanings. Since they “are [formally] indistinguishable from words” (Sager 1998/99, after: (Kageura, 2002)), we treat terminology as a part of the lexicon.

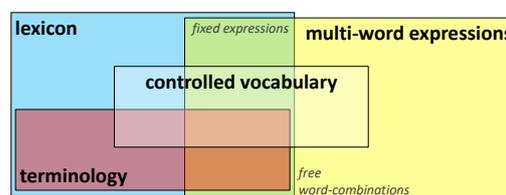


Figure 1: Relations between lexicon, terminology, multi-word expressions and controlled vocabulary.

⁵ In Germanet such MWEs are called ‘artificial’.

⁶ These are specialists that invent new scientific terms, their discussion how to define terms is the important part of scientific activity. On the contrary, ordinary language has no father and evolves spontaneously.

In Fig. 1 we present the relationships between *lexicon* (blue rectangle), *terminology* (red) and *word-combinations* (yellow). By the white one we mark the *controlled vocabulary*.

The controlled vocabulary could be found in thesauri (like *Eurovoc*), ontologies (like *SUMO*) and in subject headings systems (like *Library of Congress Subject Headings*, *LCSH*, or *MeSH*). It consists of specialist terms, ordinary words, multi-word LUs and free word-combinations, sometimes it uses plural forms representing a given category. An important feature of a controlled language is its avoidance of semantic ambiguities:

“Word or phrase indexing and symbolic surrogation systems require some sort of controlled vocabulary – an artificially constructed language in which the ambiguities of natural language are reduced or, ideally, eliminated. A controlled vocabulary is an organized list “of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search.” Controlled vocabularies have two primary objectives: (1) to represent concepts systematically and (2) to facilitate comprehensive searching of a body of information.” (Wallace, 2007)

It is worth to emphasise that *term* is used not only in the meaning ‘a unit piece of terminology’, but also in a broader sense. It may denote every single label/lemma (word or MWE), being an entry of an ontology, a thesaurus, a wordnet or any other lexical resource. All kinds of language expressions from Fig. 2.2 could be described by this word. In this paper, if we use *term* in its broader sense, we will write it down with the plus mark in a superscript (so, *term*⁺), and if we want to refer to the narrower sense (‘terminology unit’), we will write it without a plus (*term*).

plWordNet has concentrated on the Polish *lexicon*, avoiding free word combinations and proper names. Our definition of multi-word LUs points to the phenomena of lexicalisation and terminologisation (Maziarz et al., 2015).

3 Lexical resources vs. plWordNet

Polish vocabulary outside plWordNet could be found in many electronic lexical resources. We describe them below in three groups: (1) subject headings systems, (2) controlled vocabulary thesauri (of the EU, UN and US), and (3) *Wikipedia*.

3.1 Subject headings

There are five available subject heading systems comprising Polish terms⁺, and the biggest one is the Polish National Library Subject Headings.

Polish National Library Subject Headings (PNLSH) is a descriptor system based on the model of Library of Congress Subject Headings. It has reached circa 100K subject terms⁺ and still grows. PNLSH makes use of MARC 21 format, like LCSH.

MeSH, Medical Subject Headings, is the US National Library of Medicine’s controlled vocabulary for medicine. Polish translation was prepared by Main Physicians’ Library in Warsaw, Poland. It gives 28K Polish terms⁺. MeSH is mapped onto LCSH, Snomed or US National Agricultural Library Thesaurus.

Universal Decimal Classification (UDC) core was published on CC-BY-SA licence and translated into Polish by Polish National Library. The UDC core itself is linked to LCSH and through it to Dewey Decimal Classification (DDC) and MeSH.

Sternik is yet another subject headings system designed by Polish National Library. Housing terminology of bibliography and cataloguing, it gives also translations to English. It is equipped with the associative relation *related term*, definitions and alternative labels. Unfortunately, *Sternik* is isolated and has no links to external resources.

Digizaurus is a small thesaurus carefully designed by Polish Digitalization Inter-Museum Group *DigiMuz* for museum collection description in the field of *material*. It comprises 0.6K terms⁺ organised into taxonomy (obtainable in SKOS). Digizaurus is also an isolated resource, like Sternik.

resource	licence	terms ⁺	links
PNLSH ^m	NC	~100K	20K
MeSH ^{m,s}	NC	28K	10K
UDC ^s	CC-BY-SA	2.5K	0.5K
Sternik	sim. to CC-BY	1.7K	—
Digizaurus ^s	CC-BY-NC	0.6K	—

Table 2: Subject headings systems for Polish. The label “terms⁺” denotes Polish labels in each vocabulary, “links” describes an approximate number of mapping instances to external resources (for all terms⁺, including Polish), “NC” means ‘non-commercial’, the letter *s* in superscript marks resources available in SKOS RDF format, *m* represents MARC 21 format.

3.2 Thesauri

IATE, InterActive Terminology for Europe, is a large thesaurus developed collectively by the community of translators and institutions of the EU. It comprises 8.6 million terms⁺ in 24 languages. Polish vocabulary numbers 72K terms⁺.

Eurovoc is an open licence thesaurus describing activities of the EU. It provides terminology in 26 languages, also in Polish (10K terms⁺). Eurovoc has mappings to multiple other thesauri (given in SKOS), inter alia: Agrovoc, Gemet, LCSH, STW Thesaurus for Economics or UNESCO Thesaurus.

Agrovoc was created by Food and Agriculture Organization (FAO) of the United Nations. It is pretty well linked to many external resources, among them to Eurovoc, Gemet, Rameau, STW, Geonames, Thesos and 16 open datasets related to agriculture. Polish translation was done by Central Agricultural Library and comprises 29K terms⁺.

Gemet, GEneral Multilingual Environmental Thesaurus, was developed by European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA). It contains multilingual environment terminology (5K Polish terms⁺) and is a reference thesaurus in this field.

resource	licence	terms ⁺	links
IATE ^s	sim. to CC-BY	72K	>100K
Agrovoc ^s	CC BY-NC-SA	29K	50K
Eurovoc ^s	sim. to CC-BY	10K	10K
Gemet ^s	sim. to CC-BY	5K	7K

Table 3: Polish controlled vocabularies in thesauri.

3.3 Wikipedia

Wikipedia.pl and their byproducts – YAGO or dBpedia — comprise hundreds of thousands of Polish terms⁺. The whole vocabulary is structured with Wikipedia category system. YAGO expanded this system merging it with WordNet. *Wikipedia* is developed by the community of volunteers.

resource	licence	terms ⁺	links
Wikipedia	CC-BY-SA	~1M	>100K

Table 4: *Wikipedia* comprises most Polish terms⁺.

4 Linking Potential

All these lexical resources are interlinked, composing a quite complex resource net. We want to

find a path through it in order to establish mappings between them and plWordNet. We will exercise two main formats: SKOS and MARC 21.

4.1 Formats and alignment

Most resources described in this paper are recorded in SKOS RDF and in MARC 21 (for subject headings). Other relevant formats e.g., of WordNet, of Wikipedia, of dBpedia and of YAGO, will not be discussed, due to space limit.

SKOS RDF. Simple Knowledge Organization System⁷ provides “specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web.” and uses the Resource Description Framework (RDF). In SKOS RDF we have following types of information:

- Concepts: “units of thought – ideas, meanings, or (categories of) objects and events”.
- Concept groups - *schemes* (thesauri or microthesauri grouping concepts) and *collections* (smaller groups of concepts).
- Labels: expressions used in a natural language to refer to concepts. One label is *preferred*, all the others are *alternative* forms.
- Notes: describes concepts in various ways, for instance, *definitions* are verbal descriptions of term⁺'s meaning.
- Semantic relations: describe concepts in the net of semantically closest concepts. Relations *broader* and *narrower* link concepts which are hierarchically super-/subordinate or in part/whole relation.
- Mapping links between a parent thesaurus and external resources are encoded with *.*Match* relations: *exactMatch* links strict equivalents, *closeMatch* links to a less precise counterpart in one external resource, *broadMatch/narrowMatch* points to the external concept which has broader/narrower extension, *relatedMatch* denotes other semantic relations – they are crucial in our task.

MARC 21. MARC (MACHine-Readable Cataloging) 21 is a data format (ISO 2709) used for cataloguing and bibliographic description. It is used

⁷<https://www.w3.org/2004/02/skos>

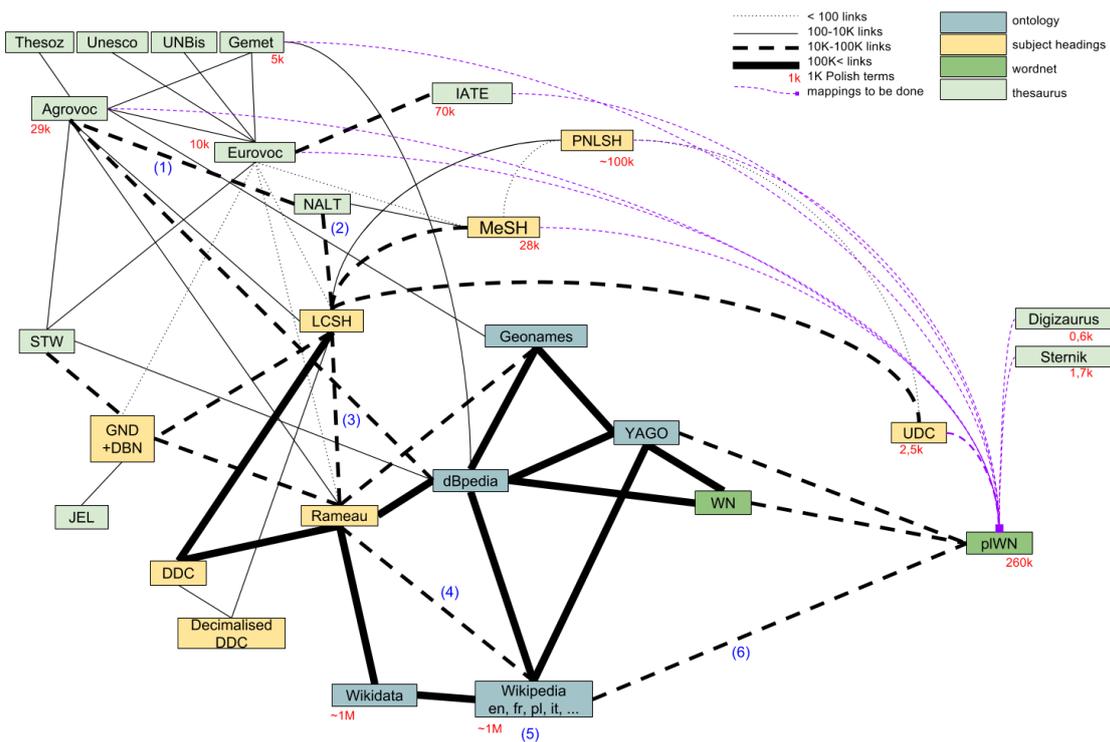


Figure 2: Linking potential of the existing lexical resources – Polish perspective.

by the Library of Congress in its famous subject headings that makes it popular. MARC provides various *fields* of which the most important for us are:

- Field 080 provides counterparts from UDC, while 082 links to DDC.
- Fields 150 and 450 gives preferred and alternative labels (respectively).
- Field 550 lists all internal semantic relations within a given subject headings system.
- Field 650 gives equivalents in distinct resources: “0” stands for LCSH, “2” – MeSH.

4.2 Vocabulary ‘propagation’

Existing mappings between lexical resources give an opportunity not only to align Polish vocabulary between two separate thesauri, but also to provide translations for not-translated terms⁺. Thesauri lacking Polish labels may be equipped with Polish equivalents. Let us call it vocabulary *propagation*.

We plan to propagate the vocabulary iteratively. At first, we will use direct links between resources to label equivalents with Polish labels. Then we

are going to use such translated lexical resources to translate resources that are linked to them. Thus Polish vocabulary would spread across the net of lexical resources. In each step we will proceed only with translations of direct equivalents.

Direct equivalents. Let us look at existing Eurovoc - STW Thesaurus for Economics and Eurovoc - Gemet mappings (see Tab. 5 and Fig. 2). In Eurovoc SKOS RDF we find 2262 *skos:exactMatch* links to STW and half as many to Gemet. Some of them have Polish labels in Eurovoc. STW does not, and Gemet does. Consider the Polish label *prawo pracy* ‘labour law’ in Eurovoc, its concept (ID: 557) has the exact match in STW (labelled *labour law*) and the exact match in Gemet (labelled with *prawo pracy*).

mapping	relation type	instances
Eurovoc-STW	<i>exactMatch</i>	2262
Eurovoc-STW	<i>closeMatch</i>	369
Eurovoc-Gemet	<i>exactMatch</i>	1294

Table 5: Mappings from Eurovoc to STW & Gemet through direct links.

In step 1 we give Polish labels to all concepts that have an exact or close match in a mapping from any labelled with Polish terms⁺ thesaurus.

Indirect equivalents. To exemplify how we plan to establish indirect links let us discuss the case of a Polish label for ‘blood protein disorders’ in Agrovoc (ID: c 969): *Zaburzenia białek krwi* (preferred label⁸). Since we may link the label to the National Agricultural Library Thesaurus (NALT) concept ‘blood protein disorders’ (ID: 18150), we may also take advantage of NALT-LCSH mapping existence (cf. Tab. 6). The concept has the exact equivalent in LCSH *Blood protein disorders* (ID: sh 85015013).

mapping	relation type	instances
Agrovoc-NALT	<i>exactMatch</i>	26520
NALT-LCSH	<i>exactMatch</i>	8501
NALT-LCSH	<i>closeMatch</i>	2755

Table 6: Mappings from Agrovoc to US National Agricultural Library Thesaurus (NALT) & from NALT to LCSH through direct links.

Even longer paths. We may go with the Agrovoc even beyond LCSH. In Fig. 2 one may find a possible way from Agrovoc to plWordNet (marked with blue numbers): Agrovoc –1→ NALT –2→ LCSH –3→ Rameau –4→ Wikipedia francophone –5→ Polish Wikipedia –6→ plWordNet. Let us trace the whole path with the concept ‘blood pressure’ from Agrovoc (ID c 967).

(1) The concept has the Polish label *Ciśnienie krwi* (prefLabel; the alternative label *Obniżone ciśnienie*, lit. ‘low blood pressure’, is not considered here). It points to NALT ‘blood pressure’ (ID: 18146) via *exactMatch*. (2) NALT ‘blood pressure’ then is matched with LCSH ‘Blood pressure’ (ID: sh 85015010), again with the *exactMatch* relation. (3) From LCSH we jump right to French National Library subject headings *Rameau* and ‘Pression artérielle’ (ID: cb11976295t). The *closeMatch* was used here.⁹ (4) Now we go with *exactMatch* to French Wikipedia to the article *Pression artérielle*¹⁰ and then (5) to Polish Wikipedia article *Ciśnienie tętnicze* (=‘artery

⁸Please, note that – according to SKOS guidelines – only preferred labels are linked by the *exactMatch* relation.

⁹Please note that: (a) the blood pressure is usually measured in arteries, (b) *closeMatch* is supposed to serve well only on short distances (one link, see SKOS definition).

¹⁰https://fr.wikipedia.org/wiki/Pression_artérielle

pressure¹¹.) (6) Since plWordNet is widely linked to Polish Wikipedia with *exactMatch*, we may finally establish link from Agrovoc ID: c 967 *Ciśnienie krwi*, *blood pressure* to the plWordNet synset {ciśnienie tętnicze 1}.

The above example raises the question on the quality of such long chains. The longer the path is, the more probable the relation is distorted. Is *ciśnienie krwi* ‘blood pressure’ a real synonym of *ciśnienie tętnicze* ‘arterial pressure’? Fortunately, we do not have only one way to choose from a given resource to plWordNet. Thanks to the mapping between plWordNet and Princeton WordNet our path bifurcates. We may choose a route from the WordNet through ontologies YAGO and dBpedia to Rameau. This gives us rare occasion to verify different links and check their consistency.

4.3 Hybrid approach

When the iterated process of vocabulary propagation is done, we will have some Polish terms⁺ introduced into different lexical resources, as well as, many matching relation instances. Of course, links to plWordNet synsets are of special importance and the whole process will focus on them.

Prompt algorithm. The next step will be running an algorithm giving suggestions to linguists. It takes into account the already established links as constraints. We plan to utilize the implementation of relaxation labelling algorithm (used successfully in plWordNet-WordNet mapping (Kędzia et al., 2013)). The algorithm can handle also linking isolated resources (like Sternik or Digizaurus).

Assessing quality of the mapping. The automatic algorithm will suggest potential links. We may expect more than 100K new terms⁺, so assessing quality of the automatic mapping will be a challenge. Mappings from small resources (e.g. Gemet) could be checked fully by plWordNet editors, and manual checking of the mappings of isolated thesauri (Digizaurus and Sternik) is a must. However, automatic matching from larger resources, like Polish Wikipedia or PNLISH, will be too big for a complete manual verification. The proposed process is presented in Fig. 3.

After checking and correcting automatically generated links, linguists will also check lexical-

¹¹https://pl.wikipedia.org/wiki/Ciśnienie_tętnicze

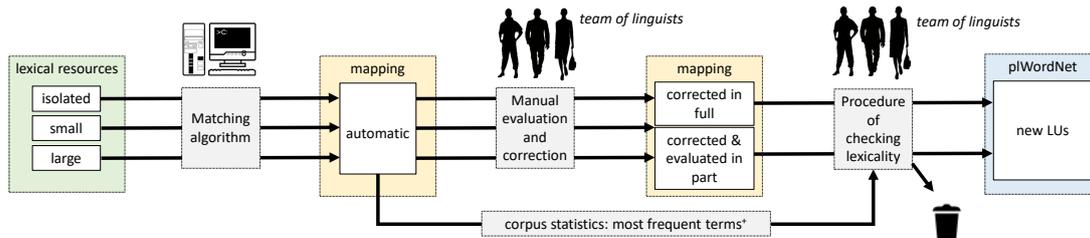


Figure 3: Semi-automatic mapping lexical resources onto pWordNet. The matching relation verification will be done in full (for small and isolated thesauri) or in part (for large resources). Linguists may check also lexicality of all verified in the preceding phase terms⁺ plus some of high corpus frequencies.

ity of terms⁺ taken from isolated or small lexical resources, and a sample of terms⁺ from large resources together with the most frequent ones. We estimate that verification of 1K automatic links and assessing their lexicality will take altogether one person-month, e.g. preparing the mapping of Sternik, would take two person-months, while Agrovoc circa 30 person-months. In order to remain consistent with most of our thesauri (Agrovoc, Digizaurus, Eurovoc, Gemet, IATE, MeSH and UDC) relation types from the SKOS format will be utilized. Linguists will choose semantically closest counterparts from pWordNet, whether they will be exact or close equivalents (*exactMatch*, *closeMatch*), or synsets which have broader or narrower meaning (*broadMatch*, *narrowMatch*).

Listing 1: Introducing terms⁺ into pLWN

```

0: X is a term+ (in a fixed sense).
1: Can X serve as a noun in a sentence?
  Y: next, N: end
2: Is X a proper name? Y: end, N: next
3: Is X already introduced into pLWN?
  Y: end, N: next
4: Is X a plurale tantum?
  Y: goto 6, N: next
5: Is X a plural form? Y: end, N: next
6: Is X a MWE? Y: next, N: introduce X
7: Is a conjunction / comma a part of X?
  Y: end, N: next
8: Is X semantically compositional?
  Y: next, N: introduce X
9: Does X belong to terminology?
  Y: introduce X, N: next
10 Does X exhibit syntactic irregularity?
  Y: introduce X, N: end

```

next means ‘go to the next step of the procedure’, **goto** denotes jumping to the specific step, **end** = ‘X is not a lexical unit’, **introduce** = ‘add a term⁺ to pWordNet’, **term⁺** denotes either a word or a MWE being a part of a lexical resource.

Introducing LUs into pWordNet. The mapping will give us a unique opportunity to expand pWordNet with new LUs. This will be done in two phases. Firstly, we will check it at the same time as the matching relation accuracy evaluation. Secondly, we will test those terms⁺ that are frequent in a reference corpus.

As we have shown in Sec. 2.2, many terms⁺ occurring in lexical resources are not lexicalised. Among them there are entries containing *conjunctions*, *commas*, being *free* word-combinations and *proper names*, or given in *plural*. We propose the following algorithm designed for pWordNet editors (Listing 1) to assess a given term⁺ as a LU.

The 10 filtering rules help sifting through non-lexicalised language expressions. At the end, lexicalised terms⁺ are introduced into pWordNet.

5 Perspectives

The presented overview and mapping method show a great potential in building a very large network of resources around pWordNet. The network can be even more expanded with LOD utilising the existing high quality manual mapping of pWordNet onto WordNet. The primary application will be improvement of a wordnet-based WSD that works better with larger and denser network. Next, it will be a basis for a method of the automated assignment of descriptive keywords to texts and will support extraction of keywords from texts. Both methods will be first used in automated semantic indexing of digital research repositories, and next in different applications in Digital Humanities and Social Sciences. For such applications possibility of finding associations between texts and specialist terms is crucial and can be done via the created complex network.

Acknowledgment Works funded by the Polish

Ministry of Science and Higher Education within CLARIN-PL Research Infrastructure.

References

- [Currás2010] Emilia Currás. 2010. *Ontologies, Taxonomies and Thesauri in Systems Science and Systematics*. Chundos Publishing, Oxford.
- [Fellbaum2011] Christiane Fellbaum. 2011. Wordnet (and why it's not an ontology). In Adam Pease, editor, *Ontology: A Practical Guide*, pages 71–73. Articulate Software Press, Angwin, CA, USA.
- [Guarino et al.2009] Nicola Guarino, Daniel Oberle, and Steffen Staab. 2009. What is an ontology? In Steffen Staab and Ruder Studer, editors, *Handbook on Ontologies*. Springer, second edition.
- [Hirst2009] Graeme Hirst. 2009. Ontology and the lexicon. In *Handbook on ontologies*, pages 269–292. Springer.
- [Kageura2002] K. Kageura. 2002. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Terminology and lexicography research and practice. J. Benjamins Pub.
- [Kędzia et al.2013] Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies*. to appear.
- [Maziarz et al.2013] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. Int.l Conf. on Recent Advances in Natural Language Processing*, Hissar, Bulgaria.
- [Maziarz et al.2015] Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2015. A procedural definition of multi-word lexical units. In *RANLP*, pages 427–435.
- [Maziarz et al.2016] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.
- [Mitkov and Matsumoto2004] Ruslan Mitkov and Yuji Matsumoto. 2004. *Handbook Of Computational Linguistics*, chapter Lexical Knowledge Acquisition. Oxford University Press.
- [Müller2015] Peter O. Müller, 2015. *Multi-word expressions*. De Gruyter Mouton.
- [Pease2011] Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press, Angwin, CA.
- [Prévot et al.2010] Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari. 2010. Ontology and the lexicon: a multi-disciplinary perspective. In *Ontology and the Lexicon. A Natural Language Processing Perspective*. Cambridge University Press.
- [Roussey et al.2011] Catherine Roussey, Francois Pinet, Myoung Ah Kang, and Oscar Corcho, 2011. *An Introduction to Ontologies and Ontology Engineering*, pages 9–38. Springer London, London.
- [Sprenger2003] Simone Sprenger. 2003. *Fixed expressions and the production of idioms*. Ph.D. thesis, Max Planck Instituut voor Psycholinguïstiek.
- [Studer et al.1998] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: Principles and methods.
- [Svensén2009] Bo Svensén. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.
- [Völker et al.2007] Johanna Völker, Pascal Hitzler, and Philipp Cimiano, 2007. *Acquisition of OWL DL Axioms from Lexical Resources*, pages 670–685. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Vossen2004] Piek Vossen, 2004. *Handbook Of Computational Linguistics*, chapter Ontologies. Oxford University Press.
- [Vrandečić2009] Denny Vrandečić, 2009. *Ontology Evaluation*, pages 293–313. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Wallace2007] Danny P. Wallace. 2007. *Knowledge Management: Historical and Cross-disciplinary Themes*. Libraries Unlimited.
- [Wright and Budin2001] S. E. Wright and G. Budin. 2001. Handbook of terminology management: Application-oriented terminology management. John Benjamins, Amsterdam and Philadelphia.
- [Zaško-Zielińska et al.2015] Monika Zaško-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP'2015*, pages 721–730, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- [Zgusta1967] Ladislav Zgusta. 1967. Multiword lexical units. *Word*, 23(1-3):578–587.

Investigating English Affixes and their Productivity with Princeton WordNet

Verginica Barbu Mititelu

Romanian Academy Research Institute for Artificial Intelligence

Bucharest, Romania

vergi@racai.ro

Abstract

Such a rich language resource like Princeton WordNet, containing linguistic information of different types (semantic, lexical, syntactic, derivational, dialectal, etc.), is a thesaurus which is worth both being used in various language-enabled applications and being explored in order to study a language. In this paper we show how we used Princeton WordNet version 3.0 to study the English affixes. We extracted pairs of base-derived words and identified the affixes by means of which the derived words were created from their bases. We distinguished among four types of derivation depending on the type of overlapping between the senses of the base word and those of the derived word that are linked by derivational relations in Princeton WordNet. We studied the behaviour of affixes with respect to these derivation types. Drawing on these data, we inferred about their productivity.

1 Introduction

Affixes productivity, i.e. their use to create new words, can be studied on a corpus or on lists of words, in particular on dictionaries. Working with a corpus has several advantages over working with a dictionary: words are seen “in action” (i.e. one can see in what contexts they are used, in what forms, with what frequency, etc.); one can find words that are not recorded in dictionaries, either because they are brand new creations or because they are obtained in a (highly) regular way by a very productive word formation rule; frequencies can be counted for either types or tokens. However, we chose Princeton WordNet (PWN) (Fellbaum, 1998) version 3.0 for studying the productivity of English affixes. We wanted to test

whether affixes productivity is influenced by the number of senses of the base form and of the derived word that are semantically unrelated. PWN has several characteristics that make it appropriate for our investigation. It contains quite a large number of words (155,287 lemmas) organized according to their senses (thus reaching 206,941 word-sense pairs)¹. PWN also displays lexical density: “all” senses of a word are included; this is a great asset for our experiment, which is run at the word sense level.

The hypothesis of our study is that the meaning of the derived word is compositional, being a function of the meaning of the base word and of the affix(es) contained (other authors (Plag, 1999) formulate this as a function of the meaning of the rule and of the base). Whenever no semantic resemblance can be found between the two (in other words, derived words have an idiomatic meaning rather than a compositional one – see Bauer et al. (2013)) we do not consider them a derived-base pair of words. Nevertheless, we presume that the original meaning(s) of the derived words is/are (a) compositional one(s), whereas the idiomatic one(s) is/are the result of a semantic evolution in independence of the semantic evolution of its base word.

2 Related work

There are two lines of research interesting as background for our experiment: one has to do with the study of affixes productivity, and the other concerns the derivational morphology studies in connection to PWN or with other wordnets, each of them detailed in a separate subsection in what follows.

¹The data are taken from <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>.

2.1 Affixes productivity

An affix is a morpheme that is attached to a word in order to create a new word, process known as derivation. Not all affixes in a language are productive to the same extent: some are more productive than others, while others may show no productivity at all; still others may cease being productive for some time and may get “reactivated” afterwards. Productivity is studied in synchrony: from one period to another one can notice differences in the productivity of the same affix, as said before.

Word formation processes, derivation included, are never totally unrestricted (Plag, 1999). Several factors have been discussed with respect to their influence on affixes productivity. On the one hand, there are both linguistic and non-linguistic ones; on the other hand, they show the interdependence of the various subsystems of the language (Aronoff, 1976). These factors are: morphological restrictions on the base word, semantic coherence (Aronoff, 1976), paradigmatic factors (van Marle, 1985), lexical government, lexical listing, phonological factors (Aronoff, 1976; Baayen, 1992), phonotactics (Hay and Baayen, 2003), etymology of the base word (Bauer et al., 2013), parsing (i.e. decomposition in perception) (Hay and Baayen, 2002), type and token frequency (Baayen, 1992), contextual appropriateness (Burgschmidt, 1977), socio-economic status of the language user and his/her attitude towards linguistic phenomena (Baayen, 1992), “fashion” (Plag, 1999).

2.2 Derivational morphology and wordnets

Several wordnets (American (Fellbaum et al., 2009), Czech (Pala and Hlaváčková, 2007), Bulgarian (Koeva, 2008; Dimitrova et al., 2014; Koeva et al., 2016), Romanian (Barbu Mititelu, 2012), among others) have gone beyond their original structure and included, between pairs of literals, new relations, derivational in nature: the connected literals are the base and the derived words, of course considered with their respective meaning (from the synset to which they belong). Such relations reflect both the formal connection between the two literals (i.e. one is created from the other by means of derivation, that is by adding an affix to it) and the semantic connection: the derived literal has a compositional meaning, in which one can recognize the meaning of the base word and the

contribution of the affix. Either manually or automatically, the pairs are identified and labeled using various sets of relation names. Such relations are identified for certain parts of speech (as is the case in Bulgarian (Koeva et al., 2016), Croatian (Koeva, 2008) or American wordnets, among others) or all of them (e.g., Polish (Piasecki et al., 2012) and Romanian (Barbu Mititelu, 2012), among others) and are labeled differently from one wordnet to the other, although some overlaps exist.

In the projects enriching wordnets with such relations there has been interest in making these resources richer and more useful for various applications (Barbu Mititelu, 2013).

3 The experiment

In this section we present an experiment in which we extracted the pairs of base - derived word from PWN and assigned them to a different class according to the way their senses are related by a derivational relation.

3.1 Aim

The hypothesis we wanted to test here and that had not been touched upon in any previous study that we are aware of is whether the number of senses the base word and the derived word, the proportion of them being interlinked and/or the semantic evolution of the derived word independently from the base are factors that could influence affixes productivity.

3.2 Data preparation

Among the relations marked in PWN v. 3.0 there are several that link pairs of derivationally related words: `derivat` (linking nouns to their noun, verb or adjective roots, verbs to their noun or adjective roots, adjectives to their noun, verb or adverb roots, and adverbs to their adjective roots), `derived_from` (linking adverbs derived from adjectives), `pertainym` (linking adjectives to their noun roots). We extracted all pairs of words linked by the first two relations mentioned. The last one (`pertainym`) was disregarded because it usually doubles the relation `derivat`, i.e. it links words that are usually also linked by the `derivat` relation, as in the following example: the adjective *academic* in its first sense establishes two relations with the noun *academia*: one is `derivat` and the other one is `pertainym`.

We extracted 77,939 pairs of words (base -

derived word) between which there is either a `derivat` or a `derived_from` relation. However, some of them are duplicates: for example, the adjective *scarce* is related to the nouns *scarcity* and *scarceness* by means of the relation `derivat`; in their turn, both nouns are linked to the adjective *scarce* by means of the relation `derivat`. Thus, we eliminated duplicates in the data and were left with 40,632 pairs. We added 73 pairs which involved participles linked to their base verbs by means of the relation `participle`: for example, *avenged* (marked as adjective) is linked to the verb *avenge* by means of the relation `participle`.

Further cleaning of the data was done in order to eliminate dialectal duplicates: words belonging to the same synsets and that differ in the spelling with *-ise* or *-ize*, on the one hand, and words containing the *-ou-* or the *-o-* sequence, on the other hand: examples: *equalise* - *equalize*; *discolouration* - *discoloration*. Only one of the pairs was kept, in each case. The former type of duplicates occurred 81 times in the data, while the latter occurred 306 times.

Thus, the list we focused on for annotation contained 40,318 pairs of base - derived words, including all parts of speech in PWN.

3.3 Data annotation

For all these pairs we automatically extracted the affix(es). The base and the derived words were compared as strings of letters and the difference found between them was checked against a list of English affixes containing 26 prefixes and 54 suffixes. In case the string was found in that list, it was considered an affix and marked as such in the annotation. Otherwise, manual intervention (by one linguist) was necessary for identifying the affix(es) or their combination in case of parasynthetic derivation (i.e. by means of both a prefix and a suffix) or successive derivation. During the manual inspection of the pairs we also identified pairs that are in no derivational relation at all: *inappropriate* and *wrongness*, *immunology* and *allogeneic*, etc. They were eliminated from the data. Another situation is that of words like *skepticism* - *skeptical*: they are both created from the same root, *skeptic*, each with a different suffix: *-ism* and, respectively, *-al*, so they are not derived one from the other. Such pairs were also disregarded, just like cases of a similar type: *atheism* - *atheis-*

tic, where one can recognize the Greek elements *a-* and *theos*, but the former is borrowed from French (where the word was obtained by adding the suffix *-isme* to the Greek elements) and the latter is derived in English by adding the suffix *-ic* to the French borrowing *athéiste* (itself derived by adding the suffix *-iste* to the Greek *atheos*). Thus, the total number of annotated pairs was 30,018.

For all these pairs we identified the affix, we extracted from PWN the number of senses each of the literals in the pairs has and the number of derivational relations established between the two literals. Afterwards, we counted:

- the number of senses with which the base word participates in the derivational links with the derived words
- their percent in the total number of senses of the base word
- the number of senses the derived word participates in the derivational links with the base
- their percent in the total number of senses of the derived word.

It is important to note that the numbers representing the number of derivational relations established between the two literals, the number of senses with which the base word participates in derivational links with the derived word, and the number of senses with which the derived word participates in the derivational links with the base need not be identical. Let us consider the following pair: *buzz* - *buzzer*. The verb base word has the following senses:

- *buzz:1* - make a buzzing sound
- *buzz:2* - fly low
- *buzz:3* - be noisy with activity
- *buzz:4* - call with a buzzer

The derived noun has the following senses:

- *buzzer:1* - a push button at an outer door that gives a ringing or buzzing signal when pushed
- *buzzer:2* - a signaling device that makes a buzzing sound

The four derivational relations established between the two words are as follows:

- *buzz:1 - buzzer:1*
- *buzz:1 - buzzer:2*
- *buzz:4 - buzzer:1*
- *buzz:4 - buzzer:2*

There are four derivational relations between the two words, but, whereas all senses of the derived word enter these relations, only two out of the four senses of the base participates to them.

Another step in the annotation was the automatic identification of the derivation type, as we will explain below. We automatically counted the number of senses specific to the base word, i.e. not establishing links with the derived word, the number of senses specific to the derived word, and the ratio between the senses specific to the derived word and those specific to the base word.

Four types of derivation were identified as types of sets intersection. Whenever **all** senses of the derived word are linked to **some** of the senses of the base word, we mark the pair as being of the **R** type: see the pair *buzz - buzzer* above. When **some** senses of the derived word are derivationally linked to **all** of the senses of the base word, we mark the pair as being of the **D** type: see *restitute - restitution*: the base verb has the following senses:

- *restitute:1* - give or bring back
- *restitute:2* - restore to a previous or better condition

The derived noun has the following senses:

- *restitution:1* - a sum of money paid in compensation for loss or injury
- *restitution:2* - the act of restoring something to its original state
- *restitution:3* - getting something back again

The derivational relations established between the two words are as follows:

- *restitute:2 - restitution:2*
- *restitute:1 - restitution:3*

Both senses of the base are linked to some of the senses of the derived word.

In case of identical sets, which means that there is no sense of the base word that is not derivationally linked to any of the senses of the derived word

and vice versa, there is no sense of the derived word that is not linked to any of the senses of the base word, we mark the pair as being of the **RD** type: see the pair *explore - exploration*: the base verb has the following senses:

- *explore:1* - inquire into
- *explore:2* - travel to or penetrate into
- *explore:3* - examine minutely
- *explore:4* - examine (organs) for diagnostic purposes

The derived noun has the following senses:

- *exploration:1* - to travel for the purpose of discovery
- *exploration:2* - a careful systematic search
- *exploration:3* - a systematic consideration

The derivational relations established between the two words are as follows:

- *explore:1 - exploration:3*
- *explore:2 - exploration:1*
- *explore:2 - exploration:3*
- *explore:3 - exploration:2*
- *explore:3 - exploration:3*
- *explore:4 - exploration:2*

All senses of both words are involved in these six derivational links between them.

When at least one sense of the derived word is linked to at least one sense of the base word, and there is at least one sense of the derived word not linked to any sense of the base word and at least one sense of the base word not linked to any sense of the derived word, we mark the pair as being of the **I** type: see *perform - performance*: the base verb has the following senses:

- *perform:1* - carry out or perform an action
- *perform:2* - perform a function
- *perform:3* - give a performance (of something)
- *perform:4* - get (something) done

The derived noun has the following senses:

- *performance:1* - a dramatic or musical entertainment
- *performance:2* - the act of presenting a play or a piece of music or other entertainment
- *performance:3* - the act of performing; of doing something successfully; using knowledge as distinguished from merely possessing it
- *performance:4* - any recognized accomplishment
- *performance:5* - process or manner of functioning or operating

There are only two derivational relations established between the two words, involving only a couple of their senses:

- *perform:1* - *performance:3*
- *perform:3* - *performance:1*

All the other senses of the two words remain derivationally unrelated.

For each affix (or combination of affixes) we calculated the frequency of the different types of derivation (R, D, RD, I) to which it participates in PWN (see subsection 4.2 below for the interpretation of these data).

4 Results and their linguistic significance

There are several results of this undertaking. One of them is the list of pairs extracted from PWN and enriched with information as described above. We discuss the others in the subsections below.

4.1 Derivation types

The total number of occurrences of the derivation types is 30,018. The most frequent one is the RD type - 12,792 occurrences. The second most frequent one is the R type (11,043 occurrences). They are followed, at long distance, by type I (4,267 occurrences) and type D (1,916 occurrences).

The highest frequency of the RD type shows that most of the derived words share the meanings of their base. However, there is also a large number of cases when the derived word is “semantically less rich” than its base word - see the high number of occurrences of type R.

Much less frequent (4,267) is the case of pairs in which the two words have both meanings in common (type I), and an independent semantic evolution. This is the case of pairs such as *dust* - *duster*. The former has the following meanings:

- *dust:1* - remove the dust from
- *dust:2* - rub the dust over a surface so as to blur the outlines of a shape
- *dust:3* - cover with a light dusting of a substance
- *dust:4* - distribute loosely

The latter has the meanings:

- *duster:1* - a windstorm that lifts up clouds of dust or sand
- *duster:2* - a loose coverall (coat or frock) reaching down to the ankles
- *duster:3* - a piece of cloth used for dusting
- *duster:4* - a pitch thrown deliberately close to the batter

Only *dust:1* is derivationally related to *duster:3*. The other meanings remain semantically distant.

We should note that types R and RD may contain false positives examples, because in wordnets there is no distinction between polysemous words and homographs of the same part of speech: they are both recorded as different senses of the same literal.

The least frequent (1,916) is the case of derived words that develop new meanings (after derivation) (type D): consider the adjective *amphibious* derived from *amphibia*. Besides the meaning “relating to or characteristic of animals of the class Amphibia”, which clearly links it to the base (having the meaning “the class of vertebrates that live on land but breed in water; frogs; toads; newts; salamanders; caecilians”), the derived word has developed another meaning (“operating or living on land and in water”), which applies to various semantic types of nouns, as the examples in PWN show: “amphibious vehicles”; “amphibious operations”; “amphibious troops”; “frogs are amphibious animals”, in complete independence from the base.

In terms of affixes productivity, only types D and I are interesting: we can think of the new

meanings of the derived words in PWN as hapax phenomena (i.e., the words occurring only once in PWN) in a corpus. Consequently, following (Baayen, 1992), who proved that the number of hapax legomena instances of words derived with a certain affix in a corpus is suggestive of that affix productivity, we can consider affixes involved in these two types of derivation to be productive ones (see the next subsection).

4.2 Affixes and types of derivation

Having annotated the type of derivation pertinent to each pair, we can test if affixes manifest any affinity with these derivation types.

A first remark on the data is that affixes rarely tend to belong to only one derivation type. We looked at the ten most frequent ones in our data. They are:

- *-ness* - 3,730 occurrences;
- *-er* - 3,100 occurrences;
- *-ly* - 2,953 occurrences;
- *-ion* - 2,469 occurrences;
- *-ing* - 2,102 occurrences;
- *-ation* - 1,546 occurrences;
- *-ic* - 1,290 occurrences;
- *-ity* - 1,186 occurrences;
- *-al* - 1,011 occurrences;
- *-ist* - 805 occurrences.

Their distribution according to the four types of derivation is rendered in Figure 1 below. All these affixes participate in all four types of derivation, even if to a different extent. We can note that the RD type is predominant for most affixes, except for *-ing*, *-ly* and *-er*, which tend to participate in derivations of type R.

Type R of derivation tends to be realized by the affixes *-ly*, *-er*, *-ness*, *-ing*, as obvious in Figure 2. Type RD is realized by the affix *-ness* to the highest extent. Type D is more frequently realized by the affix *-ion*, almost three times more often than the next frequent affix for this derivation type, namely *-ation*. Type I is realized mostly by the suffixes *-er* and *-ion* and, to a lesser and comparable extent, by the other suffixes in the top 10 most frequent ones in our data.

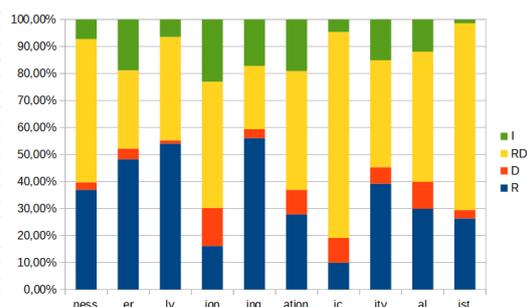


Figure 1: The 10 most frequent affixes and the frequency of the types of derivation to which they participate.

Little correlation can be noted between the affixes realizing the D and I types of derivation. Besides the prevalence of the suffix *-ion* with both types, nothing else strikes us when comparing the two.

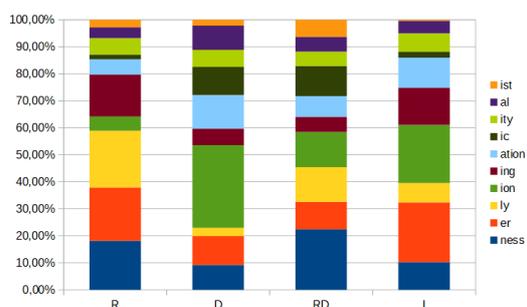


Figure 2: The four types of derivation and the affixes involved.

4.3 Affixes productivity

We compared the data we obtained with the statistical data about affixes provided by Hay and Baayen (2002). They report on a corpus-based research: their calculations “are based on a set of words extracted from the CELEX Lexical Database (Baayen et al., 1995)”. We noted a correlation of their results with the PWN-based data obtained by us.

Firstly, the frequency of affixes is similar in the two experiments: looking only at the most frequent ones, the following affixes occur on both lists: *-er*, *-ly*, *-y*, *-ness*, *-al*, *-ic*, *-ity*, *-able*. Hay and Baayen (2002) also report a high frequency of the suffixes *-like* and *-less*. The former has only one occurrence in our data, whereas the latter is completely absent: words derived with *-less* (such

as *harmless*, *speechless*, etc.) are not derivationally related in PWN to their respective bases.

Secondly, comparing the number of hapax legomena for individual affixes in the corpus-based experiment with the sum of the frequency of D type and I type derivations for the same affixes in the PWN-based experiment, we also notice similarities between data: the most productive affixes, from both perspectives, are: *-er*, *-y*, *-ly*, *-ness*. Other very productive ones are: *-or*, *-able*, *-an*. They all display a high number of hapaxes in the corpus and, respectively, high number of total occurrences in derivations of types D and I.

5 Conclusions and future work

A mature resource, PWN can be used, besides in language-enabled applications, in linguistic studies of various types. Our experiment is grounded in the assumption that derivation is a relation between word senses rather than between words as sets of meanings. This relation manifests in a formal and semantic way: formally, one word (the derived one) in the relation is obtained from the other (the base word) (usually) by adding some linguistic material (an affix); semantically, the meaning of the derived word is compositionally obtained from the meaning of the base word and of the affix(es) it contains. PWN follows this assumption and, thus, offers the perfect environment for testing the hypothesis that affixes that are involved in deriving words that develop meanings independently from their base word are morphologically productive ones. As shown above, this seems to be the case.

We have also presented here, based on the data extracted from PWN and annotated, information about affixes frequency in general and, in particular, their frequency depending on four types of derivation defined ad hoc, thus their tendencies to participate in one type or another of derivation.

However, as obvious from the discussion in this paper, the degree of coverage and of correctness of the derivational links in PWN varies from one affix to the other. It is straightforward that this fact has an impact on our research. Nevertheless, we could not evaluate it for this presentation of results.

As further work, we could also check if PWN granularity, already proved to be too fine, is reflected in the way derivation is marked in the network: for this, we would look, for each derived literal, at the number of derivational links each of

its senses establishes with its base word.

Other aspects of affixes study that can be extracted from further processing the data we now have are: affixes capacity of allowing for the inheritance by the derived word of the meaning(s) of the base word (calculated as the percent of senses of the base word that are linked to the derived word), their capacity of allowing sense evolution (calculated as the percent of senses specific to the derived word) and the ratio of the derived word specific senses and of the base word specific senses.

The semantic types of the base words to which one affix can attach is another line of research possible to be explored with our data.

Our experiment could be repeated for another language for which there is quite a large wordnet, in whose development the implementation of as many senses of a word as possible was an objective.

Acknowledgments

I would like to express my gratitude to Prof. Harald R. Baayen for the insightful discussions we had about this topic.

References

- Mark Aronoff. 1976. *Word Formation in Generative Grammar*. Cambridge, MA, London, England: MIT Press.
- Harald Baayen. 1992. *Quantitative aspects of morphological productivity*. In G. E. Booij and J. van Marle (eds.), *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers:109–149.
- R.H. Baayen, R. Piepenbrock, and L. Gulikens. 1995. *The CELEX lexical database (release 2) cd-rom*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Verginica Barbu Mititelu. 2012. *Adding morpho-semantic relations to the Romanian Wordnet*. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*:2596–2601.
- Verginica Barbu Mititelu. 2013. *Increasing the Effectiveness of the Romanian Wordnet in NLP Applications*. *Computer Science Journal of Moldova*, vol. 21, no. 3(63):320-331.
- Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press.
- Ernst Burgschmidt. 1977. *Strukturierung, Norm und Produktivität in der Wortbildung*. In H. E. Brekle

- and D. Kastovsky (eds.). *Perspektiven der Wortbildungsforschung*. Bonn: Bouvier Verlag.
- Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. *Coping with derivation in the Bulgarian WordNet*. In *Proceedings of the Seventh Global WordNet Conference (GWC 2014)*:109–117.
- Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. *Putting semantics into WordNets “morphosemantic” links*. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland*. [Reprinted in: *Responding to Information Society Challenges: New Advances in Human Language Technology*. Springer Lecture Notes in Informatics, volume 5603:350-358.
- Jennifer Hay and Harald Baayen. 2002. *Parsing and Productivity*. In G. E. Booij and J. van Marle (eds.). *Yearbook of Morphology*. Dordrecht: Kluwer Academic Publishers:203–235.
- Jennifer Hay and Harald Baayen. 2003. Phonotactics, Parsing and Productivity. *Italian Journal of Linguistics*, 1:99–130.
- Svetla Koeva. 2008. Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, 359–368.
- Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Tsvetana Dimitrova, and Maria Todorova. 2016. *Automatic Prediction of Morphosemantic Relations*. In *Proceedings of the Eighth Global WordNet Conference (GWC 2016)*:168–176.
- Karel Pala and Dana Hlaváčková. 2007. *Derivational relations in Czech WordNet*. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*:75–81.
- Maciej Piasecki, Radosław Ramocki, and Marek Maziarz. 2012. *Recognition of Polish Derivational Relations Based on Supervised Learning Scheme*. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*:916–922.
- Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: De Gruyter.
- Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, Vol 0, No 1:111–142.
- Jaen van Marle. 1985. *On the Paradigmatic Dimensions of Morphological Creativity*. Dordrecht: Foris.

Mapping WordNet Instances to Wikipedia

John P. McCrae

Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
john@mccr.ae

Abstract

Lexical resources differ from encyclopaedic resources and represent two distinct types of resource covering general language and named entities respectively. However, many lexical resources, including Princeton WordNet, contain many proper nouns, referring to named entities in the world yet it is not possible or desirable for a lexical resource to cover all named entities that may reasonably occur in a text. In this paper, we propose that instead of including synsets for instance concepts PWN should instead provide links to Wikipedia articles describing the concept. In order to enable this we have created a gold-quality mapping between all of the 7,742 instances in PWN and Wikipedia (where such a mapping is possible). As such, this resource aims to provide a gold standard for link discovery, while also allowing PWN to distinguish itself from other resources such as DBpedia or BabelNet. Moreover, this linking connects PWN to the Linguistic Linked Open Data cloud, thus creating a richer, more usable resource for natural language processing.

1 Introduction

Princeton WordNet (Fellbaum, 2010; Miller, 1995, PWN) and Wikipedia, especially in machine readable form such as DBpedia (Lehmann et al., 2015), are two of the most widely used resources in natural language processing. The nature of these resources is distinct, with WordNet constituting a lexicon of words in the English language and Wikipedia being an encyclopedia describing entities in the world. This means that WordNet should contain all the common nouns, verbs, adjectives and adverbs and Wikipedia should contain

the proper nouns referring to notable entities in a text. However, in fact there is a significant overlap between these two resources as Wikipedia contains pages for abstract general concepts, such as “play”¹, while PWN contains many proper nouns for concepts such as *Paris*, for which PWN has four synsets for the city in France (i83645), the city in Texas (i84698), the mythological prince (i86545) and a plant (i102495). In the case of WordNet, the choice of which proper nouns to include has had certain biases, for example there are many synsets for cities in the United States, e.g. *Paterson, New Jersey* (i84527), but not for *Kawasaki*, a city in Japan that is ten times larger. If however, PWN were to expand to include more proper nouns, it would lead to a much larger resource that would overlap significantly in its coverage with DBpedia. In fact, there have been several attempts to automatically create such a resource, most notably BabelNet (Navigli and Ponzetto, 2012) and UBY (Gurevych et al., 2012), however these resources have to rely on automatic alignment of the concepts. Instead, we propose that the concepts for named entities can be mapped to Wikipedia and that these concepts can thus be removed or replaced with links in future versions of PWN. Since PWN is created by careful manual effort, it is clear that an automatic mapping would not be compatible with the nature of PWN. Instead, as a principal contribution of this paper, we present the first manually created mapping between PWN instances and Wikipedia articles. This could be further used to link PWN to other resources including WikiData and GeoNames as well as help in the automatic translation of parts of WordNet.

In this paper, we first define the scope of the problem, in particular in terms of the number of instances and proper nouns that exist in PWN and

¹[https://en.wikipedia.org/wiki/Play_\(activity\)](https://en.wikipedia.org/wiki/Play_(activity))

their distribution. We then review some existing work on mapping PWN and Wikipedia instances. We present our method of linking, that uses Wikipedia categories to propose an alignment between sets of concepts simultaneously and the tool we created based on this that allows our annotators to quickly map the concepts between one resource and another. Finally, we present the results of our annotation, in particular in terms of the total effort and work required to create this mapping and conclude with some discussion and analysis of the results.

2 On Proper Nouns in WordNet

Princeton WordNet is a lexicon, that consists of a graph of *synsets*, which are collections of words that are synonymous, linked by a number of properties. All words in a synset have the same part-of-speech, however unfortunately there is only a single category for nouns and in fact synsets may contain a mixture of proper and common nouns, e.g., *Caterpillar,cat* (i51642). The links in the graph are of different types and the link `instance_hyponym` links a synset to a concept that is an *instance of* (Miller and Hristea, 2006), giving a limited set of proper nouns that we can systematically identify. There are in total 7,742 synsets in PWN which are instance hypernyms of 946 synsets and these will be the main focus of our work. Of these nearly all contain words starting with a capital letter, and of the 16 that don't, can be explained as follows: 7 are not capitalized for orthographic reasons, e.g., *al-Muhajiroun*, 6 should be capitalized but are not in WordNet, e.g., *pampas*, 2 should not be instance hypernyms but instead normal hypernyms *isle,islet* (i85598) and *sierra* (i86184) and 1 *church mouse* (i48540) is likely erroneous. As such, we can say that the set of synsets that are marked as instance hypernyms of a concept are all named entities in the world. However, there are many other synsets that contain one or more capitalized word as an entry and it is clear that we are not capturing all the proper nouns in PWN. In particular, there are a large number of capitalized words that refer to names of species or other terms in the Linnaean Taxonomy, e.g., *Felis catus* or *genus Hydrangea* and these are not instances of another synset and often share a synset with common nouns, e.g., *domestic cat,house cat,Felis domestics,Felis catus* (i46594). In addition,

there are several other large categories of proper nouns that are not captured by this approach especially beliefs, e.g., *Buddhism* (i79765) and languages, e.g., *German,High German,German language* (i73125). However, simply using the capitalization to detect proper nouns produces a lot of false positives, including acronyms and terms including a proper noun such as *Scotch terrier, Scottish terrier, Scottie* (i46443). As such, for this work we have focussed only on the synsets which are instances of synsets, as these are the terms that seem to be most encyclopedic in their content. A breakdown of the major synsets is given in Figure 1, and as we can see the major categories are (i35562), which is named people, (i35580), which is named places. A few other categories that have large number of entities include rivers and other geological features ((i85104),(i85439) and (i85674)), gods (i86570), events, especially wars (i35586), social groups, such as terrorist organizations (i79103) and books (i69848).

3 Related Work

The goal of mapping WordNet to Wikipedia has been recognized as an important one, however most of the focus has so far been on the automatic creation of mappings between the two resources, and this has led to the creation of wide-coverage lexicons that are useful for NLP applications but cannot act as a gold standard for NLP in the same way that WordNet does. The most notable such resources is BabelNet (Navigli and Ponzetto, 2012), whose mapping of WordNet to Wikipedia is based on the use of a word-sense disambiguation algorithm, where contexts are created for the Wikipedia and WordNet entities by means of using the surrounding synsets and the article texts. A second step then selects the highest scoring mapping based on structuring the Wikipedia page content using WordNet relations. The authors report a maximum F-Measure of 82.7% with a precision of 81.2%, showing that while BabelNet is a high-quality resource, it cannot be considered a gold standard. This method improved on a previous approach by these authors (Ponzetto and Navigli, 2009), which used the taxonomic structure of the resources. Another method to link WordNet and Wikipedia has been through Personalized Page Rank (Agirre and Soroa, 2009), which was first attempted as a method for linking these re-

i35545 - entity (7742)

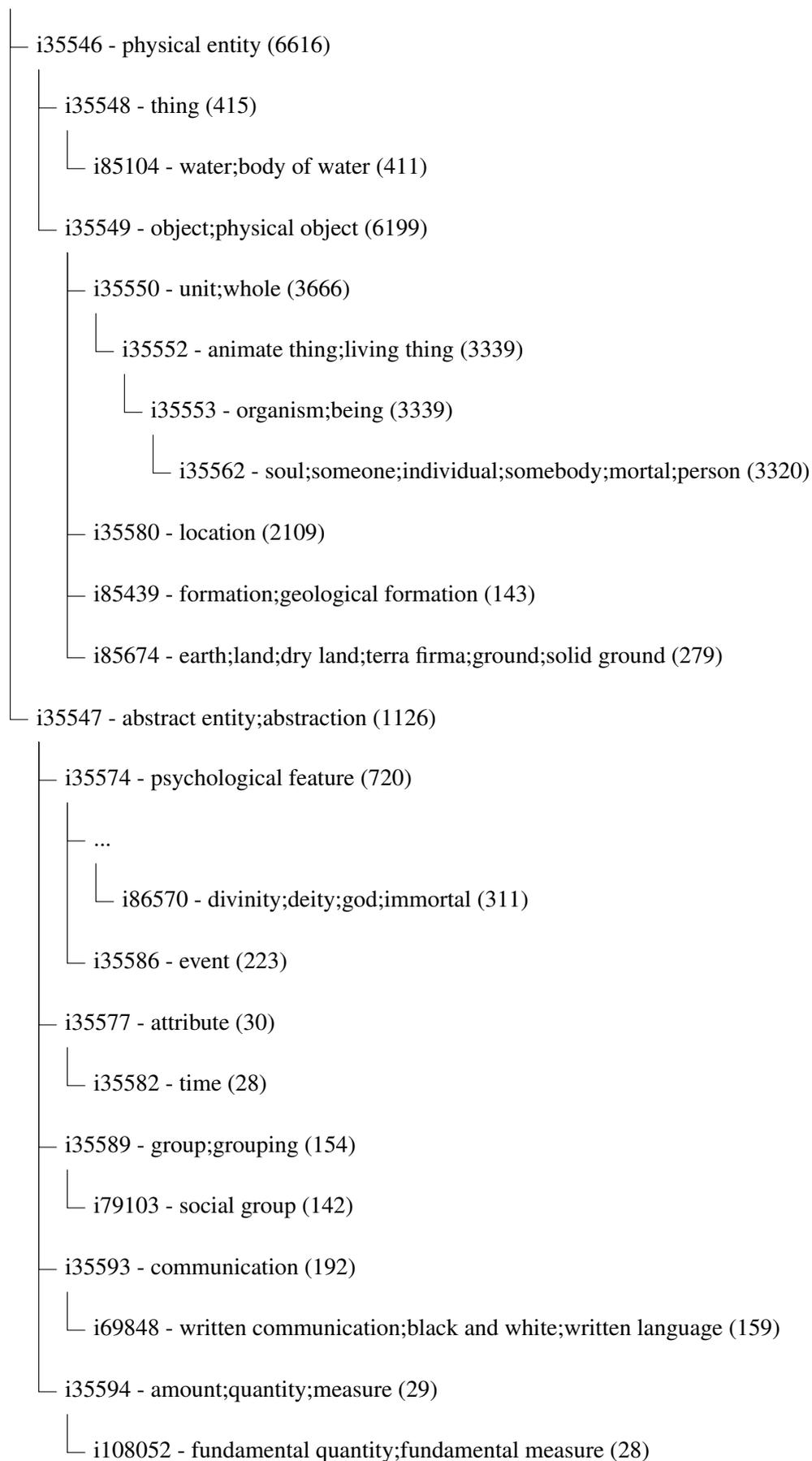


Figure 1: The most frequent hypernyms of instances in Princeton WordNet

sources in (Toral et al., 2009) and then was further improved by (Niemann and Gurevych, 2011), by the introduction of “thresholds”. Niemann and Gurevych’s methodology forms the basis of the UBY resource (Gurevych et al., 2012). Finally, Fernando and Stevenson (Fernando and Stevenson, 2012) proposed using semantic textual similarity methods and showed results that obtained an F-Measure of 84.1% outperforming Ponzetto and Navigli’s approach. Notably, this work also created a gold standard of Wikipedia-WordNet mappings that can be used for evaluation of further approaches to linking. However, this mapping is only of 200 words and as such is not on the same scale as the resource introduced in this paper.

Another large-scale resource that has been constructed by combining WordNet and Wikipedia is Yago (Suchanek et al., 2008; Suchanek et al., 2007), which created an ontology of concepts created from Wikipedia categories. This showed a very high accuracy in the mapping of concepts (97.7%), however this does not deal with the actual entities as in this work.

WordNet has also been linked to a number of other lexical resource by a variety approaches, including SemCor (Mihalcea and Moldovan, 2000), where texts were annotated with WordNet synset identifiers and this was used as a basis to create links to other resources including FrameNet (Baker et al., 1998, FN) and VerbNet (Schuler, 2005, VN), which were linked in (Shi and Mihalcea, 2005). Another linking was created by the SemLink (Palmer, 2009; Bonial et al., 2013), also based on the annotation of a corpus with PWN, FN and VN. Finally, mappings have also been proposed between WordNet and Wiktionary², a free dictionary from the Wikimedia Foundation, in works such as (McCrae et al., 2012) and (Meyer and Gurevych, 2011).

4 Mapping WordNet to Wikipedia

Our goal is to create a large manual mapping between a subset of Princeton WordNet and Wikipedia, however simply identifying this subset and starting annotation is not a suitable approach as looking up each WordNet synset in Wikipedia and recording the results would be a slow and dull process. We could try to improve this by matching the lemmas of WordNet entries to the titles of Wikipedia articles, but this would have a very

low coverage as the article title for a Wikipedia article must be unique so often includes specific disambiguating terms. To expand the coverage of this we consider a WordNet lemma to match a Wikipedia article if it matches the title ignoring case before the first comma or parentheses or any page that redirects to this article. Thus, we would match the lemma “Paris” to the page titles “Paris”, “Paris, Texas” and “Paris (Mythology)”. In addition, we also included information from disambiguation pages, as collected by DBpedia (Lehmann et al., 2015)³. This method captures most of the mappings as only 77 WordNet synsets have no candidates in Wikipedia, however it also creates significant ambiguity with an average of 21.6 candidates for each synset. For these reasons, we try to resolve these differences by suggesting category mappings, inspired by (Suchanek et al., 2008).

4.1 Unambiguous Category Matches

We start by considering all pairs of WordNet instance synsets and Wikipedia articles as $W = \{s_i, a_j\}$. Let all hypernyms of a synsets be the set of $H(s_i)$ and let all categories for a Wikipedia article by $C'(a_j)$. We also consider all categories of categories and all categories of those categories to create a list of categories $C(a_j)$, as the categories for some articles can be very narrow. The set of mappings between non-instance synsets and Wikipedia categories is created as follows:

$$M = \{h, c | \exists \{s_i, a_j\} \in W : h \in H(s_i) \wedge c \in C(a_j)\}$$

This creates a very large number of mappings and we wish to choose which mappings are most suitable, thus we create a score to rank them. We use two main constraints to do this, firstly, we note that short lemma matches tend to be quite ambiguous, e.g., “Paris, Texas” is less ambiguous than “Paris”, and secondly, we notice that mappings that create a lot of duplicate matches are challenging to annotate. Firstly, we define $l(s_i, a_j)$ as the follows, where $L(s_i, a_j)$ is the set of matching terms between the WordNet instance and the Wikipedia article, $t(l)$ gives the length (number of tokens) of this matching terms in this mapping and α is a constant:

³In particular the file disambiguations.en.ttl.gz

²<http://en.wiktionary.org>

$$l(s_i, a_j) = \sum_{l \in L(s_i, a_j)} t(l) - \alpha.$$

Secondly, we generate a set of proposed mappings based on a hypernym, $h \in H(s_i)$ and a Wikipedia category $c \in C(a_j)$ as follows

$$P(h, c) = \{(s, a) | h \in H(s) \wedge c \in C(a) \wedge L(s, a) \neq \emptyset\}$$

We say that a pair (s, a) is *unambiguous* in $P(h, c)$ if there is no distinct element $(s', a') \in P(h, c)$ such that $s = s'$ or $a = a'$. Finally, we score a mapping as follows:

$$s(h, c) = \sum_{(s, a) \in P(h, c)} \sigma(s, a)$$

$$\sigma(s, a) = \begin{cases} l(s, a) & \text{if } (s, a) \text{ is unambiguous} \\ & \text{in } P(h, c) \\ -\beta & \text{otherwise} \end{cases}$$

For parameters we chose $\alpha = 1$, as this allows us to ignore mappings created from single tokens and $\beta = 10$ as this provided a good trade-off between allowing some ambiguity in the mappings. In fact, the first 2,500 entries were annotated with a higher β value, but it became clear that this was too strict so we permitted more ambiguity in the mapping.

4.2 Annotation Tool

In order to create the annotations a tool was created to show the proposed mappings, which is depicted in Figure 2. This tool shows all the proposed category mappings and then all the individual instances and Wikipedia articles that will be linked. For each WordNet instance the definition in WordNet is given and for the Wikipedia article, its first paragraph is given. For each case, we selected whether the mapping was valid and then submitted the proposed mapping. The system allows two extra actions, “Reject”, which is the same as unselecting all mappings and submitting the form and “Reject Wikipedia Category”, which removes all mappings involving this Wikipedia category. This option was introduced as some Wikipedia categories were clearly not likely to map to any synsets in Wikipedia⁴.

⁴An example is https://en.wikipedia.org/wiki/Category:Timelines_of_cities_in_France

5 Resource and Evaluation

We used the above described methodology to annotate the vast majority of the mappings (7,582 mappings), while the remaining 239 synsets had no good candidates in Wikipedia, principally due to spelling variants and this includes the 77 synsets with no candidates and other synsets for which the category approach did not work. These remaining 239 synsets were then mapped directly (on a spreadsheet). We also used this pass to sort the links into the following types:

Exact The WordNet synset and Wikipedia article exactly describe the same entity.

Broad The Wikipedia article describes several things, of which the entity described by the WordNet synset is only one of. An example of this is the Wikipedia article for the “Wright Brothers”⁵, which is linked broader to two WordNet synsets for each brother. In this case, Wikipedia redirects “Orville Wright” and “Wilbur Wright” to this article.

Narrow The opposite of ‘broad’, i.e., the WordNet synset describes multiple Wikipedia articles. An example is *Rameses*, *Ramesses*, *Ramses* (i96663) defined as “any of 12 kings of ancient Egypt between 1315 and 1090 BC”⁶, while each is a separate Wikipedia article.

Related The Wikipedia article does not describe the WordNet synset but something intrinsically linked to it, and the lemmas of the WordNet synset have redirects to this article. For example *Hoover*, *William Hoover*, *William Henry Hoover* (i95579) is mapped to “The Hoover Company” describing the company he founded. Wikipedia also redirect “William Hoover” to this article.

Unmapped A small number of entities in WordNet were not possible to map to Wikipedia, either because the synset was not in Wikipedia (this was the case for many terrorist organizations), the description and name did not match anything in Wikipedia (for a

⁵https://en.wikipedia.org/wiki/Wright_brothers

⁶This also an error as there are only 11 Egyptian pharaohs named Ramesses

Proposed Mapping

WordNet

wn31-00001740-n entity

that which is perceived or known or inferred to have its own distinct existence (living or nonliving)

Wikipedia

People by country of descent

Reject

Reject Wikipedia Category

WordNet	Wikipedia	Accept
(wn31-10953409-n) Princess of Wales, Lady Diana Frances Spencer, Diana, Princess Diana English aristocrat who was the first wife of Prince Charles; her death in an automobile accident in Paris produced intense national mourning (1961-1997)	Diana, Princess of Wales <i>Diana, Princess of Wales (Diana Frances; née Spencer; 1 July 1961 – 31 August 1997), was the first wife of Charles, Prince of Wales, who is the eldest child and heir apparent of Queen Elizabeth II. Diana was born into a family of British nobility with royal ancestry as The Honourable Diana Spencer. She was the fourth child and third daughter of John Spencer, Viscount Althorp and the Honourable Frances Roche. She grew up in Park House, situated on the Sandringham estate, and was educated in England and Switzerland. In 1975, after her father inherited the title of Earl Spencer, she became Lady Diana Spencer.</i>	<input checked="" type="checkbox"/>
(wn31-10953680-n) Duchesse de Valentinois, Diane de Poitiers French noblewoman who was the mistress of Henry	Princess Charlotte, Duchess of Valentinois <i>Princess Charlotte of Monaco, Duchess of Valentinois (Charlotte Louise Juliette Grimaldi de Monaco; 30 September 1898 – 15 November</i>	<input checked="" type="checkbox"/>

Figure 2: The Annotation Tool used to create the mappings

Exact	Broad	Narrow	Related	Unmapped
7,582	54	21	30	59

Table 1: The size of the resource by type of link

few place names) or the synset was not something that would generally be in Wikipedia, e.g., different names for gods, such as *Jupiter Fidius, Protector of Boundaries* (i86982)

We used the following heuristic to help with this mapping. If the Wikipedia page title exactly matched one of the lemmas or the Wikipedia article was of the form “X, Y” or “X (Y)” and X was one of the lemmas and Y occurred in the definition of the synset, we accepted it as an exact match⁷. For example, this allowed us to easily validate the mappings for the Wikipedia articles “Paris” (the capital of France), “Paris, Texas” and “Paris (mythology)”. All other mappings (1,733) were manually assigned one of the above cate-

⁷As an aside, this heuristic of matching the differentiating part of the title to the WordNet definition may have been quite effective for establishing mappings in Section 4.1, but was not considered until most of the mapping was completed. In this paper, we focus on the construction of the resource and describe the methodology we followed.

gories. As a result of this mapping process we also detected 56 errors (0.7%) and improved 11 mappings, by which we mean that we changed a broader/narrower link to an exact link. For example, the synset *Downing Street* (i83390), was moved from “10 Downing Street” to “Downing Street”. The complete size of each of these categories is given in Table 1, in a few cases a wordnet synset was mapped using “narrower” to multiple Wikipedia articles thus the 7,742 entities created 7,746 links.

5.1 Improvements to Princeton WordNet

In the process of creating the mappings between PWN and Wikipedia, we closely studied a section of Princeton WordNet and thus found a large number of errors within the resource. As such we submitted a report to the developers of Princeton WordNet detailing the following errors⁸:

- Two synsets were identified to be duplicates (referring to the same concept).
- One synset was suggested to be split

⁸This document may be viewed at https://docs.google.com/document/d/lyn-UurCoeuKk_OwRzDaj1dYW88k2l0ymD7YtBIiVlCM

- 17 lemmas with typos were detected
- Two links were found to be incorrect
- Four synsets described concepts for which no reference could be found outside of PWN
- 41 definitions were found to be factually inaccurate, this was mostly due to the year that a person was born in or died in not being correct.
- We suggest 1,062 new synset members to be added to existing synsets. These were derived from the Wikipedia page titles and so represent standard well-attested variants of existing names. These primarily consist of variations of names, e.g., “University of Cambridge” is the official name for *Cambridge, Cambridge University* (i51397), but in some cases are more significant, e.g., *Seaward’s Folly* (i41225) is more commonly known as the “Alaska Purchase”.

5.2 Resource

The mapping has been created and is made available from the following URL⁹. In addition, the mapping will be contributed to the Global WordNet Index (Bond et al., 2016; Vossen et al., 2016) and as a mapping to the DBpedia project¹⁰. In this case, we provide an RDF file that links the Global WordNet ILI URIs with DBpedia URIs. The mapping is made available under a CC-Zero license to enable its re-use in as many places as possible. The source code for tools used in this project are available on GitHub¹¹.

6 Conclusion

We have presented a new mapping of all the instances in WordNet to Wikipedia articles. This represents the largest gold standard mapping for tasks such as link discovery (Nentwig et al., 2017) and is likely to be a basic resource for many tasks in natural language processing. For the future development of Princeton WordNet as a resource, this mapping can form the basis by which PWN can distinguish itself from an encyclopedia, by replacing the instance links with direct links to

⁹<http://jmccrae.github.io/wn-wiki-instances/ili-map-dbpedia.ttl>

¹⁰<http://github.com/dbpedia/links>

¹¹<https://github.com/jmccrae/wn-wiki-instances>

Wikipedia. Moreover, by linking to Wikipedia articles, we can further link to many other resources, for example it is only a matter of changing the URL to find a DBpedia entity that can be used to find machine readable information about the data. Furthermore, all Wikipedia articles are now linked to WikiData entities, so we can easily find that *Paris, City of Light, French capital, capital of France* (i83645) is linked to WikiData entity Q90¹² and then this can give us identities in many other databases including GeoNames (2968815), OpenStreetMap (71525) and even the official Twitter account (@Paris). Finally, it is worth noting that Wikipedia and Wikidata also contains links to these concepts in other languages, and as such, this linking can create a partial translation of a section of WordNet. As such, this transforms WordNet into a richer linked resource that can be part of the Web of Linguistic Linked Open Data (McCrae et al., 2016).

Acknowledgments

This work was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

- [Agirre and Soroa2009] Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.
- [Baker et al.1998] Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.
- [Bond et al.2016] Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.
- [Bonial et al.2013] Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising sem-link. In *The GenLex Workshop on Linked Data in Linguistics*.
- [Fellbaum2010] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

¹²<https://www.wikidata.org/wiki/Q90>

- [Fernando and Stevenson2012] Samuel Fernando and Mark Stevenson. 2012. Mapping WordNet synsets to Wikipedia articles. In *Proceedings of The Eighth International Conference on Language Resources and Evaluation*, pages 590–596.
- [Gurevych et al.2012] Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. UBY: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.
- [Lehmann et al.2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.
- [McCrae et al.2012] John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*, pages 25–34.
- [McCrae et al.2016] John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. 2016. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *10th Language Resource and Evaluation Conference (LREC)*.
- [Meyer and Gurevych2011] Christian M Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *IJCNLP*, pages 883–892.
- [Mihalcea and Moldovan2000] Rada Mihalcea and Dan Moldovan. 2000. Semantic indexing using WordNet senses. In *Proceedings of the ACL-2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*, pages 35–45. Association for Computational Linguistics.
- [Miller and Hristea2006] George Miller and Florentina Hristea. 2006. WordNet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.
- [Miller1995] George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- [Navigli and Ponzetto2012] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
- [Nentwig et al.2017] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. 2017. A survey of current link discovery frameworks. *Semantic Web*, 8(3):419–436.
- [Niemann and Gurevych2011] Elisabeth Niemann and Iryna Gurevych. 2011. The people’s web meets linguistic knowledge: automatic sense alignment of Wikipedia and Wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 205–214. Association for Computational Linguistics.
- [Palmer2009] Martha Palmer. 2009. Semlink: Linking propbank, verbnnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. Pisa Italy.
- [Ponzetto and Navigli2009] Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *IJCAI*, volume 9, pages 2083–2088.
- [Schuler2005] Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.
- [Shi and Mihalcea2005] Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. *Computational linguistics and intelligent text processing*, pages 100–111.
- [Suchanek et al.2007] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.
- [Suchanek et al.2008] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A large ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.
- [Toral et al.2009] Antonio Toral, Oscar Ferrández, Eneko Agirre, Rafael Munoz, Informatika Fakultatea, and Basque Country Donostia. 2009. A study on linking Wikipedia categories to WordNet synsets using text similarity. In *Proceedings of the 2009 Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 449–454.
- [Vossen et al.2016] Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.

Mapping WordNet Concepts with CPA Ontology

Svetla Koeva, Tsvetana Dimitrova, Valentina Stefanova, Dimitar Hristov

Institute for Bulgarian Language

Bulgarian Academy of Sciences

svetla, cvetana, valentina, dimitar@dcl.bas.bg

Abstract

The paper discusses the enrichment of WordNet data through merging of WordNet concepts and Corpus Pattern Analysis (CPA) semantic types. The 253 CPA semantic types are mapped to the respective WordNet concepts. As a result of mapping, the hyponyms of a synset to which a CPA semantic type is mapped inherit not only the respective WordNet semantic primitive but also the CPA semantic type.

1 Introduction

The paper presents/discusses an effort on enriching the data in WordNet and the links between WordNet concepts through expansion of the number of noun semantic classes through/by mapping the WordNet data (Miller et al., 1990) with the data in another resource – the Pattern Dictionary of English Verbs (PDEV) (Hanks, 2004; Hanks, 2005; Hanks, 2008).

WordNet synsets are classified into semantic primitives (also called semantic classes). Verbs and nouns are distributed into more elaborate classes (Miller et al., 1990), with corresponding labels (noun.person, noun.animal, noun.cognition; verb.cognition, verb.change, etc.) being assigned to them. The information about semantic primitives has been used in a number of efforts to verify/test and enrich semantic relations between noun and verb synsets (such as the type of morphosemantic relations – Agent, Undergoer, Instrument, Event, etc. – that link verb/noun pairs of synsets that contain derivationally related literals) (Fellbaum, 2009).

The semantic classification of WordNet nouns and verbs is consistent and useful for many language processing tasks. However, the natural language understanding and generation requires a precise and granular prediction for the set of concepts

that could saturate the arguments of a verb. Consider the verb {read:5} 'interpret something that is written or printed' and its sentence frame *Somebody* —s *something*. Obviously, not every noun classified as *noun.person* will/can be selected/collocate by/with the verb {read:5} as its subject and not every noun that is not classified as *noun.person* can be the object of the verb. Therefore, we assume that the WordNet noun semantic classes can be further specified in order to correlate more precisely with the verb-noun selecting requirements. To sum up, although the information is readily available in WordNet, not all useful information is explicitly accessible.

In this paper, we present an effort at mapping the WordNet concepts with the Corpus Pattern Analysis (CPA) semantic types that are part of the Pattern Dictionary of English Verbs (PDEV). PDEV is built on the basis of the lexicocentric Theory of Norms and Exploitations (Hanks, 2013) and exploits the CPA mechanism to map meaning onto words in text. PDEV consists of verb patterns and semantic types of their nominal arguments organized within the so-called CPA ontology.

Our goal is then twofold: to identify the concept or the set of concepts to which a given CPA semantic type corresponds and to explore the structures of the two hierarchies: WordNet semantic primitives and CPA semantic types.

The paper is organized as follows: in section 2, we present our motivation for the work before discussing different attempts at semantic classification of nouns in section 3. Section 4 briefly presents the CPA ontology, while section 5 outlines some issues with the WordNet noun hierarchy. The effort at mapping the CPA semantic types and WordNet concepts is discussed in section 6, with a comparison between the two structures in section 7 and some preliminary conclusions; our plans for future work are given in section 8.

2 Motivation

There are many examples, such as in (1) where the sentence frame in (1a) signals that the verb can have both human and non-human subject argument. Further, (1c), which has a definition comparable to (1a), leaves only non-human subject argument. In addition, the non-human subject arguments both in both (1b) and (1c) may both be specified as animate.

(1)

a. {purr:1, make vibrant sounds:1} 'indicate pleasure by purring; characteristic of cats'

Something —s; *Somebody* —s

b. {moo:1, low:4} 'make a low noise, characteristic of bovines'

Something —s

c. {meow:1, mew:1} 'cry like a cat; the cat meowed'

Something —s

Noun semantic primitives cannot be employed for detailed selectional restrictions on arguments because their organization is too general and some semantic classes can be missing or inappropriate. For example, the sentence frames in (2) do not specify that the verbs can be combined with nouns like *idea* (noun.cognition), *result* (noun.communication), *victory* (noun.event) but cannot co-occur with nouns such as *stone*, *table*, *sky*, etc.

(2)

{achieve:1, accomplish:2, attain:4, reach:9} 'to gain with effort'

Somebody —s *something*

Something —s *something*

Somebody —s *that CLAUSE*

To find a match between nouns and verbs, we hypothesize that verb hypernym/hyponym trees combine verbs with similar or equivalent semantic and syntactic properties.

Further, it can be tested whether verb synsets combine with noun classes that can be identified within the WordNet structure if a more detailed classification of nouns (which further specifies the semantic classes) – in line with the CPA semantic types ontology – is provided. Here, we present our work on mapping the WordNet concepts and the CPA semantic types.

Previous work on mixing resources and enriching the information on semantic and syntactic behavior of verbs encoded in WordNet builds upon resources – one or more than one – that use (Levin, 1993)'s verb classes (Dorr, 1997; Korhonen, 2002; Green et al., 2001). Proposals involve mixing up information from WordNet and Longman Dictionary of Contemporary English (Dorr, 1997; Korhonen, 2002); VerbNet (also based on Levin's classes) and FrameNet (Shi and Mihalcea, 2005); and VerbNet and PropBank (Pazienza et al., 2006). To the best of our knowledge, however, WordNet concepts and CPA ontology have not been mapped and compared yet, and below we propose such an effort.

3 Semantic classes of nouns

Although WordNet nouns are classified in a number of classes labeled by semantic primitives, numerous linguistic works argue that nouns have referential value and cannot be reduced to a set of primitives.

(Wierzbicka, 1986) claims that most (prototypical) nouns identify a certain kind of entity, a concept, but positively and not in terms of mutual differences. Thus, the function of a noun is to single out a certain kind of entity and its meaning cannot be reduced to any combination of features though it may be described using features.

In numerous works, (Wierzbicka, 1984; Wierzbicka, 1985) enumerates features such as shape, size, proportions, function, etc. that can be used in definitions of objects but in a semantic formula, these features have to be subordinated to a general taxonomic statement. For example, in conceptual representation of count/mass nouns, (Wierzbicka, 1988) motivates 14 classes of language terms, with each class being conceptually motivated by the following factors: (A) perceptual conspicuousness (depending on the use of aggregates); (B) arbitrary divisibility (whether the entity can be divided into portions of any size which are still classified as the original entity, e.g., *machine* vs. *butter*); (C) heterogeneity (whether the entities making a group are of the same or different kind); and (D) how humans interact with the entity (whether they can be seen as individuals or not, e.g., *rice* vs. *pumpkin*).

Additional efforts on noun classification are based on distribution of nouns in corpora and information (cues) from the context to extract information

about the noun (lexical) classes, description and their behaviour.

To test the plausibility of the distributional hypothesis, Hindle (1990) attempts at quasi-semantic classification of nouns observing similarity of nouns based on distribution of subject, verb, object in a corpus. This distributional hypothesis defines reciprocally most similar nouns or reciprocal nearest neighbours – a set of substitutable words, many of which are near synonyms, or closely related.

(Bel et al., 2012) propose a cue-based automatic noun classification in English and Spanish which uses previously known noun lexical classes - event, human, concrete, semiotic, location, and matter. The work is based mainly on (Harris, 1954)'s distributional hypothesis and markedness theory of the Prague Linguistic School, and assumes that lexical semantic classes are properties of a number of words that recurrently co-occur in a number of particular contexts (Bybee, 2010). They use aspects of linguistic contexts where the nouns occur as cues – namely, predicate selectional restrictions (verbal and non-verbal elements such as adjectives and nouns they combine with), grammatical functions, prepositions, suffixes – that represent distributional characteristics of a specific lexical class.

(Bel et al., 2007) work on the acquisition of deep grammatical information for nouns in Spanish using distributional evidence as features and information about all occurrences of a word as a single complex unit. These effort employs 23 linguistic cues for classifying nouns according to an HPSG-based (Head-driven phrase structure grammar) lexical typology (namely the lexicon of an HPSG-based grammars developed in the LKB (Linguistic Knowledge Builder) platform for Spanish). Grammatical features that conform to the cross-classified types are used as they are considered a better level of generalization than the type. These are namely: *mass* and *countable*; plus three additional types for subcategorization: *trans* (nouns with thematic complements introduced by the preposition *de*); *intrans* (noun that has no complements); *pcomp* (where the complements of the noun are introduced by a bound preposition). The combination of features corresponds to the final type.

Our effort as presented here is based on comparison of the semantic primitives of the nouns in

WordNet and the semantic types within the CPA ontology as used in PDEV, in order to outline the directions for further specifying the WordNet semantic classes.

4 CPA ontology

PDEV framework relies on semantic categories called semantic types, which refer to properties shared by a number of nouns that are found in verb pattern (argument) positions. Semantic types are formulated when they have been repeatedly observed in patterns and are organized into a relatively shallow ontology (up to 10 sublevels for some types) – a portion of the ontology – under the type [Liquid] is exemplified on Fig. 1.

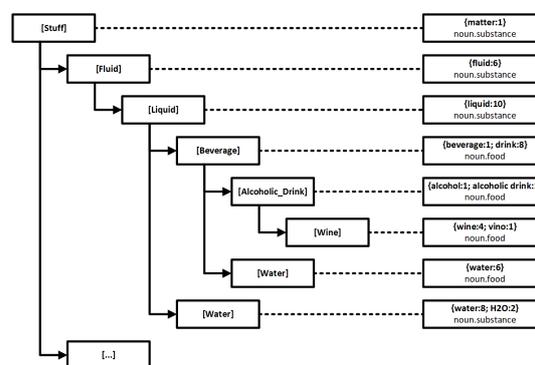


Figure 1: Part of the CPA ontology

On the other hand, some concepts are classified taking into account different properties, such as with drinks – [Beverage] is classified as both [Physical Object] [Inanimate] [Artifact] and [Physical Object] [Inanimate] [Stuff] [Fluid] [Liquid]. As in other ontologies, each semantic type inherits the formal property of the type above it in the hierarchy (Cinkova and Hanks, 2010).

The CPA ontology is language dependent: there are senses of verbs such as *bark* or *saddle* that evoke [Dog] or [Horse] as semantic types because in English there are many words that denote horses and dogs, but there are no verbs that require a distinction between jackals and hyenas, so these are not semantic types (Cinkova and Hanks, 2010).

Though a semantic type usually involves more members than are actually observed in a given pattern position, some words are preferred to others with specific patterns. Therefore, an appropriate level in the ontology should be chosen (the very abstract types such as [Anything] are usually too broad). Thus, the patterns often involve alternative semantic types and not a category, as in the

pattern of the verb *eat*: *[Human] or [Animal] or [Animate] eats ([Physical Object] or [Stuff])*. The alternative larger type can involve types from different levels of the ontology but also can be a type and its supertype. The latter instances are found when a semantic type is predominantly observed in a given pattern position, even if the higher type is also found in the same position.

One of the main indicators of the reliability of semantic types is the fact that they are corpus-driven – they are formulated on the basis of real examples encountered in corpora. Although the semantic types represent cognitive concepts that play a central role in the way words are used, they remain abstract notions as they are not linked to sets of concrete concepts and their lexical representations. Mapping CPA with WordNet will provide sets of concepts and their lexical representations linked to the CPA semantic types.

In addition, in CPA, a single lexical item or a small group of lexical items (called lexical set) that fulfill a role in the clause are included in the verb patterns but not within the ontology (as in: *[Fish] breathes (through gills); [Human] or [Animal] breathes air or dust or gas or [Vapour] (in)*). However, for a precise semantic analysis small sets of lexical items should be represented within the ontology, which implies that the WordNet is the best candidate for full representation of the semantic types ontology.

5 WordNet noun hierarchy

Noun synsets in WordNet are organized into 26 semantic classes (the so-called semantic primitives (Miller et al., 1990)), namely nouns denoting humans (noun.person), animals (noun.animal), plants (noun.plant), acts or actions (noun.act), feelings and emotions (noun.feeling), spatial position (noun.location), foods and drinks (noun.food), etc.

The synsets labeled noun.Tops are the top-level synsets in the hierarchy, the so-called unique beginners for nouns. Thus, the noun synsets are divided into (sub-)hierarchies under the unique noun.Tops labeled synset {entity:1} which has three hyponyms – two unique beginner synsets {physical entity:1} and {abstraction:1; abstract entity:1} and a noun.artifact labeled hyponym {thing:4}. Each of these synsets instantiates a sub-hierarchy. Some of the hyponyms in these sub-hierarchies are also unique beginners. The

hyponyms of the {physical entity:1} synset are:

{thing:1} – noun.Tops containing hyponyms labeled as noun.object;
 {object:1; physical object:1} – noun.Tops, containing hyponyms that are noun.objects and noun.artifacts;
 {causal agent:1; cause:1; causal agency:1} – noun.Tops, containing as hyponyms synsets labeled noun.person, noun.phenomenon, noun.state, noun.object, and noun.substance;
 {matter:1} – noun.substance, containing hyponyms that are noun.substance and noun.object;
 {process:1; physical process:1} – noun.process, with hyponyms marked as noun.process and noun.phenomenon;
 {substance:7} – noun.substance (a sole synset).

Hyponyms of the {abstraction:1; abstract entity:1} synset are (all of these have hyponyms of various semantic class):

{psychological feature:1} – noun.attribute;
 {attribute:1} – noun.attribute;
 {group:1; grouping:1} – noun.group;
 {relation:1} – noun.relation;
 {communication:1} – noun.communication;
 {measure:7; quantity:1; amount:1} – noun.quantity;
 {otherworld:1'} – noun.cognition;
 {set:41} – noun.group.

Though, the basis of classification of certain entities may seem straightforward, it is possible for different entities to inherit information for their features from different (sub-)hierarchies and to have more than one hypernyms, as in (3):

(3)
 {person:1; individual:1; someone:1; somebody:1; mortal:1; soul:1}
 hypernym: {organism:1; being:1}
 hypernym: {causal agent:1; cause:1; causal agency:1}
 (.....)
 hypernym: {physical entity:1}

Additionally, however, there is the EuroWordNet top ontology which contains 63 semantic primitives (Vossen, 1999). The ontology is designed to help the encoding of WordNet se-

semantic relations in a uniform way. The 1st Order Entities are distinguished in terms of main ways of conceptualizing or classifying a concrete entity (Pustejovsky, 1995): Origin, Form, Composition, and Function. Further, Origin is further divided into Natural and Artifact, and Natural – into Living, Plant, Human, Creature, Animal and so on. The 2nd Order Entity is any static situation (property, relation) or dynamic situation, while the 3rd Order Entity is any unobservable proposition which exists independently of time and space (idea, thought).

The WordNet Noun Base Concepts (the most important meanings representing the shared cores of the different WordNets) were classified according to the 1st Order Entity, as follows (Vossen et al., 1998):

(4)

Artifact	{article:1}
Building+Group+Artifact	{establishment:2}
Building+Group+Object+Artifact	{factory:1}

The classification into more than one higher category is a promising approach which is partially followed in our current work.

6 Mapping CPA ontology and WordNet noun hierarchy

We mapped the WordNet noun synset hierarchy onto the semantic type hierarchy in the CPA ontology by matching the CPA semantic types with WordNet synsets and choosing those that are the most probable (and populated) ones, with non-exhaustive results (i.e., many concepts that can be classified under one semantic type, may be not matched under the chosen synsets and left out). Two independent annotators worked on this task and the cases of annotators disagreement were validated by a third one.

Out of 253 instances of matching (one semantic type to one, two, three or more WordNet concepts), there were 46 cases of disagreement between the two annotators; the third annotator worked only on the matches with disagreement, and proposed a new match in 10 instances (in the other cases, the third annotator accepted one of the two choices of the first two annotators; synsets for mapping were selected after an following agreement between the three annotators – in some cases, all suggestions were accepted as matching

options, while in other cases, the annotators agreed on some of the suggestions).

The following general principles were obeyed:

- The WordNet semantic primitives are always preserved.
- New semantic primitives borrowed from the CPA ontology (further called complementary semantic primitives) are supplied added in addition to the WordNet semantic primitives.

To coordinate their work, the annotators agreed on the following:

- The highest appropriate WordNet synset is chosen.
- If necessary, more than one WordNet synset is selected, – in such cases the union of the subtrees is accepted.
- All available PDEV patterns and corpus examples were checked observed to compare them with the WordNet hyponyms belonging to a chosen synset.

As a result of the mapping, the hyponyms of a synset to which a CPA semantic type is mapped, inherit not only the respective WordNet semantic primitive but also the CPA semantic type, as well. For example, all hyponyms of the WordNet synset {location:1} a point or extent in space are classified into with the semantic primitive noun.location. All hyponyms (such as *fact*, *example*, *evidence*, etc.) of the synset {information:2} knowledge acquired through study or experience or instruction mapped with the CPA semantic type [Information] inherit not only the WordNet semantic primitive (noun.cognition) but also the more specific type [Information]. This allows to better prediction for the words connectivity and thus to achieve better results in semantic parsing, word sense disambiguation, language generation and related tasks. The 253 CPA semantic types are mapped to the respective WordNet concepts (synsets) as follows: 199 semantic types are mapped directly to one concept, i.e., [Permission] is mapped to {permission:2} approval to do something, semantic primitive noun.communication; [Dispute] is mapped to {disagreement:2} the speech act of disagreeing or arguing or disputing, semantic prime noun.communication; 39 semantic types are

mapped to two WordNet concepts, i.e., [Route] is mapped to {road:2; route:4;} an open way (generally public) for travel or transportation semantic primitive noun.artifact, and {path:3; route:5; itinerary:3} an established line of travel or access, semantic primitive noun.location; 12 semantic types are mapped to three concepts; 2 semantic types are mapped to four concepts; and 1 semantic type is mapped to five concepts.

Automatic mapping of the hyponym synsets to the inherited CPA semantic types was performed. In the cases where a semantic type and its ancestor were both mapped to the same synset, the ancestor was removed. 82,114 WordNet noun synsets were mapped to the 253 semantic types of the CPA ontology, resulting in 172,991 mappings. As a number of semantic types are classified using different properties, some synsets were mapped to more than one instance of a semantic type, e.g., {phase:6; stage:10} was mapped to both [Abstract_Entity] [Time_Period] and [Abstract_Entity] [Resource] [Asset] [Time_Period]. As these are considered the same concepts, duplicates were removed, leaving 171,359 mappings. The resulting data is available online¹, marked with the XML tag CPA in the WordNet noun synsets.

7 Comparison between WordNet and CPA hierarchies

On the top levels, some classes show a fit between the semantic type and the top level synset, e.g., [Entity] and {entity:1} with subtypes [Abstract_Entity] and {abstract entity:1}, in the most cases the match is not on the same level of the respective hierarchies. For example, [Event] matches {event:1}, but [Event] is on the same level as [Abstract_Entity] in the CPA hierarchy, while {event:1} is linked to the noun.Tops {abstract entity:1} via {psychological feature:1}. Further, [Group] is on the same level as [Entity] but in WordNet {group:1, grouping:1}, which is also noun.Tops, is a hyponym of {abstract entity:1}. Nevertheless, from the fact that not each CPA semantic type can be mapped to one synset, it is clear that the respective nodes in the WordNet hierarchy represent semantic classes and their hyponyms inherit the semantic specifications of the specific semantic class.

If we assume that the concepts are divided into {abstract entity:1} and {physical entity:1} in

¹http://dcl.bas.bg/PWN_CPA/

WordNet, the types in CPA hierarchy will be marked as follows (we match the CPA subtypes in the respective subhierarchies with probable noun synset(s), which are linked to either of the two noun.Tops; some types below involve subtypes that are matched to WordNet concepts that can be traced back to both {abstract entity:1} and {physical entity:1}) – see on Fig. 2.

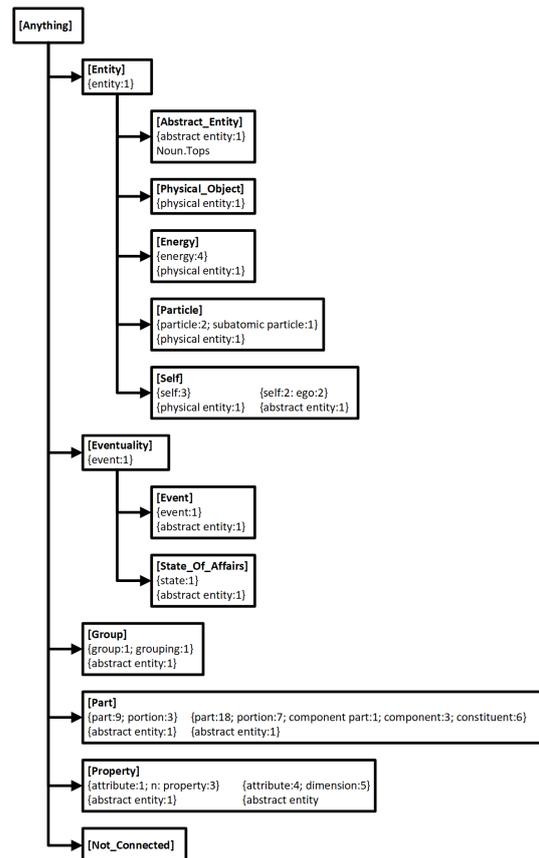


Figure 2: Matching

The matched synsets may be on different levels, and in (5), we exemplify some of the subtypes of the [Artifact] which is a subtype of [Inanimate] under [Physical Object]:

(5)

a. CPA semantic type has two (or more) possible mappings in WordNet, where the synsets belong to different hypernymy paths:

[Artwork]

{artwork:1; art:4; graphics:2; nontextual matter:1} ← {visual communication:1} ← {n: communication:1} ← {abstraction:1; abstract entity:1}

{product:2; production:5} ← {n: creation:3} ←
{artifact:1; artefact:1}

[Food]

{food:1; nutrient:1} ← {substance:2} ←
{matter:1} ← {physical entity:1}
{food:3; solid food:1} ← {solid:18} ←
{matter:1} ← {physical entity:1}

b. The WordNet synset to which a CPA semantic type is mapped has two hypernyms:

[Drug]

{drug:3} ← {agent:6} ← {causal agent:1;
cause:1; causal agency:1} ← {substance:2} ←
{physical entity:1}

c. Semantic types that are on the same level in the CPA ontology, are on different levels in WordNet:

[Musical Instrument]

{musical instrument:1; instrument:6} ←
{device:2} ← {instrumentality:1; instrumen-
tation:3} ← {artifact:1; artefact:1}

[Weapon]

{weapon:1; arm:6; weapon system:1}
← {instrument:5} ← {device:2} ←
{instrumentality:1; instrumentation:3} ←
{artifact:1; artefact:1}

d. Semantic types that are on the same level in the CPA ontology, are direct hypernyms/hyponyms in WordNet i.e., {beverage:1} is a hyponym of {food}

[Beverage]

{beverage:1; drink:8; drinkable:2; potable:2}
← {food:1; nutrient:1} ← {substance:2} ←
{matter:1} ← {physical entity:1}

[Food]

{food:1; nutrient:1} ← {substance:2} ←
{matter:1} ← {physical entity:1}
{food:3; solid food:1} ← {solid:18} ←
{matter:1} ← {physical entity:1}

The following general conclusions can be drawn:

There were certain discrepancies or errors in the

CPA hierarchy as with [Smell] – an attribute – which is included as a subtype of [Vapour] together with [Air] and [Gas] (physical forms of substance); and [Blemish] – again more of an attribute or a result – which is on the same level as [Artifact], [Location], [Structure], [Stuff], etc.

A mismatch was also observed in the hypernym/hyponym structure under the top-level concepts as not every of their hyponyms instantiates another hypernym/hyponym tree (for example {otherworld:1} has no hyponyms, and the notion of cognition is spread throughout both the CPA ontology and WordNet).

New semantic primitives borrowed from the CPA ontology were added to the WordNet structure as complementary semantic primitives and with this the information about co-occurrences between verbs and nouns belonging to particular word classes was enriched and more information encoded/expressed within the WordNet semantic network became explicit.

8 Future work

We plan to automatically assign the PDEV patterns to the WordNet verb synsets and to compare PDEV patterns and WordNet sentence frames. Further, we intend to work on the elaboration of general sentence frames to describe the semantic and syntactic properties of all verb synsets grouped in the verb hypernym/hyponym trees. Testing the semantic compatibility between the general sentence frames and the WordNet semantic primitives (both original and complementary) over corpora examples will help us further elaborate general sentence frames and complementary semantic primitives.

Acknowledgments

The work is done within the project *Towards a Semantic Network Enriched with a Variety of Relations* – DN 10/39/2016, financed by the National Scientific Fund of the Republic of Bulgaria. We would like to thank three anonymous reviewers for their helpful comments, as well as the 9th Global WordNet Conference participants.

References

Anna Korhonen. 2002. Assigning verbs to semantic classes via wordnet. *Proceedings of the 2002 Workshop on Building and Using Semantic Networks*,

- Volume 11, Association for Computational Linguistics.
- Anna Wierzbicka. 1986. What's in a noun? (or: How do nouns differ in meaning from adjectives?) *Studies in Language*, International Journal sponsored by the Foundation Foundations of Language, 10.2 (1986): 353–389.
- Anna Wierzbicka. 1984. Cups and mugs: Lexicography and conceptual analysis. *Australian Journal of Linguistics*, 4.2 (1984): 205–255.
- Anna Wierzbicka. 1985. *Lexicography and Conceptual Analysis*. Karoma, Ann Arbor.
- Anna Wierzbicka. 1988. *The semantics of Grammar*, Volume 18. John Benjamins Publishing.
- Patrick Hanks. 2004. Corpus Pattern Analysis. *Proceedings of Euralex*. Lorient, France.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4), 271–322.
- Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving verbal and compositional lexical aspect for NLP applications. *Proceedings of the Eighth Conference of European Chapter of the Association for Computational Linguistics*, 151–158.
- Christiane Fellbaum, Anne Osehrson, and Peter E. Clark. 2009. Putting semantics into WordNets morphosemantic links. *Proceedings of the Third Language and Technology Conference, Poznan, Poland*, [Repr. in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics], vol. 5603, 350–358.
- Donald Hindle. 1990. Noun classification from predicate-argument structures. *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 4 (Winter 1990), 235–312.
- James Pustejovsky. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.
- Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. *Computational linguistics and Intelligent Text Processing*. (2005), 100–111.
- Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2006. Mixing wordnet, verbnet and propbank for studying verb relations. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*.
- Nuria Bel, Sergio Espeja, and Montserrat Marimon. 2007. Automatic acquisition of grammatical types for nouns. *Human Language Technologies 2007: Proceedings of the North American Chapter of the ACL Companion Volume*, 5–8.
- Nuria Bel, Lauren Romeo, and Muntsa Padr. 2012. Automatic lexical semantic classification of nouns. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1448–1455.
- Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique applique*, 10:2.
- Patrick Hanks. 2012. Mapping meaning onto use: a Pattern Dictionary of English Verbs. *AACL 2008*, Utah.
- Patrick Hanks. 2013. *Lexical Analysis*. Cambridge, MA: MIT Press.
- Piek Vossen. 1999. EuroWordNet General Document. http://www.globalwordnet.org/ewn/general_document.ps; [October 2017]
- Piek Vossen, Graeme Hirst, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic*. Kluwer Academic Publisher.
- Rebecca Green, Lisa Pearl, Bonnie J. Dorr, and Philip Resnik. 2001. *Lexical Resource Integration across the Syntax-semantics Interface (No. LAMP-TR-069)*. Maryland University College Park Institute for Advanced Computer Studies.
- Silvie Cinkova and Patrick Hanks. 2010. Validation of Corpus Pattern Analysis – Assigning pattern numbers to random verb samples. At: http://ufal.mff.cuni.cz/spr/data/publications/annotation_manual.pdf [October 2017]
- Zellig Harris. 1954. Distributional structure. *Word*, 10(23): 146162.

Improving Wordnets for Under-Resourced Languages Using Machine Translation

Bharathi Raja Chakravarthi, Mihael Arcan, John P. McCrae

Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland

bharathi.raja@insight-centre.org,
mihael.arcan@insight-centre.org, john@mccr.ae

Abstract

Wordnets are extensively used in natural language processing, but the current approaches for manually building a wordnet from scratch involves large research groups for a long period of time, which are typically not available for under-resourced languages. Even if wordnet-like resources are available for under-resourced languages, they are often not easily accessible, which can alter the results of applications using these resources. Our proposed method presents an *expand* approach for improving and generating wordnets with the help of machine translation. We apply our methods to improve and extend wordnets for the Dravidian languages, i.e., Tamil, Telugu, Kannada, which are severely under-resourced languages. We report evaluation results of the generated wordnet senses in term of precision for these languages. In addition to that, we carried out a manual evaluation of the translations for the Tamil language, where we demonstrate that our approach can aid in improving wordnet resources for under-resourced Dravidian languages.

1 Introduction

As computational activities and the Internet creates a wider multilingual and global community, under-resourced languages acquire political as well as economic interest to develop Natural Language Processing (NLP) systems for these languages. In general, creating NLP systems requires an extensive amount of resources and manual effort, however, under-resourced languages lack in both.

Wordnets are lexical resources, which provide a hierarchical structure based on synsets (a set of one or more synonyms) and semantic features of

individual words. Wordnets can be constructed by either the *merge* or the *expand* approach (Vossen, 1997). Princeton WordNet (Miller, 1995; Fellbaum, 2010) was manually created within Princeton University covering the vocabulary in English language only. Then, based on the Princeton WordNet, wordnets for several languages were created. As an example, EuroWordNet (Vossen, 1997) is a multilingual lexical database for several European languages, structured in the same way as Princeton’s WordNet. The Multiwordnet (Pianta et al., 2002) is strictly aligned with Princeton WordNet and allows to access senses in Italian, Spanish, Portuguese, Hebrew, Romanian and Latin language. Many others have followed for different languages. The IndoWordNet (Bhattacharyya, 2010) was compiled for eighteen out of the twenty-two official languages of India and made available for public use. It is based on the *expand* approach like EuroWordNet, but from the Hindi wordnet, which is then linked to English. On the Global WordNet Association website,¹ a comprehensive list of wordnets available for different languages can be found, including IndoWordNet and EuroWordNet etc.

This paper describes the effort towards generating and improving wordnets for the under-resourced Dravidian languages. Since studies (Federico et al., 2012; Läubli et al., 2013; Green et al., 2013) have shown significant productivity gains when human translators post-edit machine translation output rather than translating text from scratch, we use the available parallel corpora from multiple sources, like OPUS,² to create a machine translation system to translate the wordnet senses in the Princeton WordNet into the mentioned under-resourced languages. Translation tools such as Google Translate,³ or open source SMT systems such as Moses (Koehn et

¹<http://globalwordnet.org/>

²<http://opus.lingfil.uu.se/>

³<http://translate.google.com/>

al., 2007) trained on generic data are the most common solutions, but they often result in unsatisfactory translations of domain-specific expressions. Therefore, we follow the idea of Arcan et al. (2016b), where the authors automatically identify relevant sentences in English containing the WordNet senses and translate them within the context, which showed translation quality improvement of the targeted entries. The effectiveness of our approach is evaluated by comparing the generated translations with the IndoWordNet entries, automatically and manually, respectively. This paper reports our first outcomes in improving wordnet for under-resourced Dravidian languages such as Tamil (ISO 639-2: tam), Telugu (ISO 639-2: tel) and Kannada (ISO 639-2: kan).

2 Related work

Scannell (2007) describes the start of the creation of a resource for the Irish language using the Web as a resource for NLP approaches. This work started by creating a resource for Irish language using the Web as a resources for NLP. Since 2000, the author and his collaborators developed many resources like monolingual corpora, bilingual corpora and parsers etc, for many under-resourced languages, but they did not cover all languages in the world. A six-level typology was proposed by Alegria et al. (2011) that separated languages into six levels. According to the authors, except for top ten languages in the world all the other languages are under-resourced languages. The third and fourth level languages are the languages which have some resource on the internet. These six level typologies is a relative definition for the under-resourced language, but still can be useful for our study of under-resourced languages.

IndoWordNet covers official Indian languages, from the major three families: Indo-Aryan, Dravidian and Sino-Tibetan languages. In general, Indian languages are rich in morphology and each of the three language families has different morphology structure. It was compiled for eighteen out of the twenty-two official languages and made publicly available.⁴ Similarly to EuroWordNet it is based on the *expand* approach, but the central language is Hindi, which is then linked to English. The IndoWordNet entries are updated frequently. For the Tamil language, Rajendran et al. (2002) proposed a design template for the Tamil wordnet.

⁴<http://www.cfilt.iitb.ac.in/indowordnet/index.jsp>

In their further work (Rajendran et al., 2010), they emphasize the need for an independent wordnet for the Dravidian languages, based on EuroWordNet. This is due the observation that the morphology and lexical concepts of these languages are different compared to other Indian languages. The authors have combined the Tamil wordnet and wordnets in other Dravidian languages to form the IndoWordNet.

Mohanty et al. (2017) built SentiWordNet for the Odia language, which is one of the official languages of India. Being an under-resourced language, Odia lacks proper machine translation system to translate the vocabulary of the available resource from English into Odia. The authors have created SentiWordNet for Odia using resources of other Indian languages and the IndoWordNet. Although the IndoWordNet structure does not map directly to the SentiWordNet, instead synsets are matched. The authors used these for translation from source lexicon to target lexicon. Aliabadi et al. (2014) have created a wordnet for the Kurdish language, one of the under-resourced languages in western Iranian language family. They have created Kurdish translation for the “core” wordnet synsets (Vossen, 1997), which is a set of 5,000 essential concepts. They used a dictionary to translate its literals (words), adopted an indirect evaluation alternative in which they look at the effectiveness of using KurdNet for rewriting Information Retrieval queries. Similarly, the work by Horváth et al. (2016) focuses on the semi-automatic construction of wordnet for the Mansi language, which is spoken by Mansi people in Russia, an endangered under-resourced languages with a low number of native speakers. The authors have used the Hungarian wordnet as a starting point. With the help of a Hungarian-Mansi dictionary, which was used to create possible translations between the languages, the Mansi wordnet was continuously expanded.

Previous works did lots of manual effort to create wordnet-like resources, which was funded by public research for a long period of time. However, IndoWordNet is not complete and biased towards Hindi, because the authors created a Hindi-Tamil bilingual dictionary, rather than a wordnet. As explained in Rajendran et al. (2010), the morphology and lexical concepts of Dravidian languages are different from Hindi, which illustrates that the IndoWordNet may not be the most suitable resource to represent the wordnet for the targeted Dravidian languages.

To evaluate and improve the wordnets for the targeted Dravidian languages, we follow the approach of Arcan et al. (2016b), which uses the existing translations of wordnets in other languages to identify contextual information for wordnet senses from a large set of generic parallel corpora. We use this contextual information to improve the translation quality of WordNet senses. We show that our approach can help overcome drawbacks of simple translations of words without context.

3 Background

Our specific aim of this work is to generate and improve wordnets for under-resourced languages. For our task we chose the *expand* approach and automatically translated the Princeton WordNet entries within a disambiguate context to obtain entries for the Dravidian languages.

3.1 Dravidian languages

Dravidian languages, a family of languages spoken primarily in the Southern part of India and also spread over South Asia. The Dravidian languages are divided into four groups: South, South-Central, Central, and North groups. Dravidian morphology is agglutinating and exclusively suffixal. Words are built from small elements called morphemes. Two broad classes of morphemes are stems and affixes. Words are made up of morphemes concatenated based on the grammar of language. Tamil language is also a free word-order language. Due to the nature of morphology, the noun phrase and verb phrase may appear in any permutation and still able to produce same sense of the sentence (Steever, 1987).

The four major literary Dravidian languages are Tamil, Telugu, Malayalam, and Kannada. Tamil, Malayalam, and Kannada fall under the South Dravidian subgroup, whereby Telugu belongs to the South Central Dravidian subgroup (Vikram and Urs, 2007). All the four languages have official status in Government of India and use their own unique script. Outside India, Tamil also has official status in Sri Lanka and Singapore. Tamil script is descended from the Southern Brahmi script and has 12 vowels, 18 consonants and one aytam (special sound). The Telugu script is also descendant of the Southern Brahmi script. It has 16 vowels and 36 consonants, which are more in number than those of Tamil alphabets. The Kannada and Telugu scripts are most similar and often considered as a regional variant. The Kannada

script is used to write other under-resourced languages like Tulu, Konkani and Sankethi. In the Kannada language, the derivation of words is either by combining two distinct words or by affixes. Different to Tamil, Kannada and Telugu inherits some of the affixes from Sanskrit.

3.2 Machine Translation

Statistical Machine Translation (SMT) systems assume that we have a set of example translations $(S^{(k)}, T^{(k)})$ for $k = 1 \dots n$, where $S^{(k)}$ is the k^{th} source sentence, $T^{(k)}$ is the k^{th} target sentence which is the translation of $S^{(k)}$ in the corpus. SMT systems try to maximize the conditional probability $p(t|s)$ of target sentence t given a source sentence s by maximizing separately a language model $p(t)$ and the inverse translation model $p(s|t)$. A language model assigns a probability $p(t)$ for any sentence t and translation model assigns a conditional probability $p(s|t)$ to source / target pair of sentence. By Bayes rule

$$p(t|s) \propto p(t)p(s|t) \quad (1)$$

This decomposition into a translation and a language model improves the fluency of generated texts by making full use of available corpora. The language model is not only meant to ensure a fluent output, but also supports difficult decisions about word order and word translation (Koehn, 2010). We used the Moses (Koehn et al., 2007) toolkit that provides end-to-end support for the creation and evaluation of machine translation system based on BLEU (Papineni et al., 2002) score. There are two major criteria for automatic SMT evaluation: completeness and correctness, which are considered by BLEU, an automatic evaluation technique, which is a geometric mean of n-gram precision. BLEU score is language independent, fast, and shows good correlation with human evaluation campaigns. Therefore we plan to use this metric to evaluate our work.

3.3 Available Corpora for Machine Translation

This section describes the data collection and the pre-processing process steps. The English-Tamil parallel corpus, which we used to train our SMT system is collected from various sources and combined into a single parallel corpus. We used the EnTam corpus (Ramasamy et al., 2012), which was pre-processed from raw Web data to become a sentence-aligned corpus. The parallel corpora

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
Number of tokens	7,738,432	6,196,245	258,165	226,264	68,197	71,697
Number of unique words	134,486	459,620	18,455	28,140	7,740	15,683
Average word length	4.2	7.0	3.7	4.8	4.5	6.0
Average sentence length	5.2	7.9	4.6	5.6	5.3	6.8
Number of sentences	449,337		44,588		13,543	

Table 1: Statistics of the parallel corpora used to train the translation systems.

contains text from the news domain,⁵ sentences from the Tamil cinema articles⁶ and the Bible.⁷ For the news corpus, the authors downloaded web pages that have matching file names in both English and Tamil. For the cinema corpus, all the English articles had a link to the corresponding Tamil translation. The collection of the Bible corpus followed a similar pattern. We also took the English-Tamil parallel corpora for six Indian languages created with the help of Mechanical Turk for Wikipedia documents (Post et al., 2012). Since the data was created by non-expert translators hired over the Mechanical Turk, it is of mixed quality. From the OPUS website, we have collected the Gnome, KDE, Ubuntu and movie subtitles (Tiedemann, 2012). We furthermore manually aligned Tamil text Tirukkural,⁸ and combined all the parallel corpora into a single corpus. We first tokenized sentences in English and Tamil and then true-cased only the English side of the parallel corpus, since the Tamil language does not have a casing. Finally, we cleaned up the data by eliminating the sentences whose length is above 80 words.

To obtain the parallel corpora for Telugu and Kannada, we used the corpora available on the OPUS website. The same pre-processing procedure was followed for Telugu and Kannada language, since both languages are close to the Tamil language. The Table 1 shows the statistics of the parallel corpora for the three language pairs. From this table we can see that the English-Tamil parallel corpus is much larger than for the other language pairs. On the other hand, the number of sentences for English-Kannada is very small. Once we have obtained the parallel corpus, we created the SMT systems for the English-Tamil, English-Telugu, and English-Kannada language pairs.

We define the following set of data:

- Development set: Randomly selected 2000 sentences from the parallel corpus as devel-

opment set is used to measure the system performance of the phrase-based translation model.

- Test set: A blind set of 1000 sentence randomly chosen from parallel corpus that is used to test the system. There is no overlap between these set of data.
- Training set: A larger size parallel corpus that is used to train the phrase-based translation model. It is remaining corpus after development and test are extracted.

In this work, we focus on three languages from Dravidian family namely, Tamil, Telugu, and Kannada. This is mainly due to available parallel corpora and we believe that this method can be extended for other under-resourced languages without much effort.

3.4 Resource Scarceness

There are few resources, which can be used to automatically create a wordnet for under-resourced languages. One way to cross the language barrier is with the help of machine translation. As with any machine learning methods, SMT tends to improve translation quality when using a large amount of training data. That is, if the training method sees a specific word or phrase multiple times during training, it is more likely to learn a correct translation. SMT suffers due to the scarcity of parallel corpora, Dravidian word order and the morphological complexity attached to the language. For the Dravidian languages when translating from or to English the translation models suffer because of syntactic differences while the morphological differences contribute to data sparsity. In contrast, small corpora used for training lead to incomplete word coverage, which may cause the out-of-vocabulary (OOV) issues.

Besides the resource scarceness, another issue observed with the corpus for Dravidian languages was code-switching contents in the data. Code-switching is an act of alternating between elements of two or more languages, which is prevalent in

⁵<http://www.wsws.org/>

⁶<http://www.cinesouth.com/>

⁷<http://biblephone.intercer.net/>

⁸<http://www.projectmadurai.org/>

	Original	Non-Code mixing
English→Tamil	20.29	20.61
English→Telugu	28.81	28.25
English→Kannada	14.64	14.45

Table 2: Automatic translation evaluation of the of 1000 randomly selected sentences in terms of the BLEU metric.

multilingual countries (Barman et al., 2014). With English being the most used language in the digital world, people tend to mix English words with their native languages. That might be the case in other languages as well.

4 Methodology

The principle approaches for constructing wordnets are the *merge* approach or the *expand* approach. In the *merge* approach, the synsets and relations are built independently and then aligned with WordNet. The drawbacks of the *merge* approach are that it is time-consuming and requires a lot of manual effort to build. On the contrary in the *expand* model, wordnet can be created automatically by translating synsets using different strategies, whereby the synsets are built in correspondence with the existing wordnet synsets. We followed the *expand* approach and created a machine translation systems to translate the sentences, which contained the WordNet senses in English to the target language

4.1 Training Machine Translation parameters

In the following section, we takes as a baseline a parallel text, that has been aligned at the sentence level. To obtain the translations, we use Moses SMT toolkit with of baseline setup with 5-gram language model created using the training data by KenLM (Heafield, 2011). The baseline SMT system was built for three language pairs, English-Tamil, English-Telugu, and English-Kannada. The test set mentioned in Section 3.3 was used to evaluate our system. From Table 1 and Table 2 we can see that size of the parallel corpus has an impact on the BLEU score for test set which is evaluation criteria for the translation model.

4.2 Context Identification

Since manual translation of wordnets using the extend approach is a very time consuming and expensive process, we apply SMT to automatically

translate WordNet entries into the targeted Dravidian languages. While an domain-unadapted SMT system can only return the most frequent translation when given a term by itself, it has been observed that translation quality of single word expressions improves when the word is given in an disambiguated context of a sentence (Arcan et al., 2016a; Arcan et al., 2016b). Therefore existing translations of WordNet senses in other languages than English were used to select the most relevant sentences for wordnet senses from a large set of generic parallel corpora. The goal is to identify sentences that share the same semantic information in respect to the synset of the WordNet entry that we want to translate. To ensure a broad lexical and domain coverage of English sentences, existing parallel corpora for various language pairs were merged into one parallel data set, i.e., Europarl (Koehn, 2005), DGT - translation memories generated by the *Directorate-General for Translation* (Steinberger et al., 2014), MultiUN corpus (Eisele and Chen, 2010), EMEA, KDE4, OpenOffice (Tiedemann, 2009), OpenSubtitles2012 (Tiedemann, 2012). Similarly, wordnets in a variety of languages, provided by the Open Multilingual Wordnet web page,⁹ were used.

As a motivating example, we consider the word *vessel*, which is a member of three synsets in Princeton WordNet, whereby the most frequent translation, e.g., as given by Google Translate, is *Schiff* in German and *nave* in Italian, corresponding to i60833¹⁰ ‘a craft designed for water transportation’. For the second sense, i65336 ‘a tube in which a body fluid circulates’, we assume that we know the German translation for this sense is *Gefäß* and we look in our approach for sentences in a parallel corpus, where the words *vessel* and *Gefäß* both occur and obtain a context such as ‘blood vessel’ that allows the SMT system to translate this sense correctly. This alone is not sufficient as *Gefäß* is also a translation of i60834 ‘an object used as a container’, however in Italian these two senses are distinct (*vaso* and *recipiente* respectively), thus by using as many languages as possible we maximize our chances of finding a well disambiguated context.

4.3 Code-mixing

Code-switching and code-mixing is a phenomenon found among bilingual communities all

⁹<http://compling.hss.ntu.edu.sg/omw/>

¹⁰We use the CILI identifiers for synsets (Bond et al., 2016)

	English-Tamil		English-Telugu		English-Kannada	
	English	Tamil	English	Telugu	English	Kannada
tok	0.5% (45,847)	1.1% (72,833)	2.8% (7,303)	4.9% (12,818)	3.5% (2,425)	9.0% (6,463)
sent	0.9% (4,100)		3.1% (1,388)		3.4% (468)	

Table 3: Number of sentences (sent) and number of tokens (tok) removed from the original corpus.

Source sentence: “இப்போது, நான் அதை loving.”
 Transliteration: :lppōtu, nāṇ atai loving
 Target sentence: “Right now, I'm loving it.”

Source sentence: “முன்னிருப்பு GNOME பொருள்”
 Transliteration: :Munniruppu GNOME poru|
 Target sentence: “Default GNOME Theme”

Figure 1: Examples of Code-mixing in Tamil-English parallel corpus. In the first example the verb *loving* is code-mixed in Tamil. In Second Example the noun *GNOME* is code-mixed.

over the world (Ayeomoni, 2006; Yoder et al., 2017). Code-mixing is mixing of words, phrases, and sentence from two or more languages with in the same sentence or between sentences. In many bilingual or multilingual communities like India, Hong Kong, Malaysia or Singapore, language interaction often happens in which two or more languages are mixed. Furthermore, it increasingly occurs in monolingual cultures due to globalization. In many contexts and domains, English is mixed with native languages within their utterance than in the past due to Internet boom. Due to the history and popularity of the English language, on the Internet Indian languages are more frequently mixed with English than other native languages (Chanda et al., 2016).

A major part of our corpora comes from movie subtitles and technical documents, which makes it even more prone to code-mixing of English in the Dravidian languages. In our corpus, movie speeches are transcribed to text and they differ from that in other written genres: the vocabulary is informal, non-linguistics sounds like *ah*, and mixing of scripts in case of English and native languages (Tiedemann, 2008). Two example of code-switching are demonstrated in Figure 1. The parallel corpus is initially segregated into English script and native script. All of the annotations are done using an automatic process. All words from a language other than the native script of our experiment are taken out on both sides of corpus if it occurs in native language side of the parallel corpus. The sentences are removed from both sides if the target language side does not contain native

script words in it. Table 3 show the percentage of code-mixed text removed from original corpus. The goal of this approach is to investigate whether code-mixing criteria and corresponding training are directly related to the improvement of the translation quality measured with automatic evaluation and manual evaluation. We assumed that code-mixed text can be found by different scripts and did not evaluate the code-mixing written in the native script or Latin script to write the native language as was done by (Das and Gambäck, 2013)

5 Evaluation

The most reliable method to evaluate the wordnet is a manual evaluation, but a manual evaluation of whole the WordNet is time consuming and very expensive. Therefore, we did the automatic evaluation of the our translations and measured the precision. In order to determine the correctness of our work, we have furthermore randomly taken 50 WordNet entries for manual evaluation on these entries.

5.1 Automatic Evaluation

In this paper, we have compared our result to the IndoWordNet. Once the translation step the of disambiguated context, containing the target entries, was finished, we use the word alignment information to extract the translation of the WordNet entry. Since several disambiguated sentences per WordNet entry were used, we took the translations for each context and then combined the results to count the most frequent one. The top 10-words entries were compared to the IndoWordNet for the exact match.

We took precision at 10, precision at 5, precision at 2, and precision at 1. We did this comparison for the all the three languages, i.e. Tamil, Telugu, and Kannada. As an additional experiment, we removed the code-mixing part of the corpus and created a new translation system, which was used again to translate the same WordNet entries. The table 4 shows the result of the automatic evaluation of the translation of the entries into the Targeted Dravidian languages. The table shows the precision at the different level of

		English→Tamil			
		P@10	P@5	P@2	P@1
original corpus		0.120	0.109	0.083	0.065
non-code mixed		0.125	0.115	0.091	0.073
		English→Telugu			
		P@10	P@5	P@2	P@1
original corpus		0.047	0.046	0.038	0.028
non-code mixed		0.047	0.045	0.038	0.027
		English→Kannada			
		P@10	P@5	P@2	P@1
original corpus		0.009	0.010	0.008	0.005
non-code mixed		0.011	0.011	0.009	0.007

Table 4: Results of Automatic evaluation of wordnet with IndoWordNet Precision at different level denoted by P@10 which means Precision at 10.

the translations, based on the translation model, generation from the original corpus and non-code mixed corpus. Non-code mixed often outperforms the baseline in terms of precision, whereby the difference is less visible in Telugu language. This is likely due to the short sentences in the Telugu corpus. These differences in the precision are significant in the manual evaluation of Tamil tests with 50 samples. The wide difference between manual and automatics evaluation can be explained in part by different forms. Table 4 shows an example of how our system differs from the baseline SMT system and how it benefits the wordnet translation. This is a clear evidence that an SMT without code-mixing described above achieves an improvement over the baseline without using any additional training data. However, it has been shown in Arcan et al. (2016b) that better performance on WordNet translation can be achieved, if the corpora contained a sufficient amount of parallel sentences. Their translation evaluation based on the BLEU metric on unigrams (similar to precision at 1, P@1), showed a range between 0.55 and 0.70 BLEU points, for the well resourced languages, like Slovene, Spanish, Croatian and Italian. Restricting the task to a small data set tends to hurt the translation performance, but it can be useful to aid in the creation or improvement of new resources for the under-resourced languages.

5.2 Manual Evaluation

In order to be able to evaluate our method in contrast to stand-alone approaches, we manually evaluated our method in comparison with IndoWordNet entries. To select the sample for manual evaluation,

	Original	Non-Code mixing
Agrees with IWN	18%	20%
Inflected Form	12%	22%
Transliteration	4%	4%
Spelling variant	2%	2%
Correct, but not in IWN	18%	24%
Incorrect	46%	28%

Table 5: Manual Evaluation of wordnet creation for Tamil language compared with IndoWordNet (IWN) at precision at 10 presented in percentage.

we proceeded as follows: we randomly extracted a sample of 50 wordnet entries from the WordNet. First, each of these 50 wordnet entries were compared to the IndoWordNet for the exact match. Subsequently, regardless of this decision, each of the 50 wordnet entries were evaluated and classified according to its quality. The classification is the following:

- **Agrees with IndoWordNet** Exact match found in IndoWordNet.
- **Inflected form** The root of a word is found with a different inflection, which can make the translation correct but imprecise.
- **Transliteration** The word is transliterated, which can be caused by the unavailability of the translation form in the parallel corpus, since some words are used in transliteration because of foreign words.
- **Spelling Variant** Since our data in day to day language of Tamil and IndoWordNet is skewed towards classical sense of language. Our method produces the Spelling Variant which can be caused by wrong or misspelling of the word according to IndoWordNet.
- **Correct, but not in IndoWordNet** IndoWordNet is large and it covers eighteen languages, but it lacks some wordnet entries for the Dravidian languages. We verified we had identified the correct sense by referring to the wordnet gloss.
- **Incorrect** This error class can be caused due to inappropriate term or mistranslation.

The examples in the Figure 2 list the Tamil translation wordnet in our experiment. Neither the word nor its translation has appeared in the training corpus therefore, the SMT system cannot translate the word and chooses to produce the word in English. On the other side, these examples may produce some insights into the word.

ILI code	Gloss	IWN	Meaning	Translation	Meaning	Comments
14647235-n	any of several compounds containing chlorine and nitrogen; used as an antiseptic in wounds	நைட்ரஜன்	nitrogen	நைதரசன்	nitrogen	Spelling variant
01026095-v	give the name or identifying characteristics of; refer to by name or some other identifying characteristic	பெயரிடு	name, identity	பெயர்	name	Inflected form, different part-of-speech
00461782-n	a game in which balls are rolled at an object or group of objects with the aim of knocking them over or moving them	பந்து	ball	பௌலிங்	bowling	Correct translation, sense missing in IWN
04751305-n	noticeable heterogeneity	பல்வேறு	diverseness, diversity	பல்வேறு	diverseness, diversity	Agrees with IWN
01546111-v	be standing; be upright	தூக்கு	to lift	நிற்க	to stand	correct translation, sense missing in IWN

Figure 2: Examples of the manual evaluation of Tamil wordnet entries in comparison to the IndoWordNet (IWN).

We should note that this evaluation was carried out for both, original, uncleaned, corpus as well as cleaned corpus (non-code mixing). We observed that the cleaned data produce better results compared to the original data which have many code-mixing entries. From the table 5, we can see that there is a significant improvement over the inflected form and correct but not found in IndoWordNet categories. This shows that our method can help to improve the wordnet entries for under-resourced languages.

6 Discussion

While our automatic evaluation results are a little disappointing, and this is perhaps unsurprising in the context of under-resourced languages as there is very little a data availability for these language, our manual evaluation shows that this is far from reality. Evaluating using a resource such as IndoWordNet is always likely to be problematic as the resource is far from complete and does not claim to cover all words in the Dravidian languages studied in this paper. Moreover, IndoWordNet is overly skewed to the the classical words of these languages, but the majority our parallel corpus is day to day conversation texts. Despite the low precision in determining the exact match to the IndoWordNet, our technique yields 48% for precision at 10 in manual evaluation, although the automatic evaluation considering pre-

cision at 10 gave only 12%. Our method relays on IndoWordNet for evaluation but IndoWordNet is biased over one particular language, which is Hindi. The resulting wordnet entries, though noisy, is suitable for aiding wordnet creation for under-resourced languages.

The handling of code-mixing in this paper appears to improve the quality of the proposed translation, outperforming the baseline results of wordnet entries once code-mixed was removed from data. Thus we believe that the method presented here still applicable to resource creation of under-resourced languages.

7 Conclusion

In this paper we showed the challenges in building wordnet for under-resourced languages and presented that our method can aid the creation or improvement of wordnets for under-resourced languages. We experimented with available data to created SMT systems for three Dravidian languages and used those as a baseline. To improve the results we removed the code-mixed terms from the corpus. Our results indicated that the proposed removing of code-mixed text from the corpus results in gains for the wordnet entries with limited data.

8 Acknowledgements

This work was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

- Iñaki Alegria, Xabier Artola, Arantza Diaz De Ilaraza, and Kepa Sarasola. 2011. Strategies to develop language technologies for less-resourced languages based on the case of basque.
- Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building KurdNet, the Kurdish WordNet. In *Proceedings of the 7th Global WordNet Conference (GWC'14)*, pages 1–6.
- Mihael Arcan, Mauro Dragoni, and Paul Buitelaar. 2016a. Translating ontologies in real-world settings. In *Proceedings of the 15th International Semantic Web Conference (ISWC-2016)*, Kobe, Japan.
- Mihael Arcan, John P McCrae, and Paul Buitelaar. 2016b. Expanding wordnets to new languages with multilingual sense disambiguation. In *International Conference on Computational Linguistics (COLING-2016)*, Osaka, Japan.
- Moses Omoniyi Ayeomoni. 2006. Code-switching and code-mixing: Style of language use in childhood in Yoruba speech community. *Nordic Journal of African Studies*, 15(1):90–99.
- Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 13–23.
- Pushpak Bhattacharyya. 2010. Indowordnet. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.
- Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016. Columbia-Jadavpur submission for emnlp 2016 code-switching workshop shared task: System description. *EMNLP 2016*, page 112.
- Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text: the last language identification frontier? *TAL*, 54(3):41–64.
- Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odijk, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.
- Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.
- Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.
- Csilla Horváth, Ágoston Nagy, Norbert Szilágyi, and Veronika Vincze. 2016. Where bears have the eyes of currant: Towards a mansi wordnet. In *Proceedings of the Eighth Global WordNet Conference*, pages 130–134.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT.
- Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.
- Samuel Lübli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. *Machine Translation Summit XIV*, page 83.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

- Gaurav Mohanty, Abishek Kannan, and Radhika Mamidi. 2017. Building a sentiwordnet for odia. *EMNLP 2017*, page 143.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Int’l Conference on Global WordNet*.
- Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics.
- S Rajendran, S Arulmozi, B Kumara Shanmugam, S Baskaran, and S Thiagarajan. 2002. Tamil WordNet. In *Proceedings of the First International Global WordNet Conference. Mysore*, volume 152, pages 271–274.
- S Rajendran, G Shivapratap, V Dhanlakshmi, and KP Soman. 2010. Building a wordnet for Dravidian languages. In *Proceedings of the Global WordNet Conference (GWC 10)*. Citeseer.
- Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for English-Tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.
- Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.
- Sanford B Steever. 1987. Tamil and the dravidian languages. *The world’s major languages*, pages 725–746.
- Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union’s highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.
- Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.
- Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- T. N. Vikram and Shalini R. Urs, 2007. *Development of Prototype Morphological Analyzer for the South Indian Language of Kannada*, pages 109–116. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Piek Vossen. 1997. Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.
- Michael Miller Yoder, Shruti Rijhwani, Carolyn Penstein Rosé, and Lori Levin. 2017. Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages. *ACL 2017*, page 73.

Semantic Feature Structure Extraction from Documents Based on Extended Lexical Chains

Terry Ruas

University of Michigan - Dearborn
Dearborn, MI, USA
truas@umich.edu

William Grosky

University of Michigan - Dearborn
Dearborn, MI, USA
wgrosky@umich.edu

Abstract

The meaning of a sentence in a document is more easily determined if its constituent words exhibit cohesion with respect to their individual semantics. This paper explores the degree of cohesion among a document's words using lexical chains as a semantic representation of its meaning. Using a combination of diverse types of lexical chains, we develop a text document representation that can be used for semantic document retrieval. For our approach, we develop two kinds of lexical chains: (i) a multilevel flexible chain representation of the extracted semantic values, which is used to construct a fixed segmentation of these chains and constituent words in the text; and (ii) a fixed lexical chain obtained directly from the initial semantic representation from a document. The extraction and processing of concepts is performed using WordNet as a lexical database. The segmentation then uses these lexical chains to model the dispersion of concepts in the document. Representing each document as a high-dimensional vector, we use spherical k-means clustering to demonstrate that our approach performs better than previous techniques.

1 Introduction

Since the late 1980's, when there was a burst of research in dimensional reduction techniques (Dumais et al., 1988), information retrieval has been concerned with semantics. Since multimedia entities, including text, have multiple meanings, examining the context in which they appear became of significant importance to their overall disambiguation.

An important example of this in the natural language processing community was the formulation of lexical chains (Morris and Hirst, 1991). A lexical chain is a contiguous portion of text which has semantic cohesion. Such chains are of

variable length and have been used throughout the intervening years in many ways; e.g., for semantic characterization of the underlying document, for question and answers tasks, for document summarization, and for clustering documents into semantically uniform groups.

In this paper, we propose two new types of lexical chains based on semantic representation: the first is called Flexible-to-Fixed Lexical Chains (Flex2Fix) and the second, Fixed Lexical Chains (FixLC). In the first, these representations follow and extend the model proposed by (Ruas and Grosky, 2017), transforming their flexible lexical chains into fixed structures, which are later transformed into vectors of semantic values. In the second, we build fixed lexical structures directly from their initial semantic value.

First, we start by identifying the most suitable semantic representation for each word, considering their context. Second, we use these semantic abstractions and find the flexible lexical chains in a document. This approach extracts and builds cohesive sequences of ideas with respect to the semantic value shared among words in a dynamic way. Third, we develop an approach to transform flexible lexical chains into fixed lexical chains. All these chains are used to construct a vector representation corresponding to a document's semantic structure. This is done to represent the document's semantic value at a higher level of abstraction. We also investigate how fixed lexical chains obtained directly from the document's semantic representation perform against traditional approaches (e.g. Bag-of-Words (BOW)) and the derived fixed structures from the flexible ones.

The remainder of this paper appears as follows. Section 2 reviews existing work in lexical chains and provides additional information on our technique. In Section 3, we present our methodology and proposed algorithms for content-based retrieval using lexical chains. Section

4 concerns the experimental validation of our approach, while in Section 5, we offer some final considerations and potential future work.

2 Related Work

The term *lexical chains* was first proposed by (Morris and Hirst, 1991) as an extension of *lexical cohesion*, a concept introduced by (Halliday and Hasan, 1976). A text in which many of its words are semantically connected often produces a certain degree of continuity in its ideas, providing good cohesion among its words. Lexical cohesion is more likely to occur between words close to each other in a text, especially those contiguously ordered. The sequence of related words, tied by a common semantic affinity is classified as a lexical chain (Morris and Hirst, 1991).

The use of lexical chains in document processing and analysis (e.g. text similarity, word disambiguation, document clustering) has been widely studied in the literature. In (Barzilay and Elhadad, 1997; Silber and McCoy, 2000), lexical chains are used to summarize texts. The former extracts and classifies lexical chains and discovers significant sentences to represent documents from them. The latter proposes a linear-time algorithm for constructing the lexical chains that will capture the meaning of a text. Some authors use WordNet (WN) to improve the search and evaluation of lexical chains. (Budanitsky and Hirst, 2001; Budanitsky and Hirst, 2006) compare several measurements of semantic distance and relatedness using lexical chains in conjunction with WN. (Moldovan and Novischi, 2002) studies the use of lexical chains for finding topically related words. This is done considering the glosses for each synset in WN. (Hotho et al., 2003) explores the benefits of using WN to improve document clustering based on an explicit matching between terms found in the text and the lexical database. (McCarthy et al., 2004) presents a methodology to categorize and find the most predominant synsets in untagged texts using WN. In (Sedding and Kazakov, 2001), WN is used for document clustering, exploring the benefits of incorporating hypernyms and synonyms into their approach. In (Pedersen et al., 2004), an application developed in Perl is proposed to calculate the relatedness of concepts via WN through different measures of similarity. (Guo and Diab, 2011) hypothesizes that if the semantics of words are known in advance, it is possible

to get a better statistical inference concerning a document's overall idea.

In more recent works, (Navigli, 2009) presents an extensive study in the Word Sense Disambiguation (WSD) arena, in which he proposes an unsupervised WSD algorithm based on generating Spreading Activation Networks (SANs) from senses of a thesaurus and the relations between them. (Meng et al., 2013) explores the theory behind state-of-the-art techniques for semantic similarity measures in four main categories: path length-based, information content-based, feature-based, and hybrid measures. (AlAgha and Nafee, 2014) proposes an approach to improve document clustering by exploring the semantic knowledge offered by Wikipedia. The authors discuss this hypothesis, comparing the results using WN and Wikipedia, claiming that the latter is more robust. In (Pradhan et al., 2015) several measures of similarity (e.g. normalized Google distance, normalized compression distance, cosine distance, latent semantic similarity) are applied to categorize sentences, words, paragraphs and documents according to their lexical and semantic similarities. In (Bär et al., 2015) an extensive study about available text similarity measures is done as part of semantic evaluation, and for the detection of text reusability. They argue that text similarity cannot be considered as a static and absolute notion. Instead, one should carefully define in which levels and perspectives two documents are similar or not. (Wei et al., 2015) combines lexical chains and WN to extract a set of semantically related words from texts and then uses them for clustering. Their approach uses an ontological hierarchical structure and relations to provide a more accurate assessment of the similarity between terms for WSD. In (Tekli, 2016), they conduct a comprehensive review of the methods related to XML-based semi-structured semantic analysis and disambiguation. Although focused in the XML arena, this work provides an overview of the semantic disambiguation field, as well. They cover traditional WSD methods and potential application scenarios that could benefit from them (e.g. data clustering, semantic-aware indexing) while discussing current on-going challenges in the area.

Although extensively studied, the concept of lexical chains still has much to be explored. Besides the fact that each idiom has its own identity, most of the presented work either relies solely on statistical approaches (e.g. *tf-idf*, BOW) or focuses on one aspect of word relatedness. Some research groups focus their efforts on exploring

algorithms and tools to calculate distances between several entities, such as words, paragraphs, synonyms and lexical chains. A few rely on annotated text and/or machine learning techniques to extract semantic-like features from documents. Others expand the set for each word, considering their immediate synonyms or hypernyms to improve corpus or query. The ones inspecting lexical chains, build them using the words individually, or often using some common/direct synonym. Although these are interesting approaches, they are only focused on the word itself, leading to an alternative BOW representation. They still do not explicitly consider the context of a given word in relation to its location or surroundings in the text. Semantic and contextual aspects are difficult to track, but are important aspects of effective human communication. In the last eleven years, the interest in these topics and their contributions to traditional approaches have been increasing among distinct scientific communities. For example, (Grosky and Ruas, 2017) examined the research conducted in the multimedia arena, consisting of 2,872 items (e.g. papers, journals, reports) in the last 11 years, and found an increasing number of publications exploring *semantics* and *contextual* aspects in different areas, pointing to a trend in these areas.

Our approach contributes to this topic by expanding the notion of WSD, considering all synsets of a given word, including the influence among them. Furthermore, our chains are produced by using the most suitable synset for a word, which is a result of the evaluation of its contiguous neighbors (context). In addition, our lexical chains consider all desired hypernyms in WN, given a certain threshold, which can be adjusted to obtain higher (more general) or lower (less general) semantic representations.

3 Building Extended Lexical Chains

As stated by (Morris and Hirst, 1991) (Budanitsky and Hirst, 2006), there are multiple categories in which lexical chains can be classified. These concepts are explored in our approach through WN by using synsets and hypernyms. WN is a lexical database that provides a complex structure of how words and their meanings are related. The following is a small summary of the main terms necessary to understand our work using extensible lexical chains and WN (Fellbaum, 1998):

- *Synonym* – a one-to-many mapping from concepts to words;
- *Synset* – a set of cognitive synonyms (one or more) of a given word that share a common concept;
- *Synset ID* – an ID that represents the entire synset;
- *Sense* – the elements in each synset;
- *Hypernym* - a general abstraction of synset, corresponding to a *kind-of* relation;
- *Lowest Common Subsumer* – is the most specific synset in the hypernym hierarchy which is an ancestor of the given synsets;
- *Root* – initial synset in WN, called *entity*.

A *synset* is a set of synonyms (one or more) for a given word, while *hypernyms* are sets of more general synsets. For example: pug and bulldog are each a kind of dog. A mammal is a generalization of dog, and so on.

Our model consists of exploring documents through their lexical structure. This will be provided by evaluating the semantic value of each word in a text (Ruas and Grosky, 2017). The main idea can be divided into four major tasks: (i) Document Extraction Process, (ii) Best Synset Disambiguation Module, (iii) Lexical Synset Chain Extraction Module and (iv) Distributed Semantic Mapping.

In (i), we select the documents to be processed and clean the data, eliminating noise, such as stopwords, special characters, punctuation, and html tags, among others. In this paper, the source of data was a set of webpages from Wikipedia, so an enhanced stopwords' list had to be used. Once the documents are preprocessed, we filter only those words that have a synset match in WN. If a word in the text has no match in WN it will not contribute to the formation of lexical chains, so we assume that it can be discarded.

3.1 Best Synset Disambiguation Module

In (ii), the Best Synset Disambiguation Module is a subroutine that applies and extends the concept of WSD, but considers the synsets extracted from w_i , w_{i-1} and w_{i+1} . WSD is the problem in which one must decide which synset is better suited for a word in a sentence, given that this word has multiple meanings and each one of these meanings may be affected by other nearby words. Most works in the lexical chains arena try to build these structures by considering only the words within the document, while some use an auxiliary annotated corpus for learning. Others have used the most common synset for each

word (first synset provided by WN for each word) as well as keeping track of word pair occurrences and their distribution in a document. Our approach considers the influence of immediate neighbors for each word w_i , evaluated using all synsets available in WN, for the word itself as well as for its hypernyms. For each word w_i , with $i=1,2,\dots,n$, there are 0 or more synsets available in WN. In our experiments, only the nouns existing in WN are considered, so nouns not present in WN are discarded. The current version of WN used in this paper (3.1) has approximately 117,000 synsets, divided into four major categories: 81,000 noun synsets, 13,600 verb synsets, 19,000 adjective synsets, and 3,600 adverb synsets. Since the number of nouns comprise almost 70% of all information available, we choose to work with this category of synsets (Fellbaum, 2010). In addition, nouns allow us to use interesting relationships between synsets, such as hypernyms.

We represent the best synset ID (BSID) of a word w_i by analyzing the effects of its predecessor (w_{i-1}) and successor (w_{i+1}), called *Former-SynsetID*(w_i) (FSID(w_i)) and *Latter-SynsetID*(w_i) (LSID(w_i)), respectively. FSID(w_i) and LSID(w_i) are selected based on the score obtained by all possible combinations between all synsets of the pairs (w_i, w_{i-1}) and (w_i, w_{i+1}). The synsets for w_i with the highest score value in comparison with w_{i-1} and w_{i+1} will be represented by FSID(w_i) and LSID(w_i) respectively. We use Jiang & Conrath’s algorithm, which is an information content-based measure used to calculate the similarity between two synsets. This value is obtained by calculating the distance of two synsets (c_1, c_2), as shown in Equation 1 (Jiang and Conrath, 1997; Meng et al., 2013),

$$dis_{jiang}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2IC(lcs(c_1, c_2)) \quad (1)$$

where c_1 and c_2 represent the synsets for word 1 and word 2; $IC(c_k)$ is the information content calculated for c_k and $lcs(c_1, c_2)$ is the lowest common subsumer (hypernym) of synset c_1 and synset c_2 . In our implementation, the information content is provided by the *ic-semcor.dat*¹ file, which is based on the *cntlist* file distributed with WN 3.0. The semantic similarity score is calculated for all synsets available for each word evaluated. Finally, every word will hold two prospective synsets (FSID(w_i) and LSID(w_i)), which represent the synsets with the highest score (except

the first and last word of the document). These will be used to produce the BSID for (w_i). There are other measures (e.g. Lin (Lin, 1998), Hirst (Hirst and St-Onge, 1998), Resnik (Resnik, 1995), Wu & Palmer (Wu and Palmer, 1994)), besides Jiang & Conrath, that can be used to calculate the relatedness of two synsets. They are divided into four main categories: path based, IC based, feature based, and hybrid methods (Meng et al., 2013).

Jiang & Conrath’s algorithm was chosen because of its execution time and robustness, since it considers the IC of the synsets. In addition, according to (Budanitsky and Hirst, 2001) Jiang & Conrath’s measure outperformed other known techniques used for semantic similarity. Further experiments using different IC files (e.g. BNC, Treebank, Brown, Shaks) in comparison with path, feature, and hybrid approaches are still necessary to improve our current findings. More details about Jiang & Conrath and the others algorithms can be found in (Jiang and Conrath, 1997; Meng et al., 2013). The latter provides a small survey about the most popular WSD algorithms available as well.

After the FSID and LSID for each word w_i has been found, it is necessary to find the BSID for the given word w_i . For this task, we use the *Best Synset Disambiguation Algorithm* (BSD) (Ruas and Grosky, 2017), which will identify what is the BSID, using as input parameters LSID and FSID. Three cases are considered prior to its selection: (a) if FSID(w_i) and LSID(w_i) are equal, then BSID(w_i) = FSID(w_i) = LSID(w_i); (b) the lowest common subsumer of FSID(w_i) and LSID(w_i), given a depth threshold; and (c), if (b) produces an empty set, the deepest synset among FSID(w_i) and LSID(w_i) is chosen. In case both have the same depth, one is chosen randomly. In (b), we used the depth of 6 (root being the initial point) as the limit to look for common hypernym extraction. This value was obtained by experimental tests considering factors like: execution time, diversity of synsets, diversity of chains, specificity of synsets, and others. This algorithm mitigates the fact that words with multiple meanings (polysemy) might have an unstable representation, by performing a two-level disambiguation process. In the first level, we apply known WSD techniques to obtain prospective pairs of synsets with the highest score, considering the context of each word. The second level extends the concept of WSD to synsets (BSD). More details about this algorithm are explained in (Ruas and Grosky, 2017).

¹ <http://wn-similarity.sourceforge.net/>

The identification of the BSID for each term w_i considers its surroundings, so the most suitable semantic representation can be used to construct our lexical chains. In (Ruas and Grosky, 2017), BSID and flexible lexical chains have been used to suggest keywords that represent the main concepts embedded in a document.

As we traverse the graph in WN for the lowest common subsumer (hypernyms) extraction (b), we consider the first hypernym on each level, for each synset. Since WN organizes its synsets from most to least frequent usage, and we are generalizing the concepts as we move towards the root, it is only natural that we extract a hypernym that will provide the most diffused element with respect to its frequency in the lexical database. In other words, the first hypernym in every upper level will provide greater probability of an intersection with another synset when we build our lexical chains.

3.2 Lexical Chain Extraction Module

Once all words have their BSID selected, we start building our lexical chains in a two-phase subroutine called Lexical Synset Chain Extraction Module. To the best of our knowledge, this module (iii) is introducing two novel contributions. First, the extension of flexible chains into fixed structures to better represent the semantic values extracted from these synsets, and second, we construct parametrized fixed lexical chains, considering the BSID representation obtained in Section 3.1.

We use the *Flexible Lexical Chains Algorithm* (FlexLC) (Ruas and Grosky, 2017), which extracts lexical chains, evaluating if a new word, represented by the BSID(w_i), or its hypernyms, present lexical cohesion among themselves and the current chain under construction. If the evaluated synset has semantic affinity with the chain being constructed, then this new synset is incorporated to the chain. Otherwise, a new chain must be initialized so that the next semantic representation can be captured.

The idea behind the algorithm presented in (Ruas and Grosky, 2017) is quite simple. As long as synsets have a common meaning (even a more general one), they will be part of the same set (chain), otherwise a new set must be created. To illustrate the FlexLC algorithm, consider the sentence “*the dog and the cat run with the child and her mom in the park, this Summer*”. After cleaning the data and applying the BSD algorithm, we only keep the BSIDs for the words that have a match in WN, producing the following list $\{dog,$

$cat, child, mom, park, summer\}$. The chain starts with BSID(dog) and evaluates BSID(cat), both of which have the hypernym “*carnivore*” in common, so BSID(“*cat*”) is added to the chain and BSID(*carnivore*) is set as the ID for the current chain under construction. Next, BSID(*carnivore*) is evaluated with BSID($child$), which has the hypernym “*organism*” in common. BSID($child$) is then added to the current chain and BSID(*organism*) is set as its new ID. Next, the other BSIDs are processed following the same idea. Since the hypernym *organism* (ID for the chain under construction) is also shared by BSID(mom), the latter BSID is also added to the chain. However, BSID($park$) and BSID($summer$) do not share any common synset with the current chain, or themselves, other than WN’s root (entity). In that case, they will have their own chain, resulting in the following structure $\{\{dog, cat, child, mom\}, \{park\}, \{summer\}\}$, where *organism*, *park* and *summer* represent, respectively, each flexible chain. More details about FlexLC algorithm is available in (Ruas and Grosky, 2017).

After all FlexLC are produced, we convert these flexible chains into fixed structures (Flex2Fix) to reduce the high dimensionality caused by the number of single-synset-chains produced in the previous step. We also want to mitigate the problem of two or more long flexible chains being separated by one single-synset-chain occurrence.

Each flexible chain in this step will have an ID (FlexLCID) that will be assigned to all component words (w_i) of this chain. For example, consider the flexible chain $\{\{dog, cat, puppy\}, \{park\}, \{summer\}, \{dog, cat, puppy\}\}$ represented by the IDs $\{\{animal\}, \{park\}, \{summer\}, \{animal\}\}$. These IDs are propagated to the BSIDs of the original chain, resulting in a new one with the following structure $\{\{animal, animal, animal\}, \{park\}, \{summer\}, \{\{animal, animal, animal\}\}\}$, which will be processed into fixed structures. In this project, we divide the FlexLCIDs in sets of 4 units, so considering our example, the new chains would have the following construction $\{\{animal, animal, animal, park\}, \{summer, animal, animal, animal\}\}$. Both, the first and the second chain, have the synset *animal* as the dominant one, causing the ID for these fixed chains to be recalibrated to $\{\{animal\}, \{animal\}\}$. In our experiments, using the chunk size of 4 provided the most diverse set of chains. Since the chains are originated from the FlexLC in this algorithm, there will not be a

common synset shared between different chains within our threshold, so we do not need to traverse WN for hypernyms again. Therefore, to track the dominant synset in each fixed chunk is enough. Figure 1 shows in detail the Flex-to-Fixed algorithm (Flex2Fix), while Figure 2 is a pictorial representation of the process itself.

```

for each word occurrence  $w_i$  in  $D$ , where  $i=(1, \dots, n)$ :
  set  $\text{synset}(w_i) = p$ , where  $w_i$  occurs in  $\text{FLC}(k)$  and  $p = \text{FLCID}(k)$ 
split  $D$  into fixed-sized chunks of  $k$  words each
for each chunk  $c_{w_j}$ , where  $j=(1, \dots, N\text{Chunks})$ :
  let  $w_{j,1}, \dots, w_{j,k}$  be the word occurrences in chunk  $c_{w_j}$ 
  let  $\theta_j = \langle \text{synset}(w_{j,1}), \text{synset}(w_{j,2}), \dots, \text{synset}(w_{j,k}) \rangle$ 
  represent chunk  $c_{w_j}$  by the dominant synset of  $\theta_j$ ,  $\text{dominant}(c_{w_j})$ 
  if  $\theta_j$  has no dominant synset, choose one randomly
return  $\text{dominant}(c_{w_j})$  for  $j=1, \dots, N\text{Chunks}$ 

```

Figure 1. Flex2Fix Algorithm.

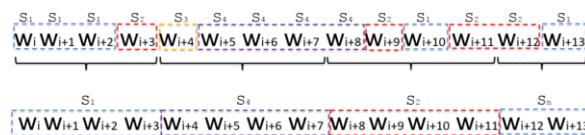


Figure 2. Flex2Fix Process.

In this paper, we also propose a variation to build fixed size chains called *Fixed Lexical Chains Algorithm* (FixLC), which is derived directly from the BSIDs found in Section 3.1. Differing from the previous algorithm (Flex2Fix), this does not use any pre-processed lexical chains, as its construction is entirely based on the BSIDs for each word. We develop this technique to compare which lexical chain structure would present better results, the one derived from FlexLC or obtained directly from BSIDs (FixLC). The latter “forces” a fixed dimensionality in the size of each chain from BSID’s, so we will need to consider the hypernyms in each fixed chunk.

The main idea behind the FixLC algorithm is to divide the BSIDs, for every document, in chunks of size n (c_n), and evaluate what is the synset that best represents each one of them these chunks. As in the previous approach (Flex2Fix), the size of 4 synsets was chosen, so both techniques could be better compared. For each chain c_n , we extract all hypernyms (including the initial synsets) from all the BSID in each chunk and select the dominant synset to represent the entire chain. If there is no dominant BSID, we select the deepest one in the chain. In case there are more than one, the choice for its representative synset is done randomly, since all of them could represent the given chain.

It is important to mention that hypernyms beyond a certain threshold are not considered in our approach. The closer to the root we get, the more common our synsets become, contributing poorly to the semantic diversity of a chain. Therefore,

hypernyms with depth below 5 (Hotho et al., 2003) are discarded. Figure 3 illustrates the FixLC algorithm in details.

```

select the set  $\lambda(c_d)$  of all hypernyms from each chain  $c_d$ 
select the set  $\beta$  of all synsets that appear in at least half of  $c_d$ 
if ( $\beta = \phi$ ) then  $\delta = \lambda(c_d)$  else  $\delta = \beta$ 
perform cut-off items in  $\delta$  based on a chosen limit, producing  $\epsilon$ 
if ( $\epsilon = \phi$ ) then  $\Omega = \delta$  else  $\Omega = \epsilon$ 
extract the set  $\alpha$  of all maximally occurring synsets in  $\Omega$ 
select the set  $\gamma$  of maximally deepest synsets in  $\alpha$ 
return a random synset from  $\gamma$  as representing the chain  $c_d$ 

```

Figure 3. Fixed Lexical Chains Algorithm (FixLC).

3.3 Semantic Dispersion

To explain (iv), we consider a document d . For each $1 \leq i \leq N\text{Synsets}$, we define $h(d,i)$ to be the histogram of relative distances (between 0 and 1) between consecutive occurrences of syn_i in document d . For this process, the number of bins of $h(d,i)$ and $h(e,j)$ will be the same for any 2 documents d, e , and synsets i, j . Also, for $h(d,i)$, if syn_i does not occur in document d , then the histogram consists of all 0’s. Document d is then represented by the normalized concatenation of $h(d,\text{syn}_1), h(d,\text{syn}_2), \dots, h(d,\text{syn}_{N\text{Synsets}})$.

We note that synsets occurring once present a problem, so we treat them in two ways: we either ignore them or not. To make sure these issues were covered, we explored three variations, considering the distances of synsets for each kind of chain (FlexLC, FixLC and Flex2Fix): (i) ignoring single occurrences of synsets, (ii) single occurrences of synsets have distance 0 from themselves and (iii) not ignoring single occurrences and treating all synsets as having a 0 relative distance from themselves. An example for each approach is shown in Table 1, which uses a 4-bin histogram for the same vector of 4 synsets illustrating (i), (ii) and (iii). For every synset, each histogram bin is initialized to 0.

Each bin is represented by a half-closed, half open set of relative distance ranges. Bin 1 corresponds to the set $[0,0.25)$, bin 2 to the set $[0.25,0.5)$, bin 3 to the set $[0.5,0.75)$, and bin 4 to the set $[0.75,1)$. Since each distance occurring in a synset string of length n is at most $n-1$, the largest relative distance possible is $(n-1)/n$, which approaches 1 as $n \rightarrow \infty$. Synsets which do not occur in a string, will have 0’s in all bins. In a nutshell, what our approach does is to characterize the spatial distribution (dispersion) of synsets in a document, using a histogram to keep track of those synsets by their relative distances. We note that using relative distances levels the representation playing field for all sizes of documents and treats them equally.

Map Type	Sequence of Synsets	Raw Distances	4-Bin Histogram Representation
I	S ₁ S ₂ S ₂ S ₄ S ₂ S ₃ S ₁	S ₁ <6>S ₂ <1,2>S ₃ <0>S ₄ <0>	<0,0,0,1> <1,1,0,0> <0,0,0,0> <0,0,0,0>
II	S ₁ S ₂ S ₂ S ₄ S ₂ S ₃ S ₁	S ₁ <6>S ₂ <1,2>S ₃ <0>S ₄ <0>	<0,0,0,1> <1,1,0,0> <1,0,0,0> <1,0,0,0>
III	S ₁ S ₂ S ₂ S ₄ S ₂ S ₃ S ₁	S ₁ <0,6>S ₂ <0,1,2>S ₃ <0>S ₄ <0>	<1,0,0,1> <2,1,0,0> <1,0,0,0> <1,0,0,0>

Table 1. Example of Mapping Distribution of Synsets in a 4-Bin Divided Document.

4 Experiments

To evaluate the proposed approaches, we used a corpus of 30 distinct documents from Wikipedia². These are distributed equally in three major categories: *dogs*, *computers* and *sports*. The *html* files of these pages were saved and parsed, so stopwords (e.g. “a”, “an”, “the”) could be removed. One might point out the small number of documents that comprise our corpus, in comparison with datasets used by statistical approaches in document similarity. However, we are proposing a semantic approach, in which every word has all its synsets examined by our algorithm. For our synset experiments, the number of synsets in our term/document matrix ranged between 1284 and 7490. In addition, the documents considered in this paper have, on average, 7,200 words each, which can produce a considerable dataset to process.

As explained in Section 3, during the preprocessing step we only maintain the nouns for each document having a synset match in WN. These steps help to remove features that do not contribute to our approach. By the end of this phase, our corpus has a total of approximately 216K words, of which 68K (nouns) have a match in WN. Table 2 shows in detail the documents/words used.

Wikipedia Category	Number of Documents	Number of Words	Nouns Matched in WN	Avg. of Nouns in WN (%)
Dogs	10	48,650	16,239	34.37
Computers	10	79,332	24,331	31.11
Sports	10	88,532	28,266	32.38
Total	30	216,514	68,836	32.62

Table 2. Wikipedia Dataset Details.

After all datasets are properly cleaned, we extract the BSID representation (Section 3.1), which is used as a base for all our lexical chains scenarios: FlexLC, FixLC and Flex2Fix. Once all flexible lexical chains are extracted from the documents, they are used to map into a fixed lexical

chain structure and to create the corresponding vector representations. We also derive FixLC directly from BSID vectors, using a fixed chain size, as shown in Section 3.2.

In our experiments, we validated our various approaches by performing a clustering task, using 256 bins for our synset-based techniques. As mentioned previously, we had documents from 3 major categories, so we performed a variant of *k-means* clustering for $k=3$ clusters and evaluated the resulting clustering using both the *Adjusted Rand Index* and the *Mean Individual Silhouette* values. The former metric is a measure of similarity between two clusters. We compared the derived clusters to the 3 ground truth clusters, consisting of all the *dog* documents, all the *computer* documents, and all the *sport* documents. The latter metric sees how well the clusters are designed, determining whether documents in the same cluster are close together, while documents in different clusters are far apart.

We used *spherical k-means clustering* (Hornik et al., 2012), as this technique uses the cosine distance (Han and Karypis, 2000) rather than Euclidean distance, and which has shown good results in clustering documents.

To validate the proposed algorithm, we also designed, implemented, and extended traditional approaches for document similarity, such as: BOW with all words (minus the stop-words) in the documents (BOWR), BOW with only matched nouns in WN (BOWN), BOW with the first synset match (most commonly used by other researchers) in WN (BOWS) and BOW with the BSID (BOWB) extracted from the BSD algorithm. Since the traditional approaches are variations of counts, only one bin is considered for these histograms. Table 3 provides a summary of all experiments performed. Figure 4 shows a scatter plot of these results. These results show that various permutations of our general approach worked better than others, and that four of our approaches stand out as better than the others.

² <https://doi.org/10.7302/Z26W980B>

Label	Algorithm	Adjusted Rand Index	Mean Individual Silhouette
A	Pure Flex--Method III	1	0.1908
B	Pure Flex--Method II	1	0.1775
C	BOW-N--Nouns in Wordnet	1	0.1757
D	BOW-B--Best Synsets	1	0.1686
E	Flex-2-Fixed--Method I	0.8981704	0.3964
F	Flex-2-Fixed--Method III	0.8981704	0.3878
G	BOW-R--Raw Words	0.8981704	0.1591
H	Flex-2-Fixed--Method II	0.8066667	0.3578
I	BOW-S--WordNet First Synset	0.6671449	0.1542
J	Pure Flex--Method I	0.6590742	0.1826
K	Pure Fixed--Method I	0.6044735	0.2137
L	Pure Fixed--Method III	0.5165853	0.2734
M	Pure Fixed--Method III	0.40252	0.2743

Table 3. Adjusted Rand Index and Mean Individual Silhouette.

The following observations are quite apparent:

- Three out of the four results with perfect clustering are from our techniques. Two of these perfect clusterings use flexible chains (considering their variations) while the third perfect clustering results from the methodology of finding the best synset representation for a document.
- The only perfect clustering result which is on the Pareto front (not dominated by another result), is the one which uses the third approach (iii) for extracting flexible chains.
- The clustering with the maximum silhouette value results from our first approach (i) to our technique for extracting Flex2Fix. This clustering is also on the Pareto front.
- The only clusterings on the Pareto front result from our techniques.

5 Final Considerations and Future Work

In this paper, we explored how extracted semantic features can aid in document retrieval tasks. Furthermore, we presented several contributions on how these features can be extracted to form more robust lexical chains. First, we explored the notion of WSD and how to represent words, considering the effect of their immediate neighbors in their meaning (BSD). Second, a new methodology to transform variable length size semantic chains (FlexLC) into fixed parametrized structures is proposed through the Flex2Fix algorithm. Third we proposed an algorithm to derive FixLC directly from semantic representations. Also, three variations of how to calculate the relative distance of those chains were explored. To establish a comparison with the proposed approaches, we compared them with traditional

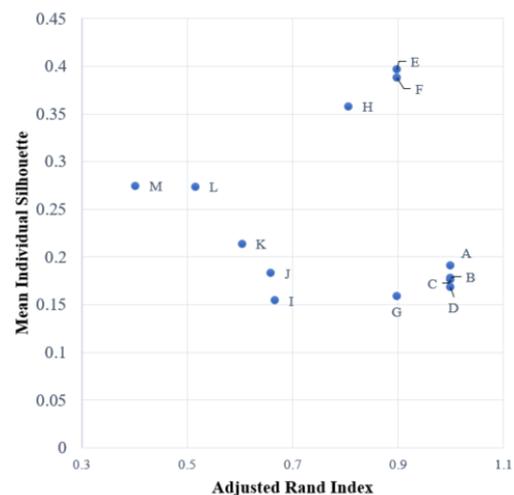


Figure 4. Scatter plot for Table 3 data.

ones, such as BOW and its variations (R/N/S/B). The comparisons showed that several of our approaches were the best performers.

Even though our model presents good results, we only touched the many possibilities that semantic features can offer in document retrieval and analysis. In future work, we intend to extend current algorithms for a more accurate representation of synsets and more solid lexical chains (fixed and flexible). In addition, we can also explore the effects of different WSD algorithms (e.g. Palmer, Leakcock & Chodorow, Lin, Resnik, Li) in the BSID choice and the construction of lexical chains (Meng et al., 2013). Other interesting linguistics challenges can be explored through the use semantic content extraction, such as: authorship identification, authorship profiling, clustering by concept structure, document summarization through concepts, and many other questions. The use of concepts, indeed, brings an interesting set of options that demands more time invested, so that its full potential can be reached.

References

- Iyad AlAgha and Rami Nafee. 2014. An Efficient Approach For Semantically-Enhanced Document Clustering By Using Wikipedia Link Structure. *International Journal of Artificial Intelligence & Applications*, 5(6):53–62.
- Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2015. Composing Measures for Computing Text Similarity. Technical report, Darmstadt.
- Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, 17(48):10–17.
- Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, 2(12):29–34.
- Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Journal Computational Linguistics*, 32(August 2005):13–47.
- Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285, New York, New York, USA. ACM Press.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. volume 71. MIT Press, Cambridge.
- Christiane Fellbaum. 2010. Theory and Applications of Ontology: Computer Applications. *Media*(2000):231–243.
- William I. Grosky and Terry L. Ruas. 2017. The Continuing Reinvention of Content-Based Retrieval: Multimedia Is Not Dead. *IEEE MultiMedia*, 24(1):6–11, January.
- Weiwei Guo and Mona Diab. 2011. Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions. In *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 552–561, Edinburgh.
- M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in english*. Longman Group, London.
- EH Han and George Karypis. 2000. Centroid-based document classification: Analysis and experimental results. *Principles of Data Mining and Knowledge Discovery*.
- Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet - An Electronic Lexical Database*(April):305–332.
- Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. 2012. Spherical k-Means Clustering. *Journal of Statistical Software*, 50(10):1–22.
- Andreas Hotho, Steffen Staab, and Gerd Stumme. 2003. Wordnet improves Text Document Clustering. In *In Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544.
- Jay J Jiang and David W Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics*(Rocling X):19–33, September.
- Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. *Proceedings of ICML*:296–304.
- Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, page 279–es, Morristown, NJ, USA. Association for Computational Linguistics.
- Lingling Meng, Runqing Huang, and Junzhong Gu. 2013. A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1):1–12.
- Dan Moldovan and Adrian Novischi. 2002. Lexical Chains for Question Answering. *Coling*:1–7.
- Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.
- Roberto Navigli. 2009. Word sense disambiguation. *ACM Computing Surveys*, 41(2):1–69, February.
- Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity. In *Demonstration Papers at HLT-NAACL 2004 on XX - HLT-NAACL '04*, number July, pages 38–41, Morristown, NJ, USA. Association for Computational Linguistics.
- Nitesh Pradhan, Manasi Gyanchandani, and Rajesh Wadhvani. 2015. A Review on Text Similarity Technique used in IR and its Application. *International Journal of Computer Applications*, 120(9):29–34.
- Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. , 1, November.

- Terry Ruas and William Grosky. 2017. Keyword Extraction Through Contextual Semantic Analysis of Documents. In *Proceedings of the 9th International Conference on Management of Emergent Digital EcoSystems*, Bangkok. ACM Press (To Appear).
- Julian Sedding and Dimitar Kazakov. 2001. WordNet-based Text Document Clustering. *Proceedings of the 3rd Workshop on ROBust Methods in Analysis of Natural Language Data*(1999):104–113.
- H Gregory Silber and Kathleen F McCoy. 2000. Efficient Text Summarization Using Lexical Chains. *Proceedings of the ACM Conference on Intelligent User Interfaces*:252–255.
- Joe Tekli. 2016. An Overview on XML Semantic Disambiguation from Unstructured Text to Semi-Structured Data: Background, Applications, and Ongoing Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1383–1407, June.
- Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. 2015. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4):2264–2275, March.
- Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* -, pages 133–138, Morristown, NJ, USA. Association for Computational Linguistics.

Toward A Semantic Concordancer

Adam Pease

Articulate Software
San Jose, USA

apease@articulatesoftware.com

Andrew K. F. Cheung

Hong Kong Polytechnic University
Hong Kong SAR, China

andrew.cheung@polyu.edu.hk

Abstract

Concordancers are an accepted and valuable part of the tool set of linguists and lexicographers. They allow the user to see the context of use of a word or phrase in a corpus. A large enough corpus, such as the Corpus Of Contemporary American English, provides the data needed to enumerate all common uses or meanings.

One challenge is that there may be too many results for short search phrases or common words when only a specific context is desired. However, finding meaningful groupings of usage may be impractical if it entails enumerating long lists of possible values, such as city names. If a tool existed that could create some semantic abstractions, it would free the lexicographer from the need to resort to customized development of analysis software.

To address this need, we have developed a Semantic Concordancer that uses dependency parsing and the Suggested Upper Merged Ontology (SUMO) to support linguistic analysis at a level of semantic abstraction above the original textual elements. We show how this facility can be employed to analyze the use of English prepositions by non-native speakers.

We briefly introduce concordancers and then describe the corpora on which we applied this work. Next we provide a detailed description of the NLP pipeline followed by how this captures detailed semantics. We show how the semantics can be used to analyze errors in the use of English prepositions by non-native speakers of English. Then we provide a description of a tool that allows users to build seman-

tic search specifications from a set of English examples and how those results can be employed to build rules that translate sentences into logical forms. Finally, we summarize our conclusions and mention future work.

1 Introduction

Concordancers¹ enable the linguist to see the context of use of words or phrases. This is valuable in understanding how a word can have different senses, or in finding rules or exceptions for collocations. One issue for the linguist using such tools is that many linguistic constructions are *patterns* or types, rather than literal collections of words. We “take a pill” but “eat a muffin”, we “play music” but “draw a picture”, “fly a plane” but “drive a car” or “pilot a boat”. For each of the nouns, a class or group determines the verb (such as “medicine”, “2-D art” or “aircraft”), but enumerating those possibilities is cumbersome. A computational linguist could develop customized analysis software, but no general purpose tool fit for this task appears to exist. We have developed software that allows the linguist to specify dependency relations and semantic types, based on a formal ontology, that can alleviate the need to enumerate large numbers of alternative strings of search terms with a conventional concordancer.

2 Corpora

To motivate development of this software we have two use cases. The first case is in analysis of corpora for classes of errors in usage that are common for non-native speakers of English. We chose to look at a small corpus of translated speech and analyze it for these classes of errors. In this way, we can provide specific feedback to translators on

¹such as http://www.antlab.sci.waseda.ac.jp/antconc_index.html and <https://www.lex tutor.ca/conc/eng/>

what problems to avoid in the future. To augment this work, we also examined a larger and broader corpus of non-native English usage, in order to help validate the utility of the tool on a corpus that has more, and more obvious, usage errors. We begin with a corpus of legal judgments translated from Chinese into English.

Judgments translated from Chinese into English are essential to the rule of law in Hong Kong. Hong Kong is the only common law jurisdiction where Chinese and English languages are used alongside each other in the judicial system (Cheng and He, 2016). Judgments form an essential part of common law. Because the majority of the population is Chinese speaking, court cases are sometimes heard in Chinese. Judgments in these Chinese cases are written in Chinese. Judgments of cases with jurisprudence value are translated into English. These translated judgments may be used in the future by legal professionals who are not necessarily familiar with the Chinese language. Translated English judgments were downloaded from the Hong Kong Judiciary website² to build the Hong Kong translated English judgments corpus.

Non-native speakers can find it challenging to use English prepositions properly. Compared to English, Chinese is a verb heavy language. The Chinese language has significantly fewer prepositions than the English language does. Unlike English, Chinese sentences without prepositions are grammatically correct and comprehensible (Shih, 2012). Chinese speakers, even with good English language abilities, may not be as sensitive to the use of prepositions when using the English language. Therefore, one of the challenges facing Chinese speakers when translating into English is the accurate use of prepositions.

After removing titles, headings and other incomplete sentences in the legal corpus, we arrived at 8818 sentences in suitable for further processing by our semantic concordancer.

To broaden our study, we also examined the Cambridge Learner Corpus (CLC)³ (Yannakoudakis et al., 2011), which has a greater number of English usage errors and is roughly twice the size of our legal corpus, at 16068 lines of text, also ignoring titles and headings.

Our second use case is in validating linguistic

²<http://www.judiciary.hk/en/index/>

³<https://www.illexir.co.uk/datasets/index.html>

patterns and creating rules to translate language to logical forms, for which we employ two large corpora of native English writing. These are the Corpus Of Contemporary American English (COCA) (Davies, 2008) and 2722 articles from Wikipedia converted to plain text⁴.

3 NLP Pipeline

Our work relies upon the Stanford CoreNLP (Duchi et al., 2011) pipeline, which is free and open source, and either the top performing system or at least state of art on each element of its pipeline. The system is structured as a series of *annotations* on tokens. Each annotator builds up annotations on the textual input.

To illustrate the pipeline, let's take a particular example.

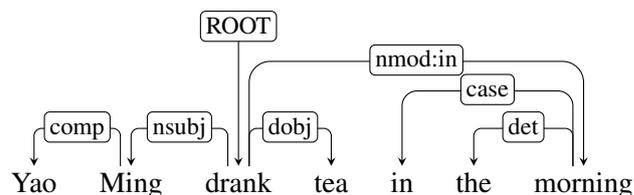
- (1) Yao Ming drank tea in the morning.

The Stanford Named Entity Recognizer (Finkel et al., 2005) identifies linguistic references to things like names and dates. It results in the following markings of our example (where "O" is a tag for "other", meaning not a named entity)

- (2) PERSON PERSON O O O TIME
Yao Ming drank tea in the morning

We have added a multi-word phrase recognizer to the CoreNLP pipeline that uses WordNet and SUMO as dictionaries. Matching multi-word elements are reduced to a single token, so "Yao Ming" or "July 23rd" will become a single token with a class membership in SUMO (Human or Day respectively here).

Dependency parsing (Chen and Manning, 2014) abstracts parse trees into a set of linguistic relations that are as independent of language as possible. We have the following dependency graph for example 1:



Note that dependencies as a data structure can also be represented as just a list of triples.

⁴<http://www.evanjones.ca/software/wikipedia2text.html>

```

root (ROOT-0, drank-2)
compound (Ming-2, Yao-1)
nsubj (drank-3, Ming-2)
dobj (drank-3, tea-4)
case (morning-7, in-5)
det (morning-7, the-6)
nmod:in (drank-3, morning-7)

```

CoreNLP lacks a module for determining word senses so we have utilized our existing system from (Pease and Li, 2010). This process normally addresses just nouns, verbs, adjectives and adverbs. Determining named entities is done in the NER system described earlier. WSD annotations are shown as example 3, and definitions for some different senses of “tea” are shown in table 1.

```

(3) Yao Ming drank      tea      in the
    .      .      201170052 107933274 . .
    morning
    115165289

```

These IDs are for WordNet 3.0 (Fellbaum, 1998) (with the part of speech number prepended) and they have been manually linked to the Suggested Upper Merged Ontology (SUMO)⁵ (Niles and Pease, 2001; Pease, 2011). Since the original mapping effort in 2002, tens of thousands of synsets have been remapped to more specific SUMO terms as they have been defined. In particular, several thousand have been remapped in 2017 alone. The current statistics for the mappings are shown in Table 2. Note that a small number of adjectives and adverbs have not been mapped.

Instance mappings are from a SUMO term to a particular instance synset in WordNet, such as SUMO’s *Battle* mapping to WordNet’s “Battle of Britain”. *Equivalence* mappings are close but informal equivalences, such as the mapping between SUMO’s *Cloud* and WordNet’s synset 109247410 “a visible mass of water or ice particles suspended at a considerable altitude.” *Subsuming* mappings are between specific WordNet synsets and more general SUMO terms, such as “Meniere’s.disease” and SUMO’s *DiseaseOrSyndrome*. Of note is that recently, with the growth of SUMO in several domains, we increasingly have need for what we might term a “subsumed-by” relation, where a SUMO term is more specific than any available WordNet synset, as is the case with the new ontologies of Law and Weather. This relation is likely to appear in a future release of the mappings.

⁵<http://www.ontologyportal.org>

We also augment the WordNet lexicon with lexical entries provided in the ontology for each new term, such as the string “mono crystalline” being associated with the recently-added SUMO term *MonoCrystalline*.

To perform word sense disambiguation, we rely on WordNet SemCor (Landes et al., 1998), a corpus of manually-marked word senses, indexed to the WordNet semantic lexicon, and annotated on the Brown Corpus of English (Kucera and Francis, 1967). For each word sense, we create a table counting the frequency of co-occurring words in the corpus. We use a frequency threshold so that low-frequency senses that have little co-occurrence data aren’t influenced by random small amounts of data. One criticism of WordNet has been that it makes some overly fine distinctions among word senses (Snow et al., 2007). We use the SUMO-WordNet mappings to collapse senses that map to the same term in the ontology. Note however that this grouping is much more fine grained than the coarse-grained aggregation to categories done in SemEval-17 on OntoNotes (Pradhan et al., 2007b), so that fewer (if any) meaningful distinctions in sense are lost. This approach has the added benefit of increasing the statistical significance of some of the merged co-occurrence relationships. This approach however does not perform as well as some recent effort in WSD that employ machine learning, such as (Zhong and Ng, 2010). When tested on the OntoNotes corpus (Pradhan et al., 2007a) we achieve roughly 66% accuracy, which approaches the score (stated at 72% in (Brown et al., 2010)) for inter-annotator agreement on fine grained senses. Since we cannot assume a particular domain, accuracies are likely to be lower than the best results of other reported studies (Zhong et al., 2008). However, it is likely that more training data from a wider set of corpora⁶ will help improve performance.

We augment Stanford dependency parses with SUMO terms. Continuing the example above, we add the triples

```

sumo (Drinking, drank-3)
sumo (Morning, morning-7)
sumo (Tea, tea-4)

```

While SUMO does have a taxonomy, it also has definitions in a higher order logic that explain, in a computable way, the meaning of each term. So,

⁶<https://github.com/getalp/LREC2018-Vialetal>

sense key	words	definition
107575510	tea, teatime	a light midafternoon meal
107933274	tea	a beverage made by steeping tea leaves in water
107932841	tea, tea_leaf	dried leaves of the tea shrub

Table 1: Word senses (definitions and word lists shortened from WordNet)

	instance	equivalence	subsuming
noun	9,570	6,505	67,914
verb	0	971	13,204
adjective	730	596	14,832
adverb	57	119	3,222
total	10,357	8191	99,172

Table 2: SUMO-WordNet mapping statistics (117,720 total synsets mapped)

for the example of `Drinking` we have logical axioms such as

```
(=>
  (attribute ?A Thirsty)
  (desires ?A
    (exists (?D)
      (and
        (instance ?D Drinking)
        (agent ?D ?A))))))
```

that states that being `Thirsty` implies a desire to drink something. Axioms such as this are more specific and detailed than entailment links and can enable further logical reasoning.

We have linked the Stanford 7-class NER model to SUMO types, which allows us to assert

```
sumo (Human, Yao_Ming-1)
```

from the NER output shown in example 2.

We also employ Stanford’s `SUTime` (McClosky and Manning, 2012) to recognize temporal expressions. If we have the slightly modified example

(4) Yao Ming drank tea in July.

we would add the clauses.

```
month(time-1, July)
time(drank-3, time-1)
```

Although the current semantic concordancer system does not employ logical deduction, the information captured would allow us to use SUMO’s temporal axioms and its associated `E Theorem Prover` (Pease and Schulz, 2014) to do simple temporal reasoning, and further expand the possibilities of searching for semantic patterns to include relative periods like “before June” or “during 2016” and return sentences that meet those constraints rather than a literal pattern of words.

4 Semantic Concordance

Concordancers are very useful for checking intuitions with respect to language usage. Searching on a word or phrase provides samples of usage in context. But not all language patterns are strict phrases. Idioms can have insertions (Minugh, 2007), such as “drop in the bucket” being modified to “drop in the proverbial bucket” or “drop in the fiscal bucket” but not “He put a drop of water in the bucket”. Being able to search a dependency parse for a grammatical pattern rather than a literal string or even a string with wildcards may be a useful tool.

Some patterns of usage are selected with respect to the types of participants in a phrase, rather than particular words. These can be quite specific. For example, if a linguist wants to examine usage of the preposition “in” in its physical, rather than temporal sense, an exhaustive number of searches would be required to enumerate physical words and phrases and temporal words or phrases. However, given that we have dependency parse forms and SUMO terms we can search for patterns such as:

```
nmod:in(?X, ?Y), sumo(?C, ?Y),
isSubclass(?C, TimePosition)
nmod:in(?X, ?Y), sumo(?C, ?Y),
isSubclass(?C, Object)
```

To carry on with example 1, note how the first pattern involving `TimePosition` above matches with the clauses of the example, and the variables are bound to `?X=drank-3`, `?Y=morning-7` and `?C=Morning`.

```

root (ROOT-0, drank-3)
det (morning-7, the-6)
nmod:in (drank-3, morning-7)
sumo (Human, Yao.Ming-1)
sumo (Drinking, drank-3)
sumo (Morning, morning-7)
names (Yao.Ming-1, "Yao")
dobj (drank-3, tea-4)
case (morning-7, in-5)
sumo (Tea, tea-4)
names (Yao.Ming-1, "Ming")
nsubj (drank-3, Yao.Ming-1)

```

While WordNet noun synsets could be used to capture common classes of words, SUMO provides extra utility when searching for groups of verbs. For example, one “looks for” or “searches for” something in order to find it and some language learners may omit the preposition. In each case there is a mapping to SUMO’s *Searching*, but no common hypernym for those WN 3.0 senses (201315613 and 202153709, respectively).

Because WSD and dependency parsing are not always correct, it is necessary to review results rather than simply tabulating them. Also, language is flexible, and what constitutes “correct” usage is more like correspondence to a preponderance of use than a strict rule in many cases.

5 Preposition Errors

We looked for common errors in preposition usage⁷ in our corpora of non-native English. The first error type that was searched for was the use of prepositions with times of day (see example 5), where “night” is an exception.

- (5) ... in the morning ...
 * ... at the morning ...
 ... in the evening ...
 * ... at the evening ...
 ... at night ...
 * ... in night ...

We can state the (ungrammatical) dependency pattern

```

nmod:at (?X, ?Y), sumo (?C, ?Y),
isSubclass (?C, TimeInterval)

```

One sentence found in the corpus was example 6,

- (6) “We usually have lessons at the morning, till afternoon.”

This sentence has the augmented dependency parse of

```

root (ROOT-0, have-3)
nsubj (have-3, We-1)
advmod (have-3, usually-2)
dobj (have-3, lessons-4)
case (morning-7, at-5)
det (morning-7, the-6)
nmod:at (lessons-4, morning-7)
case (afternoon-10, till-9)
nmod:till (have-3, afternoon-10)
sumo (SubjectiveAssessmentAttribute, usually-2)
sumo (EducationalProcess, lessons-4)
sumo (Morning, morning-7)
sumo (Afternoon, afternoon-10)

```

Other examples of linguistic errors in the corpus found by matching dependency patterns are

- (7) * I’ve been working here since five years.
 * If Tang Dan-dan was also manipulated as was the applicant, she should have arrived at Hong Kong as scheduled.

6 Query Composition

One of the challenges in using this tool is that it requires some knowledge of dependency parsing and SUMO. To address this, we have created a component that find the common structure of several sentences and returns a dependency parse for that common structure. That specification can then be used to search for other sentences that match the pattern. In this way, the linguist simply has to prepare several sentences that illustrate a common construction and let the system do the work to state the commonality in a formal language.

Take for example the following two sentences

- (8) John kicks the cart.
 (9) Susan pushes the wagon.

which produce the following respective augmented dependency parses -

```

root (ROOT-0, kicks-2)
det (cart-4, the-3)
names (John-1, "John")
sumo (Wagon, cart-4)
sumo (Kicking, kicks-2)
nsubj (kicks-2, John-1)
dobj (kicks-2, cart-4)
attribute (John-1, Male)
sumo (Human, John-1)

```

⁷<http://blog.oxforddictionaries.com/2017/01/preposition-mistakes-for-english-learners/>

```

root (ROOT-0, pushes-2)
det (wagon-4, the-3)
names (Susan-1, "Susan")
attribute (Susan-1, Female)
sumo (Pushing, pushes-2)
sumo (Human, Susan-1)
dobj (pushes-2, wagon-4)
nsubj (pushes-2, Susan-1)
sumo (Wagon, wagon-4)

```

We can then produce their common, unified abstraction as follows, in which labels with question marks denote variables -

```

root (ROOT-0, ?B)
det (?D, ?C)
names (?A, ?E)
attribute (?A, SexAttribute)
sumo (Motion, ?B)
sumo (Human, ?A)
dobj (?B, ?D)
nsubj (?B, ?A)
sumo (Wagon, ?D)

```

Note that the expression can be verified to unify with the original dependency parses, using the following substitutions for sentence 8 as an example.

```

?A=John-1
?B=kicks-2
?C=the-3
?D=cart-4

```

A linguist who does not have the facility to write dependency parses or use SUMO can simply use the resulting expression as a “black box” search input to the concordancer. A future version of the system could even have an option to hide it entirely, thereby performing a form of semantic search.

7 Semantic Rewriting

The Semantic Concordancer is an intermediate result from efforts to translate language into logic. We are extending prior work on the Controlled English to Logic Translation (Pease and Li, 2010) to use modern parsing techniques with Stanford’s CoreNLP instead of a restricted English grammar.

When the semantics of sentences are fully captured it opens up opportunities for deductive reasoning that goes beyond simple retrieval of previous sentences. It also creates the possibility to vet utterances for contradictions with known facts about the world, thereby allowing a system to exclude faulty parses based on world knowledge.

For example, the simple sentence 8 above becomes the following first-order logic sentence with SUMO terms -

```

(exists (?John-1 ?cart-4 ?kicks-2)
  (and
    (agent ?kicks-2 ?John-1)
    (attribute ?John-1 Male)
    (names ?John-1 "John")
    (patient ?kicks-2 ?cart-4)
    (instance ?cart-4 Wagon)
    (instance ?kicks-2 Kicking)
    (instance ?John-1 Human) ) )

```

The process of accomplishing this is what we call Semantic Rewriting, and is based on previous efforts called Transfer Semantics or Packed Rewriting (Crouch, 2005; Crouch and King, 2006). It involves the iterative application of production rules to dependency parses. In the case of sentence 8 this involves execution of just two rules (along with a simple mechanical listing of the types of terms with instance and generation of the name of “John” as a male human from a common name database) -

```

dobj (?E, ?Y) ==> (patient (?E, ?Y)).
line 1041 : {?E=kicks-2, ?Y=cart-4}

nsubj (?E, ?X), sumo (?A, ?E),
isSubclass (?A, Process), sumo (?C, ?X),
isSubclass (?C, Agent) ==> (agent (?E, ?X)).
line 1063 :
{?X=John-1, ?A=Kicking, ?C=Human,
?E=kicks-2}

```

The first rule is a general default that if we have no more specific pattern, the direct object in a sentence becomes the “patient” in a SUMO expression. The second rule is more interesting. It states that if the grammatical subject of a Process is an Agent (rather than some inanimate object) then we generate a SUMO agent relationship between the entity and the process.

While creating a few simple rules of this sort is easy, as the rule set grows and the remaining rules become more complex, authoring them through introspection become impractical. The Query Composition tool described above provides a principled way to create patterns by example, which form the left hand side of a Semantic Rewriting rule. The Semantic Concordancer then becomes useful as a way to validate the prevalence of a particular pattern of language use in a large corpus.

8 Conclusions and Future Work

The software is available open source at <https://github.com/ontologyportal> and has been used on a practical pilot project in analysis of non-native English. We expect to apply it further to more systematic studies in this area as

well as others. The implementation is in Java, using the H2 database⁸. All the words in each sentence and terms in dependency parses are indexed, so all semantic processing occurs at the time the database is built, rather than when a query is run. After sentences and dependencies matching a bag of terms are returned, a simple unification algorithm attempts to match the dependency parse literals with the dependency parse query, similar to a Prolog-style unification algorithm (Baader and Snyder, 2001). This enables the system to scale well to the requirements of modern large corpora.

We are employing the Semantic Concordancer and its associated Query Composition tool to create and validate semantic rules that translate language into logical expressions.

The system will be available by the time of GWC2018 on a server at <https://nlp.ontologyportal.org:8443/sigmanlp/semconcor.jsp>.

References

- Baader, F. and Snyder, W. (2001). Unification theory. In *Handbook of Automated Reasoning (in 2 volumes)*, pages 445–532.
- Brown, S. W., Rood, T., and Palmer, M. (2010). Number or nuance: Which factors restrict reliable word sense annotation? In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.
- Chen, D. and Manning, C. D. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of EMNLP 2014*.
- Cheng, L. and He, L. (2016). Revisiting judgment translation in hong kong. *Semiotica*, (209):59–75.
- Crouch, R. (2005). Packed rewriting for mapping semantics to KR. In *Proc. 6th Int. Workshop on Computational Semantics*, pages 103–114, Tilburg.
- Crouch, R. and King, T. H. (2006). Semantics via f-structure rewriting. In *Proceedings of LFG06*, pages 145–165.
- Davies, M. (2008). The corpus of contemporary american english (cocca): 520 million words, 1990-present. available online at <https://corpus.byu.edu/cocca/>.
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*.
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.
- Kucera and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.
- Landes, S., Leacock, C., and Tengi, R. (1998). Building semantic concordances. In Fellbaum, C., editor, *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication. MIT Press, Cambridge (Mass.).
- McClosky, D. and Manning, C. D. (2012). Learning constraints for consistent timeline extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Minugh, D. C. (2007). The filling in the sandwich: internal modification of idioms. *Language and Computers*, 62(1):205–224.
- Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Welty, C. and Smith, B., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.
- Pease, A. (2011). *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.
- Pease, A. and Li, J. (2010). Controlled English to Logic Translation. In Poli, R., Healy, M., and Kameas, A., editors, *Theory and Applications of Ontology*. Springer.
- Pease, A. and Schulz, S. (2014). Knowledge engineering for large ontologies with sigma kee 3.0. In Demri, S., Kapur, D., and Weidenbach,

⁸www.h2database.com/

- C., editors, *Automated Reasoning: 7th International Joint Conference, IJCAR 2014, Held as Part of the Vienna Summer of Logic, VSL 2014, Vienna, Austria, July 19-22, 2014. Proceedings*, pages 519–525. Springer International Publishing.
- Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007a). Ontonotes: a unified relational semantic representation. *Int. J. Semantic Computing*, 1(4):405–419.
- Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (2007b). Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07*, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Shih, C. L. (2012). A corpus-aided study of shifts in english-to-chinese translation of prepositions. *International Journal of English Linguistics*, 2(6).
- Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to merge word senses. In *EMNLP-CoNLL*, pages 1005–1014. ACL.
- Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
- Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations, ACLDemos '10*, pages 78–83, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Zhong, Z., Ng, H. T., and Chan, Y. S. (2008). Word sense disambiguation using ontonotes: An empirical study. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1002–1010.

Using OpenWordnet-PT for Question Answering on Legal Domain*

Pedro Delfino

FGV Direito Rio and EMap/FGV

Bruno Cuconato

EMAp/FGV

Guilherme Paulino-Passos

COPPE/UFRJ and IBM Research

Gerson Zaverucha

COPPE/UFRJ

Alexandre Rademaker

IBM Research and EMap/FGV

Abstract

In order to practice a legal profession in Brazil, law graduates must be approved in the OAB national unified bar exam. For their topic coverage and national reach, the OAB exams provide an excellent benchmark for the performance of legal information systems, as it provides objective metrics and are challenging even for humans, as only 20% of its candidates are approved. After constructing a new data set on the exams and doing shallow experiments on it, we now employ the OpenWordnet-PT to verify whether using word senses and relations we can improve previous results. We discuss the results, possible future ideas and the additions to the OpenWordnet-PT that we made.

1 Introduction

Automatic analysis of legal content offers opportunities for improving the effectiveness of legal actors, transparency of the system and, ultimately, the welfare of the public. As law is practiced with language itself, linguistic approaches are invaluable. This focus on language and higher demand for precision created by a technical domain makes it natural to try to grow upon and evaluate the performance of a lexical-semantic resource, such as wordnets, in this area.

One task for legal technology is question answering: an automatic way of determining the right answer to a question presented in natural language form (Mitkov, 2005). An ideal legal question answering system would take a question in natural language and a corpus of all legal documents in a given jurisdiction, and would return both a correct answer and its legal foundation (answer justification), i.e., which sections

The authors would like to thank João Alberto de Oliveira Lima for introducing us to the LexML resources.

(or articles) of which norms provide support for the answer. Considering lack of knowledge about facts, incompleteness, inconsistency or disagreements between sources of law, an ideal system would generate each possible answer with corresponding arguments, explanations and confidence value. Since such a system is still far from our current capabilities, as the results of recent evaluation tasks such as ResPubliQA (Peñas et al., 2010) has shown, we started with a simpler task.

In Brazil, even after graduating from Law school, it is required that one is approved in the OAB exam in order to practice a legal profession. The “Ordem dos Advogados do Brasil” (Order of Attorneys of Brazil, OAB) is the professional body of lawyers in Brazil. The first stage of the exam is a multiple-choice test. We are interested in investigating the performance of simple methods in answering this test correctly, and providing justifications for its answers. We measure the impact of the usage of an open lexical resource such as wordnet, and also promote its expansion into the legal domain by demand. In particular, we use FreeLing (Carreras et al., 2004b) for linguistic analysis, and evaluate specially the usage of the word sense disambiguation (WSD) module (Padró et al., 2010), which in Portuguese, uses openWordnet-PT (de Paiva et al., 2012) (OWN-PT). We find that the system does not improve considerably over the performance of our previous effort (Delfino et al., 2017); however, this might be because of missing concepts and relations in OWN-PT, which in turn render some of Freeling’s processing inaccurate.

In Section 2 we present the data-set we created and made available for experimentation. In Section 3 we discuss our previous experiments with the data-set, while in Section 4 we describe the tools and resources we employed for our current experiment: Freeling, OWN-PT, and the word sense disambiguation algorithm UKB (Agirre and Soroa, 2009). In Section 5 we describe the meth-

ods used in our experiments and then discuss its results in Section 6. Finally, we conclude and debate future works in Section 7.

2 The OAB Exams data set

Among other responsibilities, OAB is responsible for the regulation of the legal profession in the Brazilian jurisdiction. One of the key ways of regulating the legal practice is through the “Exame Unificado da OAB” (unified bar examination), required for enrolling at OAB, which is mandatory to practice law.

In order to be approved in the OAB exam, candidates need to be approved in two stages. The first phase consists of multiple choice questions, while the second phase involves free-text questions. Since 2012, the first phase has 80 multiple choice questions and each question has 4 alternatives. Candidates are asked to choose the correct alternative and in order to be approved, candidates need at least a 50% performance. Historically, the exam has had a global 80% failure rate, with the first stage being responsible for eliminating the majority of the candidates (Amorim and Tebechrani Neto, 2016).

Thus, the first stage of the OAB exams provides an excellent benchmark for the performance of a system attempting to reason about the law. That is, passing the OAB exam would signal that the system has acquired important aspects of legal knowledge, up to a level comparable to a human lawyer. In trying to build such a system, it was necessary to create the appropriate data sets, which includes not only the questions and answer keys in machine readable format, but also the legal literature involved (Delfino et al., 2017).

In previous work (Delfino et al., 2017), we have obtained the PDF files of the all the previous OAB exams, extracted their text, cleaned them up and made the data freely available in a public repository¹.

Along with the 1820 questions (from 22 exams) in plain text and in XML, it contains a golden set of 30 questions which were manually analyzed and annotated with the answer keys’ legal basis, i.e., which articles from which norms justify the correct answer to the question. These 30 questions are on a single subject, legal ethics.

Since 2012, the exams have revealed a pattern for which areas of Law the examination board fo-

cuses on and in which order the questions appear on the exam. Traditionally, the first 10 questions are about legal ethics, that is, the rights, the duties and the responsibilities of the lawyer in regard to Brazilian law. We have chosen to provide a golden set on legal ethics because this subject area is the simplest part of the exam with respect to the legal foundations of the questions. It also has a high frequency rate, and the highest performance rate among candidates (65%) (Amorim and Tebechrani Neto, 2016).

The key finding from analysis done in our previous work is that, usually, only one article on federal law no. 8906 was enough to justify the answer to the legal ethics questions (15 questions). Less often, in four questions, the justification was in “Regulamento Geral da OAB” (OAB General Regulation), or on the “Código de Ética da OAB” (OAB Ethics Code, 7 questions). Three other questions were justified by two articles in law no. 8906 each, and one question only in case law from the Superior Court of Justice about an article from the law no. 8906. Federal law no. 8906 has 89 articles, while the OAB general regulation has 169 articles, and the OAB ethics code has 66 articles.

2.1 Brazilian law texts

Another critical component of our data set is Brazilian legal norms in machine-readable format. This resource is essential for employing legal knowledge in answering the exam questions.

For the experiments made on the golden set, we needed the three normative documents (see Section 2) in a machine readable format. Moreover, we needed the documents in a format that preserved the original internal structure of the documents, i.e., the sections, articles, and paragraphs.

In order to obtain this data, we employed a legal document parser,² provided by the LexML project (de Oliveira Lima and Ciciliati, 2008). The LexML is a joint initiative of the Civil Law legal system countries seeking to establish open standards for the interchange, identification and structuring of legislative and court information. The goal is to convergence the national standards to international standardization of some instruments, such as URN-LEX, the use of XML formatting standards and the exchange of its metadata.

¹<http://github.com/own-pt/oab-exams>

²<https://github.com/lexml/lexml-parser-projeto-lei>

The LexML parser, still in beta, receives as input a DOCX³ file with the norm and outputs it in XML format, using the tags and the structure following the conventions of the LexML schema (de Oliveira Lima and Ciciliati, 2008). We had to make minor modifications in the three documents before submitting them to the parser; the XML files produced and the modifications made are available in our repository.

3 The previous work

Question answering in legal domain is a hard problem. In the last ResPubliQA evaluation task, the only system that dealt with Portuguese texts, the Priberam system, has the worst performance among the competitors, obtaining only 0.56 in the C@1 score (Peñas et al., 2010).⁴

In (Fawei et al., 2016) the authors report a textual entailment study on the US Bar exam material. In the experiment, the authors treat the relationship between the question and the multiple-choice answers as a form of textual entailment. Answering a multiple choice legal exam is a more feasible challenge, although it is still a daunting project without restrictions on the input form. That is the reason we have chosen in (Delfino et al., 2017) to restrict the domain to a single section of the OAB exams: legal ethics, one which is governed by only a few legal norms. In (Delfino et al., 2017), we conducted 3 experiments in question answering (section 5). In the first experiment, they tried to find the right answer between the multiple-choice alternatives. The last 2 were in shallow question answering (SQA), a form of question answering where a system retrieves documents that justify the already provided answer. They have adapted the methodology described in (Monroy et al., 2008; Monroy et al., 2009) to answer multiple-choice exams instead of closed-ended answers.

A range of issues on the texts of the questions of the exams was identified. Many of the problems are similar to the ones found in the US bar exams and described by (Fawei et al., 2016). For instance, some questions do not contain an introductory paragraph defining a context situation for the question. Instead of that, they have only meta comments (e.g. “assume that...” and “which of the following alternative is correct?”) followed by the

³The Microsoft Word editor format, commonly used for Brazilian legal documents.

⁴We were not able to obtain the article describing the Priberam system.

choices. Some questions are in a negative form, asking the examinee to select the wrong option or providing a statement in the negative form such as “The collective security order **cannot** be filed by...”. Moreover, some questions explicitly mention the law under consideration, others do not. Many questions present a sentence fragment and ask for the best complement among the alternatives, also exposed as incomplete sentences.

Even in the presence of such problems, our results in this previous work were not bad, given our system’s simplicity. But our initial approach also had its shortcomings: it could not distinguish successfully between two almost identical alternatives which differed only by few words (such as an alternative and its negation), nor could it treat related words in an appropriate manner. The former problem may require deep linguistic processing of the texts for properly obtaining the meaning of the utterances, while the latter can be partly tackled by the use of lexical resource such as the OWN-PT, as is done in this paper.

4 Freeling, OpenWordnet-PT and Word Sense Disambiguation

Freeling is an open source language processing library developed at the TALP research center⁵ (Carreras et al., 2004a; Padró and Stanilovsky, 2012). It has support for many languages, including English, Portuguese, among others. It implements modules for tokenization, sentence splitting, morphological analysis, part-of-speech tagging, word sense disambiguation, parsing and other tasks. FreeLing distribution includes linguistic data for the supported languages provided by many different projects and collaborators: morphological dictionaries, gazettes, lexical-semantic resources etc. Particularly, for Portuguese, its word sense disambiguation (WSD) module relies on OWN-PT, an open freely available wordnet for Portuguese (de Paiva et al., 2012).

FreeLing implements a pipeline-based approach. After tokenization, sentence split and the morphological analysis and part-of-speech tagging, the user can choose to execute the WSD module to search for senses in Wordnet matching the lemma and part-of-speech tag of each word or multi-word expression. Every possible sense is returned and may be weighted by the sense disambiguation module. The disambiguation is

⁵<http://nlp.cs.upc.edu/freeling/>

an implementation of the UKB algorithm (Agirre and Soroa, 2009), an unsupervised graph-based method which uses Personalized PageRank to select the right sense of each word in a lexical database such as OWN-PT.

Before running our experiment, we did a preliminary survey on the coverage of OWN-PT for the OAB corpus – a proxy for the legal domain as a whole. In Princeton Wordnet (PWN) (Fellbaum, 1998), the synset [08441203-n, *law/jurisprudence: the collection of rules imposed by authority.*] is a general concept about law, and is linked to hundreds of synsets via the `classifiesByTopic` relation. This suggests that PWN already covers (synset-wise) the relevant context, but it remained to be investigated whether such synsets are properly translated in OWN-PT with the relevant words, and if the existent concepts indeed encompass notions used in the Brazilian legal context, as legal jargon can be language and cultural dependant.

In order to further evaluate the coverage of the legal domain in OWN-PT we have taken a simple approach: after running Freeling on our corpus, we have listed the most common words whose senses Freeling could not find. We then proceeded to add them to OWN-PT. Some synsets did not seem to exist yet, such as one for “cartório” (notary office).⁶ Other synsets existed, but the word at hand was not included in it, as in [06532763-n, *nulidade: nullity*]. Other cases were those of relations that did not exist in OWN-PT; if present, these relations would improve the results of the UKB algorithm. One such relation that we included in OWN-PT was the nominalization (morphosemantic link) between [00664276-v, *comprovar: authenticate*] and [06855035-n, *comprovação: authentication*]. In the end, since we focused only on the possible improvements to our immediate purpose, we have added to OWN-PT two synsets, eight semantic and lexical relations, and 25 words.

After running our experiment (to be described in the next sections), we also reevaluated the legal domain coverage in OWN-PT. To do so we looked at the difference between the questions answered and justified correctly by our previous system (Delfino et al., 2017) and the present one. One observation is that even when the WSD was not

⁶We will make the data available as part of the OWN-PT distribution available at <http://wnpt.br/1cloud.com/wn/>.

done correctly, as when a Portuguese word that should be in the synset [06532095-n, *ato: legal act*] was assigned to the synset [00037396-n, *act: as in action*], these mistakes were consistent, so that terms in both legal norm and OAB question had been given the same senses. Surely, that is not the most desirable outcome, but at least does not impose a problem for our experiments.

The question below and the first article from law no. 8906 following it illustrate cases where Wordnet resources are helpful and a more shallow approach could fail. Even though article and question alternative are related, this relation is not captured by our previous algorithm, because it does not take into account anything but the equivalence of tokens. Using OWN-PT, we can exploit the relationship between the action (sign, “visar”) and the result of the action (signature, “visto”).

Constitutive acts and contracts of legal persons, in order to be registered regarding the legal practice statute, must: [...] C) contain the lawyer’s [...] **signature**. (17th ed. OAB exam, question 2)

§ 2º The constitutive acts and contracts of legal persons can only be registered in the competent bodies, under a penalty of invalidity, when **signed** by lawyers. (law no. 8906, article 1)

In the example above, however, OWN-PT was missing the words “visar” and “visto” in the appropriate synsets: [00996485-v, *sign, subscribe: mark with one’s signature*] and [06404582-n, *signature: your name written in your own handwriting*]. These missing senses, of course, had to be created before being properly linked by the morphosemantic link *result*.

During our evaluation, we also had to make some changes in the Freeling dictionary, some adjectives and their lemmas and part-of-speech tags were introduced. An important attribute of this approach is that it propagates. Extending the Wordnet and giving the right senses for some words can improve the classification of other words that were not changed directly due to correct part-of-speech tagging and adequate linking between senses, tasks which depend on neighboring words. The missing words, synsets and links in OWN-PT is both a problem and an opportunity: in order to make better use of OWN-PT for the task at

hand one must further extend it to the legal domain (Sagri et al., 2004).

5 Experiment Setup

The original idea for the experiment was inspired by (Monroy et al., 2008), and it runs as follows: one collects legal norms in a corpus and preprocesses them performing tasks such as converting text to lower case, eliminating punctuation and numbers and removing stop-words. After that, the articles of the norms are represented as Term Frequency - Inverse Document Frequency (TF-IDF) vectors in a Vector Space Model (VSM) (Manning et al., 2008). In (Delfino et al., 2017), we have adapted this method to deal with exam questions with multiple choice alternatives. In the present article, we relied on Freeling to incorporate more linguistic processing in our pipeline.

We use the Freeling tokenizer, sentence splitting, morphological analyzer (POS tagging and lemmatisation), and the WSD modules to assign OWN-PT synsets, with a weight value (normalized in order to sum 1), to each token or sequence of tokens. For an input text we thus have a list of key-value pairs (s, w) with a sense key and a weight value, in contrast to a simple list of tokens, as we had in the previous experiment.

The intuition behind TF-IDF is that the more similar two text fragments are, the lesser is the distance between them. As the articles of the norms are not lists of tokens anymore, we have adapted the TF-IDF definition to deal with the weights assigned to each synset, as Equation 1 shows.

$$\begin{aligned} \text{TFIDF}_{s,w,d} &= \text{TF}_{s,w,d} \text{IDF}_{s,w,D} & (1) \\ \text{TF}_{s,w,d} &= \frac{f_{s,w,d}}{\sum_{s' \in d} f_{s',w',d}} \\ \text{IDF}_{s,w,D} &= \log \left(\frac{|D|}{\sum_{d \in D} w^{\mathbb{1}(w < 1)} \mathbb{1}(s \in d)} \right) \end{aligned}$$

where $f_{s,w}$ is the sum of each occurrence of sense s weighted by w . Here $\mathbb{1}_X$ is the characteristic function for X : 1 if X is true and 0 otherwise. An intuitive explanation is that, for TF, we count the weighted occurrence as a “continuous occurrence”, instead of boolean, where the degree of occurrence is the weight of the sense. For IDF, if the sum in a document is higher than 1, then it counts as an occurrence, which is counted only once. Otherwise, it counts only according to the weight received.

A directed graph is then created, with a node for each article of the used norms. This is the base graph, used for answering all questions. When provided a question-answer pair, our system processes the question statement and the alternatives in the same way as it does to the articles in the base graph: turning them into a list of (s, w) pairs. It then turns them into TF-IDF vectors using IDF values from the document corpus.⁷ The statement node is connected to every article node, and each article node is then connected to every alternative node. In this we differ once more from (Monroy et al., 2008), as we have no need for heuristic rules for splitting the questions.

The edges are given weights whose value is the inverse cosine similarity between the connected nodes’ TF-IDF vectors. The system then calculates the shortest path between question statement and answer item using Dijkstra’s algorithm, and returns the article that connects them as the answer justification. Unlike (Monroy et al., 2008) our graph structure does not allow for more than one node connecting statement and alternative, as we knew from previous analysis that questions were usually justified by only a single article. Figure 1 illustrates the types of graphs we construct for each question.

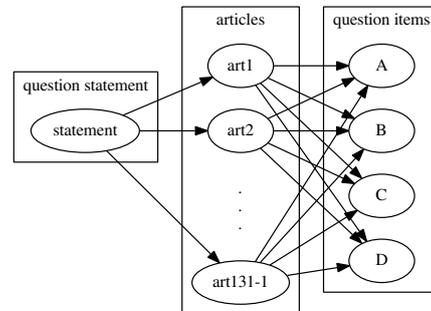


Figure 1: If a A is the number of article nodes, we then have $5A$ edges (as we have one statement and four alternatives).

6 Results

Using the method described in section 5, we conducted two experiments. As we explained in section 2, our golden answer set was manually created

⁷This means that if a sense occurs in the question statement or alternative but not in the legal norm corpus, its IDF value is 0.

	QA	QA+J	J
word system	12	12	18
synset system	14	11	17

Table 1: Experiments results, number of right answers out of the 30 question-answer pairs from the golden data.

by one of the authors and it consists of 30 questions from eleven different editions of the OAB exam associated to the article on the appropriate norm that justify the answer of the question. Table 1 presents the results comparing the current system (“synset system”) to the previous system (Delfino et al., 2017) (“word system”).

Our first experiment aimed to evaluate the main task (QA): choosing the right answer at the multiple choice problem, given the questions and the laws (all three normative documents related to the legal ethics area). The performance of the synset system was of 14 questions, against 12 in the word system. If we require not only correct answer, but a correct justification as well, experiment (QA+J), the synset system achieves 11 correct answers, while the word system scores the same 12 correct answers.

In some cases, both systems would find the correct justification article for the correct answer, but would pick as their putative answer another (incorrect) item, because it had a shorter path. Other times, they would not be capable of deciding between two (or more) answer items, as they all had a shortest path of the same length. The following exam question is a sample case where this statistical approach to question answering is defective:

Concerning the expiration of punitive disciplinary infractions, choose the right alternative. [...] A) The punitive aim in regard to disciplinary infractions expires after **five** years [...] B) The punitive aim in regard to disciplinary infractions expires after **three** years [...]

(15th ed. OAB exam, question 4)

These two options differ by only one word (the number of years until expiration), and coincidentally both are in the text of the article which justifies the answer key. In the synset system, as “three” and “five” are both hyponyms of [13741022-n, digit: one of the elements that col-

lectively form a system of numeration], this difference shouldn’t interfere with WSD of the other words. This gives us almost the same distance between the question statement and these two answer choices, and the system is incapable of choosing between them. A similar situation arises when one answer item makes a statement and another item denies this statement:

[question statement] [...] A) does **not compel** him to pay the agreed upon legal fees. [...] B) does **compel** him to pay the agreed upon legal fees. [...]

(18th ed. OAB exam, question 1)

In a question like this a system can only systematically report a correct answer if it has a higher-level understanding of the texts at hand: no bag-of-words model will suffice.

Although results in the first and second experiments may be humble, we then considered shallow question answering. As our approach tries to find not only the correct answer, but to find through a justification, it’s reasonable to evaluate the ability to find the correct justification given the correct answer to the question. Therefore in our third experiment (J) the system’s task was to determine which article (considered every law it has seen) justified the (already given) answer to the sentence. For each question in our golden set, we again added its statement and correct answer as nodes connected to all article nodes in the graph (see Figure 1). The word system was able to find 18 while the synset system found 17.

The overall results are not very impressive, although they are not bad as well. Using part-of-speech tagging and word sense disambiguation in order to improve the use of TF-IDF does not solve important difficulties, such as compositional understanding, pragmatics, etc. Nevertheless, the contributions to OWN-PT can be seen as a benefit by itself and will be valuable in the future planned experiments. These contributions may also improve the synset system to the point that it outperforms the word system noticeably.

7 Conclusion and Future Works

We tested the coverage and improved OWN-PT with terms from the Legal Domain. We also presented a new data set with all Brazilian OAB exams and their answer keys jointly with three Brazilian norms in LexML format. Furthermore,

we also reported our findings in the course of constructing a system to pass in the OAB exams. We obtained reasonable results considering the simplicity of the methods employed and the limited golden data available.

For the next steps, many other ideas can be tested. The TF-IDF VSM approach was devised as a baseline for the next phases of the project. Even so, we can still explore variations on that approach with lemmas and edges between articles, considering that 10% of our golden set includes more than one article as justification. Moreover, such approach can be combined with other methods, following classical ideas such as (Hobbs, 1986), since it seems to be sufficient for solving many questions. In another direction, we need to increase the size of the golden set. Using crowdsourcing websites to obtain more justifications from humans or crawling data from websites dedicated to discussions about the OAB exams is likewise a possibility.

Many different proposals for encoding laws in a machine readable format are available. Why no single standard have been largely adopted yet? We aim to explore the best candidates for the remaining normative documents that we will need to cover all areas of the OAB exams. We can consider ideas used in the data preparation of the ResPubliQA editions (Peñas et al., 2010).

Other techniques for textual entailment could be used as well for the task of answering multiple choice questions. Given the legal information (such as statutes, regulations and case law) as background knowledge, inferring the correct answer would amount to selecting the item which is entailed by the question statement and background knowledge (in case of multiple entailed answers, the one with highest confidence). The results of the experiments presented here clearly show that we need ‘deep’ linguistic processing to capture the meaning of natural language utterances in representations suitable for performing inferences. That will require the use of a combination of linguistic and statistical processing methods, possibly using leveraging experiences from (Quaresma and Rodrigues, 2005). In (Delfino et al., 2017) we begin to explore the use of the logic called *i*ALC (de Paiva et al., 2010; Haeusler et al., 2010). *i*ALC can be used to represent legal knowledge and it may help in the next steps of our project.

We may also explore recent advances in statistical relational learning, specially combining probabilistic and logical methods for semantic tasks, such as done by (Beltagy, 2016; Beltagy et al., 2013). This approach uses syntactical parsing in order to construct a logical form, which is given probabilistic semantics, weighted by linguistic resources (e.g. Wordnet). Using probabilistic logics (such as Markov Logic Networks (Richardson and Domingos, 2006) and Probabilistic Soft Logic (Kimmig et al., 2012)) allows a semantic with clear support for vagueness and ambiguity, as well for a integrated use of lexical resources, hand-coded rules and information learned from the data itself. The base of this approach is general: logical forms could be encoded in different formalisms, such as *i*ALC or others intermediary semantic representation formats such as AMR (Banarescu et al., 2013), if suitable probabilistic semantics could be given.

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In Alex Lascarides, Claire Gardent, and Joakim Nivre, editors, *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 33–41. The Association for Computer Linguistics.
- Flávia Alfenas Amorim and Gabriel Dib Tebechrani Neto. 2016. Exame da ordem em números. Technical report, Fundação Getulio Vargas and Conselho Federal da Ordem dos Advogados do Brasil. <http://hdl.handle.net/10438/18493>.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.
- Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garette, Katrin Erk, and Raymond J. Mooney. 2013. Montague meets markov: Deep semantics with probabilistic logical form. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA.*, pages 11–21. Association for Computational Linguistics.
- I. Beltagy. 2016. *Natural Language Semantics Using Probabilistic Logic*. Ph.D. thesis, Department

- of Computer Science, The University of Texas at Austin, December.
- Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004a. Freeling: An open-source suite of language analyzers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.
- Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004b. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.
- João Alberto de Oliveira Lima and Fernando Ciciliati. 2008. LexML brasil: versão 1.0. available at <http://projeto.lexml.gov.br/>, December.
- Valeria de Paiva, Edward Hermann Hausler, and Alexandre Rademaker. 2010. Constructive description logic: Hybrid-style. In *Proc. HyLo'2010*.
- Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. In Martin Kay and Christian Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 353–360. Indian Institute of Technology Bombay.
- Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. 2017. Passing the brazilian oab exam: data preparation and some experiments. under review.
- Biralatei Fawei, Adam Z Wyner, and Jeff Pan. 2016. Passing a USA national bar exam: a first corpus for experimentation. In *Language Resources and Evaluation*, pages 3373–3378.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Edward Hermann Haeusler, Valeria de Paiva, and Alexandre Rademaker. 2010. Using intuitionistic logic as a basis for legal ontologies. In *Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques*, pages 69–76, Fiesole, Florence, Italy. European University Institute.
- Jerry R. Hobbs. 1986. Overview of the tacitus project. In *Strategic Computing - Natural Language Workshop: Proceedings of a Workshop Held at Marina del Rey, California, May 1-2, 1986*.
- Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pages 1–4.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177.
- Ruslan Mitkov. 2005. *The Oxford handbook of computational linguistics*. Oxford University Press.
- Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2008. Using graphs for shallow question answering on legal documents. In *MICAI 2008: Advances in Artificial Intelligence: 7th Mexican International Conference on Artificial Intelligence, Atizapán de Zaragoza, Mexico, October 27-31, 2008 Proceedings*, pages 165–173. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2009. NLP for shallow question answering of legal documents using graphs. *Computational Linguistics and Intelligent Text Processing*, pages 498–508.
- Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2473–2479. European Language Resources Association (ELRA).
- Lluís Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February. Global Wordnet Conference 2010, Narosa Publishing House.
- Anselmo Peñas, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, Corina Forascu, and Cristina Mota. 2010. Overview of respubliqa 2010: Question answering evaluation over european legislation. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*.
- Paulo Quaresma and Irene Rodrigues. 2005. A question-answering system for portuguese juridical documents. In *10th international conference on Artificial intelligence and law*, pages 256–257. ACM.
- Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1):107–136, Feb.
- Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. 2004. Jur-WordNet. In Petr Sojka, Karel Pala, Pavel Smrz, Christiane Fellbaum, and Piek Vossen, editors, *Global Wordnet Conference*.

Implementation of the Verb Model in plWordNet 4.0

Agnieszka Dziob

Maciej Piasecki

Wrocław University of Science and Technology

agnieszka.dziob|maciej.piasecki@pwr.edu.pl

Abstract

The paper presents an expansion of the verb model for plWordNet – the wordnet of Polish. A modified system of constitutive features (register, aspect and verb classes), synset and lexical relations is presented. A special attention is given to the proposed new relations and changes in the verb classification. We discuss also the results of its verification by application to the description of a relatively large sample of Polish verbs. The model introduces a new class of relations, namely non-constitutive synset relations that are shared among lexical units, but describe, not define synsets. The proposed model is compared to the entailment relations in other wordnets, and the description of verbs based on valency frames.

1 Introduction

plWordNet 3.0 emo (Maziarz et al., 2016) describes 17,391 Polish verb lemmas by 31,834 *lexical units*¹ (LUs), and 75,643 relations. Thus, a very significant subset of Polish verbs has been covered. These numbers are also much higher than in any other wordnet, including Princeton WordNet (henceforth, PWN) (Fellbaum, 1998a). Nevertheless, plWordNet (plWN) 3.0 achieved the coverage of only ~30% of the verbs with the frequency >10 (57,969 in total²) in the *plWordNet Corpus*, i.e. 4 billion words³ corpus of Polish. plWN 3.0 verbs represent only 58.9% of 29,532 verbs described in SGJP (Saloni et al., 2015) - the most comprehensive morphological dictionary of Polish. Due to a very large size of plWN Corpus

this number can be a good predictor of the expected coverage in NLP applications of plWN. It could be higher. The relation density for verbs in plWN 3.0 emo is high, but several verb lexico-semantic relations are rather infrequent⁴.

(Dziob et al., 2017) presented a significantly modified, new model for the description of verbs in plWN. Our goal was to apply this model in expanding plWN 3.0 by a couple of thousand Polish verb lemmas, verify the proposed relation definitions in editing practice, both from the qualitative and quantitative point of view, as well as to propose some improvements and generalisations.

2 Verb Model in Brief

The system of lexico-semantic relations proposed for verbs in plWordNet 4.0 (Dziob et al., 2017) is based on the plWN 2.0 model. (Maziarz et al., 2011). A pair of relations: *hypernymy* and *hyponymy* organise verbs into a hierarchy. This differentiates plWN from PWN, in which hypernymy and *troponymy* are used (Fellbaum, 1998b), but is close to the models of EuroWordNet (Vossen, 2002) and GermaNet (Kunze, 1999).

Felbaum (1998b) argued against verb hyponymy that verbs differ from nouns and it is not possible to adapt a hyponymy test to them:

An x is a y.

As a consequence, troponymy in PWN “represents a special case on entailment: pairs that are always temporally coextensive and are related by entailment” (Fellbaum, 1998b). In plWN temporal co-extensiveness is expressed by two verb relations: hypernymy and meronymy, see Sec. 4. Fellbaum (1998b) defined troponymy as a *manner relation* and illustrated with a substitution test:

To V1 is to V2 in some particular manner.

¹ Lexical unit is a triple: a lemma, Part of Speech and sense id.

² However, some substantial number of these verbs can result from the errors of the morphological guesser.

³ plWN Corpus 10.0 includes: ICS PAS corpus (Przepiórkowski, 2004) National Corpus of Polish (Przepiórkowski et al., 2012), Corpus of Rzeczpospolita (Weiss, 2008), Polish Wikipedia, and a large amount of texts selected from Internet with automated quality check; duplicates were automatically removed.

⁴ See <http://plwordnet.pwr.edu.pl/wordnet/stats>

A test proposed for verb hyponymy in pLWN 2.0 correlates with the PWN troponymy test (Maziarz et al., 2011):

to X(inf) is to Y(inf) in a special way, somehow.

where the expression *a special way, somehow* represents a manner which is an intrinsic element of the situation definition. In order to cover this part of the definition in an explicit way *manner* relation was proposed (Dziob et al., 2017), which can be paraphrased: *X-ować to robić coś Y-owo* 'To X is to do something in an Y way'.

pLWN 1.0 included both relations: hyponymy and troponymy. However, the former was a synset relation⁵, while the latter was defined only for LUs and strictly related to the prefix derivational associations between members of aspectual pairs. Derwojedowa (et al., 2007) argues that there is a large group of verbs in the Polish language that are derived from such verbs that seem to be their hypernyms (i.e. expressing more general meaning than their derivatives), but of different aspect. Because it was assumed that verbs in the same hypernymy branch have the same aspect, cf (Maziarz et al., 2011), Derwojedowa (et al., 2007) proposed to use troponymy to link such verb hyponymy-like pairs in which elements differ in aspect and express some semantic addition. The use of troponymy was finally abandoned, also because its definition was very significantly different than in PWN. Instead, in order to link verbs associated by prefixal derivation such that one has a narrower meaning than the other, *secondary aspectuality* relation was introduced (Maziarz et al., 2011). It links, e.g., perfective: accumulative, distributive, and delimitative verbs with their imperfective derivational bases, like in the case of *posiedzieć* 'to keep sitting for a while' ↔ *siedzieć* 'to sit_{imp}'.

In addition to hyponymy, which organises verbs into hierarchies, there are several more relations in pLWN that describe relationships between situations, namely: *presupposition*, *preceding*, *meronymy/holonymy*, *inchoativity*, *causality*, *processuality* and *state*.

Presupposition is close to the logical presupposition, expresses temporal backward relation, and signals the necessary occurrence of one situation before the other, e.g. *żywy_{Adj}* 'alive' ← *umrzeć_{Verb}* 'to die'⁶.

Preceding is also a temporal backward relation signalling an usual, but not necessary occurrence of one situation before the second one, it can be considered as a 'weaker variant of presupposition', e.g. *siedzieć* 'to sit' or *leżeć* 'to lie' ← *wstać* 'to stand'),

Verb **meronymy/holonymy** (not automatically reverse) express co-occurrence of two situations in the same time period, e.g. *chrapać* 'to snore' ← *spać* 'to sleep', cf (Dziob et al., 2017).

Inchoativity links verbs representing the beginning of a situation and this situation, e.g. *zakochać się* 'fall in love' → *kochać* 'love'.

Causality describes the relation between LUs representing two situations where the first (represented by a verb) results in the second (represented by V, N, Adj or Adv), e.g. *zablokować* 'to lock' → *blokada* 'lock'.

Processuality links a verb LU and a noun, adjective or adverb representing a state resulting from the situation represented by the verb, e.g. *zmienić się* 'to change' → *inny* 'different'.

Multiplicativity is a relation emphasising an aspect of repetition in the verb meaning. It signals that some situation is repeated several times or an action performed on several objects. Multiplicativity is divided into two subtypes:

- *distributivity* (perfective) representing multiple performance, e.g. *nakupić* 'to buy many things' → *kupić* 'to buy_{perf}', and
- *iterativity* (imperfective) representing multiple repetitions, e.g. *czytywać* 'to read_{imp} many times' → *czytać* 'to read_{imp}'.

State connects state verbs with nouns, adjectives and adverbs describing states, e.g. *czzerwienić się* 'to be red' → *czzerwony* 'red'.

The next group of relations links verbs with LUs describing conditions in which situations occur.

Circumstance was introduced for pLWN 4.0 to link a verb representing a situation with a noun LU which is the semantic head of a prepositional phrase used to express conditions in which this situation occurs, e.g. *dopłynąć* 'to swim_{perf} to some point/place' → *brzeg* 'a bank'.

Manner, added for pLWN 4.0 links a verb LU with an adverb representing a manner in which an action is performed or a state happens, e.g. *popracować* 'to work a little' → *trochę* 'a little'.

⁵ pLWN model is based on LUs as basic building blocks. All relations are defined for LUs and synset relations are notational abbreviations for relations shared among LUs belonging to the two linked synsets, cf (Maziarz et al., 2013).

⁶ In pLWN 4.0 model many verb relations were expanded to cross-categorical relations, see (Dziob et al., 2017)

Object and **subject**, introduced for pLWN 4.0, link a verb LU with noun LUs representing, respectively, an object, e.g. *obuć* ‘to put on shoe’ → *but* ‘a shoe’, and subject, e.g. *oźrebić się* ‘to foal’ → *klacz* ‘a mare’. Such noun LUs must typically occur as intrinsic elements of semantic definitions (e.g. in dictionaries) of verbs that are linked to them.

All the relations mentioned so far are synset relations, as they are shared among LUs belonging to the same synset. All of them, except *circumstance*, *manner*, *object* and *subject*, are constitutive relations, i.e. relations defining synsets. Synonymy is defined in pLWN on the basis of sharing constitutive relations by LUs, cf (Maziarz et al., 2013). The set of constitutive relations determines the structure of a wordnet.

The above listed four relations are meant to be a tool for expanded characterisation of verb meanings (e.g. for WSD). They are defined in a less strict way and do not express necessary constraints. To limit their excessive proliferation, we included sanity conditions in their definitions: if there are more than three possible instances of such a relation per one synset, than we resign from adding this relation to this synset at all. Thus, this verb characterising relations are not meant to be a tool for identifying different lexical meanings and are not constitutive relations. For instance, *jechać* ‘to ride’ can be linked by *circumstance* to *pojazd* ‘a vehicle’ or *zwierzę* ‘an animal’, but because of this we do not want to differentiate between two different meanings of *jechać*. However, as these relations are mostly shared among LUs, we represent them as synset relations. They initiate a new class of wordnet relations: supporting, non-constitutive synset relations.

As it was already mentioned, the identity of aspect is a fundamental rule in linking verbs in the hypernymy structure and, as a consequence, in grouping them into synsets. Two main aspects are morphologically distinguished⁷ in Polish: *perfective* and *imperfective*. There is also a set of ~150 bi-aspectual verbs with the same lemma for both aspects (or ambiguous with respect to aspect) (Mędak, 1997), e.g. *nobilitować* ‘to ennoble’. In Slavic linguistics, it is used to describe the difference between the two aspects as the difference in the perspective of a subject perceiving a given situation: imperfective verb describes the situation

as lasting, while perfective describes it as finished, and besides this difference there is no other difference in the meaning of the two verbs of an aspectual pair, cf (Młynarczyk, 2004; Laskowski, 1998).

However, Młynarczyk (2004) argues that although such a definition of the aspectual verb pair is not controversial, this binary distinction does not originate from the language system as such, but it is caused by the prefixation. The derivational prefixes express semantic information beyond the mere change of the aspect. This correlates with the two types of aspectual lexico-semantic relations introduced in pLWN 2.0 (Maziarz et al., 2011): *pure* and *secondary* aspectuality - both defined as lexical relations (i.e. for LUs, not shared).

The former links pure aspectual pairs, i.e. such that two verbs in two different aspects do not differ in their meanings⁸, e.g. *czytać_{impf.}* ‘to read_{impf.}’ ↔ *przeczytać_{perf.}* ‘to read_{perf.}’. *Secondary aspectual* verb LU pairs are such that they express different aspects and share their derivational basis or the second is derived from the first, but the meaning of the second LU is modified beyond the aspectual difference in relation to the first, e.g. *czytać_{impf.}* ↔ *poczytać_{perf.}* ‘to read a little’, cf (Dziob et al., 2017).

The rest of verb lexical relations stay the same in pLWN 4.0 as in pLWN 2.0 model (Maziarz et al., 2011). The set encompasses (see also Tab. 2): **role inclusion** - a semantic association signalled by derivation of verbs from nouns - which expresses information similar to semantic roles, e.g. *bronować* ‘to harrow’ ← *brona* ‘a harrow’, *pieprzyć* ‘to pepper’ ← *pieprz* ‘a pepper’, *niańczyć* ‘to nurse’ ← *niańka* ‘a nanny’; **derivationality** representing verb links signalled by derivation, but without clear enough semantic character and not yet covered by more specific relations e.g. *hamletyzować* ‘to vacillate, to consider something pointless’ → Hamlet (PN, Shakespeare’s hero); and **antonymy** (with two subtypes), which is in pLWN a lexical relation (Piasiecki et al., 2009) and is not a constitutive relation (Maziarz et al., 2013).

PWN verb relations link only verbs (Fellbaum, 1998b), in similar way to GermaNet (Kunze, 1999). In pLWN, following EuroWordNet (Vossen 2002) verb LUs can be linked to all PoS. Modification of the verb part of pLWN 4.0 model

⁷ I.e. A verb lemma encodes its aspect, it is not inflected with respect to aspect.

⁸ However, more precisely, we should say that they do not significantly differ in their meanings beyond the information expressed by the aspect change.

was inspired by relations for adjectives and adverbs from pLWN 3.0, cf (Maziarz et al., 2016a, 2016b). The verb relations expanded to cross-categorical relations include: **processuality** (e.g. *anarchizować się* ‘to become_{Imp} anarchic’ → *anarchista* ‘anarchist’ / *anarchiczny* ‘anarchic’ / *anarchicznie* ‘anarchically’), **causality** (e.g. *zmienić* ‘to change’ → *zmiana* ‘a change’ / *inny* ‘different’ / *inaczej* ‘other’), **presupposition** (e.g. *całość* ‘a whole’ / *cały* ‘whole’ ← *podzielić się* ‘to divide itself’; *jasno* ‘brightly’ ← *ściemnić* ‘to dim’), **preceding** (e.g. *dobry* ‘good_{adj}’ / *zły* ‘bad_{adj}’ / *dobrze* ‘good_{adv}’ / *źle* ‘bad_{adv}’ ← *pogorszyć się* ‘to worsen’; *mąż* ‘a husband’, *żona* ‘a wife’ ← *rozwiść się* ‘to get divorced’), **state** (e.g. *jaśnieć* ‘to shine’ → *jasny* ‘bright’, *jasno* ‘brightly’; *królować* ‘to reign’ → *król* ‘a king’), cf (Dziob et al., 2017). This expansion resulted in a significant increase of their frequency in pLWN, see Sec. 6.

3 Semantic Classes

The pLWN 2.0 top part of the verb hypernymy structure consisted of artificial synsets expressing verb semantic classification originating from 7 classes of Laskowski (1998): processes, actions, acts, accidents, activities, events, states, were defined on the basis of (Vendler, 1967). This classification resulted in a large number of subclasses that constrained the rest of the verb hypernymy structure.

This classification system was sophisticated and potentially useful in applications, but appeared to be very hard to be applied consistently by wordnet editors (Dziob et al., 2017), especially as the verb classes constrain verb relations in pLWN. After analysis of the editing practice and the obtained results, the classification was simplified with only two main classes left in pLWN 4.0: *state* and *dynamic* verbs. This basic division corresponds to the general linguistic tradition, cf e.g. (Vendler, 1967; Comrie, 1989, Paduceva, 1996), Polish, e.g. (Karolak, 2001; Grzesiak, 1989), and also EWN. Vossen (2002) defines dynamic verbs as:

“specific transition from one state to another (bounded in time) or a continuous transition perceived as an ongoing temporally unbounded process,”

while static verbs as

“in which there is no transition from one eventuality or situation to another, i.e. they are non-dynamic”.

pLWN 4.0 uses similar definitions for both classes, but more attention is given to detailed characterisation of subgroups of the general classes and formulation of paraphrase-based descriptions for them. As a result, state verbs in pLWN 4.0 include verbs representing: 1) **localisation** (in space): *X jest gdzieś, ma jakieś położenie, jest w jakiejś pozycji*; ‘X is somewhere, has some location, is in a location’, e.g. *znajdować się* ‘to be in some place’, *sit* ‘siedzieć’, *otaczać* ‘to surround’; 2) **possession of permanent material features**, e.g. weight or volume (*X jest jakiś, jakoś, ma jakąś cechę, coś na stałe* ‘X possesses some feature, something permanent’; e.g. *jaśnieć* ‘to shine’, *mierzyć* ‘~to be of particular size’), 3) **relationships** between entities, both material and non-material (*X pozostaje w relacji do czegoś* ‘X stays in a relation to something’; e.g. *składać się* ‘to comprise’, *należać* ‘to belong’), 4) **mental states**, emotional, sense experience (*X odczuwa coś, doświadcza czegoś* ‘X feels something, experiences sth.’; e.g. *kochać* ‘to love’, *być przy nadziei* ‘be pregnant’, *istnieć* ‘to exist’), and also the 5) group which includes all other verbs that do not express **dynamics of situation** (i.e. do not represent a change from situation X to Y).

Dynamic verbs in pLWN 4.0 are perfective verbs: 1) **distributive** (to do something by many agents or in relation to many objects, e.g. *przebadać* ‘to examine many people’), 2) **accumulative** (to do something to such an extent that it is enough; e.g. *ubawić się* ‘to amuse itself’), 3) **perdurative** (to be doing something during limited time; e.g. *przemieszkać* ‘to live during some period in a place’), 4) **delimitative** (to be doing/happening for some time or to some extent; e.g. *pomieszkać* ‘to live for short time in a place’); and also 5) **action verbs** a) all perfective and bi-aspectual, b) imperfective derivatives of accumulative, delimitative, perdurative, and distributive verbs (representing changing situations), c) imperfective derivatives of semelfactive verbs (i.e. representing punctual or instantaneous events), d) imperfective causative verbs e.g. *rozśmieszać* ‘to make_{Imp} someone laughing’), e) **processive** (*X staje się czymś, jakoś* ‘X becomes sth, somehow’; e.g. *starzeć się* ‘to become_{Imp} gradually old’), f) **inchoative** (*X zaczyna się, zaczyna coś robić* ‘X is starting, begins doing sth’; e.g. *położyć się* ‘to lie down’), g) **limitative** (*X przestaje być czymś, jakimś, jakoś, przestaje coś robić* ‘X stops being sth, somehow, stops doing sth.’; e.g. *wybarwiać się* ‘to lose_{Imp} colour’) and h) all other imperfective verbs that represent situation changing due to actions of entities involved (e.g. *iść* ‘to walk’).

The subclass definitions (summarised above) are formulated in an operational way, on the basis of several substitution tests. They are referred to in relation definition and support linguists in editing. Thus, semantic class is a constitutive feature, together with stylistic register and aspect. Semantic subclasses of dynamic verbs are clearly connected to several relations that are characteristic for this class, namely: processuality, causality, inchoativity and multiplicativity. Only state verbs can participate in state relation. Other types of relations occur in both verb classes.

Verb classification is expressed by a hierarchy of *artificial LUs* (represented by singleton synsets) as in (Maziarz et al., 2011). Class assignment is done by placing a verb in an appropriate hypernymic branch, as hyper/hyponymy and synonymy (due to relation sharing) requires equality of semantic classes.

Semantic subclasses clearly refer to well-known linguistic classifications of verbs, e.g. (Levin, 1993; Fellbaum, 1998) and support wordnet editors in building hypernymic trees on the basis of semantic properties of verbs. The reduction of the number of classes (from 7 to 2) should facilitate identification of only real verb meanings and prevent introduction of non-natural and too fine-grained meanings.

4 Entailment

Verb entailment relation plays an important role in PWN and GermaNet, which is defined by Fellbaum (1998b) as:

“the relation between two verbs V_1 and V_2 that holds when the sentence Someone V_1 logically entails the sentence Someone V_2 .”

In addition, Fellbaum (1998b) introduces four subtypes of entailment. In pLWN a more fine-grained division of the spectrum of verb relations is proposed, see the comparison in Table 1.

We can notice a different perspective on situations co-occurring in the same time period. In PWN it is always represented by troponymy, which is defined as a kind of entailment (see Sec. 2), while in pLWN temporal co-occurrence of situations is covered by verb meronymy. In pLWN 2.0 a dedicated subtype of sub-situation meronymy was used (Maziarz, et al., 2011) (plus *associated situation* subtype), e.g., *komunikować się* ‘to contact’ and *zadawać się* ‘to associate with sb’

- communication is a part of a relationship, but they are different situations. Verb meronymy is necessary after troponymy has been excluded from pLWN and partially exchanged with hyponymy. We observed that the distinction between sub-situation and associated situation subtypes was too subtle in practice. Thus, verb meronymy in pLWN 4.0 does not have subtypes and is described by the following test:

Jeśli coś/ktoś X-uje, to na pewno jednocześnie Y-uje, bo X-ować można tylko Y-ując.

‘If sb./sth. is X-ing, then it/he is surely Y-ing, as X-ing is possible only if Y-ing is performed’.

Examples: *lunatykować* ‘to sleepwalk’ → *spać* ‘to sleep’, *nakopać się* ‘to kick so long, to be enough of it’ → *kopać* ‘to kick’.

EWN entailment	+Temporal inclusion		-Temporal inclusion	
	Co-extensiveness -troponymy	Proper inclusion	Backward presupposition	Cause
pLWN	Hyponymy, meronymy	Mero- nymy	Presupposition, preceding	Cause

Table 1: Temporal relations in PWN vs pLWN

On basis of the experience from the work on adverbs in pLWN 3.0, most verb relations of pLWN 4.0 allow for linking verbs with other PoS, including adverbs (Dziob et al., 2017). The system of relations for adverbs was derived from the one of adjectives in pLWN 3.0 (Maziarz et al., 2016b) that simplified extension of verb relations; e.g., a *processuality* link to an adjective or adverb is identified by the following tests:

X-ować to stawać się / stać się Y-owym

X-ing means to be becoming/to become Y-like.

e.g. *ochłodzić się* ‘to become cool / cooler’ → *chłodny* ‘cool’)

X-ować to stać się / stawać się Y_{Adv}-owo

X-ing to be becoming / to become Y_{Adv}

e.g. *ochłodzić się* ‘to become cool / cooler’ → *chłodno* ‘chilly’

5 Relations Signalled by Derivation

Derivational prefixes of verbs are important semantic signal in Polish. So far, verb prefixes have been only selectively and implicitly described as correlated with relations signalled by derivations. Although, we have not yet studied this issue in a

⁹ English gloss suggests that only verbs for which progressive forms exist can be used in this relation, but this limitation does not exist in Polish.

Table 2. Verb lexico-semantic relations in the plWordNet 4.0 model (first synset relations)

Relation	POs	Example	v3.1	G%
inter-register synonymy	V-V	<i>pieprzyć się</i> [vulgar] ‘to have sex’ → <i>uprawiać seks</i> ‘to have sex’	2529	25.4
hyponymy	V-V	<i>nadgryźć</i> ‘to chew a little’ → <i>ugryźć</i> ‘to chew’	29433	29.8
meronymy	V-V	<i>gryźć</i> ‘to chew’ is an integral part of situation <i>jeść</i> ‘to eat’	2311	-18,3
holonymy	V-V	<i>jeść</i> ‘to eat’ is a typical situation including <i>gryźć</i> ‘to chew’	3156	9.3
manner	V-Adv	<i>nadgryźć</i> ‘to bite a little’ → <i>trochę</i> ‘a little’	651	<i>new</i>
inchoativity	V-V, N	<i>urodzić się</i> ‘to be born’ → <i>żyć</i> ‘to live’	482	19.6
processuality	V-N, Adj, Adv	<i>ocieplać się</i> ‘to get warmer’ → <i>ciepły</i> ‘warm (adj)’, <i>ciepło</i> ‘warm (adv)’	1137	56.0
causality	V-V, N, Adj, Adv	<i>ocieplać</i> ‘to grow warm’ → <i>ocieplać się</i> ‘to get warmer’, <i>ciepły</i> ‘warm (adj)’, <i>ciepło</i> ‘warm (adv)’	3091	74.3
presupposition	V-V, N, Adj, Adv	<i>dodać</i> ‘to add’ presupposes <i>istnieć</i> ‘to be’ (no subject’s identity presupposition)	261	56.3
preceding	V-V, N, Adj, Adv	<i>rozwiść się</i> ‘to divorced’ precedes [to be] <i>żona</i> ‘a wife’ or <i>mąż</i> ‘a husband’ (subject’s identity preceding)	571	241.9
multiplicativity	V-V			
- iterativity		<i>jadać</i> ‘~to eat from time to time’ → <i>jeść</i> ‘to eat’	144	9.1
- distributivity		<i>popodgrzewać</i> ‘~to warm up many things’ → <i>podgrzać</i> ‘to warm up’	419	39.6
state	V-V, N, Adj, Adv	<i>dłużyć się</i> ‘to drag’ → <i>długi</i> ‘long (adj)’, <i>długo</i> ‘long (adv)’	176	89.2
subject	V-N	<i>ankietować</i> ‘to poll’ → <i>ankieter</i> ‘pollster’	221	<i>new</i>
object	V-N	<i>ankietować</i> ‘to poll’ → <i>ankietowany</i> ‘polled’	187	<i>new</i>
circumstance	V-N	<i>ankietować</i> ‘to poll’ → <i>kwestionariusz</i> ‘a questionnaire’	66	<i>new</i>
aspectuality	V-V		33351	25.6
- pure		<i>nadgryźć</i> ‘to bite _{perf} a little’ - <i>nadgryzać</i> ‘to bite _{imperf} a little’		
- secondary		<i>nadgryźć</i> ‘~to chew _{perf} a little’ - <i>gryźć</i> ‘to chew _{imperf} ’		
derivationality	V-V, N, Adj, Adv	<i>ocieplać</i> ‘to get warmer’ → <i>ciepły</i> ‘warm’	396	40.9
antonymy	V-V	<i>odezwać się</i> ‘to said’ - <i>przemilczać</i> ‘to left unsaid’	2530	7.6
- complementary		<i>rozbierać</i> ‘to undress’ - <i>ubierać</i> ‘to dress’		
- proper				
converseness	V-V	<i>implikować</i> ‘to imply’ - <i>wynikać</i> ‘to result’	134	19.6
role inclusion	V-N		1793	32.1
- subject		<i>gospodarować</i> ‘to farm’ ← <i>gospodarz</i> ‘a farmer’		
- instrument		<i>betonować</i> ‘to concrete’ ← <i>beton</i> ‘a concrete’		
- result		<i>filetować</i> ‘to fillet’ ← <i>filet</i> ‘a fillet’		
- location		<i>magazynować</i> ‘to store’ ← <i>garaż</i> ‘a store’		
- object		<i>lajkować</i> ‘to give a like’ ← <i>lajk</i> ‘a like’		
- time		<i>ucztować</i> ‘to feast’ ← <i>uczta</i> ‘a feast’		
- indefinite		<i>litować się</i> ‘to have pity’ ← <i>litość</i> ‘pity’		

systematic way, some associations between prefixes, meanings and lexico-semantic relations became visible.

Prefixes *do-*, *wy-* can signal situations in which an agent is accomplishing his goal, e.g. *dojść* ‘to have reached sth’, *dokopać się* ‘to have dug down

to sth’, *wysiedzieć* ‘to have continued sitting until sth happened’, *wyczekać* ‘to have continued waiting ...’. They express a relation to a goal or an end that are often implicit.

Another example is a set of prefixes expressing

a kind of manner relation in the case of delimitative verbs: *po-* and *do-*. Concerning the first, *po-* prefix means to do a little, e.g. *posiedzieć* ‘to sit a little’ (*siedzieć* ‘to sit’), *popoglądać* ‘to watch a little’ (*oglądać* to watch). The prefix *do-* signals more advanced or intensive situation, e.g. *doszkolić się* ‘to improve qualifications’ (*szkolić się* ‘to learn by himself’), *dogęszczać* ‘to thicken more (a mixture, substance etc.)’ (*zagęszczać* ‘to make thicker’).

Verbs derived by prefixes are linked by *secondary aspectuality*, e.g. *wysiedzieć* ‘to have continued sitting’ – *siedzieć* ‘to sit’ or by more specific relations, e.g. inchoativity. However, secondary aspectuality is intentionality vague, only slightly more informative than fuzzynymy, and is a way of registering LU pairs requiring deeper investigation in future. A more in depth exploration of derivational verb prefixes focused on enrichment of wordnet relations is a very interesting task to be undertaken in the future.

6 Implementation

plWN 3.0 includes 17,391 verb lemmas described by 31,834 LUs that should cover all meanings of the verbs. As it was declared earlier, one of the goals for plWN 4.0 is a significant expansion of the verb database. Following the corpus-based development scheme, a set of 8,000 most frequent verbs in the plWN corpus was selected that were lacking in plWN 3.0. With the help of the word2vec (Mikolov et al., 2013) model based on plWN Corpus, the selected verbs were clustered in packages of ~100 verbs each. Each package is intended to cover a limited number of topics and to be a unit of work assigned to a linguist.

So far, the number of verb lemmas in plWN has been increased to 19,272 i.e. by 11%. In parallel, we have updated the verb hypernymy structures and verb relations to a large extent. This enabled us to observe the changes triggered by the new verb model. Tab. 2 present statistics for the relations and changes in relations.

We can notice that the modification of the model resulted in the increased frequency of the following relations: processuality, causality, presupposition, inchoativity, state. In the same time the number of verb meronymy instances has decreased but this could be expected due to the more stricter definition and the remove of the ambiguous division into two subtypes (this ambiguity led to too far going interpretations).

7 Verb Model vs Valency Lexicon

A high quality valency dictionary with good coverage is an indispensable resource for many NLP applications. Unfortunately, its construction is very laborious and costly. plWN model defines a rich system of verb relations. The question is to what extent it can supplement a valency lexicon? Marantz (1981) argues that semantic roles are indispensable in the description of the predicate-argument structures, e.g. the *agens* role refers to the logical subject of a predicate, while the theme and patiens roles to the logical objects.

A clear reference made in the plWN verb model to the syntactic-semantic relations is aimed at improving richness of LU descriptions following Apresjan (2000) who argues that a dictionary should provide description of co-occurrences of lexico-semantic and syntactic features. In Czech WordNet (Pala et al., 2004) valency frames are added to synsets. However, we assumed in plWN that syntactic valency is not a constitutive feature of verb LUs, and does not need to be shared by synset members, so is not used to define synsets. It could be described on the level of LUs, but this is in fact done in Walenty (Hajnicz et al., 2017), a large valency lexicon of Polish. Thus, syntactic valency is not expressed in plWN, a semantic lexicon, and there are no plans for introducing it. So, this part is clearly missing, but verb arguments which are mentioned in relation definitions can be implicitly expressed in the lexico-semantic relations. As a consequence, quite a lot of information about semantic restrictions on valency arguments is hidden in plWN relations. It is partial and selective, but still can be useful.

Three relations introduced in plWN 4.0 directly evoke structure relations, namely: *subject* (referring to the semantic agent role), *object* (patient role) and *circumstance*, whose detection is based on prepositional phrases, which can correspond to other roles, for example location, result, time. As it was said in Sec. 2, *subject*, *object* and *circumstance* relations (*manner* does not link nouns) are not constitutive relations, but emphasise selected aspects of LU meanings that are common to the whole synset, and in the same time relate these aspects to the syntactic structure, e.g. *circumstance* links *brzeg* ‘a shore’ with *dobijać* ‘to reach a shore’ informs also that one of the *dobijać* predicate arguments represents location. In a similar way *object* relation links *usypiać* ‘to put down, to put to sleep, to euthanize’ with *zwierzę* ‘an animal’ and signals that one of the arguments repre-

sents animal or its hyponym. The guidelines instruct to find for these relations nouns that are located on relatively high levels of the hypernymy to describe the meaning of the verb LU, not its collocational behaviour. Linguists are also required to check if most of the hyponyms of the selected target noun fulfil the tests for this relation. In the same time the target noun should not be located too high in order to preserve meaningfulness of the link, i.e. LUs from the top level of the hypernymy hierarchy should be avoided, e.g. *byt* ‘an entity’, *istota* ‘a being’).

In Walenty semantic description is based on *selectional preferences*: “lexico-semantic dependencies between a unit which is a predicate of an utterance and units that are its arguments, that determine what kind of notions can co-occur on the subsequent valency arguments” (Hajnicz et al., 2017). Because Walenty frames have been built in relation to the pLWN LUs, selectional preferences of the Walenty entries tend to be correlated with pLWN synsets. Hajnicz (et al., 2017) aims at encompassing by selectional preferences all hyponyms of a given synset, e.g. for *rżeć* ‘to neigh’ there are two semantic frames: selectional preferences of the first restrict *agent* (“Initiator”) to *koń* ‘a horse’ (pLWN: *koń 1* ‘a horse’) and in the second to *człowiek* ‘a man’ characterising the second meaning of *rżeć* as ‘to laugh producing sound resembling neighing’. Selectional preferences in Walenty are chosen according to the frequency, i.e. in the case of *rżeć* ‘to neigh’ the editor decided that the constraint *koń* ‘a horse’ for the agents is enough frequent to be expressed in the frame; in addition, all hyponyms of *koń* ‘a horse’, e.g. *pegaz* ‘Pegasus-like’, *gniadosz* ‘a bay’, but also derivatives, i.e. diminutives e.g. *konik* ‘~a little horse’ and augmentatives, e.g. *konisko* ‘~a large, not pretty horse’ are included in the preferences. pLWN describes the *subject* link between *rżeć* ‘to neigh’ and *koniowate 1* ‘an equine’, because also zebras or giraffes are neighing (at least in Polish) and they belong to equines taxonomy together with *koń* ‘a horse’. These links can be further interpreted by explicit derivational links.

Semantic valency information can be also found in lexical relations: *role* (N-V, describing deverbal nouns) *role inclusion* (V-N, verbs derived from nouns). Both relations have 7 subtypes: agents, instrument, product, location, patients, time and indefinite subtype (Maziarz et al., 2011) that refer to thematic roles of Fillmore (1968), on the one side and to the studies on the semantics of deverbal nouns in the Polish literature, cf

(Wróbel, 2001). Both relations tell something about the selectional preferences.

For instance *solić 1* ‘to salt’ is a hyponym of *przyprawiać 2* ‘to spice’ and means ‘to spice with salt’ and is linked with *sól* ‘salt’ by *role_inclusion:instrument* as a verb derived from a noun - a tool name. The expression *solić solą* ‘to salt with salt’ is redundant and incorrect, but one can say *przyprawiać solą* ‘to spice with salt’, where *przyprawiać 2* is linked by *role_inclusion:instrument* to *przyprawa 1* (‘a spice’); *przyprawiać* ‘to spice’ can be done by salt or different spices - cohyponyms and cousins of *sól 1* ‘salt’. Another example can be *bokser* ‘a boxer’ linked by *role:agens* to *boksować 1* ‘to box’ (its derivational basis), which is a hyponym of *bić 4* ‘to hit, to beat’. The expression *bokser boksuje* is redundant but *bokser bije* ‘a boxer is beating’ is correct. Thus, the combination of *role/role inclusion* and verb and noun hypernymy can be used to draw conclusions about selectional preferences of the verb arguments.

Relations defined on the level of synsets go beyond the derivational associations. During the work on pLWN 4.0 we have realised that a lot of valuable semantic knowledge is not covered by strictly derivationally motivated relations. Analysis of fuzzynymy from pLWN 3.0 showed that semantic associations visible in derivations can be cautiously generalised, i.e. in a way based on strict procedure, substitution tests and guaranteeing good consistency among editors.

8 Conclusion

We presented an expanded verb model for pLWN, including modified constitutive features, and synset and lexical relations. Non-constitutive synset relations were introduced. They are shared among LUs in a synset, characterise important aspects of verb meaning, but are not necessary constraints for defining synsets. They seem to be a good tool for the inclusion of knowledge valuable for wordnet applications, e.g., WSD. The proposed model was verified and slightly amended on the basis of its application to a large sample of Polish verbs. The first statistical data showing the results of the proposed changes were discussed. We showed that the proposed system of relations provides information about entailment and selectional preferences. Open issues are: the relation between the defined lexico-semantic relations and relations between verb valency frames, and the extent of automatization in identification of the selectional preferences on the basis of the relations.

Acknowledgements

Works funded by the Polish Ministry of Science and Higher Education within CLARIN-PL Research Infrastructure.

Reference

- Iurii D. Apresian. 2000. *Systematic lexicography*. Oxford University Press on Demand, Oxford.
- Joan Bresnan. 1982. *The mental representation of grammatical relations*, volume 1. The MIT Press, Cambridge.
- Agnieszka Dziob, Maciej Piasecki, Marek Maziarz, Justyna Wieczorek, and Marta Dobrowolska-Pigoń. 2017. Towards Revised System of Verb Wordnet Relations for Polish. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, Galway, Ireland, 19-20 June.
- Christiane Fellbaum (ed.). 1998a. *WordNet: An electronic lexical database*. MIT Press, Cambridge.
- Christiane Fellbaum. 1998b. *A semantic network of English verbs*. In: Christiane Fellbaum (ed.), *WordNet: An electronic lexical database*. MIT Press, Cambridge.
- Charles J. Fillmore. 1968. *The case for case*. In Emmon Bach, and Robert T. Harms. (ed.). *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York.
- Jane B. Grimshaw. 1990. *Argument structure*. The MIT Press, Cambridge.
- Renata Grzegorzczkova. 1990. *Wprowadzenie do semantyki językoznawczej*. [Introduction to linguistic semantics]. PWN, Warszawa.
- Renata Grzegorzczkova. 2008. *Wstęp do językoznawstwa* [Introduction to polish linguistics]. PWN, Warszawa.
- Romuald Grzesiak. 1983. *Semantyka i składnia czasowników percepcji zmysłowej*, Zakład Narodowy im. Ossolińskich, Wrocław-Warszawa-Kraków.
- Robert T. Harms, and Emmon Bach (ed.). 1968. *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York.
- Elżbieta Hajnicz, and Bartłomiej Nitoń. 2017. Instrukcja dostępu do słownika walencyjnego Walenty za pośrednictwem programu Słowl. Institute of Computer Science PAS, URL: http://clarin-pl.eu/wp-content/uploads/2017/05/instrukcja_uzytkownika_Walentego.pdf.
- Stanisław Karolak. 2001. *Od semantyki do gramatyki. Wybór rozpraw*. [From semantics to grammar. The choice of papers]. Sławistyczny Ośrodek Wydawniczy, Warszawa.
- Claudia Kunze. 1999. *Semantics of verbs within GermaNet and EuroWordNet*. In *Proceedings of 11th European Summer School in Logic, Language and Information*. Utrecht.
- Roman Laskowski. 1998. *Kategorie morfologiczne języka polskiego—charakterystyka funkcjonalna* [Morphological categories of Polish-functional characteristic]. In Renata Grzegorzczkova, Henryk Wróbel, Roman Laskowski (ed.), *Gramatyka współczesnego języka polskiego. Morfologia 1* [Grammar of Polish language. Morphology 1]. PWN, Warszawa.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, Chicago.
- Alec Marantz. 1981. *On the nature of grammatical relations*. PhD Thesis. Massachusetts Institute of Technology.
- Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz, Joanna Rabeiga-Wiśniewska, and Bożena Hojka. 2011. Semantic relations between verbs in polish wordnet 2.0. *Cognitive studies*, 11:183-200.
- Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769-796.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stanisław Szpakowicz, and Paweł Kędzia. 2016a. *plWordNet 3.0-a Comprehensive Lexical-Semantic Resource*. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, Osaka, Japan.
- Marek Maziarz, Stanisław Szpakowicz, and Michał Kaliński. 2016b. *Adverbs in plWordNet: Theory and Implementation*. In *Proceedings of the 8th International WordNet Conference — GWC 2016*, Bucharest, Romania, 27-30 January.
- Stanisław Mędrak. 1997. *Słownik form koniugacyjnych czasowników polskich* [A dictionary of polish verbs patterns], Universitas, Kraków.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. CoRR, vol. abs/1301.3781.
- Anna K. Młynarczyk. 2004. *Aspectual pairing in Polish*. PhD Thesis. Utrecht University, Utrecht.
- Elena V. Padučeva. 1996. *Semantičeskie issledowanija: Semantika vremeni i vida v russkom jazyke; Semantika narrativa*. Škola Jazyki Russkoj Kultury.
- Karel Pala, and Pavel Smrž. 2004. Building czech wordnet. *Romanian Journal of Information Science and Technology*, 7(2-3):79-88.

- Maciej Piasecki, Bartosz Broda, and Stanisław Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.
- Adam Przepiórkowski. 2004. *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS.
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk. (ed.). 2012. *Narodowy Korpus Języka Polskiego* [National Corpus of Polish]. PWN, Warszawa.
- Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. 2015. *Słownik gramatyczny języka polskiego*. [Grammatical dictionary of Polish]. 3rd edition. URL: <http://sgjp.pl/>.
- Zeno Vendler. 1967. *Verbs in Times*. In Z. Vendler, *Linguistics and Philosophy*, Ithaca, New York, Cornell University Press.
- Dawid Weiss. 2008. *Korpus Rzeczypospolitej* [Corpus of text from the online edition of "Rzeczpospolita"]. Unpublished. URL: <http://www.cs.put.poznan.pl/dweiss/rzeczpospolita>.
- Henryk Wróbel. 2001. *Gramatyka języka polskiego* [Grammar of the Polish language]. Spółka Wydawnicza "Od Nowa", Kraków.

Public Apologies in India - Semantics, Sentiment and Emotion

Sangeeta Shukla, Rajita Shukla

Bennett University, Tech Zone 2, Greater Noida, (U.P.) India
sangeeta.shukla, rajita.shukla { @bennett.edu.in }

Abstract

This paper reports a pilot study related to public apologies in India, with reference to certain keywords found in them. The study is of importance as the choice of lexical items holds importance which goes beyond the surface meaning of the words. The analysis of the lexical items has been done using interlinked digital lexical resources which, in future, can lend this study to computational tasks related to opinion mining, sentiment analysis and document classification. The study attempts an in-depth psycholinguistic analysis of whether the apology conveys a sincerity of intent or is it a mere ritualistic exercise to control and repair damage.

Keywords: apology, sorry, regret, apologize, WordNet, SentiWordNet, WordNet-Affect, corporate apologies, corporate communication

1 Introduction

Public apologies, as a tool to repair damage and manage reputation, have been used by organizations and individuals frequently the world over. The dynamics of speech act of apologizing are very different from that of written apologies. Written apologies are not supported by the nonverbal elements of communication. The remorse on the face, the earnestness in the voice, the intent in the gestures are all absent in the written apologies. The words stand alone to convey the guilt, remorse, regret and forbearance. The tone and tenor of writing can thus play an important role in either leading the customers to take a forgiving stance to the organization or rejecting it as a ritualistic gimmick.

Communication researchers agree that the oral and written language differ significantly in their communication impact. While the speech act has been analyzed in detail, not much attention has been paid to the written word. Specifically, in the Indian context, there is very little research on public apologies. This paper aims at making an analysis about the semantics, sentiment and emotion of written apologies delivered digitally in India by using three inter-linked digital lexical resources, namely, WordNet¹, SentiWordNet² and WordNet-Affect³ respectively. The paper limits itself to the analyses of a set of selected keywords found in these apologies. To the best of our knowledge, this is the first such study. Our hypothesis is that the choice of lexical items plays an important role in conveying the intent of the writer in a public apology and the sentiments and emotions associated with an apology expression can go beyond the surface meaning of the word.

Roadmap

Section 2 deals with the related work. Section 3 discusses apologies in the digital media and such apologies in India. Section 4 outlines the methodology followed in the study. Section 5 is presents the analysis with reference to WordNet, SentiWordNet and WordNet-Affect. Section 6 contains the overall discussion. Section 7 discusses the future work.

2 Related Work

Linguistic analysis of social discourse, using digital lexical resources and related software, has been an upward trend in the recent past. WordNet has been used for marking the event profile of news articles as a function of verb type (Klavans, 1998). An Adversary-Intent-Target (AIT) model has been developed which is based

¹ <http://wordnet.princeton.edu/>

² <http://sentiwordnet.isti.cnr.it/>

³ <http://wdomains.fbk.eu/wnaffect.html>

on an Ontology for the Analysis of Terrorist Attacks (Turner et al, 2011). DICTION 5.0 text analysis master variable, CERTAINTY has been used to analyze top management language for signals of possible deception (Craig et al, 2013). A viable approach to sentiment analysis of newspaper headlines has been developed by using linguistic techniques and a broad-coverage lexicon (Chaumartin, 2007).

From the point of view of communication study, most of the research on public apologies is focused on apology as a speech act (e.g. Edmondson, 1981; Fraser, 1981; Holmes 1990; Blum-Kulka et al.1989; Olshtain and Cohen 1983; Owen, 1983; Trosborg, 1987). The studies are based on two perspectives. The first is from the point of view of the offended party (Lee & Chung, 2012) and the second sees apology from the point of view of the offender (Darby & Schlenker, 1989; Goffman, 1971; Hearit, 1994, 1996, 1997, 2010; Schlenker & Darby, 1981).

Although an emphasis has been laid on the different nature and aspects of written and spoken discourse (Halliday (1989, 2007, Tillmann, 1997, Aijmer and Stenström, 2004, Wikberg, 2004, Nelson, Balass and Perfetti 2005, Biber, 2006, Miller, 2006, McCarthy and Slade, 2007 and Wichmann, 2007, Chafe, 1992), not much attention has been paid to the written word. Moreover, research on the written apology delivered via the digital medium needs further analysis.

3 Apologies in the Digital Media

The practice of tendering an apology as a means of acknowledging and compensating for failure is an ancient one. Etymologically, the word apology is derived from the Greek *apo* (away, off, absolve) and *logia* (speech) and should be differentiated from the word *apologia*.

Corporations the world over have used public apologies effectively for multiple purposes - as a tool for damage control, for defending their position in a particular situation and also for conveying their commitment to all stakeholders. Due to the advent of e-commerce companies and the increasing reach of the social media companies have their finger on the pulse of public sentiment constantly. Minor events and lapses go viral within a few minutes. The word of mouth is now faster than it was ever before.

The digital medium differs from ordinary face to face communication in many ways: it requires a select choice of words to express the apology,

it can be stored and retrieved at a later date, and, it becomes a quasi-legal document. The art of apologizing is a powerful marketing tool that can induce trust on the one hand and fuel mistrust on the other, if poorly managed.

3.1 The Indian Context

Culturally, saying sorry does not come easy to Indians and more so to Indian business and political leaders. This hesitation can perhaps be linked to the fact that in India a public apology is seen as an admission of guilt (Maddux et al, 2012). On the other hand it is a common occurrence in countries like Japan and Hong Kong, where the corporate apology is an expression of eagerness to repair damage and relationships and does not imply guilt (ibid). In the past, the speech act of apology was almost absent from the repertoire of Indian corporates and public figures (Kaul et al,2015). Even written apologies were very few and were offered only when there was a strong demand from different sections of society.

However, the new generation e-commerce companies seem to be heralding an attitudinal change in this corporate practice. This could be due to the increasing digital customer base for India Inc. India's internet user base has grown to 324.95 million in September 2015, a 27.73% YOY growth (TRAI, 2016). On social media platforms situations can escalate rapidly, breaking down the traditional barriers of time, location, and gatekeepers of information (Kaul et al, 2015). Thus, in stark contrast to the past, we see a spate of apology e-mails, tweets and blog posts being offered by e-commerce players (ibid). Figure 1 shows the rising trend of apologies being given publicly in the written digital media, with a sharp increase from the year 2016 to 2017.

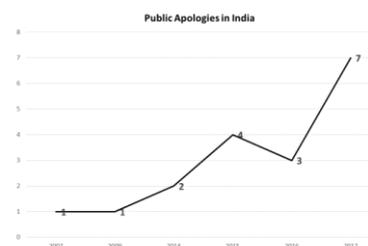


Figure 1: Graph showing rising trend of public apologies in India

Since the practice of offering a public apology is relatively new for Indian businesses, it is to be understood that an apology not delivered effectively rather than mitigating the damage, can escalate the damage done. In this context, it is important to analyze the lexical choice made in these apologies and the implications thereof.

4 Methodology

The research design is qualitative and is based on an analysis of a self-built corpus. The following steps were followed as part of the methodology.

- Corpus Collection
- Keyword Selection
- Determination of POS of keywords
- Determination of the correct sense of the keywords
- Analysis using WordNet, SentiWordNet and Wordnet-Affect.

4.1 Corpus collection

The study uses a self-built corpus. Since the phenomenon of public apologies is relatively recent in India, we could only access a corpus of 18 apologies available in the digital public domain, offered during 2007-2017. The corpus is in the English language as it is the second official language in India. It is the lingua franca spoken amongst a wide proportion of the population and has about 125 million speakers, which is, country-wise, the second highest in the world, only below United States of America⁴. We employ a close reading approach (Amernic et al., 2007) for the analysis.

All of the selected apologies were delivered in India, by Indians so as to understand any cultural implication of the communication. All of these were offered by senior executives of the company or prominent public personalities in India. Of these two were electronic mails, seven were letters, four were blog posts, four were tweets out of which two are related to the same event, and one was a media statement. Out of the 18 apologies, 11 were given by individual(s) in a role, 3 were given by organizations and 4 were given by individuals. The gender-wise distribution of the apology givers is 14 males and 4 females. The apologies selected have been assigned a code number for easy reference.

These apologies are listed below, with the name of the company, the year and a short context.

1. **Infosys (2007)** - Narayana Murthy, founder of one of India's leading technology companies, Infosys, apologized after being accused of making rude comments about India's national anthem.
2. **Satyam (2008)** - Letter written by Ramalinga Raju (the then chairman of India's IT Company Satyam Computer Services) on 30 September to the board of directors of Satyam Computer Services Limited informing them about his company's corporate fraud.
3. **Flipkart (2014)** - E-mail from Sachin Bansal and Binny Bansal founders of Flipkart, a leading retail e-commerce company in India, apologized to disgruntled shoppers after technical glitches during their 'The Big Billion Day' sale on October 7.
4. **Uber India (Dec. 2014)** - Days after it was banned following the rape of a woman by an Uber driver, in New Delhi, India, the global cab booking firm sent out apology mail to its customer.
5. **Myntra 1 (2015)** - Myntra, an e-commerce company in India, apologised to its customers via e-mail for the technical glitches faced during a mega-sale.
6. **ScoopWhoop (2015)** - Editor-in-Chief of ScoopWhoop, an internet media and news company from India, apologised after it carried an insensitive article on a massive earthquake that hit parts of Nepal and India.
7. **Lenskart (2015)** - Bansal & Chaudhary, co-founders, Lenskart, apologised on the company's behalf, when the company sent out an SMS offer which referred to the massive earthquake that struck India and Nepal in poor taste.
8. **AIB (2015)** - AIB (All India Bakchod Comedy Company), a comedy group of India, offered an unconditional apology to the Auxiliary Bishop of Bombay and the community for any offence caused to the christian community by their jokes.
9. **Myntra 2 (2016)** - An apology was posted on Myntra's blog by Shamik Sharma, CTO, Myntra, for inundating customers' phones with notifications due to technical lapse.

⁴https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

10. **Amazon India (2016)** - Amit Agarwal, Vice President and Country Manager, Amazon India, apologized to the Indian External Affairs Minister for hurting Indian sentiment by selling doormats with Indian tricolour on them.
11. **Axis Bank (2016)** - After two Axis Bank managers in New Delhi were accused of being involved in money laundering, Shikha Sharma, CEO Axis Bank, sent an e-mail letter by to all Axis Bank customers to address the issue.
12. **PETA (2017)** - PETA India CEO, Poorva Joshipura wrote an apology to the Indian actor, Suriya, when the latter issued a legal notice to PETA for calling his voice in favour of Jallikattu as a promotional strategy for his upcoming film 'C3'.
13. **Member of Parliament's Apology (2017)** - A Member of Parliament, Ravindra Gaikwad, courted controversy after thrashing an Air India employee. He expressed regret in a letter to Civil Aviation minister.
14. **Tech Mahindra Layoff audio clip controversy 1 (2017)** - In an audio recording that went viral on social media, a female HR executive of Tech Mahindra, a leading IT company of India, was heard telling an employee to resign by 10 am the next day. Shortly afterwards, Vice-chairman of Tech Mahindra, Vineet Nayyar, apologized on the matter.
15. **Tech Mahindra Layoff audio clip controversy 2 (2017)** - Following the Vice-chairman's apology, Mahindra Group Chairman, Anand Mahindra and Tech Mahindra CEO CP Gurnani also came out to apologize on Twitter on the same matter.
16. **Film actor, Priyanka Chopra's apology, (2017)** – Film actor apologized after she addressed the northeastern state of India, Sikkim, as troubled with insurgency and troubling situations, while talking about her Sikkimese production.
17. **Indigo, Domestic airline company, apology (2017)** – A domestic airline company apologized after a video clip, which went viral, which showed the airline staff assaulting a passenger named Rajeev Katiyal.
18. **Air India, National airline company, apology, (2017)** – The airline apologized after an Indian classical singer, Shubha Mudgal, took to Twitter after her Air India business class ticket from Mumbai to Goa was changed to economy class without any prior notice.

4.2 Keyword Selection

After the selection of documents for analysis, a list of keywords was prepared independently by the authors and then compiled. As traditionally held, an apology consists of five major parts (Cohen et al, 1981). These are the following:

- a. **Expression of apology** – using Illocutionary Force Indicating Device (IFID), which is an explicit expression which directly conveys the writer's remorse. (Blum-Kulka et al, 1989).
- b. **Explanation or an account** (e.g. I missed the bus)
- c. **Acknowledgment of responsibility for the offense** (e.g. It's my fault)
- d. **Offer of repair/redress** (e.g. I'll pay for your damage)
- e. **Promise of forbearance** (e.g. I'll never forget it again)

It was decided to conduct a focused analysis of a few selected IFIDs. The four that were selected were - *sorry*, *regret*, *apologize* (*apologizes* and *apologizing*) and *apology* and are termed as keywords henceforth. It was decided to exclude other IFIDs such, *forgive*, *forgiveness*, *excuse*, *afraid*, *pardon* for this study. These selected words were then marked in the corpus.

Figure 2 below shows the frequency of the keywords in the selected apologies. As is evident from the Figure, the adjective *sorry* has the highest occurrence (12) as compared to the other three, keywords – *apology* (including *apologies*), *apologize* and *regret* (both as verb and noun), which are in the range of 7, 6 and 8 each respectively.

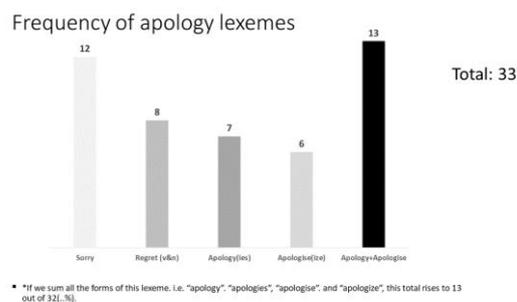


Figure 2: Frequency of Keywords

4.3 Determination of POS of Keywords

To correctly determine the part of speech of the keywords, the sentences where they occurred were put through an online Part-of-speech tagger⁵. This was found to be necessary as some keywords could belong to more than one category. The output of the tagger marked the words *apology* and *regret* as *NN1* (singular common noun), the words *apologies* and *regrets* as *NN2* (plural common noun), the words *apologize* and *regret* as *VV0* (base form of lexical verb), the words *apologizes* and *regrets* as *VVZ* (-s form of lexical verb), the word *apologizing* as *VVG* (-ing participle of lexical verb) and the word *sorry* as *JJ* (general adjective).

4.4 Determination of Keyword Senses

For the determination of the correct sense of the keywords, we put the sentences where the keywords occur in an online sense disambiguator⁶. Sense determination was done as the keywords were found to be polysemous. The senses thus determined were mapped to the senses in English WordNet (3.1). The selected senses are mentioned in the analysis of the keywords in section 5.

5 Analysis

A three-fold analysis of the selected keywords was done. The semantics of the words was studied by using WordNet. In dialogue acts such as apologizing, thanking, or expressing sympathy, affective language is often employed to represent and convey psychological attitudes (Novielli et al, 2013). Also, there is what is called a 'heartfelt apology' as against 'routine apology' (Owen, 1983). Hence, it was decided to further explore the sentiments and emotions associated with the keywords. The sentiments were studied using SentiWordNet and the emotion labels were determined through WordNet-Affect. The analysis and conclusions thus drawn are presented below.

5.1 Semantic Analysis using WordNet

A semantic analysis of the selected keywords was done using WordNet (3.1). We used semantic relations such as hypernymy, troponymy and entailment (Fellbaum, 1998) to find the implications that the keywords may have, as far as their communicative goals are concerned.

5.1.1 Verb – Apologize and Regret

The main aspect of an apology lies in the verb that the tenderer chooses to use. We do an analysis of the two verbs, *apologize* and *regret*, using WordNet, the former being an explicit performative verb (Austin, 1975). The selected sense of the verb *apologize* is defined as *-to acknowledge faults or shortcomings or failing*. Its semantic relation of entailment is *admit, acknowledge*, which means to *declare to be true or admit the existence or reality or truth of*. One of its troponym is to *concede, profess, confess* which is defined as to *admit (to a wrongdoing)*. The superordinate concept of this chain is the verb *think, cogitate, cerebrare* which is defined as *-to use or exercise the mind or one's power of reason in order to make inferences, decisions, or arrive at a solution or judgments*. Thus, it is clear from the semantic hierarchy that to *apologize* is to undergo a logical thought process, the natural entailment of which is to admit to a wrong. Once the wrongdoing is admitted the natural consequence should be to take responsibility and offer amends. For instance, apology number 2 says- *I sincerely apologize to all Satyamites and stakeholders*. This is a clear admission of wrongdoing.

The selected concept of the verb *regret* is defined as to *feel remorse for, feel sorry for or be contrite about*. Its inherited hypernymy is to *feel, experience*, which is defined as to *undergo an emotional sensation or be in a particular state of mind*. Thus, to *regret* is to undergo a feeling by the offender about the wrongdoing. In the corpus apology number 10, the Amazon India letter states, *To the extent that these items offered by a third-party seller in Canada offended Indian sensibilities, Amazon regrets the same*.

5.1.2 Adjective – Sorry

Adjectives are primarily used for modification of nouns. They have lexical organization and

⁵ Free CLAWS WWW tagger, accessed January 15, 2017, <http://ucrel.lancs.ac.uk/claws/trial.html>, tag set C6.

⁶ <http://babelfy.org/>

semantic properties that are not shared by other modifiers and are unique to them (Miller et al, 1993). The selected sense of the adjective *sorry* in WordNet has the gloss as *feeling or expressing regret or sorrow or a sense of loss over something done or undone*. The see also relation for this is the adjective *penitent, repentant*, which means *feeling or expressing remorse for misdeeds*. Thus, the underlying semantic connotation of the word is a feeling or an emotional state.

An example of this is the sentence in the apology number 3 which states- *We are truly sorry for this and will ensure that this never happens again*. Here the use of *sorry* refers to the feelings expressed by the offender. In our dataset, out of the 18 communications, 7 have the use of *sorry*. In these 7 letters it is used 12 times.

5.1.3 Nouns – Apology and Regret

The nouns are organized as an inheritance system in WordNet (Fellbaum, 1998). Under this system there is a sequence of levels, a hierarchy, in which the lower levels inherit the features of the top levels, plus have at least one distinguishing feature. The two semantic relations of interest in the present study are hypernymy and hyponymy (Fellbaum, 1998). The selected sense of the noun *apology* has the gloss *-an expression of regret at having caused trouble for someone*. It has *acknowledgement* as its direct hypernymy, which is defined as *a statement acknowledging something or someone*. From the communicative perspective this acknowledgment is a precursor to the expectation of some sort of reparation or compensation on the part of the offended. In the corpus, the apology number 7, has the sentence, *We would like to tender an unconditional apology to the society at large and especially to the affected families and to everyone whom we have offended*. This is an unequivocal expression of apology and shows that tenderers do not want to make any excuses for their wrongdoing.

The gloss of selected sense of the noun *regret* is *sadness associated with some wrong done or some disappointment*. The direct hypernymy of this is the concept of *sadness* which is *emotions experienced when not in a state of well-being*. This is followed by the concept of *feeling or the experiencing of affective and emotional states*. Thus the hypernymy relation makes it clear that *regret* is a kind of feeling associated with

sadness. From a communicative point of view, it is simply an expression of an emotion on the part of the tenderer of the apology and not necessarily expression of remorse or liability. For example, in apology number 13, the Member of Parliament states, *I write to convey my regrets for the unfortunate incident that took place on 23rd March 2017 in the Air India flight No. AI 852, seat No.1F*. Given that the writer only uses the noun *regret*, it can be implied that the writer feels sad about the incident but not necessarily repentant. However, it is important to look at the results of SentiWordNet and WordNet-Affect to understand the implications and underlying emotions and sentiments before arriving at any further conclusions.

5.2. Keywords in SentiWordNet

The study of the sentiment associated with the keywords is done using SentiWordNet (3.0), a lexical resource which assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity (Stefano et al, 2010). The task of finding the sentiments of the words in an apology as expressed in online forums can be put to a rich set of applications (Esuli and Sebastiani, 2007). As for public apologies these tasks can range from tracking readers’ opinions about the sincerity of the communication to customer relationship management.

The selected synsets of the keywords were searched for in SentiWordNet. The sentiment scores of each of them were recorded and the results were analyzed. Table 1 shows the sentiment scores for positivity, negativity and objectivity for each of the keywords.

Keywords	PosScore [0,1]	NegScore [0,1]	ObjScore [0,1]
Sorry (Adjective)	0.125	0.75	0.125
Apology (Noun)	0.375	0.5	0.125
Regret (Verb)	0.25	0	0.75
Regret (Noun)	0.125	0.625	0.25

Apologize/ Apologise (Verb)	0	0	1
-----------------------------------	---	---	---

Table 1: SentiWordNet Scores of Keywords

In the analysis of the sentiments associated with keywords, of particular interest are the objective scores. The verb *apologize* has the highest objective score (1.0). Its negative and positive scores are zero. The high ObjScore (Objective Score) of one (1.0) implies that this verb does not convey any sentiment. In a public apology act, this could entail that when an organization or person renders an apology it distances itself from the event or issue and takes an objective position. Similarly the next highest ObjScore is for *regret* as a verb (0.75). Thus, both verbs - *apologize* and *regret*- do not connect with the negative sentiments associated with the act of an apology.

The highest NegScore (Negative Score) is for the adjective *sorry* (0.75), followed by the noun *regret* which has a NegScore of 0.625. The strong negative connotation of the adjective *sorry* could help the writer to convey his genuine feeling of remorse and hence should be preferred by the writer to connect with the reader at an emotional level. Since adjectives are the words that carry the most notions of sentiment, their use in the apology can carry the sentiment most effectively. This implies that the adjective *sorry* carries the highest sentimental load to convey the feeling associated with act of apology.

Interesting is the comparison between the verb *regret* and noun *regret*. While the verb *regret* has a high objective sentiment (0.75); the noun *regret* has a high NegScore (0.625). Thus, ‘*I regret*’ and ‘*with deep regret*’- can have very different sentimental connotations. The verb implying neutral sentiments of the apology giver and not connecting to remorse, guilt or culpability; the noun implying a strong sentiment connect.

5.3 Keywords in WordNet-Affect

We analyzed the results related to the keywords in WordNet-Affect (Strapparava & Valitutti, 2004; Strapparava et al., 2006), a linguistic resource for the lexical representation of affective knowledge. In this the affective concepts representing emotional state are individuated by synsets marked with the a-label EMOTION. There are also other a-labels

for those concepts representing moods, situations eliciting emotions, or emotional responses.

Using version 1.1, we searched for the keywords in the resource named *a-synsets* and found out its corresponding affective category in *a-hierarchy*. The presence of the word implied an emotion and the absence implied the lack of it. Table 2 shows the output for the keywords.

Keyword	WN-Affect 1.1
	a-synsets / a-hierarchy
Sorry (adj)	<adj-syn id="a#01102326" noun-id="n#05602279" caus-stat="stat"/> / <noun-syn id="n#05602279" categ="regret-sorrow"/>
Regret (verb)	<verb-syn id="v#01225879" noun-id="n#05602852" caus-stat="stat"/> / <noun-syn id="n#05602852" categ="repentance"/>
Regret (noun)	<noun-syn id="n#05602279" categ="regret-sorrow"/> / <categ name="regret-sorrow" isa="sorrow"/>
Apologize	no result
Apology	no result

Table 2. Output of Wordnet-Affect 1.1

Since the words *sorry*, and *regret* (both as noun and verb) are present in the resource we conclude that these words bear emotion. The affective category of the adjective *sorry* is *regret-sorrow* via the noun (n#05602279) and *regret-sorrow* is a *sorrow*. The verb *regret* has its affective category as *repentance* via noun (n#05602852), which in turn is a *compunction*. The noun *regret* has the affective category *regret-sorrow* which is a *sorrow*. Both the adjective *sorry* and the verb *regret* are stative, which means that the emotion related to these words are owned or felt by the speaker. The keywords *apology* (noun) and *apologize* (verb) were not present in WordNet-Affect and hence they can said to be devoid of any emotion.

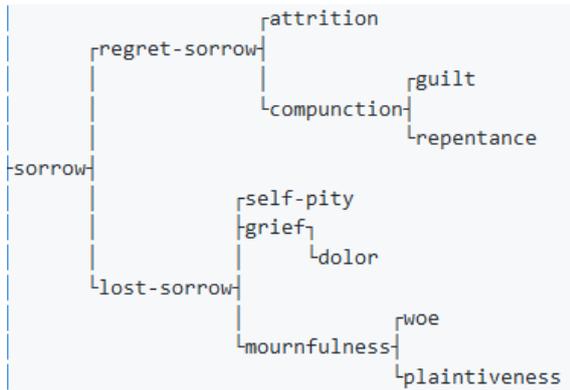


Diagram 1. Sub-tree of negative emotion *sorrow* from WordNet Domains 3.2

Thus it is clear that the emotion of the keywords found in WordNet-Affect are related to negative emotion via sadness and sorrow.

6 Discussion

In this paper we have studied a few selected keywords related to apologies, using the interlinked lexical resources, namely, WordNet, SentiWordnet and WordNet-Affect. This has given us important insights into the semantics, sentiments and emotions attached with these words and has thrown up some interesting observations which are discussed below. It is seen that semantics alone is not sufficient to give the full import of the words. The related sentiment and emotion tags provide a deeper insight into the meaning and the communicative perspective of the keywords.

First and foremost, we observed the fact that, due to a mix of factors such as greater media vigilance, and the viral nature of social media, there is certainly an increased willingness to issue public apologies in India (Kaul et.al, 2015). However, apologies available in the public domain are still limited, and so we cannot draw any generalizations from them. Hence, we can put forth certain trends and suggestions which need to be tested further on a much bigger corpus.

From the apology texts available with us, we posit that the written apology can be an effective tool for damage repair only when it crafted to communicate honest intent and a sincere tone. Thus, the words chosen should effectively convey the writer's intent.

The main observations drawn from our analysis of the keywords using WordNet, SentiWordNet and WordNet-Affect are as follows:

- **Apologize** (verb) – it is an act of cogitation, with a high objective score and no emotion label. It can be used in formal communication where emotionally laden words are to be avoided.
- **Regret** (noun) – is a kind of sadness, with a high negative score and has the emotion label of regret-sorrow and is stative. It expresses the feeling of the tenderer about the wrongdoing.
- **Sorry** (adjective) – is a kind of feeling, with a high negative score and emotion label of regret-sorrow. This keyword can be effective in situations where emotions and sentiments are strongly involved. Its use can also make the communication sound like a heartfelt apology. Also, to be noted is the fact that though the adjective *sorry* is found to be the most commonly-used form in different spoken corpora. (Harrison, 2013), yet in our data, the word *sorry* has a higher occurrence in written apologies given by individuals-in-a role and organizations. The reasons for its high occurrence in the written media in India needs to be explored further. It may be due to the very nature of the language use in social media interaction, or it could be because English is second language for Indians and poses its own compulsions on users of this language in the country.
- **Apology** (noun) – is a kind of acknowledgement, which has a high negative sentiment but no emotion label. The noun form *apologies* enable writers to distance themselves and minimise their responsibility for the offence (Harrison, 2013). When writers use this form, they may simply be following convention without consciously seeking to minimise their responsibility. Nonetheless, the established convention incorporates a distancing from the offence. Also, writers use *apologies* when they are

apologising in a role (e.g. as the representative of an organisation). When speaking personally, they use other forms, typically *sorry* (Hatipoğlu, 2005). Another possibility is that use of the noun form enables the writer to avoid the personal pronoun, creating a distance between the writer and the responsibility for the offence (ibid).

In our data, individuals have not used this form at all and of the seven occurrences of the noun form, six are by individuals as representative of an organisation. This co-relates to Harrison's finding that the word *apology/apologies* help the writers to distance themselves from the instance or event.

- **Regret** (verb) – is a kind of feeling, which has a high objective score but an emotion label of repentance. An organization or individual that is repentant of its act is less likely to repeat the transgression. An implication of this emotion label could be that the verb *regret* can imply a forbearance or even a possible reparation.

Of particular interest to us were the keywords *apology* (noun) and *regret* (verb). We compare the SentiWordNet scores and the WordNet-Affect labels of these two keywords. While emotion is defined as a relatively brief episode of response to the evaluation of an external or internal event as being of major significance. (such as angry, sad, joyful, fearful, ashamed, proud, elated, desperate), a sentiment is the positive or negative orientation that a person expresses toward some object or situation (Scherer, 2000). Thus, we can posit that the word *apology* which has no emotion label, has no or weak emotional connect, which also aligns with our conclusion about the keyword *apologize*. In contrast, the verb *regret* helps to effectively communicate the emotion of repentance. Looking at the sentiment associated with these words, we conclude that the mental attitude of the writer is more objective to the situation in using the verb *regret* while it is highly negative in the case of the usage of the word *apology*. This further implies that a high negative sentiment score means that the writer of the apology realizes the gravity of the transgression and to some extent admits to the wrong done. However,

a high objective score implies the writer taking a neutral stance to the situation and not necessarily admitting to any wrongdoing.

7 Future Work

The future plan is to make a cross-cultural analysis of written public apologies. For this purpose, the dataset will be enhanced by adding apologies from a different culture. The idea is to explore whether the linguistic aspects are affected by culture and environment. Also, we propose to validate our psycholinguistic analysis by mapping it to the readers' perception of these keywords. It will also be interesting to do a cross-lingual analysis by studying the lexical semantics of apology related words in native Indian languages.

Further, we have come across words which are being more profusely being used in written communication which were earlier thought to be part of speech acts, notably the word *sorry*. We want to understand whether this is due to the very nature of the social media where they are being used or is it because of overuse that certain words traditionally used in written media have been bleached of the sentiments and emotions attached with them, hence giving space to other words.

It is also proposed to make this study interdisciplinary by lending it to computational analysis. With an increased data set the study can be used to build a supervised sentiment analyzer using lexicons or for text categorization according to affective relevance, and opinion analysis.

Acknowledgement

We thank Dr. Sridhar Swaminathan, post-doctoral fellow at Bennett University, India, for his help in accessing WordNet-Affect.

References

- Aaron Lazare. 2005. *On apology*. Oxford University Press.
- Adam Kilgarriff. 2000. WordNet: An Electronic Lexical Database. *Language*, 76(3), 706-708.
- AIB apologises to Christians. Feb.9, 2015. Retrieved July 15, 2016 from <http://www.thehindu.com/news/cities/mumbai/aib-apologizes-to-christians/article6875094.ece>
- [Air India moves Shubha Mudgal to economy class in Mumbai-Goa flight. 2017, December 17.](https://www.hindustantimes.com/mumbai-news/air-india-moves-shubha-mudgal-to-economy-class-in-mumbai-go-a-flight-2017-12-17) Retrieved December 18, 2017, from <https://www.hindustantimes.com/mumbai-news/air-india-moves-shubha-mudgal-to-economy-class-in-mumbai-go-a-flight-offers-refund-after-her-tweet/story-8WjgBtToFGie9zXVbdtriM.html>
- Aishwarya Reganti, Tushar Maheshwari, Amitava Das, and Erik Cambria. 2017, February. Open Secrets and Wrong Rights: Automatic Satire Detection in English Text. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 291-294). ACM.
- Akio Yabuuchi. 1998. Spoken and written discourse: What's the true difference? *Semiotica*, 120(1-2), 1-38.
- Allan M. Collins and M. Ross Quillian. 1969. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2), 240-247.
- Alfonso Caramazza and Rita Sloan Berndt. 1978. Semantic and syntactic processes in aphasia: A review of the literature. *Psychological Bulletin*, 85(4), 898.
- Amazon Apologises for 'offending' Indian sentiments; writes to Sushma. 2017, January 12. Retrieved January 15, 2017, from <http://www.siasat.com/news/amazon-apologises-offending-indian-sentiments-writes-sushma-1107999/>.
- Anand Mahindra. July 7, 2017. Retrieved July 8, 2017 from <https://twitter.com/anandmahindra/status/883292235923693568>.
- Andrea Esuli, and Fabrizio Sebastiani. 2007. SentiWordNet: a high-coverage lexical resource for opinion mining. *Evaluation*, 1-26.
- Anna Trosborg. 1987. Apology strategies in natives/non-natives. *Journal of pragmatics*, 11(2), 147-167.
- Anne Wichmann. 2007. Corpora and spoken discourse. *Language and computers studies in practical linguistics*, 62(1), 73.
- Asha Kaul, Vidhi Chaudhri, Dilip Cherian, Karen Freberg, Smeeta Mishra, Rajeev Kumar, , ... and C. E. Carroll. 2015. Social Media: The New Mantra for Managing Reputation. *Vikalpa*, 40(4), 455-491.
- Bary R. Schlenker and Bruce W. Darby. 1981. The use of apologies in social predicaments. *Social Psychology Quarterly*, 271-278.
- Blum-Kulka, Shoshana, Juliane House and Gabriele Kasper. 1989. Appendix: The CCSARP coding manual. In: *Cross-Cultural Pragmatics: Requests and Apologies*. 273-294. Norwood, NJ: Ablex.
- Brian Paltridge. 2012. *Discourse analysis: An introduction*. Bloomsbury Publishing.
- Carla Bartsch. 1997. Oral style, written style, and Bible translation. *Notes on Translation-Summer Institute of Linguistics*, 11, 41-48.
- Carlo Strapparava and Alessandro Valitutti. 2004, May. WordNet Affect: an Affective Extension of WordNet. In *LREC* (Vol. 4, pp. 1083-1086).
- Carlo Strapparava, Alessandro Valitutti and Oliviero Stock. 2006, May. The affective weight of lexicon. In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 423-426).
- Chih-Hao Ku and Gondy Leroy. 2011. A crime reports analysis system to identify related crimes. *Journal of the Association for Information Science and Technology*, 62(8), 1533-1547.
- Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.
- Ciler Hatipoğlu. 2005. Do apologies in e-mails follow spoken or written norms? Some examples from British English. *Studies About Languages*, 5, 21-29.
- Clément Michard. n.d. WNAffect - A python module to get the emotion of a word. Retrieved July 12, 2017 from <https://github.com/clemtoy/WNAffect>.
- Cohen, A. D., & Olshtain, E. (1981). Developing a measure of sociocultural competence: The case of apology. *Language learning*, 31(1), 113-134.

- C.P. Gurnani. July 7, 2017. Retrieved July 8, 2017 from https://twitter.com/C_P_Gurnani/status/883275712886480896.
- Douglas Biber. 2006. *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.
- E.D.Kuhn, Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. 1991. Cross-Cultural Pragmatics: Requests and Apologies. *Language*, 67(1), 169.
- Elite Olshtain and Andrew Cohen. 1983. Apology: A speech act set. *Sociolinguistics and language acquisition*, 18-35.
- Embarrassed, upset over handful of employees: Shikha Sharma. 2016, December 18. Retrieved July 26, 2017, from <http://www.moneycontrol.com/news/business/companies/embarrassed-upset-over-handfulemployees-shikha-sharma-937258.html>
- François-Régis Chaumartin. 2007, June. UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 422-425). Association for Computational Linguistics.
- Flipkart sends apology mail to customers after its botched 'Big Billion Day Sale'. 2014, October 07. Retrieved June 23, 2016, from <http://www.news18.com/news/business/flipkart-sends-apology-mail-to-customers-after-its-botched-big-billion-day-sale-718718.html>.
- Full text: This letter Ramalinga Raju wrote uncovered the Rs 4,676 cr Satyam scam. 2015, April 09. Retrieved June 23, 2016, from <http://www.firstpost.com/business/full-text-this-letter-ramalinga-raju-wrote-uncovered-the-rs-4676-cr-satyam-scam-2190559.html>.
- Gary Chapman and Jennifer Thomas. 2008. *The five languages of apology: How to experience healing in all your relationships*. Moody Publishers.
- George A. Miller and Philip N. Johnson-Laird. 1976. *Language and perception*. Belknap Press.
- George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1), 1-28.
- George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- IndiGo on Twitter: We truly apologise. 2017, November 07. Retrieved December 15, 2017, from <https://twitter.com/indigo6e/status/927926970641297408>
- Infosys chief under fire for remark on national anthem. 2007, April 11. Retrieved October 05, 2016, from <http://www.thehindu.com/todays-paper/Infosys-chief-under-fire-for-remark-on-national-anthem/article14747977.ece>.
- In letter to users, Myntra apologises for the big app crash. 2015, June 03. Retrieved June 23, 2017, from <http://indianexpress.com/article/business/companies/in-letter-to-users-myntra-apologises-for-the-big-app-crash/>
- In Sena vs All Airlines, Threats Fly. Then, a Note Of Apology From Its MP .2017, April 06. Retrieved April 07, 2017, from <http://www.ndtv.com/india-news/as-usual-sena-threatens-misbehaves-and-airline-ban-on-its-mp-ravindra-gaikwad-to-end-1678128>.
- James W. Pennebaker, Martha E. Francis and R. J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001), 2001.
- Janet Holmes. 1990. Apologies in New Zealand English1. *Language in society*, 19(2), 155-199.
- Jessica R. Nelson, Michal Balass and Charles A. Perfetti. 2005. Differences between written and spoken input in learning new words. *Written Language & Literacy*, 8(2), 25-44.
- Jim Miller. 2006. Clause Structure in Spoken Discourse. *Encyclopedia of Language & Linguistics*, 481-483.
- Jiyang Bae, and Sun- A. Park. 2011. Socio-Contextual Influences on the Korean News Media's Interpretation of Samsung's \$847.6 Million Donation. *Journal of Public Relations Research*, 23(2), 141-166.
- John L. Austin. 1975. How to do things with words (JO Urmson & M. Sbisá, Eds.). *Harvard U. Press, Cambridge, MA*.
- Jorge Carrillo-de-Albornoz and Laura Plaza. 2013. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the Association for Information Science and Technology*, 64(8), 1618-1633.
- Judith Klavans and Min Yen Kan, 1998. Role of verbs in document analysis. In *Proceedings of the 17th international conference on Computational*

- linguistics-Volume 1* (pp. 680-686). Association for Computational Linguistics.
- Karin Aijmer and Anna-Brita Stenström, (Eds.). 2004. *Discourse patterns in spoken and written corpora* (Vol. 120). John Benjamins Publishing.
- Kay Wikberg. 2004. *English metaphors and their translation* (pp. 245-265). Philadelphia: John Benjamins.
- Keith Michael Hearit. 2010. *Crisis management by apology: corporate response to allegations of wrongdoing*. Routledge.
- Ken-ichi Ohbuchi, Masuyo Kameda, and Nariyuki Agarie. 1989. Apology as aggression control: its role in mediating appraisal of and response to harm. *Journal of personality and social psychology*, 56(2), 219.
- Khosrow Jahandarie. 1999. *Spoken and written discourse: A multi-disciplinary perspective* (No. 1). Greenwood Publishing Group.
- Klaus R. Scherer. 2000. Psychological models of emotion. *The neuropsychology of emotion*, 137(3), 137-162.
- Marion Owen. 1983. Apologies and remedial exchanges. *The Hague*. Mouton.
- Matthew D. Turner, David M. Weinberg, and Jessica A. Turner. 2011. *A Simple Ontology for the Analysis of Terrorist Attacks*.
- Merryanna L. Swartz and Masoud Yazdani (Eds.). 2012. *Intelligent tutoring systems for foreign language learning: The bridge to international communication* (Vol. 80). Springer Science & Business Media.
- Michael A. Halliday. 1989. *Spoken and written language*. Oxford: Oxford University Press.
- Michael McCarthy and Diana Slade. 2007. Extending our understanding of spoken discourse. *International Handbook of English Language Teaching*, 859-873.
- Myntra does a Flipkart, apologises over tech glitch during mega-sale. 2015, June 04. Retrieved June 15, 2015, from http://www.exchange4media.com/digital/myntra-does-a-flipkart- apologises-over-tech-glitch-during-mega-sale_60277.html.
- Nicole Novielli and Carlo Strapparava. 2013. The role of affect analysis in dialogue act identification. *IEEE Transactions on Affective Computing*, 4(4), 439-451.
- PETA-Apology-Letter-Suriya. 2017, January 27. Retrieved February 02, 2017, from <http://tamil.cinemapettai.in/peta-apologize-to-suriya/peta-apology-letter-suriya/>.
- Peter Oram. 2001. WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). Cambridge, MA: MIT Press, 1998. Pp. 423.-. *Applied Psycholinguistics*, 22(1), 131-134.
- Rebecca Hughes. 1996. *English in speech and writing: Investigating language and literature*. Psychology Press.
- Roy C. O'Donnell. 1973. Some Syntactic Characteristics of Spoken and Written Discourse. *Studies in Language Education*, Report No. 4.
- Russel Craig, Tony Mortensen, and Shefali Iyer. 2013. Exploring top management language for signals of possible deception: The words of Satyam's chair Ramalinga Raju. *Journal of Business Ethics*, 113(2), 333-347.
- Ryan Goei, Anthony Roberto, Gary Meyer and Kellie Carlyle. 2007. The effects of favor and apology on compliance. *Communication Research*, 34(6), 575-595.
- Samuel Fillenbaum and Lyle V. Jones. (1965). Grammatical contingencies in word association. *Journal of Verbal Learning and Verbal Behavior*, 4(3), 248-255.
- Sandra Harrison and Diane Ailton in Herring, S., Stein, D., & Virtanen, T. (Eds.). 2013. *Pragmatics of computer-mediated communication* (Vol. 9). Walter de Gruyter.
- Sorry, We Messed Up. 2015, April 25. Retrieved July 23, 2016, from <https://www.scoopwhoop.com/inothernews/apology/>
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010, May. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 2200-2204).
- Stephen H. Levinsohn. 2000. *Discourse features of New Testament Greek: A coursebook on the information structure of New Testament Greek* (p. 39). Dallas: SIL International.
- Suman Lee and Surin Chung. 2012. Corporate apology and crisis communication: The effect of responsibility admittance and sympathetic

expression on public's anger relief. *Public Relations Review*, 38(5), 932-934.

TRAI.2016. Indian Telecom Services Performance Indicator Report for the Quarter ending September, 2015 – Released 16 Feb 2016. Retrieved March,2016 from <http://www.trai.gov.in/release-publication/reports/performance-indicators-reports>

Trolled, Priyanka Chopra Apologises For Calling Sikkim 'Troubled By Insurgency'.2017, September 14. Retrieved October 20, 2017, from <https://www.ndtv.com/entertainment/priyanka-chopra-made-people-very-angry-by-calling-sikkim-insurgency-troubled-1750119>

Tushar Maheshwari, Aishwarya N.Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck and Amitava Das. 2017. A Societal Sentiment Analysis: Predicting the Values and Ethics of Individuals by Analysing Social Media Content. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Vol. 1, pp. 731-741).

Wallace Chafe. 1992. The flow of ideas in a sample of written language. *Discourse description: Diverse linguistic analyses of a fund-raising text*, 267-94.

We Are Sorry! 2015, April 25. Retrieved June 23, 2016, from <http://blog.lenskart.com/we-are-sorry/>

William W. Maddux, Peter H. Kim, Tetsushi Okumara, and Jeanne M. Brett. 2012, June. Why 'I'm sorry' doesn't always translate. *Harvard Business Review*. Retrieved 7 September, 2015 from <http://hbr.org/2012/06/why-im-sorry-doesnt-always-translate>.

Willis J. Edmondson. 1981. On Saying You're Sorry. *Conversational Routine. Explorations in Standardized Communication Situations and Prepatterned Speech*, Florian Coulmas (ed.). The Hague Mouton,.pp. 273-288.

Derivational Relations in Arabic WordNet

Mohamed Ali Batita

Research Laboratory in Technologies
of Information and Communication &
Electrical Engineering,
Tunisia
BatitaMohamedAli@gmail.com

Mounir Zrigui

Research Laboratory in Technologies
of Information and Communication &
Electrical Engineering,
Tunisia
Mounir.Zrigui@fsm.rnu.tn

Abstract

When derivational relations deficiency exists in a wordnet, such as the Arabic WordNet, it makes it very difficult to exploit in the natural language processing community. Such deficiency is raised when many wordnets follow the same development path of Princeton WordNet. A rule-based approach for Arabic derivational relations is proposed in this paper to deal with this deficiency. The proposed approach is explained step by step. It involves the gathering of lexical entries that share the same root, into a bag of words. Rules are then used to affect the appropriate derivational relations, i.e. to relate existing words in the AWN, involving part-of-speech switch. The method is implemented using Java. Manual verification by a lexicographer takes place to ensure good results. The described approach gave good results. It could be useful for other morphologically complex languages as well.

1 Introduction

A wordnet is a lexical database built of synsets. One synset represents one concept and contains words from the same part of speech (POS) (noun, adjective, verb, and adverb). Synsets are interconnected with different relations. But, there are no cross-part-of-speech relations. This type of relation is a link between words sharing the same stem and meaning like the verb 'eat' and the noun 'eater'. The first WordNet, Princeton WordNet (Fellbaum, 2010), was built for the English language. Since that, many wordnets has seen the light for over 160 languages¹. One of them is the Arabic WordNet (Elkateb et al., 2006)

¹Extended Open Multilingual WordNet: <http://compling.hss.ntu.edu.sg/omw/summx.html>

(henceforth, AWN) for the Modern Standard Arabic. AWN followed the development of Princeton WordNet and EuroWordNet (Vossen, 1998).

Started in 2007, researches on AWN are made to improve it. Some of them improved its contents (Boudabous et al., 2013; Saif et al., 2015). Others used it in different disciplines of the Natural Language Processing (henceforth, NLP) (Abouenour et al., 2008; Abouenour et al., 2013). Despite the greatness of these works, it clearly did not take into consideration the specificities of the Arabic language, especially, its morphological aspect.

A lexicon, like AWN, needs to have an extensive coverage, high quality, and multiple use in NLP applications (Mallat et al., 2015a) (Mallat et al., 2015b) (Ayadi et al., 2014) (Mohamed et al., 2015). Adding to that, derivational morphology provides handful information for the benefit of the NLP. As proof, Wilbur and Smith (Wilbur and Smith, 2013) showed that it can be used to calculate probabilities of semantic relatedness. Also, Sagot (Sagot, 2010) used derivational analysis to determine if an unknown word can be used to create a new one for a lexicon extension. Derivational morphology is used to extend different wordnets like Bulgarian, Serbian and Romanian WordNet (Koeva et al., 2008; Koeva, 2008; Mititelu, 2012). The aim of this pilot study is to enrich the AWN with derivational relations using a rule-based approach to extend its coverage and turns it into a more useful knowledge base.

The rule-based approach includes domain knowledge into linguistic knowledge. This yield accurate results. Yet, linguistic knowledge acquired for one NLP system may be reused to build others systems that require similar knowledge. Those approaches are based on a solid core of linguistic knowledge. They depend on hand-constructed rules from a lexicographic rather than automatically gathered from data.

This paper is structured into five sections. Section 2 is an overview of the AWN. Section 3 will provide some background on the Arabic language. In section 4, we will discuss some related works. We will also discuss the choice of the rule-based approach regarding other approaches. We will speak about our approach in details in section 5. Last but not least, we will show the obtained results in section 6.

2 Overview of the Arabic WordNet

AWN's development followed the top-down procedure. It consists of translation the Princeton WordNet's core and extending it through more specific concepts related to the Arabic culture. This procedure expands the compatibility between wordnets. It is based on manually encoding of the specific concepts. The first version V1 of AWN contains 21,813 words grouped into 9,698 synsets, 6 types of relations between those synsets corresponding to 143,715 links. The second version V2 released in 2008 containing new synsets and links. It contains 11,269 synsets, equivalent to 23,481 words, and 161,705 links equivalent to 22 types (5 of them for the interconnection between PWN and AWN) (Batita and Zrigui, 2017). Another version is freely available on the internet structured under the Lexical Markup Framework (LMF) developed by (Abouenour et al., 2013). Table 1 below will recapitulate the number of the words, synsets, and links in all the 3 versions of the AWN.

Table 1: AWN's versions.

	V1	V2	LMF
Words	21,813	23,481	60,157
Synsets	9,698	11,269	8,550
Links	143,715 (6 types)	161,705 (22 types)	41,136 (4 types)

We notice that V2 contains more synsets and fewer words than the LMF file. In one hand, V2 has 11,269 synsets related with 161,705 links, on the other hand, the LMF file 8,550 synsets related with only 41,136 links. This is not proportional to the number of the words.

There are many kinds of links in AWN V2. Table 2 displays 17 links. There are 5 others links but they are not between Arabic words. They are inter-language links. We have no interest in them.

each one with an example and if it exists in the LMF file or not.

Table 2: Links in AWN.

Link	Example	Frequency	
		V2	LMF
Has hyponym	ماء، شراب، ماء (drink, water) <i>šrāb, mā-</i>	9,347	19,806
Has derived	تعليم، تعليمي <i>tʿlym,</i> <i>tʿlym</i> (educational, education)	178	-
Related to	ملجأ، ملجأ <i>lǧʿa, mlǧʿa</i> (refuge, shelter)	4,769	-
Has holomember	اكل لحم، لواحم <i>ākl</i> <i>lḥm, lwāḥm</i> (carnivore, carnivores)	334	-
Near antonym	زيادة، نقصان <i>zyādt,</i> <i>nqṣān</i> (increase, decrease)	772	14
See also wn15	وديعة، طلبية <i>wdyʿt,</i> <i>ṭlbyh</i> (deposit, order)	166	-
Has holopart	خلية، متعضي <i>ḥlyt,</i> <i>mtʿdy</i> (cell, organism)	697	-
Has holomade of	ورقة، صفحة <i>wrqt,</i> <i>ṣfḥh</i> (paper, page)	60	-
Has subevent	وقف، قام <i>wqf, qām</i> (stand, stand up)	128	-
Category term	انسان، جسم <i>ānsān,</i> <i>ǧsm</i> (human, body)	548	-
Near synonym	أسبق، مبكر <i>ʿasbq,</i> <i>mbkr</i> (premier, early)	122	412
Be in state	اتصل، متصل <i>ātṣl, mtṣl</i> (contact, connected)	83	-
Has instance	عاصمة، القاهرة <i>ʿā-</i> <i>ṣmt, ālqāhrh</i> (capital, Cairo)	929	549
Verb group	ضرب، صدم <i>drb, ṣdm</i> (hit, bump)	142	-
Causes	حوّل، حرك، حوّل <i>ḥrk, ḥwl</i> (move, displace)	75	-
Region term	بابل، عشتار <i>bābl, ṣtār</i> (Babylon, Ishtar)	35	-

Usage term	أسبرين، إسم تجاري asbryn, ism tġāry (Aspirin, commercial name)	3	-
------------	----------------------------------------------------------------------	---	---

To clarify, the link *near synonym* is represented in the LMF file by the name *similar*² and *near antonym* by just *antonym*. The two links *has hyponym* and *has instance* are splitted into *hyponym/hypernym* and *isInstance/hasInstance* respectively.

If we can classify those links, we can say that there is two types; semantic and derivational link. Semantic links rely on words sharing some meaning. Most of the presented links are semantic like *has holo part*, *has holo made of*, *has subevent*... Only two links are morphosemantic links; *has derived* and *related to*. Not only they are morphologically but also semantically relying on words. They rely on words that share the same root but have different POS.

There is a third type of link which is morphosemantic relations. As it is claimed in (Šojat et al., 2012), there is a difference between the derivational and morphosemantic links. The derivational relations are language-specific while the morphosemantic relations are not.

3 Arabic language

As it is widely known, the Arabic language is a Semitic language which makes it different from other languages, like English or French. It is characterized by an inflectional and derivational morphology. Inflectional morphology is divided into verbal and nominal morphology. The verbal morphology bends on the aspect, the mood, the voice and the subject (person, gender, and number) of the verbs. The nominal morphology bends on the gender, the number, the state, and the case of nouns, the adjectives, and the proper nouns. The derivational morphology consists of the deverbal noun, the active participle, the passive participle and other derivations based on patterns change (Habash, 2010). All things considered, this richness provides an effective information for many NLP tasks.

Besides, Arabic is a notable language for its *nonconcatenative* morphology which is the modification of the internal structure of a word. In other

²The link *similar* exist in V2 but it is an interlanguage link.

words, it is a form of a word in which the root, usually three consonants and called *triliteral root*, is modified by adding other consonants and vowels. Generally, in Arabic, the derivation is based on three concepts. Given a *root* and a *pattern*, you can create a word *form* by applying derivational rules. This makes it difficult to automatically construct new words from a primitive root. For example, the Arabic words *دّارس* *dārs* (student) and *مدرّس* *mdrs* (teacher)³ share the same Arabic root *د - ر - س* *d - r - s* (d - r - s) which is the concept of studying. To that end, we can say that those two words are *derivationally* and *semantically* related. More details about the Arabic morphology, you can found it in (Habash, 2010).

4 Related Work

Even though derivational morphology is a numerous area of studies, we did not found many lexical resources that rely on this kind of morphology, in the Arabic language. Derivational relations enrichment started with the Turkish WordNet in 2004. Bilgin et al. (Bilgin et al., 2004) described a semi-automatic approach to add new synsets by applying derivational rules to existing words and add a morpho-semantic link between them. This type of approach is basically adding automatically suffix and prefix to a steam. Since it is automatic, manually validation is mandatory and important. the same work is done to the Czech WordNet (Pala and Hlaváčková, 2007).

Fellbaum et al. (Fellbaum et al., 2007) did not follow the same approach but instead, he added morphological relations between derived pairs of words in PWN. The derived pairs of words are recognized automatically since they share the same steam. Manual validation is also necessary. This type of relation is cross-POS (between verb and noun pairs). In 2012, the same kind of work is followed in the Romanian WordNet by Mititelu (Mititelu, 2012). The work is summarized in two steps. The first step is to create all possible combination, given 3 lists of words, prefixes, and suffixes. The second step is to validate the affectation of prefixes and suffixes, each one aside, using a set transformation rules.

The Bulgarian (Koeva, 2008), the Serbian (Koeva et al., 2008) and the Polish WordNet (Piasecki

³From now on, Arabic words will be followed by their transliteration using the transliteration system of L^AT_EX and their English translations in brackets.

et al., 2009) adopted another type of approach. Based on the alignment to the PWN, the approach consists of transforming the derivational relations existing in the PWN to each wordnet. In their case of study, Koeva et al. (Koeva et al., 2008) proposed several approaches to make the generating of new synsets and relations possible based on derivational patterns of different POS.

Outside wordnets, *Lefff* (Sagot, 2010) is a morphological lexicon for French based on the lexical framework Alexina. This framework is used with different languages to develop morphological and syntactic NLP lexicons like Persian, Sorani, Kurdish and even English. This lexicon is freely available and has a large coverage. It is constructed by merging several existing resources using semi-automatic techniques and conversion. Remaining with the same language, *VerbAction* (Tanguy and Hathout, 2002) too is a morphological resource who couples verbs with their action nouns (inspect/inspection). *VerbAgent* (Tribout et al., 2012) is the same as *VerbAction* but with agent nouns (inspect/inspector).

The available evidence seems to suggest that the development of those resources is either based on manual work or validation and/or lexical information (derivational and morphological rules). Other attempted researches are less supervised and based only on morphological information. Can et al. (Can and Manandhar, 2009) proposed an unsupervised method based on different POS to produce morphological rules. Bernhard (Bernhard, 2010) described two methods for unsupervised learning of morphological families. The first one is called *MorphoClust*. It clusters words into families using hierarchical classification methods. The second one is called *MorphoNet*. It constructs a lexical network from the words presented in *MorphoClust*. The words represent the nodes and the morphological relations represent the links between those words.

Recently in 2016, Zaghouni et al. (Zaghouni et al., 2016) have developed the AMPN, a semantic resource, based on Arabic morphological patterns. It clusters the verbs of Arabic PropBank⁴ (Kipper et al., 2008) according to their morphological patterns. Arabic verbs are studied according to their lemmas. They are defined as a combination of root and morpheme patterns of the verbs.

⁴Annotated corpus with verbal propositions and arguments.

Basically, the cited approaches rely on morphological rules. In another way, they are rule-based approaches. Each one used some morphological rules specific to its language to whether generate new words (adding prefixes and suffixes) or coupling existing words (share the same stem). The advantage of this type of approach is the analysis of the input and output of a system using linguistic knowledge. Also, it helps the language learner's to better understand the language. However, other approaches, like statistical-based or machine learning, cannot distinguish between well-formed or ill-formed input which is an issue in some NLP tasks (Shaalán, 2010).

There is a rapidly growing literature on (Shaalán, 2010), which indicates that rule-based approach for Arabic NLP tasks reported successful results. Shaalan presented 4 tools and 3 systems based on Arabic morphological and syntactic rules. The tools are about morphological analyzer/generator and syntactic analyzer/generator. The 3 systems are Machine Translation, Named Entity Recognition, and Computer-assisted Language Learning. The aim of this study is to show that the development of systems based on rule-based approach is feasible with languages like Arabic (absence of linguistic resources and difficulties of adapting tools from other languages...). All things considered, it seems reasonable to base our work on this kind of approach. Next section will describe precisely each step of our proposed approach.

5 Our Approach

Since there is a lack of derivational relations in AWN, we will attempt to add them based on the existent words in it. The suggested approach depends on lexical entries and some transformation rules. We will gather lexical entries sharing the same root into *bag of words* and we will use the rules to affect the appropriate types of derivational relations. Each rule is based on the POS and the patterns of the words. The following figure 1 shows an Arabic word with its derived forms and each with its pattern (*1, 2, and 3 in the patterns refer to the three consonants of the trilateral root*).

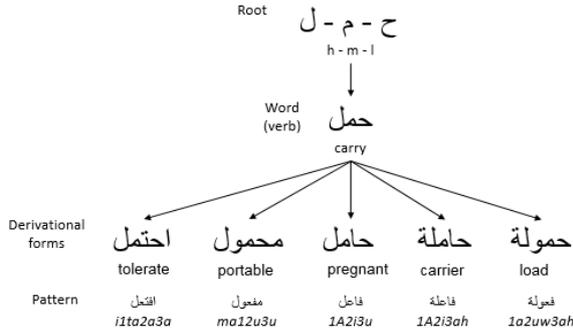


Figure 1: The derivations of the Arabic verb حمل *hml* (carry) with their patterns.

The issue under scrutiny in derivational morphology is creating new words from others. In our work, instead of creating new words we will use only words that exist in AWN and try to make a connection between them. This task involves POS switch (sometimes, it is preserved and we will see how). To give an illustration, let's look at the example in table 3. We gain from a verb a noun and from a noun another noun and an adjective.

Table 3: Derivation between part of speeches.

Verb → Noun	Noun → Adjective	Noun → Noun
كتب، كتاب <i>ktb, ktāb</i> (write, book)	كتابي، كتاب <i>ktāb, ktāby</i> (book, my book)	كتيب، كتاب <i>ktāb, ktyb</i> (book, brochure)

One can tell that there is a link between two words if (i) they belong to the AWN (ii) they share the same root and (iii) there is a rule which allows the transformation. Our method is described step by step in the next subsections.

5.1 Clustering Words into Bag of Words

First of all, we gather the words that share the same root in a so-called *bag of words*. This step is based on the root of each word in AWN. This step will help us to:

1. Eliminate the *underivatized* words like named entities ... بنز - مرسيديس - اينشتاين *ā-ynštāyn, mrsyds - bnz...* (Einstein, Mercedes-Benz...) and multiword expression,
2. Keep the *apolistic generic noun* like ... خروف، فيل *ħrwf, fyl...* (sheep, elephant...),

3. Distinguish words that share the same root but no relationship in the stage of meaning like the noun شجر *šğrun* (trees) and the verb شاجر *šāğr* (dispute),

4. Finally, verify the POS of the rest of the word, since it has an important role in our work.

Most of the Arabic nouns are derived from verbs. Verbs are categorized into their classes. First of all, we see the class of the verb if it is trilateral or not. Classes need to be indicated in each bag of words because different class means different rule to get the correct noun. To better understand the issue let us take a look at the example in table 4 of some verb forms with different classes, their verbal nouns, and examples.

Table 4: Verb forms with verbal nouns and examples.

Verb form	Verbal noun	Example
أَفْعَلَ <i>aafʿala</i> (a12a3a)	إِفْعَالٌ <i>ifʿaālun</i> (a12A3u)	أَسْلَمَ، إِسْلَامٌ <i>aslm, islām</i> (embrace Islam, Islam)
إِنْفَعَلَ <i>infʿala</i> (an1a2a3a)	إِنْفِعَالٌ <i>ānfiʿāl</i> (an1i2A3u)	إِنْقَلَبَ، إِنْقِلَابٌ <i>inqlb, inqlāb</i> (Turn over, coup)
فَعَّلَ <i>faʿala</i> (1a223a)	تَفْعِيلٌ <i>tafiyl</i> (ta12I3u)	نَفَّسَ، تَنْفِيسٌ <i>nfs, tnfys</i> (discharge, discharged)

We can notice that there is a change in verbal noun if we change the class and the form. This issue is detailed with the transformation rules in the next subsection.

5.2 Transformation Rules

As explained before, the rules are the main part of our method because they provide the existence of the relationship and its type. First, the existence of a relation between the pair of words in the same bag is determined by the set of rules in table 5.

Table 5: Transformation rules related to the POS.

No	POS switch	Type of relation	Example
1	Verb → Verb	HasDerivedVerb	أكل، تأكل <i>akl, tākḷ</i> (eat, abrade)
2	Verb → Noun	ActiveParticiple	كتب، كاتب <i>ktb, kātb</i> (write, writer)
		PassiveParticiple	كتب، مكتوب <i>ktb, mktwb</i> (write, written)
		Location	لعب، ملعب <i>lb, mlb</i> (play, stadium)
		Time	غرب، مغرب <i>grb, mgrb</i> (go west, Occident)
		Instrument	فتح، مفتاح <i>ftḥ, mftāḥ</i> (open, key)
3	Noun → Noun	HasDerivedNoun	كلب، كليب <i>klb, klyb</i> (dog, doggy)
4	Noun → Adjective	Relatedness	سياسة، سياسي <i>syāst, syā-sy</i> (politic, political)

The problem now is how we can determine the relationship between words in the same bag if it exists of course. Different POS in the same bag is the key for this. Table 5 shows the possible combination in a bag of words that one can find. With the first rules it is easy, if the pair has the same POS (which in this case is a verb) the relation should be *hasDrivenVerb* like the example shows and the same thing goes for the third and the fourth rule. The rule number 2 is a complex one. From all the nouns that you have, e.g you need to distinguish between the active and the passive participle.

The next set of rules will help us to determine all the type of relations between the nouns derived from one verb according to their forms. This will be based on the class of the verb presented in each bag. After a deep look into the behavior of the Arabic verbs and their derivations, the study ap-

pears to suggest that we should classify the verbs into two classes, trilateral, and non-trilateral verbs. The table 6 will summarize the transformation rules needed.

Table 6: Transformation rules for the relations between verbs and nouns.

Relation	Verb class	Noun Pattern	Example
ActiveParticiple	Trilateral	فاعل <i>fāʿl</i> (1A2i3u)	حمد، حامد <i>ḥmd, ḥā-md</i> (praise, praiser)
		<i>weak letter</i> ⁵ in the 2nd position → ي <i>y hamza</i>	فاح، فائح <i>fāḥ, fā-yḥ</i> (spread, Exhalant)
		<i>weak letter</i> in the 3rd position → ي <i>y ya</i>	دعا، دعي <i>dā, d-y</i> (call, caller)
	Non-trilateral	مُفْعِل <i>mufʿil</i> (mu1a2i3u)	علم، معلم <i>ʿlm, mʿlm</i> (teach, teacher)
PassiveParticiple	Trilateral	مفعول <i>mfʿwl</i> (ma12u3u)	شرب، مشروب <i>šrb, mšrwb</i> (drink drinkable)
		م <i>m</i> (m)+ the deverbal noun	قال، مقول <i>qāl, mqwl</i> (say,)said
	Non-trilateral	مفاعل <i>mfāʿl</i> (m1A2i3u)	بارك، مبارك <i>bārk, mbā-rk</i> (bless, blessed)
Location	Trilateral	مفعل <i>mfʿal</i> (ma12a3u)	طبخ، مطبخ <i>ṭbḥ, mṭbḡ</i> (cook, kitchen)
Time	Trilateral	مفعل <i>mfʿil</i> (ma12i3u)	غرب، مغرب <i>grb, mgrb</i> (go west, sundown)

Instrument -	مفعول <i>mfʻl</i> (mi12a3u)	عول، معول <i>wl, mʻwl</i> (count on, pick)
	مفعلة <i>mfʻh</i> (mi12a3h)	قلم، مقلمة <i>qlm, mqlmh</i> (prune, pen case)
	مفعال <i>mfʻāl</i> (mi12A3u)	فتح، مفتاح <i>fth, mftāh</i> (open, key)
	فعالة <i>fʻālh</i> (li2A3h)	غسل، غسالة <i>gsl, gśālh</i> Washer

To better understand the pattern transformation, you have to think of it as an algorithm. Take the example of the active participle with a trilateral verb who has a weak letter in the second position⁶, if such verb does exist in the bag of words alongside with a noun who has a *hamza* in its 3rd position then the relation between them should be made and it is a *activeParticiple* one, and so on for the rest of the nouns. The example of the *instrument* relation, if in the bag of words, a noun with the same pattern as مفعلة *mfʻh* (mi12a3h) does exist then the relation between its verb should be made.

If you look carefully, the pattern مفعال *mfʻāl* (ma12a3u) is presented with four relations, *activeParticiple*, *location*, *time*, and *instrument*. We can distinguish the *activeParticiple* by the diacritics. In our work, the diacritics are taken into consideration to affect the proper relations. Beside, AWN's words presented with diacritics. *Location*, *time*, and *instrument* are undistinguished and it is totally logic. The kind of patterns used with those relations are distinguished only in the context. Otherwise, we cannot separate them. Like the words مغرب *mḡrb* presented in the example of سافرنا إلى المغرب *sāfrnā ilā ālmḡrb* (we traveled to Morocco) and عدنا قبل المغرب *dnā qbl ālmḡrb* (we come back before sundown) with a different purpose. The first one indicates the location

⁵There are 3 weak letters in the Arabic ا، و، ي *ā, w, y* according to their positions in the root we can tell if the verb is *asimilated*, *hollow* or *defective*

⁶This type of verb is called *hollow verb*.

and the second indicates the time. After All these automatic steps we finally can to stage of validation.

5.3 Validation

The steps of the approach are validated according to a lexicographer. The rules too, they are proposed and well studied, as well as the classes of the verbs. Some irregular rules are not taken into considerations because (i) we did not found much of them in AWN or (ii) they will create a confusion with other rules. For example, with nouns, there are other rules like the dual, plural, possessive form. We did not find much of them so we decided to put a general rule for all of them (rule №3 in table 5). We suggested to only work with pertinent rules. We did not go for the automatic validation because the manual verification always leads to better results than the automatic one. It is time-consuming but when you need a better precision you have to sacrifice time.

6 Test and Evaluation

We implemented the method described in the previous section using Java. The first thing we did is cluster words sharing the same root in a bag of words. We notice that some nouns are tagged as an adverb so we verified the POS of each word. Also, some adjectives are wrongly tagged. We corrected as many as we could. We also eliminate named entities and multiword expression because they are *underivatized*. For our own good, The named entities are already tagged so we only eliminated the multiword expression. We only retained nominal, verbal, and adjectival entries. The results are presented in table 7 after the elimination and correction.

Table 7: New frequency of the words in the LMF.

POS	Frequency	New frequency
Noun	16,432	10,325
Verb	42,298	40,143
Adjective	771	498
Total number of bags	6,608	5,462

We fixed the number of bags to 5,462. Each bag has its own set of verbs, nouns, and adjectives and it is cleaned for anything that will misguide the affection of the relation in the next step.

As described in the previous section, the verb class is an important fact in the affectation of the relation. 4,275 bags contain verbs. We classified those bags according to the verb form into two classes. Table 8 shows the detailed frequency.

Table 8: Frequency of verb classes.

Verb class	number of bag of words
Triliteral	3,089
Non-triliteral	1,186

The classification will facilitate the affectation of the relation, which is our next step. All kind of relations described in table 5 was found in the bag of words. Table 9 shows the frequency of each one. Adding the 8,865 new relations to the existing ones, we got 50,001.

Table 9: Frequency of new relations.

Relation	Frequency
HasDerivedVerb	2,005
ActiveParticiple	1,347
PassiveParticiple	1,004
Location	985
Time	752
Instrument	184
HasDerivedNoun	1,784
Relatedness	804
Total	8,865

7 Conclusion

The present paper puts forward a pilot study on the derivational relations between words in Arabic WordNet. Our goal was to engage the specificity of the Arabic word's morphology to enrich the AWN with more precisely relations. Firstly, we clustered the words presented in AWN into a bag of words based on their roots. Secondly, we proposed some morphological rules based on a core of solid linguistic knowledge to identify the existence and the type of relations in each bag of words. Each rule presents the possible patterns that a word can have. Finally, we validated our work with a native speaker and a lexicographer. Our future work will be the test of this new version

of the Arabic WordNet in a system like Retrieval Information or Word Sense Disambiguation.

Acknowledgments

This work was supported by the programmer Mr. Abobakr Ahmed Bagais and the lexicographer Mr. Abouloubaba Regui. We would like to thank them for their invaluable advice and encouragement on this research work.

References

- Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2008. Improving q/a using arabic wordnet. In *Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia, December*.
- Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2013. On the evaluation and improvement of arabic wordnet coverage and usability. *Language resources and evaluation*, 47(3):891–917.
- Rami Ayadi, Mohsen Maraoui, and Mounir Zrigui. 2014. Latent topic model for indexing arabic documents. *International Journal of Information Retrieval Research (IJIRR)*, 4(1):29–45.
- Mohamed Ali Batita and Mounir Zrigui. 2017. The enrichment of arabic wordnet antonym relations. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017*.
- Delphine Bernhard. 2010. Apprentissage non supervisé de familles morphologiques: comparaison de méthodes et aspects multilingues. *Traitement Automatique des Langues*, 2(51):11–39.
- Orhan Bilgin, Ozlem Cetinoglu, and Kemal Ofizer. 2004. Morphosemantic relations in and across wordnets. In *Proceedings of the Global Wordnet Conference*, pages 60–66.
- Mohamed Mahdi Boudabous, Nouha Chaâben Kamoun, Nacef Khedher, Lamia Hadrich Belguith, and Fatiha Sadat. 2013. Arabic wordnet semantic relations enrichment through morpho-lexical patterns. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6. IEEE.
- Burcu Can and Suresh Manandhar. 2009. Unsupervised learning of morphology by using syntactic categories. In *CLEF (Working Notes)*.
- Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 22–28.

- Christiane Fellbaum, Anne Osherson, and Peter E Clark. 2007. Putting semantics into wordnets "morphosemantic" links. In *Language and Technology Conference*, pages 350–358. Springer.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- Nizar Y Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.
- Svetla Koeva, Cvetana Krstev, and Duško Vitas. 2008. Morpho-semantic relations in wordnet—a case study for two slavic languages. In *Global Wordnet Conference*, pages 239–253. University of Szeged, Department of Informatics.
- Svetla Koeva. 2008. Derivational and morphosemantic relations in bulgarian wordnet. *Intelligent Information Systems*, 16:359–369.
- Souheyl Mallat, Emna Hkiri, Mohsen Maraoui, and Mounir Zrigui. 2015a. Lexical network enrichment using association rules model. In *CICLing (1)*, pages 59–72.
- Souheyl Mallat, Emna Hkiri, Mohsen Maraoui, and Mounir Zrigui. 2015b. Semantic network formalism for knowledge representation: Towards consideration of contextual information. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 11(4):64–85.
- Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the romanian wordnet. In *LREC*, pages 2596–2601.
- Mohamed Achraf Ben Mohamed, Souheyl Mallat, Mohamed Amine Nahdi, and Mounir Zrigui. 2015. Exploring the potential of schemes in building nlp tools for arabic language. *International Arab Journal of Information Technology (IAJIT)*, 12(6).
- Karel Pala and Dana Hlaváčková. 2007. Derivational relations in czech wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 75–81. Association for Computational Linguistics.
- Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.
- Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.
- Abdulgabbar Saif, Mohd Juzaidin Ab Aziz, and Nazlia Omar. 2015. Mapping arabic wordnet synsets to wikipedia articles using monolingual and bilingual features. *Natural Language Engineering*, pages 1–39.
- Khaled Shaalan. 2010. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3):11–19.
- Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and semantic relations of croatian verbs. *Journal of Language Modelling*, (1):111–142.
- Ludovic Tanguy and Nabil Hathout. 2002. Webaffix: un outil d’acquisition morphologique dérivationnelle à partir du web. In *9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, page 254.
- Delphine Tribout, Anne-Laure Ligozat, and Delphine Bernhard. 2012. Constitution automatique d’une ressource morphologique: Verbagent. In *SHS Web of Conferences*, volume 1, pages 2517–2528. EDP Sciences.
- Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.
- W John Wilbur and Larry Smith. 2013. A study of the morpho-semantic relationship in medline. *The open information systems journal*, 6:1.
- Wajdi Zaghouani, Abdelati Hawwari, Mona Diab, Tim OGorman, and Ahmed Badran. 2016. Ampn: a semantic resource for arabic morphological patterns. *International Journal of Speech Technology*, 19(2):281–288.

Extending Wordnet to Geological Times

Henrique Muniz
EMAp/FGV, Brazil

Fabricio Chalub
IBM Research, Brazil

Alexandre Rademaker
IBM Research and EMap/FGV, Brazil

Valeria de Paiva
Nuance Communications, USA

Abstract

This paper describes work extending Princeton WordNet to the domain of geological texts, associated with the time periods of the geological eras of the Earth History. We intend this extension to be considered as an example for any other domain extension that we might want to pursue. To provide this extension, we first produce a textual version of Princeton WordNet. Then we map a fragment of the International Commission on Stratigraphy (ICS) ontologies to WordNet and create the appropriate new synsets. We check the extended ontology on a small corpus of sentences from Gas and Oil technical reports and realize that more work needs to be done, as we need new words, new senses and new compounds in our extended WordNet.

1 Introduction

Princeton WordNet (Fellbaum, 1998) works well as a dictionary and thesaurus for uses of English, as found, for instance, in newspapers and general knowledge texts, such as Wikipedia. Some attempts at extending it, for specific domains, such as Bioinformatics, Geography or Law (Smith and Fellbaum, 2004; Buscaldi and Rosso, 2008; Sagri et al., 2004; Lazari and Zarco-Tejada, 2012) have been made, but it is not clear how these extensions should be done, as different stakeholders will want to extend the basic dataset into different directions and with different tools and objectives.

The goal of our work is to describe a possible process of extension of the basic Princeton WordNet, for a restricted domain (Geological Time Periods) and to discuss issues, challenges and opportunities for other generic extensions.

One might wonder whether extensions of WordNet are really necessary. To this we say that

even for everyday language we are convinced that WordNet misses some necessary synsets. For example, there are several issues related to tokenization: words like *ping-pong*, *kickboxing*, *water-ski* and *fistfight* should appear with space, hyphens or not, in the respective synsets. They do not, which means that quite a bit of post-processing is necessary. It would be good to add many prefixes, suffixes and regular endings, which are perfectly understandable by humans, but not so much by machines, for instance *shirtless* and *localizer*, *focalizer* are not in WordNet. Also many verbs ending in *-ize*, *ise* or *ify* are not present in PWN, while being in Wiktionary, for instance *adjective*, *Africanize* or *incentify*, *girlify*.

We might also want to discuss **why** the kinds of extension of WordNet we describe in this work are useful. We offer two complementary explanations. First we want to use WordNet as a sort of “lightweight” ontology. As discussed in (Bobrow et al., 2007; de Paiva, 2011) while full comprehensive ontologies like SUMO (Niles and Pease, 2001) or Cyc (Matuszek et al., 2006) would be best for reasoning formally with the information in texts, these tend to be very ragged. They only have detailed information in the specific domains that people felt the need to complete them for. For daily words and everyday, commonsense, events they miss many concepts. Some shallow reasoning can be done on the basis of the information provided by lexical taxonomies and it seems best to cover all concepts, at the expense of being shallow than to have big gaping holes in the concepts covered.

The second explanation has to do with bootstrapping specific domain ontologies for specific domains. Even if we did have a fully comprehensive version of an open source ontology for commonsense, we would still need to complement it for specific domains like Geology and Paleontology. There are too many concepts specific to the

field that English fluent speakers have never heard of and that should not be part of a basic lexical resource for English. But these specific, say, geological concepts, need to be fitted within the taxonomic framework of a lexical knowledge base like WordNet, so that we can take advantage of the aforementioned framework. Some of us would like to use this aspect of WordNet expansion to construct Gas and Oil ontologies for supporting projects on information extraction on that industry.

In the (small) experiments we report in this paper, we discuss a very specific extension to a hopefully not very controversial domain. We want to add to WordNet specific information concerning geological time periods. The geologic time scale (GTS) is a system of chronological dating that relates geological strata of rocks (stratigraphy) to time as measured in years in Earth's history.

2 Geological Time Periods

The geologic time scale is used by geologists, paleontologists, and other Earth scientists to describe the timing and relationships of events in Earth's history. The table of geologic time spans set forth by the International Commission on Stratigraphy, which we take to be the official body for these scientists, is described in <http://www.stratigraphy.org>.

Both Wikipedia and Wiktionary have some information about geologic time periods that seem more complete than the information in WordNet. This is to be expected, as lexicographers tend to be conservative about the terms they add to their repository of the language. But to be useful, when analyzing scientific texts about geological descriptions, we need to take the newer and more specific information present in the Wiktionary and Wikipedia in consideration. This is a common pattern. For several specific domains Wikipedia and Wiktionary have more current and more specific information than WordNet. WordNet is concerned about not inflating the lexicon with terms that are clearly derived, when looked from a human perspective, (e.g. *coaly* is simply the adjective form of having to deal with *coal*) or easily compositional (like *basinward*— in the direction of a basin). Also new expressions consisting of prefixes and suffixes are not considered good material for WordNet, so WordNet has *aeon*, but not *super-aeon*.

We would like to devise and describe a process to extend WordNet for a specific domain, when we do have specific information about the domain in the shape of a well curated ontology for the domain, as well as high quality texts in the same. We use geological time periods and a small collection of papers in Petrology as a paradigmatic example of a domain specific extension.

2.1 Geological Time in WordNet

Princeton WordNet has only 28 synsets dedicated to the most well-known geological periods. All the information about geological periods is concentrated on synsets that are hyponyms of [15116283-n: *geological time, geologic time - the time of the physical formation and development of the earth (especially prior to human history)*]. Hyponyms include synsets for each of *aeon, geological era, geological period* and *epoch*. We discuss briefly the essentials on these synsets below.

The geologic time scale is organized in a hierarchical fashion. Eons (or aeons) are divided into eras. Eras contain periods that contain epochs, and finally epochs contain ages. The first three eons (Hadean, Archean, Proterozoic) are collectively referred as the Precambrian super-eon. The most recent eon, the Phanerozoic is subdivided into several periods. All of these five names of periods have their respective synsets in WordNet, but *super-eon* is not in any synset. However, geologists and paleontologists need more detail than the 28 synsets in PWN provide.

The International Commission on Stratigraphy, a sub-committee of the International Union of Geological Sciences, publishes regularly the International Chronostratigraphic Chart¹ as the current standard of the organization of the geologic timescale of the Earth. One can read about the development of the chart in (Cohen et al., 2013). As explained in that paper, geological time periods are not as well-established as one might expect. They say:

Most of the systems, series and stages were first defined from type-sections in Europe, the historical home of stratigraphy. Subsequent study of stratigraphical successions worldwide has led to a proliferation of regional units. These histor-

¹<http://www.stratigraphy.org/index.php/ics-chart-timescale>

ical units did allow Phanerozoic strata to be correlated and mapped worldwide. However, as it happened, most successive chronostratigraphic units are located in geographically separated type sections, which have more recently been shown to be separated by significant gaps or to overlap considerably. These problems, and the general lack of defined boundaries for historically established units, became serious hindrances to high-resolution correlation of geographically widespread stratigraphic successions.

A committee was tasked with producing a chart that solved the issues of conflicting and overlapping regional strata. We assume the chart and the new periods and boundaries represent the consensus between scientists working on this area. The chart mentioned above contains 176 names of geological periods. Of these only 28 are in WordNet and all but 40 are in Wiktionary. The last 11 are in Wikipedia, but not in WordNet or Wiktionary.

While the common noun *stratigraphy* is in PWN, [06118236-n: *stratigraphy - the branch of geology that studies the arrangement and succession of strata*], even the adjective *stratigraphic* is not in the database and neither is the compound *chronostratigraphic*. Presumably because these words are too specific and their meaning can be easily derived from the prefix *chronos*, meaning ‘time’ and the suffix denoting a pertainym adjective *-ic*. However, even the word *strata* (the irregular plural of *stratum*) used in the gloss, and presumably more primitive than *stratigraphy* (the study of strata) is not in WordNet, which signals clearly that PWN needs to be extended, if it is to deal with the needs of the area.

One reasonable way of extending a lexical resource in the direction of a specific field is to process a corpus of quality texts in this field and check for missing entries. This was part of our work for this experiment. But another avenue of expansion open to us, in this case, was to incorporate a domain-specific ontology created by the professionals of the area. We searched for experts and found the ISC ontology <http://resource.geosciml.org/def/voc/>, described in the next section.

We should note though that the new ontology is not a full solution to our problem. There are

many compounds and single words that acquire specific meanings within a field. Finding and creating synsets for these is also part of our challenge. Also, discovering when compounds are to be treated as multiword expressions, as opposed to compositional compounds, is a challenge, compounded by the use of abbreviations, specific to the field.

For instance, one of the main concepts of the area, the idea of a GSSP (Global Boundary Stratotype Section and Point ²), is usually called a *golden spike* in text. Anyone who is not from the field might think that a golden spike is just a compositional English compound. Seeing the expression by itself, without context, they might not know that the expression stands for “an internationally agreed upon reference point on a stratigraphic section which defines the lower boundary of a stage on the geologic time scale”, as explained.

We first discuss how to incorporate the information from an already structured ontology and then how to use corpora to improve our specific lexicon of geological time scales.

3 The ISC Ontology

The ISC ontology presents a view of the knowledge associated to the International Stratigraphic Chart. The ISC ontology contains many sub-ontologies, including the Geologic Timescale (GTS³) that would seem perfect for our uses.

In this ontology, *age*, *eon*, *epoch*, *era*, *period*, *sub-period*, and *super-eon* are sub-classes of *GeochronologicEra* (abbreviated as *GE*), which seems simply a different name for what is called ‘geological time’ in WordNet. However, there is no formally defined hierarchy between these concepts. Instead, greater emphasis is placed on the boundaries of the periods and only the approximate duration of the period is given in the chart. It is important to note that geologists qualify the units as “early”, “mid”, and “late” when referring to time, and “lower”, “middle”, and “upper” when referring to the corresponding rocks. For example, the lower Jurassic Series in chronostratigraphy corresponds to the early Jurassic Epoch in geochronology. The adjectives are

²https://en.wikipedia.org/wiki/Global_Boundary_Stratotype_Section_and_Point

³<http://resource.geosciml.org/ontology/timescale/gts.html>

capitalized when the subdivision is formally recognized, and lower case when not; thus “early Miocene” but “Early Jurassic”.

While the commission was created exactly to unify and organize the classification of both strata and geochronological periods, it appears that the work is both not finished and bound to disagreement. The above mentioned paper also says

[...] disagreement often arises, because type sections that are favoured for historical reasons may be abandoned, previously established boundary levels may be greatly changed, and in some instances historical units are replaced by different new ones.

Thus while the ontology might look very much a finished product, it seems that its contents are still subject to debate.

The boundaries between periods seem to be annotated using another ontology, the Temporal Hierarchical Ordinal Reference System model (THORS⁴), which is used to formally define the hierarchy between instances of GE. Fragments of the ISO19108:2002 standard (Geographic information – temporal schema) are also used to specify the temporal position of geochronologic boundaries.

The time interval of a GE is given in terms of its boundaries to other GEs via `thors:begin` and `thors:end`. Each boundary is a `GeochronologicBoundary` and it is temporally located via `iso19108:temporalPosition` which specifies a `iso19108:Coordinate` with a value, frame (e.g., “Ma”), and a positional uncertainty.

For example, the Maastrichtian period is defined by Wiktionary in <https://en.wiktionary.org/wiki/Maastrichtian> as “in the ICS geologic timescale, the latest age or upper stage of the Late Cretaceous epoch or Upper Cretaceous series, the Cretaceous period or system, and of the Mesozoic era or erathem”.

In the ISC ontology itself the definition is more complex. The Maastrichtian period (66–72.1 Million years) is defined using boundaries and frames (Figure 1).

⁴<http://resource.geosciml.org/ontology/timescale/thors.html>

```
Maastrichtian a GeochronologicEra ;
  rank Age ;
  begin BaseMaastrichtian ;
  end BaseCenozoic .
BaseMaastrichtian a GeochronologicBoundary ;
  temporalPosition BaseMaastrichtianTime .
BaseCenozoic a GeochronologicBoundary ;
  temporalPosition BaseCenozoicTime .
BaseMaastrichtianTime a Coordinate ;
  frame ma ;
  value "72.1" .
BaseCenozoicTime a Coordinate ;
  frame ma ;
  value "66" .
```

Figure 1: A fragment of the Maastrichtian period definition on ISC ontology

The boundary modeling should be sufficient for representing the hierarchical relationship between GEs, but ISC further defines a explicit set inclusion relationship between GEs via the `thors:member` property. Also, SKOS (Isaac and Summers, 2008) is also used to represent inclusion via `skos:narrower`, `skos:broader` along with their transitive versions, `skos:narrowerTransitive` and `skos:broaderTransitive`.

In any case a collection of 176 basic geologic period terms is easy to deal with, if the scientists are in agreement. However, we still have to deal with common nouns (e.g. *play*, *basin*, *cleats*) and compounds (e.g. *golden spike*), whose geological meanings are very different from their usual meanings. These need to be extracted from a geology corpus, similar to the one we describe in the next section.

4 A corpus of Geological Reports

The source documents for the our small experiment come from 155 randomly selected text passages relevant to petroleum systems extracted from a corpus of 1,298 publicly available English language geological reports, published by the United States Geological Survey (USGS), Geological Survey of Canada (GSC), and British Geological Survey (BGS).

The passages were segmented in 5,661 sentences (186,244 tokens) and parsed in the Universal Dependencies scheme by UDpipe (Straka and Straková, 2017) ⁵. UDpipe is a generic, off the shelf processing pipeline trained with the English corpus from the Universal Dependencies project

⁵<http://ufal.mff.cuni.cz/udpipe>

(Nivre et al., 2016). Using the model available, trained on newswire data, it does not do well on Named Entity recognition in our corpus. Our preliminary semantic pipeline looks up nouns, verbs, adjectives and adverbs in Princeton WordNet. Out of 8800 noun lemmas uncovered by UDpipe, more than half were not recognized as present in WordNet. Because the reports are describing real world geological work, the corpus is full of named entities, e.g. names of places, people and organizations that cause Named Entity Recognition to be such a hard task.

Some of these missing words are processing mistakes. For instance, the word ‘reservoirs’ was not correctly lematized to ‘reservoir’. A large proportion are named entities, people, places and organizations that WordNet is not supposed to list in any case. But a small proportion are really common words that WordNet should have, in our opinion. Finding these seems to be a positive side effect of trying to extend WordNet for specific domains.

Since our aim is not the processing of this corpus, but simply its use as a source of extra vocabulary for our extended WordNet, we decided to look at all tokens in the corpus with more than 10 occurrences, trying to decide whether they were Named Entities or not. And we assumed that the processing could be corrected, by hand, if need be. It is well-known that PWN lacks some important compounds and that the cut-off line for compounds to be lexicalized is a difficult one to decide on. Moreover, in this specific field, we do not know exactly when compounds are compositional or not. But a shallow processing of the text provides us with some 20K proper nouns, so almost 4 proper nouns per sentence. This means that NER is a very hard job, even assuming near perfect Geoname resources, which unfortunately we do not have.

5 Creating New Synsets

The language of ISC and its various ontologies is complex, and for a reason. They want to be precise, while trying to merge different standards. As we want to map all their precision into an extended version of Princeton WordNet we need a kind of a *domain specific language* (DSL) to describe new synsets. This language helps us not only to describe the new synsets we need, but also should help us localize these new synsets within the original WordNet structure.

The file format we decided to use is intended mainly for human consumption, even at the cost of a more complicated parsing routine. Redundancies are eliminated, for example there is no need to specify both sides of reflexive relations, such as hyponymy and hyperonymy. Artificial identities (ids) are avoided to make maintenance easy. Actual ids are based on the lexical units, following the ideas of the original lexicographer files for Princeton WordNet. Instead of using symbols such as @, !, etc. for relations, we use mnemonics such as `hyper` (hypernym) and `ant` (antonym). The goal is to make a standalone domain specific language – one that is usable without any accompanying integrated development environment (IDE) or other auxiliary program.

Synsets are defined by groups of lines, separated by a single empty line. Words of the synsets should have their spaces converted to underscores and repeated words in the same file should have suffixes to distinguish them, also following the original lexicographer files of PWN. For example the synset for *eon* will be written as

```
w: eon drf adj.pert:eonian
w: aeon drf adj.pert:aeonian
hyper: geological_time
g: the longest division of
geological time
```

where `drf` stands for ‘derived form’, `adj.pert` is the usual WordNet description of the pertainym adjective file and `g` stands for the ‘gloss’. Each word entry is essentially a *sense*. Links between senses are specified in the same line as the `w:` word, for example:

```
w: uptime ant downtime
```

means that ‘uptime’ and ‘downtime’ are antonyms. Semantic relations (i.e., links between synsets) are specified on lines of their own, such as the hypernym `hyper: geological_time` above.

The first word of a synset is used as its identifier. The lexicographer file filename should also be included to further disambiguate words, if necessary. This is usually the case when there are semantic links across synsets defined in different files. For example, the file `noun.location` contains the following excerpt for the synset “Brazil”:

```
w: Brazil drf adj.pert:Brazilian
hp: noun.object:South_America
```

To maintain compatibility with existing systems that already use PWN sense keys and synset ids we provide mappings between our sense ids and PWN. Similarly, mappings that link synsets and existing ontologies can also be defined.

The full set of PWN synsets extended with the nodes created for the geological time periods and the new concepts we deem necessary to understand our corpus could be called PWN_{GTS} for WordNet extended for the Geological Time Scale. In the next section we describe a toy application of the extension developed. In <http://github.com/own-pt/wordnet-dsl> we provide the PWN_{GTS} and the mappings from the new synsets to the ISC Ontology.

6 Using PWN_{GTS}

The following discussion showcases an example where a number of geochronologic entities may be referenced implicitly by the text. Consider the following sentence from our corpus:

In this chapter, the kinematic interpretation of the west Carbonate shear zone is placed in a regional context, with regard to intrusive and tectonic activity from 2740 to 2690 Ma ago.

Assuming that a parser correctly identifies the numerical range above as being 2740–2690 and the unit ‘Ma’ (for a million years), one can use our extended WordNet, creating a query to the ISC ontology that searches for entities that encompass this period of time. The SPARQL query used is in the appendix, note that such a query does not take into consideration the variance of the boundaries of time periods (modeled by the ontology). We opted to omit this feature to keep the SPARQL code simple. This natural query is not enough to uniquely disambiguate the appropriate instance that is referenced above, since the query returns three ISC entries: the Neoproterozoic era (2500–2800 Ma), the Archean eon (2500–4000 Ma), and also the Precambrian super-eon (541–4567 Ma).

While this toy example shows one possible use we envisage for very restricted forms of extension of the basic English WordNet, the larger question of evaluating such extensions beckons. From our preliminary work we can see some possibilities, which we discuss next.

7 Evaluating Extensions

It is clear that different kinds of text and different content domains play a big role in the vocabulary that lexical resources are expected to cope with. This is clear for specific content domains, such as BioInformatics, where changes are recent and newer vocabulary is being created at impressive speeds. But even for domains, such as Geology, where one might have expected the main vocabulary to have been established by the end of the 19th century, things are not as well settled as expected.

Certainly there is a need for more (open source, downloadable) online glossaries, apart from the (small) Wikipedia one⁶, the OpenLearn project⁷ and the one from USGS⁸ that has not been updated since the mid 2000’s. But it seems that the proprietary ones still have the upper hand. The American Geosciences Institute (AGI) offers their fifth revised edition of the Glossary of Geology (Neuen-dorf, 2005) as a book and as paid subscribing content online. They say that their reference tool contains nearly 40,000 entries, including 3,600 new terms and nearly 13,000 entries with revised definitions from the previous edition. None of the open source glossaries we found has as many entries as that.

One way of measuring how much we can do with the open resources online is to measure how much of the informational contents of technical reports can be gleaned by a impoverished NLP pipeline that builds bag-of-concept semantics from the sentences of the chosen corpus. In a previous experiment we have computed this kind of bag-of-concepts semantics for sentences of the corpus SICK (Marelli et al., 2014). The corpus SICK is much easier to deal with, as it was engineered to not have any named entities at all. If we had no named entities in our geological reports, we could produce concepts from SUMO (Niles and Pease, 2001) using a bare bones pipeline that transforms sentences into universal dependencies (using UDPipe), dependencies into WordNet concepts or synsets (using, say, Freeling/UKB (Agirre and Soroa, 2009) for disambiguation) and WordNet synsets into SUMO con-

⁶https://en.wikipedia.org/wiki/Glossary_of_geology

⁷<http://www.openlearn.org/openlearn/science-maths-technology/science/geology/geological-glossary>

⁸<https://geomaps.wr.usgs.gov/parks/misc/glossarya.html>

cepts (using the SUMO mappings). An example of a processed sentence is displayed in Figure 2.

The idea here is not to produce knowledge representations of the meanings of the sentences, but simply to list the expressions for which we do not have a concept. For these ‘empty concepts’ we need either geographical information or new synsets, as they correspond to either new content words (that never appeared in WordNet before, like e.g. *vitrinite* or *stratigraphic*), or new compounds (e.g. *pre-Mississippian*, *antiform* or *sub-basin*, *golden spike*) or new senses of words already in WordNet (e.g. *cleats*, *play*, *sequence*, which have completely different meanings in Geology from their usual ones). However we need to find a way of coping programmatically with named entities, for this baseline calculation to work.

Given the hardness of the NER problems in this particular kind of texts, we resorted to different open systems (with different training data and heuristics, e.g. OpenNLP⁹ and Freeling (Padro and Stanilovsky, 2012)) to try to extract most of the named entities. In this corpus apart from locations, people and entities we have many *Geological Formations*, which span counties and even states’ lines. To help debug our processing, we are experimenting with interfaces that allow linguists, computer scientists and geologists to communicate more easily <http://wnpt.brlcloud.com/demo>. We hope to improve, using subject matter experts, the number of new synsets and new senses. The manual ‘ensemble’ effort to recognize named entities we produced for this small corpus, needs to be streamlined in the future, for the work in extending other domains.

8 Conclusions

This preliminary work discusses extensions of Princeton WordNet for specific content domains. The case we considered is the well delimited domain of geological time periods. We expected it to be less controversial and to have a more established vocabulary than it turned out to have. However, we stand by our initial suggestion that specific domains require specific extensions. That these specific extensions need to be built as much as possible from open source resources, in a collaborative fashion, using as much as possible associated ontologies produced by the subject mat-

ter experts. However, a useful way to augment the specific knowledge required is to shallow process scientific texts on the specific subject (we used gas and oil technical reports) and try to extract more lexical information from them. Our small experiment with geological reports indicate that a more robust mapping of named entities is required before we can evaluate the usefulness of our new Geological Time Scale WordNet. We are working on a tool that would pre-annotate some of these geonamed entities and would facilitate the correction of the mistaken annotations.

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.
- Daniel G Bobrow, Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, and Valeria de Paiva. 2007. PARCs bridge and question answering system. *Grammar Engineering Across Frameworks*, pages 46–66.
- Davide Buscaldi and Paolo Rosso. 2008. Using geowordnet for geographical information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pages 863–866.
- K.M. Cohen, S.C. Finney, P.L. Gibbard, and J-X Fan. 2013. The ICS International Chronostratigraphic Chart. *Episodes*, 36(3):199–204.
- Valeria de Paiva. 2011. Bridges from language to logic: Concepts, contexts and ontologies. *Electronic Notes in Theoretical Computer Science*, 269:83–94.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Antoine Isaac and Ed Summers. 2008. Skos primer. Technical report, W3C. latest version available at <http://www.w3.org/TR/skos-primer>.
- Antonio Lazari and M^a Ángeles Zarco-Tejada. 2012. Jurwordnet and framenet approaches to meaning representation: a legal case study. In *Semantic Processing of Legal Texts (SPLeT-2012) Workshop Programme*, page 21.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

⁹<https://opennlp.apache.org/>

+#	text	=	Paper	and	scissors	both	cut												
+1	Paper	paper	NOUN	NN	_	5	nsubj	_	NN 06267145-n Newspaper=										
+2	and	and	CONJ	CC	_	1	cc	_	CC ? ?										
+3	scissors	scissor	NOUN	NNS	_	1	conj	_	NNS ? ?										
+4	both	both	DET	DT	_	1	dep	_	DT ? ?										
+5	cut	cut	VERB	VBD	_	0	ROOT	_	NN 00352331-n Process+										

Figure 2: Bag-of-concepts

Cynthia Matuszek, John Cabral, Michael J Witbrock, and John DeOliveira. 2006. An introduction to the syntax and content of cyc. In *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49.

Klaus KE Neuendorf. 2005. *Glossary of Geology*. Springer Science & Business Media.

Ian Niles and Adam Pease. 2001. Toward a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Lluís Padro and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. 2004. Jur-WordNet. In Petr Sojka, Karel Pala, Pavel Smrz, Christiane Fellbaum, and Piek Vossen, editors, *Global Wordnet Conference*.

Barry Smith and Christiane Fellbaum. 2004. Medical wordnet: A new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING '04*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

A Example of Query

In the query below, notice if we remove the restriction on `isc:rank Age` we get multiple hits (Maaast. [age], Cret. [period], Upper Cret.

[epoch]) since the range 67–70 is included on all of them.

```

select ?era ?rank ?vbegin ?vend
{
  ?era gts:rank ?rank ;
    thors:begin ?tb;
    thors:end ?te .

  ?tb ts:temporalPosition ?begin;
  ?te ts:temporalPosition ?end .

  ?begin ts:frame age:ma ;
    ts:value ?vbegin .

  ?end ts:frame age:ma ;
    ts:value ?vend .

  bind (2690 as ?a)
  bind (2740 as ?b)

  filter ((?a <= ?vbegin &&
    ?a >= ?vend) ||
    (?b <= ?vbegin &&
    ?b >= ?vend) )
}

```

Towards Emotive Annotation in plWordNet 4.0

Monika Zaśko-Zielińska¹, Maciej Piasecki²

¹,University of Wrocław, Wrocław, Poland

²G4.19 Research Group, Department of Computational Intelligence

Wrocław University of Technology, Wrocław, Poland

maciej.piasecki@pwr.edu.pl, monik@uni.wroc.pl

Abstract

The paper presents an approach to building a very large emotive lexicon for Polish based on plWordNet. An expanded annotation model is discussed, in which lexical units (word senses) are annotated with basic emotions, fundamental human values and sentiment polarisation. The annotation process is performed manually in the 2+1 scheme by pairs of linguists and psychologists. Guidelines referring to the usage in corpora, substitution tests as well linguistic properties of lexical units (e.g. derivational associations) are discussed. Application of the model in a substantial extension of the emotive annotation of plWordNet is presented. The achieved high inter-annotator agreement shows that with relatively small workload a promising emotive resource can be created.

1 Introduction

Since plWordNet (Maziarz et al., 2016) achieved good coverage with the version 2.0 the number of its users and applications has been quickly growing. Many users declared sentiment analysis, as their intended use of plWordNet, contrary to the lack of sentiment information in it. In response, a pilot project on emotive annotation of a selected subset of senses in plWordNet was conducted (Zaśko-Zielińska et al., 2015) which next resulted with plWordNet 2.3 emo including *emotive annotation* for more than 31,000 lexical units (word senses). This prototype emotive annotation showed its usefulness in lexicon-based sentiment analysis, but its coverage was limited and selective (i.e. around 10% of noun senses and 25% of adjective senses of plWordNet 3.0 emo).

Our goal is to develop an improved and expanded model of emotive annotation for a word-

net, and also an expanded version of the manual annotation procedures. In addition we will also present application of the model in a substantial extension of the emotive annotation of plWordNet.

2 Emotions in Wordnets

Several sentiment lexicons are available for English, but hardly any for most other languages. Chen and Skiena (2014) found 12 publicly available sentiment lexicons for 5 languages; there were none for Polish. Some sentiment lexicons were built upon Princeton WordNet (PWN), a natural starting point because of its comprehensive coverage and its numerous applications. The lexicons not based on PWN consider lemmas rather than lexical meanings or concepts.

WordNet-Affect is a selection of synsets very likely to represent “affective concepts” (Strapparava and Valitutti, 2004). A small core of 1903 lemmas was selected and described manually with “affective labels”. Next, a set of rules based on wordnet relation semantics drove the transfer of the sentiment description onto the synsets connected to the core by wordnet relations. This produced 2874 synsets and 4787 lemmas.

SentiWordNet (Esuli and Sebastiani, 2006) annotates a synset with three values from the interval $\langle 0, 1 \rangle$. They describe “how objective, positive, and negative the terms contained in the synset are”. About 10% of the adjectives were manually annotated, each by 3-5 annotators (Baccianella et al., 2010). In SentiWordNet 3.0, the automated annotation process starts with all the synsets which include 7 “paradigmatically positive” and 7 “paradigmatically negative” lemmas.¹ In the end, SentiWordNet 3.0 added automatic sentiment annotation to all of PWN 3.0.

¹good, nice, excellent, positive, fortunate, correct, superior; bad, nasty, poor, negative, unfortunate, wrong, inferior (Turney and Littman, 2003)

SentiSense (Carrillo de Albornoz et al., 2012) is also a concept-based affective lexicon, with emotion categories assigned to PWN synsets. The initial list of 20 categories, a sum of several sets including WordNet-Affect, was reduced to 14 after some work with annotators. The authors write: “the manual labelling techniques generate resources with very low coverage but very high precision”, but note that such precision can be only achieved for specific domains. The construction of SentiSense began with a manual annotation of only 1200 synsets with 14 emotions. Annotation was transferred onto other synsets using wordnet relations. The authors’ visualisation and editing tools, designed to allow relatively easy expansion and adaptation, did not add much to the resource, so every user must enlarge it further to make it really applicable.

To sum up, a wordnet may be a good starting point for the construction of a sentiment lexicon: annotation can be done at the level of lexical meanings (concepts) or lemmas. PWN appears to be a good choice due to its sense-based model and large coverage. All large wordnet-based sentiment lexicons have been built by giving very limited manual annotation to algorithms for automated expansion onto other synsets. This, however, seems to have to result in lower precision, as noted, *e.g.*, by Poria et al. (2012): “Currently available lexical resources for opinion polarity and affect recognition such as SentiWordNet (Esuli and Sebastiani, 2006) or WordNet-Affect are known to be rather noisy and limited.”

No large wordnets are available for most languages other than English. Many sentiment lexicons were created by translating sentiment-annotated PWN, *e.g.*, Bengali WordNet-Affect (Das and Bandyopadhyay, 2010), Japanese WordNet-Affect (Torii et al., 2011) and Chinese Emotion Lexicon (Xu et al., 2013). It is not clear how well annotations of that kind can be transferred across the language barrier. Moreover, as we discuss it in section 3, plWordNet’s model differs slightly from that of PWN.

Crowdsourcing has also been used to develop sentiment lexicons (Mohammad and Turney, 2013). It *can* outdo automated annotation (or automatic expansion of a manually annotated part), but the consistency of the result is low compared to manual description by trained annotators.

Unlike most of the existing methods, our aim

is a manual annotation of a substantial part of plWordNet by a team of linguists and psychologists. The manually annotated part – several times larger than other known manually created sentiment lexicons – can be an important resource on its own. It can also be a solid basis for the development of automated sentiment annotation methods for more lexical material in a wordnet. We have adopted a rich annotation model in which sentiment polarity description is combined with emotion categories.

3 Annotation Model

For the sake of compatibility with plWordNet 2.3 emo, the main assumptions and annotation scheme have been preserved without significant changes, see Sec. 3.1,3.2. However, we plan to encompass by emotive annotation all PoS in plWordNet (i.e. nouns, adjectives, verbs and adverbs) and expand it very substantially by 100,000 annotated lexical units. It goes beyond typical sentiment polarity encoding and includes: sentiment polarity, basic emotions and fundamental values. On the basis of the analysis of the results of (Zaśko-Zielińska et al., 2015) we modified the annotation guidelines for nouns and adjectives to improve annotation quality, see Sec. 4.

3.1 Main Assumptions

plWordNet has been constructed on the basis of the corpus-based wordnet development method (Piasecki et al., 2009), according to which *lexical units* (henceforth LUs) are basic building blocks of the wordnet, *e.g.* use examples for LUs can be collected and analysed in corpora, but not for synsets, linguistic lexico-semantic relations are defined for LUs, and linguistic substitution tests can be applied to LUs. Synsets and their relations are derived in plWordNet from the structure of relations linking LUs, cf (Maziarz et al., 2013). Thus, emotive annotation is naturally done on level of LUs and includes LU use examples.

The analysis of the results of (Zaśko-Zielińska et al., 2015), i.e. the model, annotated LUs and the first feedback from the applications, has brought us to the revision of that model. However, first we agree with (Zaśko-Zielińska et al., 2015), that emotive annotation is focused on those emotive properties of LUs that are revealed in situation in which the given LU is maximally detached from the interpretation context, or, from the other point

of view, the description requires as little knowledge about the context as possible. This assumption coincides with the idea of dictionary and plWordNet is undoubtedly one.

As in (Zaśko-Zielińska et al., 2015), context-independent emotive characterisation of an LU is obtained by comparing its authentic uses found in the text corpora. During the annotation process features that are common to the LU usages are isolated, while the occasional ones discarded. Validating the obtained results we search for polarisation stability that should be repeated in the collocations of the given LU. However, contrary to (Zaśko-Zielińska et al., 2015), we claim that LU's emotive polarisation determined in this way does not provide information about emotive attitudes of the speaker. We can only read what is expressed in the examples. This is a difference similar to the one between the intent and the statement function, cf (Bartmiński and Niebrzegowska-Bartmińska, 2009). Thus, while still preserving the fundamental premise of aiming at the detection of the LU characteristics outside of the context, we assume that it is not the knowledge of the speaker's emotive attitude that is being described in annotation, but the emotive characteristics that is common to the analysed expressions and salient to the recipient, i.e. an annotator. The process of averaging across different LU use examples in search for emotive features independent of the contexts, or common to different contexts, is amplified by searching for agreement of the annotators applying independently the same annotation procedure.

In (Zaśko-Zielińska et al., 2015) fundamental human values (Puzynina, 1992), see Sec. 3.2, have been also included into the emotive annotation. This is a unique solution in comparison to other wordnet-based emotive annotations. There are also important reasons to follow and expand it in our work. While emotional assessment is always associated with a kind of evaluation in the meaning of LUs, it is very often neglected that some LUs lack emotional aspect, but still are associated with a form of evaluation. Annotating of LUs with fundamental human values expressing evaluation is particularly important for the analysis of product reviews or brands (opinion mining) It helps to extend sentiment polarisation also to multiword LUs that are quite numerous in plWordNet (>54k) and many of them belong to terminology. This is especially valuable because general dictionar-

ies usually omit this type of LUs. They are often treated as a group of vocabulary without polarisation. However, it is worth to notice that in works on opinion mining in Polish texts from the economics point of view, speaker's attitude is an important factor in the analysis of product reviews This is partially possible, but does not take into account the impact of the speaker's error on the quality of the message or the beliefs of the recipient, which, as contextual information, is inherent in receiving the message (Lula et al., 2016).

3.2 Annotation Scheme

Following (Zaśko-Zielińska et al., 2015) the main distinction is between *neutrality vs polarity* of LUs. Polarised LUs are assigned the *intensity* of the sentiment polarisation, *basic emotions* and *fundamental human values*. The latter two help to determine the sentiment polarity and its intensity expressed in the 5 grade scale: *strong* or *weak vs negative* and *positive*. Annotator decisions are supported by use examples that must be included in the annotations.

Due to the compatibility (Zaśko-Zielińska et al., 2015) with other wordnet-based annotations, we use the set of eight basic emotions recognised by Plutchik (Plutchik, 1980) whose wheel shows how basic emotions can contribute to secondary emotions. It contains Ekman's six basic emotions (Ekman, 1992): *joy*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, complemented by Plutchik's *trust* and *anticipation*. As a result, negative emotions do not prevail in the set, cf (Mohammad and Turney, 2013). One LU can be assigned more than one emotion and, as a result, complex emotions can be represented by using the same eight-element set. Plutchik states (Plutchik, 1980) that his basic emotions are primary, that is, they appear first in ontogenesis and phylogenesis. So we assume that they are repetitive for all language users regardless of their age and development. Ekman, on the other hand, refers not to evolution but to intercultural nature and claims that facial expressions and underlying emotions are common to different cultures (Ekman and Friesen, 1971).

As in (Zaśko-Zielińska et al., 2015), we use the set of fundamental human values postulated by Puzynina (Puzynina, 1992), later followed in many works on lexicology and derivation. Thus we assume that the emotive state of the speaker is linked to the *evaluative attitude*.

Although, the evaluation can also be independent of emotions (Waszakowa, 1991). The set of the fundamental human values encompasses: *użyteczność* ‘utility’, *dobro drugiego człowieka* ‘another’s good’, *prawda* ‘truth’, *wiedza* ‘knowledge’, *piękno* ‘beauty’, *szczęście* ‘happiness’ (all of them positive), *nieużyteczność* ‘futility’, *krzywda* ‘harm’, *niewiedza* ‘ignorance’, *błąd* ‘error’, *brzydota* ‘ugliness’, *nieszczęście* ‘misfortune’ (all negative) (Puzynina, 1992).

3.3 Examples of Annotation

Below we present examples of complete emotive annotations for three LUs (where A1 and A2 means, respectively the first and the second annotation added, **BE** – basic emotions, **FHV** – fundamental human values, **SP** – sentiment polarity, and **Exam** – usage example):

dziad 1 gloss: “stary mężczyzna” ‘an old man’
 (**Annot.:**A1, **BE:** {*złość* ‘anger’, *wstręt* ‘disgust’}, **FHV:**{*nieużyteczność* ‘futility’, *niewiedza* ‘ignorance’}, **SP:**–s

Exam: “Stary dziad nie powinien podrywać młodych dziewczyn.”

‘An old geezer should not pick up young girls.’
 (**Annot.:**A2, **BE:** {*wstręt* ‘disgust’}, **FHV:**{*nieużyteczność* ‘futility’, *brzydota* ‘ugliness’}, **SP:**–w

Exam: “Jakiś dziad się dosiadł do naszego przedziału i wyciągnął śmierdzące kanapki z jajkiem.” ‘An old geezer joined our compartment and took out stinky egg sandwiches.’

(**Annot.:**A3, **BE:** {*wstręt* ‘disgust’}, **FHV:**{*nieużyteczność* ‘futility’, *brzydota* ‘ugliness’}, **SP:**–s

Exam: “Kilkanaście lat minęło i zrobił się z niego stary dziad.”

‘Several years have passed and he has become an old geezer’)

szalbierski 2 ‘deceitful’

(**Annot.:**A1, **BE:** {*smutek* ‘sadness’, *złość* ‘anger’}, **FHV:** {*krzywda* ‘harm’, *błąd* ‘error’ }, **SP:**–s,

Exam: “Nie chciałam brać udziału w tym szalbierskim planie, którego pomyslność zależała od stopnia naiwności nieświadomych klientów.”

‘I did not want to take part in this deceitful plan, whose success depended on the level of naiveness of the unaware clients.’)

(**A2**, **BE:** {*smutek* ‘sadness’, *złość* ‘anger’}, **FHV:** {*krzywda* ‘harm’, *błąd* ‘error’}, **SP:**–s,

Exam: “Mam szalbierski pomysł, który pomoże nam naciągnąć paru idiotów.”

‘I have a deceitful idea which might help us to con a couple of idiots.’)

wytrzymały 2 ‘enduring’

(**Annot.:**A1, **BE:**{*zaufanie* ‘trust’}, **FHV:**{*użyteczność* ‘utility’}, **SP:**+w,

Exam: “Wykonaliśmy podłogę z wytrzymałych paneli, dzięki temu od lat prezentuje się wspaniale.”

‘We made the floor from enduring panels, that is why it has been looking splendid for years’)

(**Annot.:**A2: **BE:**{*zaufanie* ‘trust’}, **FHV:**{*użyteczność* ‘utility’}, **SP:**+w

Exam: “Postanowiłem nie oszczędzać i kupić plecak z wytrzymałego materiału — przynajmniej wiem, że nie rozleci mi się po roku.”

‘I decided to not economize and to buy a backpack made of enduring material — at least I know that it will not tear apart after one year.’)

4 Annotation Procedure

The annotation is performed² by: linguists and psychologists, where each LUs is annotated by a mixed pair: one psychologist and one linguist. The annotators must follow guidelines that consist of a core common to all PoSs and detailed guidelines dedicated to each PoS. The work of annotators is coordinated and verified by a supervisor, who can access all annotations and solve disagreements³ by adding the final annotation.

The common core is similar to the procedure in (Zaśko-Zielińska et al., 2015):

Step 1 identification of LUs with *neutral* and *non-neutral* sentiment polarity;

Step 2 assignment of the basic emotions and fundamental human values;

Step 3 recognition of the LU polarity direction: negative or positive, but also *ambiguous*, if the collected use examples show both behaviours;

Step 4 assignment of sentiment polarity intensity;

Step 5 illustration of the assigned annotation by sentences representing use examples: at least

² Six persons were working on the results reported here: four linguists and two psychologists.

³ As it is presented in Sec. 5 disagreements in sentiment polarity are quite infrequent.

one sentence in the case of positive and negative LUs, and at least two example sentences for ambiguous LUs.

Each step is associated with several linguistic tests, including substitution tests and requires consulting corpus data. The detailed specification of the subsequent steps is dependant on a particular PoS. In the case of nouns see (Zaśko-Zielińska et al., 2015), the specification for adjectives proposed by us is presented in Sec. 4.2.

Annotators can returned from the later steps to the previous ones. We could observe that, e.g., assignment of fundamental human values or basic emotions can be helpful in verifying the polarity of the given LU.

For the annotation process, we use Wordnet-Loom – a wordnet editing system (Piasecki et al., 2013) – which has been extended by additional windows and database tables (to eliminate errors in the annotation representation), as well as a mechanism that separates work of individual annotators. They do not see annotation decisions of other annotators and they do not know who is the second annotator of the given LU. Moreover, annotators are rotated in the pairs in order to minimise a potential bias. This strict separation of annotators is a significant difference in relation to (Zaśko-Zielińska et al., 2015), where the second annotator was told not to take a look into the decision of the first annotator before having made his own one, but he could see it and could change his own one later. The second could report a possible error of the first one in the pilot project, but we decided to resign from this possibility and to separate them strictly. The inter-annotator agreement is on a high level, but inevitably lower than reported in (Zaśko-Zielińska et al., 2015), see Sec.5. However, we sometimes observed a tendency to too prompt classification of a LU as a neutral one. If such a decision is taken without a detailed analysis, then the annotation process is actually discontinued after the first step and any change of mind of the given annotator later along the process is impossible. To amend this potential problem we paid more attention to the detailed guidelines for Step 1, as well as to the training of annotators and verification of their work.

4.1 Nouns

As annotation of nouns was not completed in the pilot project, we also started with nouns. We used

guidelines from the pilot project. Only minor details were fine-tuned, e.g. we added a test for distinguishing diminutive formant function (Siudzńska, 2016). Formants appropriate for diminutives are not always connected with sentiment polarity. The test involves attaching three groups of adjuncts to the nouns:

- A adjuncts indicating size (e.g. expressing senses: *small, fine, young, ...*);
- B adjuncts showing positive emotions towards the person represented by the derivative or emotional bond with a person (e.g., senses: *my, our, good, loved, nice, sympathetic, unusual, modest, poor, tiny, thin, mischievous, miserable, etc.*);
- C adjuncts indicating negative emotions (e.g., *clumsy, unfulfilled, stupid, backward, lying, poor*); in this way, the sender may indicate the immaturity, helplessness of the person called by the derivative, and also show pity, irony, disregard and contempt.

Test A covers LUs like: *minutka* ‘≈a small minute’, *chwilunia* ‘≈a tiny moment’, that are related to size.

4.2 Adjectives

Annotation of adjectives started at the end of the pilot project on limited material, so the guidelines for adjectives required more substantial changes.

First annotators are reminded that adjective LUs in plWordNet have mostly more fine grained meanings than those in Polish dictionaries. Thus, all the time the annotator has to check whether he is working on the same and appropriate LU, not, e.g. deviating accidentally to another sense of the LU lemma. For this purpose annotators should check and use collocations as a tool for prompting a particular meaning. For instance *ciężki* ‘≈ heavy’ corresponds to 23 LUs, that can be distinguished (mentally or in the corpus) by different collocations, e.g.: *heavy 1* – ‘weighs a lot’ (heavy bag), *heavy 2* – ‘sluggish, slow’ (heavy steps); *heavy 8* – ‘bulky, overwhelming’ (heavy curtains), *heavy 9* – ‘thick, not transparent’ (heavy air), *heavy 12* – ‘sad’ (heavy film), *heavy 14* – ‘difficult to bear’ (heavy silence), *heavy 15* – ‘heavy with fatigue’ (heavy eyelids), *heavy 18* – ‘intense, expressive’ (heavy wine), *heavy 19* – ‘ponderous’ (feels heavy); *heavy 22* – ‘with great power

(heavy artillery), or *heavy* 23 – ‘strong, aggressive’ (heavy sound).

Step 1 Neutrality test for adjectives is related to the wordnet structure of derivational relations for adjectives, non-derived adjectives are analysed according to the noun procedure. Adjectives derived from adjectives can be skipped in **Step 1**. The rest of derived adjectives are recognised as non-neutral:

- adjectives from polarised nouns: *domowa atmosfera* ‘home atmosphere’, derived from *dom* ‘from home (as a group of people)’ in opposition to the neutral *domowy strój* ‘a casual outfit’ where *domowy* is derived from ‘home (place), ≈ ‘somebody’s flat’;
- adjectives derived from verbs, called *dispositional*, including subtypes: *potential* – expressing potentia to do something, e.g. *powtarzalny* ‘repeatable’, *habitual* emphasising that something is permanent and in large amounts, e.g. *krzykliwy* ‘≈ noisy, vociferous’ in *krzykliwe dziecko* ‘a noisy child’, *quantificational* signalling large amount or quantity, e.g. *wytrzymały* ‘hardy, inured, hardened’ in *wytrzymały człowiek* ‘a hardened man’, and positively *evaluating*, e.g. *bitny* ‘valiant’ in *bitny żołnierz* ‘a valiant soldier’.

Step 2 Assignment of emotions and values: adjectives derived from verbs by the suffix *-alny* (meaning ‘to be able to’, ‘it is possible to’) form a very characteristic group of LUs. They are not connected with emotions, but they are related to the fundamental values: utility, futility, e.g. *zmywalny* ‘such that, can be removed by washing’ in *tataż zmywalny* ‘a tattoo that can be washed out’, *egzekwowny* ‘such that can be enforced’.

Step 3 Marking LUs as negative, positive or ambiguous: this step requires especially careful identification of meanings. In order to recognise polarity we perform tests: a *congruence test*, a *discord test*, a *test of collocation* and a *test of dictionary definitions*. The way they are formulated and applied is similar to the corresponding test for nouns, see (Zaśko-Zielińska et al., 2015). However, more attention should be sometimes paid to affixes, whose semantic transparency in adjective derivatives seems to be weaker.

The congruence test not only allows to detect the LU polarity, but also helps in creating exam-

ple sentences in **Step 5** that confirm the polarity recognised earlier, e.g. for *tęskny* ‘wistful’:

positive: *Upajaliśmy się tym tęsknym, nastrojowym widokiem.*

‘We were intoxicated by this **wistful** and romantic view.’

negative: *Nie mogłam już dłużej wytrzymać tego zawodzenia i jego tęsknych pieśni.*

‘I could not bear this crooning and his **wistful** songs any longer.’

The presence of the same LU in the two opposing contexts reveals its ambiguous emotive character. The occurrence of suffixes: *-usieńki*, *-uteńki*, *-eńki* does not determine the polarisation of LUs, because it also depends on the derivation basis. Although these suffixes express a positive polarisation (Grzegorzczkowska et al., 1998), the combination with the derivation basis, which can be polarised negatively, only weakens the marking, for example: *chudzieńki* ‘≈ very thin and weak’, *pijaniusieńki* ‘≈ completely drunk, not controlling himself’.

The discord test is used to correct linguistic awareness, which is primarily focused on negative polarisation: only antonyms, e.g. *clean – dirty* shows that both elements are polarised in this pair. Often only the collocation test allows you to capture the ambiguity of the polarisation for example: for *pedantic a pedantic order vs morbidly pedantic*.

Step 4 Assignment intensity of sentiment polarity: annotators are reminded that grade forms of adjectives do not inform about the sentiment polarity intensity of the derivational basis, but they show comparison between objects or phenomena; e.g., the suffix derivative *-utki* which expresses that the described feature is not at its maximum, in the lower part of a scale, and there may be something that is even smaller than *malutki* ‘≈very small’. In comparison to it, LUs with *-uteńki* ‘≈tiny’, *-usieńki* ‘≈very tiny’ may be a cause of doubt, as their suffixes signals that some feature value is even smaller. In resolving this problem one has to remember two aspects of such a derivation process: semantic and pragmatic. Although LUs *mokrzuteńki* ‘≈completely wet’, *mokrzusieńki* ‘≈completely wet’ can be interpreted as representing some extreme values of the feature, this is rather semantic information, and the emotive aspect of this LUs is be maximised. (Bogusławski, 1991) argues that the func-

PoS	# Comp	# Sing	-s	-w	n	+w	+s	amb
N	25,919	18,574	16.62	14.64	51.59	6.05	4.23	6.87
Adj	14,817	5,392	14.87	22.59	31.39	15.03	7.50	8.62
All	40,773	24,002	15.89	17.95	43.18	9.79	5.59	7.60

Table 1: Sentiment polarity annotation of plWordNet 4.0 in progress (Comp – completed, Sing – one annotator only so far); -s, -w, n, +w, +s, amb (negative strong/weak, neutral, positive weak/strong, ambiguous) are shown in percentage points.

tion of these affixes is similar to inflection, i.e., it corresponds to grade of adjectives.

5 Intermediate Results

During the pilot project more than 31,000 LUs (19,625 noun LUs and 11,573 adjective LUs) were described in plWordNet 3.0 emo by emotive annotation (Zaśko-Zielińska et al., 2015). From that point we started the annotation process aiming at its expansion by complete emotive annotations (2+1) for around 100k more LUs. Annotations done in the pilot project including decisions of only one annotator had to be completed.

We started adding emotive annotation from noun LUs with focus on hypernymic branches that are likely to include LUs with polarised sentiment. In addition we try to distribute manual annotations across the network of synsets in such a way that it will be possible to apply an algorithm for automated spreading annotations to the rest of LUs.

The statistics describing the current state of the work are presented in Tab. 1. Only LUs annotated by two annotators are counted as completed. This number includes also completed annotations for LUs processed during the pilot project. As annotators are mixed in pairs and subsets of LUs are assigned to them in diversified ways, a large number of LUs have received so far only one annotation. As it was also the case in the pilot project, more than half of the noun LUs are annotated as neutral. However, only $\approx 30\%$ of adjective LUs are neutral contrary to almost 60% in plWordNet 3.0 emo. This difference can be caused by a much broader coverage of noun LUs, while adjective LUs were selected by (Zaśko-Zielińska et al., 2015) in a slightly accidental way (there was an ongoing plWordNet expansion work on that time).

As our annotators work completely independently, we could measure the inter-annotator agreement (IAA) with respect to the sentiment polarity using the Cohen’s Kappa measure (Cohen, 1960), see Tab. 2. Due to the large number of annotators, and simplifying a little bit, we present the

PoS	All	-s	-w	n	+w	+s	amb
All	0.78	0.77	0.78	0.82	0.74	0.73	0.65
Mrk.	0.84	0.80	0.84	–	0.89	0.80	0.86

Table 2: Inter-annotator agreement (IAA), measured in Cohen’s κ , for different types of sentiment polarity: -s, -w, n, +w, +s, amb (negative strong/weak, neutral, positive weak/strong, ambiguous). *All* describes agreement for all decisions, *Mrk* presents estimated IAA value for marked LUs only.

agreement between the first and the second decision registered in the system for LUs. LUs with at least one annotation from the pilot project were excluded from this analysis. The observed IAA values, both, 0.78 for all decisions and around 0.75 for different sentiment polarity values, are very high. The value for the neutral polarity is a value for the decision: polarised vs non-polarised in fact. It can show that the annotators are quite confident about the neutrality of the LUs, but also it can be biased by the fact that describing a LU as a neutral can be easier than by other values. This issue needs further investigation.

As the neutral annotations dominate (almost half of all decisions), we have calculated an estimated IAA value for the marked LUs only by simply taking into account LUs for which any annotator did not proposed the neutral value. The obtained values are much higher than for all decisions, so we can conclude that neutral values do not increase artificially the general IAA.

Negative sentiment polarity values dominate in annotation: 33.84% vs 15.38% in Tab. 2. This correlates with the dominance of the negative basic emotions that can be observed in the statistics presented in Tab. 3, i.e. 76.48% emotions associated with noun LUs and 70.13% with adjective LUs are negative. A similar dominance of words marked negatively could be also observed in the dictionary of the colloquial Polish language (Anusiewicz and Skawiński, 1996). For instance, if we compare two thematic fields of this dictionary, namely: act-

PoS	joy	trust	antic.	surprise	fear	disgust	sadness	anger				
N	15.17	6.74	0.96	0.65	7.66	21.78	16.77	30.27	–	–	–	–
Adj	20.95	8.01	0.54	0.37	5.31	18.56	21.56	24.71	–	–	–	–
	util.	good	truth	know.	beauty	happ.	futility	harm	ignor.	error	uglin.	misfor.
N	18.89	3.06	0.76	4.76	2.17	14.98	13.93	12.69	3.07	13.40	2.71	9.58
Adj	23.88	3.62	1.01	2.53	4.03	14.37	15.29	8.85	1.18	14.30	3.56	7.40

Table 3: Basic emotions (see Sec. 3.2) and fundamental human values (see Sec. 3.2) annotation of plWordNet 4.0 (in progress) are shown in percentage points.

ing towards somebody’s harm – enforcing some particular behaviours (id:2.3.2) and acting towards somebody’s profit (id.: 2.3.3), we can notice that the former includes 324 entries while the latter only 20. In plWordNet emo it is also characteristic that almost all emotions except *fear* are approximately frequent while *joy* is a single dominating positive emotion. This bias can be a result of assigning *joy* not only as a simple emotions, but also as a basic component of the complex emotions.

Contrary to the basic emotions, the fundamental human values are evenly distributed between the positive and negative ones, see Tab. 2: 55.38% negative values assigned to nouns and 50.57% to adjectives. There are no single fundamental human values that are substantially more frequent across the annotations, but only some of them, e.g. *prawda* ‘truth’ are significantly less frequent. Language users mostly perform evaluations of an emotional or utility (advantageous vs non-advantageous) character. They relatively infrequently assess phenomena from the rational perspective. Emotively marked LUs are more frequent in colloquial or informal communication where emotions and advantages are more important than rational thinking.

We checked also combinations of sentiment polarity values inside synsets. Almost all synsets are consistent with respect to the sentiment polarity, i.e. only ≈ 20 synsets from many thousands analysed included LUs of both positive and negative polarity, and most of them result from errors in plWordNet, e.g. too broad synsets. Synsets including marked LUs and neutral or ambiguous ones are more frequent, but perfectly compatible with the annotation guidelines. LU linked by hypernymy (via synset hypernymy) are in the vast majority of cases in the same polarity. We found only less than 700 hundred LUs linked by hypernymy per more than 70,000 analysed pairs in which both LUs were in the different polarity, among which we found only 32 $\langle -s, +s \rangle$ pairs.

6 Conclusions and Further Works

A large emotive lexicon can be an indispensable language resources for sentiment analysis and opinion mining, if it is of good coverage and quality, especially if the lexicon-based method is expanded with domain adaptation on the basis of machine learning. At least the use of the lexicon can help to improve the domain independent aspect of the method. The pilot project (Zaško-Zielińska et al., 2015) showed that with relatively small workload a promising emotive resource was be created. We presented an annotation process following this project and aiming at building a very large emotive lexicon of Polish of more than 130k manually annotated lexical units from plWordNet, i.e. on a scale incomparable to the majority of existing resources. The intended size is meant to suit the envisaged applications. A slightly modified general model and annotation guidelines were presented, together with improved specific guidelines for adjectives. Both the lexicon as well guidelines utilise the rich relation structure of plWordNet. The observed high values of the inter-annotator agreement (measured on a large sample of data according to an objective procedure) is very promising for the future applications and is a strong argument in favour of the assumed model and annotation procedure. The presented first results for nouns and adjectives, but for quite large sample, allows for collecting interesting observation that are in line with qualitative analysis in literature. We plan to complete annotation (>130k lexical units in total) of all Parts of Speech in plWordNet by the July 2018. The results will be completely open. The annotation will be extended to the rest of plWordNet by automated method (e.g. based on activation propagation or machine learning.) We plan also to compare our annotation with annotation built for English using the mapping of plWordNet onto Princeton WordNet.

Acknowledgment

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL.

References

- [Anusiewicz and Skawiński1996] Janusz Anusiewicz and Jacek Skawiński. 1996. *Słownik polszczyzny potocznej*. Wrocław.
- [Baccianella et al.2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, MT., pages 2200–2204. ELRA.
- [Bartmiński and Niebrzegowska-Bartmińska2009] Jerzy Bartmiński and Stanisława Niebrzegowska-Bartmińska. 2009. *Tekstologia*. Wydawnictwo PWN, Warszawa.
- [Bogusławski1991] Andrzej Bogusławski. 1991. Polski sufiks –utki,. *Poradnik Językowy*, 5–6:174–179.
- [Carrillo de Albornoz et al.2012] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/236_Paper.pdf.
- [Chen and Skiena2014] Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 383–389, Baltimore, Maryland, USA, June 23-25 2014. ACL.
- [Cohen1960] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- [Das and Bandyopadhyay2010] Amitava Das and Sivaji Bandyopadhyay. 2010. SentiWordNet for Indian Languages. In *Proceedings of the Eighth Workshop on Asian Language Resources*, pages 56–63.
- [Ekman and Friesen1971] P. Ekman and WV. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, Feb.
- [Ekman1992] Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200.
- [Esuli and Sebastiani2006] Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.
- [Grzegorzczkova et al.1998] R. Grzegorzczkova, R. Laskowski, and H. Wróbel, editors. 1998. *Morfologia. Gramatyka współczesnego języka polskiego*, volume T. 2. PWN.
- [Lula et al.2016] Paweł Lula, Katarzyna Wójcik, and Janusz Tuchowski. 2016. Analiza wydźwięku polskojęzycznych opinii konsumenckich ukierunkowanych na cechy produktu [feature-based sentiment analysis of opinions in polish]., *Prace Naukowe Uniwersytetu Ekonomicznego We Wrocławiu [Research Papers of Wrocław University of Economics]*, 207:p.155.
- [Maziarz et al.2013] Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.
- [Maziarz et al.2016] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.
- [Mohammad and Turney2013] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.
- [Piasecki et al.2009] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press.
- [Piasecki et al.2013] Maciej Piasecki, Michał Marcińczuk, Radosław Ramocki, and Marek Maziarz. 2013. WordnetLoom: a wordnet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232.
- [Plutchik1980] Robert Plutchik. 1980. *EMOTION: A Psychoevolutionary Synthesis*. Harper & Row.
- [Poria et al.2012] S. Poria, A. Gelbukh, E. Cambria, PeiPei Yang, A. Hussain, and T. Durrani. 2012. Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. In *IEEE 11th International Conference on Signal Processing (ICSP), 2012*, volume 2, pages 1251–1255, Beijing.
- [Puzynina1992] Jadwiga Puzynina. 1992. *Język wartości [The language of values]*. Scientific Publishers PWN.
- [Siudzińska2016] Natalia Siudzińska. 2016. *Formacje ekspresywne we współczesnym języku polskim (na przykładzie wybranych pospolitych nazw osobowych)*. Warszawa.

- [Strapparava and Valitutti2004] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.
- [Torii et al.2011] Yoshimitsu Torii, Dipankar Das, Sivaji Bandyopadhyay, and Manabu Okumura. 2011. A Developing Japanese WordNet Affect for Analyzing Emotions. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 80–86.
- [Turney and Littman2003] Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.
- [Waszakowa1991] Krystyna Waszakowa. 1991. O wartościowaniu w słowotwórstwie. *Poradnik Językowy*, 5-6:180–186.
- [Xu et al.2013] Jun Xu, Ruifeng Xu, Yanzhen Zheng, Qin Lu, Kam-Fai Wong, and Xiaolong Wang. 2013. Chinese Emotion Lexicon Developing via Multilingual Lexical Resources Integration. In *Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics CI-Cling 2013*, pages 174–182.
- [Zaśko-Zielińska et al.2015] Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP’2015*, pages 721–730, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

The Company They Keep: Extracting Japanese Neologisms Using Language Patterns

James Breen

University of Melbourne
Melbourne, Australia
jimbreen@gmail.com

Timothy Baldwin

University of Melbourne
Melbourne, Australia
tb@ldwin.net

Francis Bond

Nanyang Technological
University, Singapore
bond@ieee.org

Abstract

We describe an investigation into the identification and extraction of unrecorded potential lexical items in Japanese text by detecting text passages containing selected language patterns typically associated with such items. We identified a set of suitable patterns, then tested them with two large collections of text drawn from the WWW and Twitter. Samples of the extracted items were evaluated, and it was demonstrated that the approach has considerable potential for identifying terms for later lexicographic analysis.

1 Introduction

As the coverage of lexicons (including wordnets) improves, deciding which words should be added next becomes an issue. New words are constantly being added to languages, and existing words are not always covered by current lexical resources.

This paper reports on an investigation as to whether it is possible to identify and extract neologisms (newly created words and expressions) from Japanese text based on the language patterns in which they occur. The genesis of the project is the observation that one often encounters in Japanese text terms which the writer thinks needs some explanation, either because they are new or uncommon. This may be signalled by following the term with phrases such as *というのは* (*to iu no wa* “as for that which is said ⟨term⟩”) and *とは* (*to wa* “as for ⟨term⟩”), sometimes combined with the reading in parentheses, and then followed by an explanation. The phenomenon is well known to Japanese translators, who often

will do a WWW search for “⟨term⟩ とは”, etc. when encountering an unfamiliar term in order to identify cases where the term is being described, discussed or otherwise highlighted.

The investigation broadly breaks into two components:

- a. the identification of the sorts of language patterns used to describe, discuss, highlight, etc. terms;
- b. the extraction and evaluation of terms so targeted by those language patterns.

2 Prior Work

Research into the use of linguistic patterns in text to detect terms of interest has taken place in several contexts. In keyphrase extraction Hasan and Ng (2014) have produced a wide-ranging survey of the various techniques used in keyphrase extraction and their relative effectiveness, and Kim et al. (2013) evaluate the performance of a variety of supervised and unsupervised approaches. In term extraction, which is a major part of the broader field of terminology, usually in technical contexts (Kageura (2000)), Takeuchi et al. (2009) adapted the French ACABIT system, which detects morpho-syntactic sequences, to isolate terms in Japanese for later analysis. Le et al. (2013) used patterns of phrases to identify particular Japanese legal documents of interest. Mathieu (2013) successfully adapted a keyphrase extractor for use with Japanese, although its use was restricted to *kanji* sequences. The relationship between a text pattern and a term of interest is a form of **collocation**, i.e. lying between idiomatic expressions and free word combinations. In their survey of collocations in language processing, McKeown and Radev (2000) explore the role of the extraction of collocations in lexicography, although the focus is on the identification

of general terms rather than those which are highlighted as being of interest. Prior published research into the use of Japanese text patterns which target general terms of interest appears to be quite limited. Sato and Kaide (2010) employed a related technique for extracting English–Japanese name pairs by scanning texts for nearby occurrences of *Mr*, *Mrs*, etc. and the Japanese equivalents, e.g. *さん* (*san*).

3 Text Corpora

An essential element of the investigation is the availability of substantial quantities of Japanese text, preferably from a variety of sources. While there are number of Japanese corpora available for use in NLP work, most are actually quite small. In this study we used two text collections:

- a. the Kyoto WWW Corpus. This is a collection of 500 million Japanese sentences collected from WWW pages in 2004. The main problem is that it is getting dated, and hence what may have been neologisms at the time of its collation may well be recorded and accepted now, or have totally faded from use.
- b. Twitter text. We used a collection of 870 million Japanese text passages extracted from 2014 and 2015 Twitter data. This data provides the opportunity to see how the techniques under investigation perform with with contemporary and at times slangy text.

4 Initial Exploration

4.1 Pattern Frequencies

Initially we explored whether the text patterns typically associated with the discussion of particular terms occur in sufficient quantities to make them useful search keys by examining their frequencies in the Google Japanese *n*-gram Corpus (Kudo and Kazawa, 2007) (see Table 1).

The high-scoring *とは* is really a common form of topic marker without any particular association with new or unusual terms, and almost certainly would produce very noisy results if used as a search pattern. On the other hand *というのは*, *という言葉*, *という意味* and *の意味は* are typically associated with particular

Term	Frequency
<i>とは to wa</i> “as for”	169,756,339
<i>というのは/と言うのは to iu no wa</i> “as for the said”	19,134,679/1,207,555
<i>という言葉/ということば to iu kotoba</i> “said term”	5,360,613/167,095
<i>という意味/といういみ to iu imi</i> “said term’s meaning”	4,544,800/10,364
<i>という意味は to iu imi wa</i> “as for the said term’s meaning”	51,726
<i>の意味は/のいみは no imi wa</i> “as for the meaning of”	1,979,108/1,169

Table 1: Google *n*-gram Corpus Frequencies of Text Patterns

terms and are probably worth further investigation.

4.2 Testing Contexts of Known New Terms

We also investigated the sorts of contexts in which known new terms are being used to see if any useful additional patterns could be identified. As an initial exploration 5 terms were chosen from recent additions to the JMdict database (Breen, 2004) which had been noted as popular new words/expressions. The 5 terms were:

- *マタハラ matahara* abbreviation meaning “workplace discrimination against pregnant women”;
- *こじらせ女子 kojirase joshi* “girl who has low self-esteem”;
- *ナマポ namapo* slang for “welfare recipient”
- *美魔女 bimaajo* “middle-aged woman who looks very young for her age”
- *隠れメタボ kakure metabo* abbreviation meaning “normal weight obesity”

10 sentences for each term were extracted using a WWW search. While this is clearly a small number of samples, it emerged that there were relatively few of the *という/とは/etc.* sorts of patterns used; only four occurred a total of seven times in the 50 sentences, and quite a number of the terms being tested occurred encapsulated by some form of parentheses, either “...” (5 occurrences), 「...」

Term	Frequency
造語 <i>zōgo</i> “neologism, coinage”	232,837
新語 <i>shingo</i> “neologism, new word”	152,785
現代用語 <i>gendai yōgo</i> “neologism, recent word”	62,705
新造語 <i>shinzōgo</i> “neologism, new coinage”	3,978
言語新作 <i>genko shinsaku</i> “neologism (esp. medical)”	220
造語症 <i>zōgoshō</i> “neologism (esp. medical)”	<20
ネオロジズム <i>neorojizumu</i> “neologism”	<20
ネオレジズム <i>neorejizumu</i> “neologism”	<20

Table 2: Google *n*-gram Frequencies for Words Meaning Neologism

Term	Frequency
という造語/と言う造語 (<i>to iu zōgo</i>)	10042/491
という新語/と言う新語 (<i>to iu shingo</i>)	3140/117
という現代用語/と言う現代用語 (<i>to iu gendaiyōgo</i>)	50/<20

Table 3: Google *n*-gram Frequencies for Extended Neologism Patterns

(10 occurrences) or 『...』 (1 occurrence).¹

4.3 Explicit Neologism Labelling

We then investigated the use of terms in Japanese which can mean neologism, some of which are given in Table 2, along with their relative frequencies from the Google *n*-grams. As the first three account for almost all the usage, these were investigated further for their use in combination with the *という* and *と言う* (“as said”) patterns (Table 3).

As the frequencies for *という造語* and *という新語* looked promising, a sample of 10 sentences for each was identified via a Google WWW search. These sample sentences indi-

¹Japanese orthography uses a variety of symbols for text encapsulation, with the 「」 pair commonly used where inverted commas are used in English. Other symbols used for this include: ◇, ◈, <>, ◻, □ and ▭

cate the approach seems to have considerable promise. Quite a few relatively new terms, such as *ブロマンス* *buromansu* “bromance”, were in the samples. It is also interesting to note that all the terms referenced by the patterns were encapsulated in some forms of parentheses.

4.4 Parenthesized Kana

It has been observed that explanations of terms in Japanese are often accompanied by the reading of the term in parentheses.

To evaluate whether parenthesized readings are present in association with the sorts of language patterns under consideration, and if so whether they are in sufficient quantities to include them in the text analysis, a scan was made of the Kyoto Corpus to extract all sentences containing the patterns described above (という言葉, という造語, etc.). Approximately 2.4 million sentences were extracted, and these were analyzed to determine if they contained parenthesized strings of kana. Only 116 text lines contained “(*kana*)” patterns, and of these there was only one passage containing the “term (reading)” pattern, which indicated that this pattern was not common enough to make it worth a lot of attention.

4.5 Expansion of Linguistic Patterns

Discussions were held with several native speakers of Japanese in order to identify possible patterns which may be used with new terms. From this a number of additional patterns were identified. Some also typically followed the term in question, e.g. *xx* という言葉 を聞き *to iu kotoba wo kiki* “hearing the said word *xx*” and *xx* という不思議な *to iu fushigi na* “the said *xx* is strange/curious”.

In addition, a set of phrases which would precede a target word was identified, e.g. この頃よく聞く *xx kono goro yoku kiku*, 近頃よく聞く *xx chikagoro yoku kiku*, and 最近はやりの *xx saikin yoku kiku*, all of which mean “the often heard recently *xx*”.

This resulted in an overall set of 37 text patterns, some of which have alternative surface forms, e.g. *このごろ* and *この頃* (*kono goro*).

4.6 Initial Evaluation of the Language Patterns

The 37 text patterns were tested against the Kyoto WWW Corpus. For each pattern a sample of 20 sentences was examined in detail, with each sentence being classified into one of three groups: sentences which did not directly discuss any identifiable word or term (1); sentences which focussed on a word or term which is already established in one or more lexicons (2); and sentences which focussed on a word or term which is not in an accessible lexicon, and which warrants further investigation (3).

It was clear that some of the text patterns were quite effective in identifying text passages which focus on words or terms of interest, and in some cases the precision appeared to be quite high; in three of the sets of samples (という造語, という新語, という新しい言葉) all of the passages had such a focus, and in another five (という言葉聞き, という言葉を耳に, という言葉が話題に, という言葉がはやって, という流行語) 85% or more had that focus.

Around half of the sampled passages (349) were classified into Groups 2 and 3, and these were about evenly split between those where the target term was in parentheses (177) and those where it was not (172).

Overall the numbers of sentences extracted with the selected patterns only made up a very small proportion of the sentences in the Corpus. Of the approximately 500 million sentences the high precision patterns only extracted 2,600 sentences. When combined with lower precision patterns the numbers extracted came to about 280,000 (about 0.06%), and it was observed that most of these were from one pattern (という言葉).

5 Detailed Investigation

From the original set of 37 patterns, a set of 18 were chosen for further experimentation. The selection process was to choose those patterns which had resulted in the higher proportion of Group 2/3 being detected in the sampling.

Excluded from the original set were three of the more commonly occurring patterns: と言うのは/というの, という and といういみ/という意味. Although between them they accounted for about 80% of the of the sentence selections, they performed compara-

tively poorly in being associated with possibly useful terms. Of the chosen patterns ということば/という言葉 accounted for over 90% of the remaining extracted lines, and 最近はやりの/最近流行の/最近流行りの accounted for a further ~7%. Thus the overwhelming majority of remaining extractions come from two patterns. They are among the middle-ranking performers according to the sampling, and certainly cannot be ignored. While there are other patterns which performed considerably better in the sampling in terms of precision, the number of actual extractions associated with them is much lower.

5.1 Text Scanning and Target Term Extraction

With over a billion lines of text to examine for the presence of the language patterns a reasonably fast searching technique is desirable. The possibility of training a machine learning model was considered, however since we are dealing with a constrained set of patterns a direct pattern-matching approach is clearly more appropriate. Also the nature of the patterns lends itself to a fast character-by-character search using a search tree. The patterns being used begin with only four different characters: こ, と, 近 and 最, and initially each character in a line of text only has to be compared with them to determine whether more of the tree is to be searched. Similarly at each level of the tree only a few characters typically need to be tested.

The 500 million lines in the Kyoto Corpus had 280,574 matches with these patterns, and the 870 million tweets had 130,310 matches. The hit rate for these patterns in Twitter is thus only about 30% that of the WWW text, which is probably indicative of both the brevity of many tweets, and possibly a very different text style for longer tweets.

From the extracted lines of text, it was necessary to isolate the target terms associated with the patterns. The approach taken was:

- divide the patterns into those where the target usually precedes the pattern (these always begin with という), and those where the target usually follows (the rest).
- detect and extract text which occurred in some form of parentheses before or after the pattern. The extraction was re-

stricted to parenthesized terms beginning 3 or fewer characters before or after the pattern. This margin was to allow for the occasional punctuation characters and words such as など *nado* “et cetera”. Also it was clear that there were occasionally quite long strings of parenthesized text, typically quotations, which were not going to be considered valid lexical items, so the extraction was restricted to strings of up to 10 characters.

- c. where there are no parenthesized target strings associated with the text patterns, it is necessary to attempt to extract target terms from the text preceding or following the patterns. Inspection of a number of passages indicated that most likely candidates were made up of combinations such as noun–noun, prefix–noun, noun–suffix, adverb–noun, adjective–noun, etc. and that a reasonable heuristic would be to collect morphemes until one which typically lies on the boundary of an expression, such as a particle or a verb, was encountered.

To implement this approach, the text following or preceding the pattern was passed through the *MeCab* morphological analyzer (Kudo et al., 2004; Kudo, 2008)² operating with the Unidic morpheme lexicon (Den et al., 2007), and adjacent morphemes which met a limited set of part-of-speech (POS) attributes were aggregated

For each text collection the target term extraction as described above was run, the extracted terms were filtered against a large reference lexicon (as the aim of the investigation is to determine whether the method is extracting new or unrecorded terms), and the remaining unlexicalized extractions were sorted and aggregated to determine how often they occur. This is to enable evaluation of the hypothesis that more frequently-occurring terms are more likely to be potential lexical items. The numbers of target terms extracted from the text collections is shown in Table 4.

Some general observations that can be made about these extractions are:

- a. the extractions comprise a very small proportion of the text in the two collections. The passages extracted from the WWW

Corpus represent only 0.056% of the text and the ones from the Twitter collection only 0.015%.

- b. the *ということば/という言葉* pattern is relatively much more common in the Kyoto Corpus (0.054%) than in the Twitter collection (0.013%). The *最近流行りの/etc.* pattern is also more common in the Kyoto Corpus, but not to such a degree.
- c. the target terms are clearly less likely to be parenthesized in Twitter text, and also the target terms associated with *という...* patterns are more likely to be parenthesized than the others where the target follows the pattern.

6 Evaluation of Extracted Target Terms

The extracted terms were then categorized according to the usefulness of the term as a lexical item. This involved examining the term both in the context of the text passage(s) in which it was detected, in other text passages such as those discovered from WWW searches, and in reference material such as glossaries which were not part of the reference lexicon. From this categorization codes were assigned to the terms as follows: (A) in the reference dictionary in different surface form, e.g. partially or fully in kana instead of kanji; (B) an inflected or variant form of existing entry; (C) definitely of interest as it has the potential to be a valid lexical item; (D) other, e.g. a phrase not of particular interest; (E) corrupted text.

Also recorded was whether the occurrences of the terms were parenthesized or not, and which pattern(s) generated the extraction. (This was done for the “C” terms.)

6.1 WWW Corpus

Of the 234,733 terms extracted from this Corpus, 68,644 were not in the reference lexicon. Of these 52,277 were terms that occurred only once, and the remainder occurred multiple times (the maximum was 55 times).

A detailed analysis of 120 terms was carried out as follows: the most common 50 terms (13–55 occurrences), a sample of 20 terms which occurred 5 times each, and a sample of 50 terms which occurred once each. The categorization of the terms is shown in Table 5.

²<http://taku910.github.io/mecab/>

Source	Total lines	Extractions (Paren.)	Extractions (Non-paren.)	None extracted
WWW Corpus (all patterns)	280574	124371	110362	45841
Twitter (all patterns)	130310	37083	71995	21232
WWW Corpus (という言葉)	270553	122727	103111	44715
Twitter (という言葉)	119871	36074	64254	19543
WWW Corpus (最近流行りの)	6711	573	5653	485
Twitter (最近流行りの)	7635	314	6530	791
WWW Corpus (the rest)	3310	1071	1598	641
Twitter (the rest)	2805	696	1211	898

Table 4: Target Term Extraction Counts

Category	Top 50	5 Times (20)	Once (50)
A	15	2	0
B	6	6	1
C	18	10	3
D	8	2	46
E	3	0	0

Table 5: Categorizations of Extracted Text — WWW Corpus

Some examples of the extractions are:

- (A) がんばれ *ganbare*: *kana* form of 頑張れ “go for it!”
- (A) ガイジン *gaijin*: *katakana* form of 外人 “foreigner”
- (B) 愛している *aishiteiru*: from the verb 愛する and meaning “to be in love”
- (B) 感動した *kandōshite* — past tense of 感動する “to be moved”
- (C) ゲーム性 *gēmusei* “quality of a video game; game rating”
- (C) 共創 *kyōsō* “growing together; joint development”
- (D) シンプルイズベスト *shinpuru izu besuto* (“Simple Is Best”: pop song name)

The relatively high proportion of “C” terms in the multiply-occurring sets (36–50%) is interesting. It might seem intuitively obvious that more commonly used or discussed terms would be more likely to be potential lexical items, but it could well not have been the case. More sampling of the 2, 3 and 4 batches may be appropriate, but it seems clear that multiple occurrences of a term, at least among the terms extracted here, is a signal of its likelihood to be of interest.

6.2 Significance of Multiple Occurrences

It was noted that the three singly-occurring C extractions in Table 5 all had reasonably high counts of occurrences in the *n*-gram Corpus (258–473). That raises the question of whether the number of Corpus occurrences is linked or correlated to the usefulness of extracted terms. To test this a sample of 10 of the singly-occurring “D” terms was checked to determine the number of occurrences in the Corpus. 6 of these occurred fewer than 10 times and the others occurred 39, 52, 62 and 1,561 times respectively. Also checked were the Corpus counts of the 8 “D” terms in the “top 50” set. While they varied, they were noticeably lower than the “C” counts. This seems to indi-

cate support for a (quite reasonable) hypothesis that low overall occurrence counts are related to the usefulness of extracted terms.

As a further test of this hypothesis, a set of 2,000 of the singly-extracted terms was chosen and their overall counts in the Corpus established. About 160 of these (8%) each occurred 400 or more times. Examination of a sample of 20 of these more commonly occurring terms resulted in the following category counts: B: 1, C: 14, D: 6.

This is a very different outcome to that shown by the randomly selected singly-extracted terms, and it seems likely that a high extraction count and/or a high overall Corpus count are good indicators that an extracted term has a chance of being a term of interest. The overall Corpus count of a term may not be a particularly useful metric as it would be difficult to obtain in a general harvesting process. They are only available with the Kyoto WWW Corpus because an n -gram corpus and associated utility software are available. However a useful corpus count could well be taken from a different comprehensive corpus such as the Google n -gram Corpus.

6.3 Twitter Data

A similar analysis was carried out on the text of 2014/15 Twitter data. Some additional analysis was carried out on two aspects of this data: where the text passages were identified as “re-tweets” these were aggregated and a separate investigation made of the term to see if occurrence within a re-tweet was any different to other target terms in terms of usefulness; and since the Twitter text was associated with specific dates, an analysis was made to determine if identified terms were clustered and if so whether this was associated with greater usefulness.

6.4 Re-tweets

The fact that Twitter text contains “re-tweets”, i.e. messages repeated by Twitter users to their followers, raises a number of issues in terms of the analysis of the text. On the one hand the re-tweeting can seriously distort any analysis which attempts to use frequency information with regard to such things as extracted terms (Lu et al., 2014). On the other hand the fact that a passage is being re-

layed by Twitter users may in itself be useful in the analysis of the passage.

The actual identification of re-tweets has proved to be a significant problem as we observed that often the users make minor amendments before sending the message as though it were new; often such relays of modified tweets outnumbered the formal re-tweets.

6.5 Analysis of Re-tweets

The terms extracted from re-tweets were aggregated and ranked according to the numbers of times the tweet was repeated in order to see if greater repetition was associated with the usefulness of the extracted term. Samples of terms from the over 100 repetitions, 10 to 99 repetitions and 5 repetitions groups were selected and examined. From this examination it was concluded that the occurrence of extracted terms in re-tweets was not a strong indication of usefulness.

A similar investigation was made of a sample of multiply-occurring terms that were not in re-tweets, and as with the investigation of the extractions from the WWW Corpus discussed above, it does appear that the number of times a term is extracted is correlated with the likelihood it is of interest.

As with the WWW Corpus terms, a sample of singly-occurring terms was checked against an n -gram corpus, in this case the Google n -gram Corpus. A selection of 2,000 singly-extracted candidate terms was matched against the Corpus and a sample of 20 of the higher-ranking terms was evaluated. The results were 7 terms ranked as A or B, 5 as C and 8 as D. While this is only a small sample, it does seem to indicate that a high count in an n -gram Corpus indicates a greater likelihood that a term is of interest.

6.6 Classification of Names

In contrast to the terms identified in the WWW Corpus, a significant proportion of the terms extracted from Twitter text were names, e.g. *anime* characters, Pokemon characters, singers, etc. In hindsight there probably should have been a category for them, as they have been treated as “D” (not of interest). The fact they are being collected is an indication of the efficacy of the approach.

6.7 Issue of Parenthesized Terms

As previously described, the method for extracting possible terms involves either collecting a string of text in parentheses associated with the pattern, or collecting a string of morphemes with restricted POSs associated with the pattern. It is worth examining the relative outcomes of these two approaches to determine if there is a qualitative difference.

Of the approximately 27,000 potential terms extracted from the Twitter text, 12,650 were parenthesized and 14,348 were not parenthesized. Samples were selected from the two groups of terms and examined in detail. From this it was determined that there is no clear domination of one approach over the other.

6.8 Burstiness

As the Twitter texts have dates in their metadata it was possible to examine whether multiple occurrences were in bursts, and whether this might be associated with greater or lesser relevance. A sample of ten non-re-tweet multiply-occurring extractions ranging from 16 to 48 occurrences was examined. Of the 10, 3 were clustered into a relatively short period, e.g. a few days, and the other 7 were spread over the whole period of the data. From this it does not appear that clustered multiple occurrences of candidate terms have any particular advantages. The clustering may indicate a degree of topicality of a term, although it may lead to focus on an ephemeral term, when a greater spread of usage over time may indicate more general usage.

7 Precision and Recall

The establishment of precision and recall metrics in this area poses an interesting challenge. In terms of precision the testing reported above indicates that some patterns, e.g. という造語/という新語, are likely to result in fairly high levels, however if they result in a relatively small number of lexical items being collected it is of limited use in lexicon building. Casting a wider net and being prepared to sift results is probably a better course.

In terms of measuring recall the typical approach would be to identify how many terms-of-interest there are in a corpus, and test how often they are identified by the extraction

method. To probe this issue the 10 candidate terms examined in Section 6.8 above were tested to see how often they occurred in the text, both in and out of the extraction patterns.

In 8 of the 10 terms over half of the occurrences in the Twitter text had been identified, and in three cases over 95% were identified. The proportions identified in the WWW Corpus were noticeably lower.

8 Discussion and Conclusions

From the investigations described above, a number of conclusions can be drawn and observations made about the techniques being investigated. Among them are:

- a. it is clear that the technique is quite effective in highlighting terms suitable for further investigation, as it identifies candidates that are often very worthy of detailed examination and subsequent lexicalization.
- b. it is interesting and not a little frustrating that after all the early work in identifying useful text patterns for identifying possible terms, the outcome has been so totally dominated by two patterns, to the extent that the others may as well be ignored. Several of the other text patterns have demonstrably better precision, but their recall of useful terms is so low as to make them of little use in a practical harvesting exercise. (That is no reason, of course, to exclude them as they add little overhead to process and at the margin can improve the outcome.)
- c. the technique can clearly be enhanced by association with an n -gram corpus with frequency counts. A term, particularly one which has not been extracted often, is much more likely to be a useful candidate if it has a high n -gram count.
- d. at present we have no real indication of the recall of the techniques being investigated. Objective analysis of recall would be a major task and best left for further work.
- e. one could envisage this technique being attached to something like a Twitter feed, and passing extracted candidate terms through a frequency and n -gram analysis, and ultimately on to lexicographers for analysis.

References

- James Breen. 2004. JMdict: a Japanese-Multilingual Dictionary. In *Proceedings of the COLING-2004 Workshop on Multilingual Resources*, pages 65–72. Geneva, Switzerland.
- Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Mine-matsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.
- Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273. Association for Computational Linguistics, Baltimore, Maryland. URL <http://www.aclweb.org/anthology/P14-1119>.
- Kyo Kageura. 2000. *The Dynamics of Terminology: a descriptive theory of term formations and terminological growth*. John Benjamins.
- Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2013. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, 47(3):723–742.
- Taku Kudo. 2008. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. <http://mecab.sourceforge.net/>.
- Taku Kudo and Hideto Kazawa. 2007. Japanese Web N-gram Corpus version 1. <http://www ldc.upenn.edu/Catalog/docs/LDC2009T08/>.
- Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 230–237. Barcelona, Spain.
- Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu. 2013. Unsupervised keyword extraction for Japanese legal documents. In *26th International Conference on Legal Knowledge and Information Systems, Bologna, Italy*.
- Yao Lu, Peng Zhang, Yanan Cao, Yue Hu, and Li Guo. 2014. On the frequency distribution of retweets. In *2nd International Conference on Information Technology and Quantitative Management, ITQM 2014*.
- Jérôme Mathieu. 2013. Adaptation of a key phrase extractor for Japanese text. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l’ACSI*.
- Kathleen McKeown and Dragomir Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, chapter 21. Marcel Dekker.
- Satoshi Sato and Sayoko Kaide. 2010. A person-name filter for automatic compilation of bilingual person-name lexicons. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA), Valletta, Malta.
- Koichi Takeuchi, Kyo Kageura, Teruo Koyama, Béatrice Daille, and Laurent Romary. 2009. Pattern based term extraction using ACABIT system. *CoRR*, abs/0907.2452. URL <http://arxiv.org/abs/0907.2452>.

Lexical-semantic resources: yet powerful resources for automatic personality classification

Xuan-Son Vu^{+1*}, Lucie Flekova^{¶2*}, Lili Jiang⁺³, Iryna Gurevych^{§4}

[¶] Amazon Research Germany, Aachen

[†] Department of Computing Science, Umeå University, Sweden

[§] UKP Lab, Computer Science Department, Technische Universität Darmstadt

^{†1}sonvx@cs.umu.se, ^{¶2}lflekova@amazon.com

⁺³lili.jiang@cs.umu.se, ^{§4}gurevych@ukp.informatik.tu-darmstadt.de

Abstract

In this paper, we aim to reveal the impact of lexical-semantic resources, used in particular for word sense disambiguation and sense-level semantic categorization, on automatic personality classification task. While stylistic features (e.g., part-of-speech counts) have been shown their power in this task, the impact of semantics beyond targeted word lists is relatively unexplored. We propose and extract three types of lexical-semantic features, which capture high-level concepts and emotions, overcoming the lexical gap of word n-grams. Our experimental results are comparable to state-of-the-art methods, while no personality-specific resources are required.

1 Introduction

Automatic personality classification (APC) has been employed on *user generated content* (UGC), such as Tweets, to collect the user personality for various personalized intelligent applications, including recommender systems (Hu and Pu, 2011), mental health diagnosis (Uba, 2003), recruitment and career counseling (Gardner et al., 2012). Especially, the recommender applications benefit from knowing the personality of real as well as fictional characters (Flekova and Gurevych, 2015). For example, if a user is known to favor the personality traits displayed by the main

characters of, say, *Terminator*¹ and *Rambo*¹, then the system should automatically recommend movies with similar characters.

Currently, the performance of APC depends on how user personality is modeled and what types of personality features can be extracted. Regarding the first factor, one well-known model called Five Factor Model (Costa and McCrae, 2008) has been highly accepted as a standard model. It consists of five personality traits (i.e., extraversion, neuroticism, agreeableness, conscientiousness, openness to experience). The APC task is then formulated as a regular document classification on these five labels. To the second factor of feature extraction, the existing studies heavily depend on personality specific resources such as linguistic inquiry word count (LIWC) (Pennebaker et al., 2007). These resources, however, are rather time consuming and expensive to construct especially for minor languages (Vu and Park, 2014). Moreover, the resource construction requires expertise in both psychology and linguistic (e.g., LIWC). In contrast, it is observed that lexical-semantic features which could be extracted from the publicly available lexical resources (e.g., WordNet (Miller, 1995)) can help to improve the performance of the APC task. However, their impact on real world UGC data for APC had been relatively unexplored.

Among lexical-semantic features, sense-level features were explored in previous works (Kehagias et al., 2003; Vossen et al., 2006) with varying conclusions. In this paper, we conduct extensive experiments, aiming at obtaining a more detailed understanding of whether or not the senses can be beneficial in certain cases compared to word-based fea-

*The research by the 1st and the 2nd authors has been done during their employment at the UKP Lab, Technische Universität Darmstadt, Germany, and supported by the German Research Foundation under grant No. GU 798/14-1.

¹ Famous fiction/action movies.

tures. Broadly, we explore the use of word senses, supersenses, and WordNet sentiment features (Baccianella et al., 2010) in personality classification. Our main contributions are:

- Investigating the impact of different lexical-semantic features on APC task.
- Revealing the accumulated benefit by combining word sense disambiguation (WSD) with semantic and sentiment features in APC.
- Proposing and evaluating a feature selection method called *Selective.WSD* to improve WSD usage in APC.
- Proposing a unified framework on top of the UIMA framework² to integrate different lexical-semantic resources for APC.

The rest of this paper is organized as follows. Section 2 presents the related work and our novel contributions, as well as background knowledge of the Five Factor Model. Section 3 describes the experimental datasets. Our proposed framework and methodology are presented in Section 4. Experimental results and discussion are in Section 5. Section 6 concludes this paper.

2 Related Work and Background

Previous studies concerned the positive impact of sense-level features (i.e., using WordNet based WSD) on the performance of document classification systems (Rose et al., 2002; Kehagias et al., 2003; Moschitti and Basili, 2004; Vossen et al., 2006). Though they had different focuses, they suggest that word senses are not adequate to improve text classification accuracy. Vossen et al. (2006) report an improvement from 0.70 to 0.76 F-score while negative results have been reported by Kehagias et al. (2003). This is why supersenses, the coarse-grained semantic labels based on WordNet’s lexicographer files, have recently gained attention for text classification tasks. In this paper, we further explore the impact of these features in personality prediction.

There have been many different attempts to automatically classify personality traits from texts. However, there were not any studies

incorporating senses, supersenses, and sentiment features into the APC. Some works (Iacobelli et al., 2011; Bachrach et al., 2012; Iacobelli and Culotta, 2013; Okada et al., 2015) start from the data and seek linguistic cues associated with personality traits, while other approaches (Mairesse et al., 2007; Golbeck et al., 2011; Farnadi et al., 2016) make heavy use of external resources, such as LIWC (Pennebaker et al., 2007), MRC (Wilson, 1988), NRC (Mohammad et al., 2014), *SentiStrength*³, where they detect the correlations between those resources and personality traits.

However, the resources require the efforts of experts in psychology and linguistics, e.g., LIWC of Pennebaker et al. (2007), to construct. This constrains the available resources for APC, especially for minor languages. Thus, we aim at broadly available resources (e.g., WordNet and SentiWordNet), to benefit APC.

Close to our work, Mairesse et al. (2007) run personality prediction in both observer judgments through conversation and self-assessments using text via the Five Factor Model. They also exploit two lexical resources as features, LIWC and MRC, to predict both personality scores and classes using Support Vector Machines (SVMs) and M5 trees respectively. As for personality prediction on social network data, Golbeck et al. (2011) use both linguistic features (from LIWC) and social features (i.e., friend count, relationship status). Recently, Farnadi et al. (2016) deal with the automatic personality classification based on users social media traces, which include three of the four datasets in our study. However, similar to other studies (Mairesse et al., 2007; Farnadi et al., 2013), they mainly use the personality specific resources.

At the time of writing, the use of personality specific resources for APC has received much attention, while the impact of lexical-semantic features has been neglected. The only existing work that explores sense-level features is from Flekova and Gurevych (2015). They partially used sense-level features among others (i.e., lexical features, stylistic features, and word embedding features) for personality profiling of fictional characters. As a complement of the existing

²<https://uima.apache.org/>

³<http://sentistrength.wlv.ac.uk/>

work on automatic personality classification, the novel contributions of this paper include: (1) we present how WSD and lexical-semantic features influence personality prediction by conducting different experiments on four public datasets; and (2) we explore the accumulated impact of supersenses and sentiment features in combination with WSD.

The Five Factor Model

In personality prediction, the most influential Five Factor Model (*FFM*) has become a standard model in psychology over the last 50 years (Mairesse et al., 2007). The five factors are defined as extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience. Pennebaker and King (1999) identify many linguistic features associated with each of personality traits in *FFM*. (1) Extroversion (cEXT) tends to seek stimulation in the external world, the company of others, and to express positive emotions. (2) Neurotics (cNEU) people use more 1st person singular pronouns, more negative emotion words than positive emotion words. (3) Agreeable (cAGR) people express more positive and fewer negative emotions. Moreover, they use relatively fewer articles. (4) Conscientious (cCON) people avoid negations, negative emotion words and words reflecting discrepancies (e.g., should and would). (5) Openness to experience (cOPN) people prefer longer words and tentative expressions (e.g., perhaps and maybe), and reduce the usage of 1st person singular pronouns and present tense forms.

Table 1: A quick overview of the four datasets with the number of sentences (#Sen), the number of words (#Word), and the number of users (#Users). *Non-standard words* may be either out-of-vocabulary tokens (e.g., *tmrw* for ‘tomorrow’) or in-vocabulary tokens (e.g., *wit* for *with* in ‘I come wit you’).

Dataset	#Sen	#Word	#Users	Non-standard words
TWITTER	145.7	216.8	153	51.27%
FACEBOOK	67.1	78.3	250	23.3%
ESSAYS	48.8	15.3	2469	30.85%
YOUTUBE	41.7	29.5	404	8.05%

3 Dataset and Statistics

3.1 Dataset Overview

We conducted our experimental studies on four public datasets, three of which are from public social media platforms (i.e., Twitter, Facebook, Youtube) and the fourth one is a well-known public dataset specially for personality research. These datasets are chosen for their popularity and diversity in data size, scale of users, and writing styles.

- **TWITTER** : collected by PAN’ 15 (Stamatatos et al., 2015), it contains Tweets of 328 Twitter users in 4 languages in which only the Tweets come from 153 users written in English are selected in this study.
- **FACEBOOK** : collected through the myPersonality project ⁴ (Stillwell and Kosinski, 2015) containing status updates of 250 Facebook users with 9,917 status updates and personality labels.
- **YOUTUBE** : collected by Biel et al. (2011), it consists of a collection of behavioral features, speech transcriptions, and personality impression scores for a set of 404 YouTube vloggers. About 28 hours of video were annotated.
- **ESSAYS** : collected and analysed by Pennebaker and King (1999). It contains 2,479 essays from psychology students, who were required to write whatever came into their mind for 20 minutes. The data includes users, raw text, and gold standard classification labels.

3.2 Data Statistics

Table 1 shows the overview statistics of the four datasets. All values are normalized by the number of users in each corresponding dataset. *Non-standard words* denotes the fraction of non-standard words (unseen vocabularies in WordNet) over the total number of words in each dataset.

The statistics in Table 1 indicate that Twitter dataset has the highest value of #Sen and #Word but the lowest number of users. Moreover, the TWITTER dataset also has the highest ratio of *non-standard words*, which makes

⁴<http://myPersonality.com>

it more challenges to the APC task. All in all, these diverse characteristics benefit our results analysis on improving personality classification.

As depicted in Figure 1, we design a system based on UIMA framework⁵ for experimental studies. It contains three main processes including (1) Data Loading and Data Processing, 2) Feature Extraction, (3) Personality Classification and Evaluation. After loading data into the whole system (i.e., four datasets and lexical resources), feature extraction is performed. Afterwards, we formulate personality classification as a binary classification on each personality trait since more than one trait can be embodied in a user. We apply the SVM classifier (linear kernel) and the TF-IDF feature weighting scheme. In the evaluation, we use 10-fold cross validation, i.e., rotating the 10% test data selection over the dataset and training the SVM classifier on the 90% of not-tested data, to get accuracy scores. Since the goal of this paper is revealing the impact of different lexical-semantic features in APC, we used exactly the same classification algorithm as used in the popular work of Mairesse et al. (2007). Details about the second process of feature extraction will be described in the following subsection.

3.3 Feature Extraction

Based on our observations and the previous studies, we found that people with different personal traits have different writing styles and word usage. For example, *neurotic* and *extrovert* people use the emotion words significantly differently. *Neurotic* people use more 1st person single pronouns while less positive emotional words. And it is observed that *openness* people use more abstract concepts. Motivated by these observations, we manage to capture these personality trait differences by extracting the semantic and sentiment features.

4 Methodology

We denote four kinds of features as $F = \{WORD, SENSE, S_SENSE, SENTI\}$ where *WORD* is a set of word-level features, *SENSE* is a set of sense-level features, *S_SENSE* is a set of

⁵<https://uima.apache.org/>

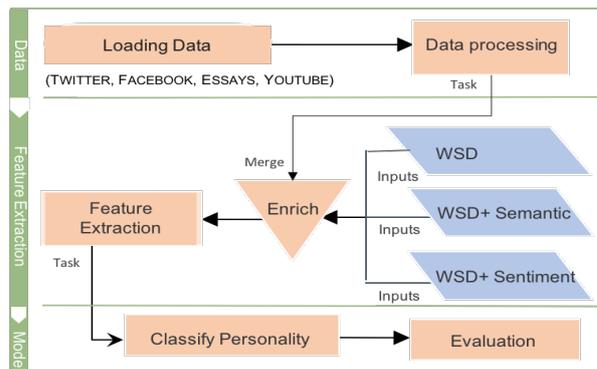


Figure 1: Workflow of the experimental pipeline.

WordNet supersense features, and *SENTI* is a set of sentiment features. (*S_SENSE*) is extracted from WordNet supersenses as a complement to *SENSE*. Regarding sense-level feature, we applied two different WordNet based WSD algorithms, SimLesk and MostFreq (Miller et al., 2013). Correspondingly, instead of *SENSE*, we have two different feature sets *WN-S-LESK* and *WN-MFS*. Thus, we finally have the feature list of $F = \{WORD, WN-S-LESK, WN-MFS, S_SENSE, SENTI\}$

Semantic Features

Regarding semantic features, we focus on extracting topic information given input texts from different people. We firstly recognize lexical knowledge by applying WordNet semantic labels⁶. For example, based on the given personal texts, after extracting word n-grams, the topic information is detected and organized in the form of *pos.suffix*. Here, *pos* denotes part-of-speech and *suffix* organizes groups of synsets into different categories (e.g., *a tiger* can be categorized into *noun.animal* and *a tree* is categorized into *noun.plant*). In this paper, DKPro Uby (Gurevych et al., 2012) is further employed to extract all above required information to represent in *pos* and *suffix* from given texts.

Sentiment Features

For sentiment features, we extracted emotional information, which are extremely important to characterize personality according to Pennebaker and King (1999). For example, neurotics use more negative emotion words

⁶<https://wordnet.princeton.edu/man/lexnames.5WN.html>

(e.g., *ugly* and *hurt*) than positive emotion (e.g., *happy* and *love*). In details, we applied the sentiment word disambiguation algorithm (i.e., SentiWordNet) to match the disambiguated word senses for each term with three scores, Positive (P), Negative (N) and Objective/Neutral (O) scores. Finally, we obtained the individually final P , N and O scores for each personal text, which were averaged by the total number of sentiment features.

4.1 Word Sense Disambiguation

Above, we have discussed and presented feature extraction for APC. However, one primary challenge in feature extraction is word sense ambiguity. To address this challenge, word sense disambiguation (WSD) is broadly applied to match the exact sense of an ambiguous word in a particular context. For word, sense, supersense, and sentiment features, it is necessary to first disambiguate the words to reduce the semantic gap.

However, due to the high ambiguity of words, it is extremely challenging to detect the exact sense in a certain context. Postma et al. (2016) showed that current WSD systems perform an extremely poor performance on low frequent senses. To address this challenge, we propose an algorithm *Selective.WSD* to reduce the side effect of WSD by finding senses of a word subset rather than all possible words in the BoW model. *Selective.WSD* is presented in Algorithm 1. The algorithm takes a word-level document as an input to return a mixture of word-level and sense-level feature list. The *wordLevelFeature(f)* function in the algorithm will return a word-level feature (e.g., bank) of a sense-level feature (e.g., bank%1) by removing the extra notation (e.g., %1). The function of *wsd.annotateSenses* in the algorithm is implemented based on DKPro WSD (Miller et al., 2013) - annotating the exact sense of a disambiguated word in a context. In the following experimental study section, we will show the impact of WSD on personality prediction.

4.2 Feature Selection

Feature selection is naturally motivated by the need to automatically select the best determinants for each personality trait. Thus, we can derive a qualitative description of the state

Procedure 1 *Selective.WSD*

Input: a word-level document.

Output: a selective mixture of word-level and sense-level feature list.

```

1: featuresL  $\leftarrow$  initialize an empty list
2: L  $\leftarrow$  topK word-level features ordered by  $\chi^2$ 
3: for sentence s  $\in$  document d do
4:   mixFeatList  $\leftarrow$  wsd.annotateSenses(s)
5:   for feature f  $\in$  mixFeatList do
6:     if wordLevelFeature(f)  $\notin$  L then
7:       f  $\leftarrow$  wordLevel(f)
8:     else
9:       f  $\leftarrow$  senseLevel(f)
10:  featuresL  $\leftarrow$   $\cup$  f
return featuresL

```

characteristics. In this way, the noisy features are filtered out. We used the χ^2 feature selection algorithm before feeding the features (i.e., word, sense, supersense, and sentiment features) to a classifier. The feature selection strategy was chosen empirically based on our preliminary experiments on training dataset, where we compared χ^2 with three other state-of-the-art feature selection methods for the supervised classification (i.e., Information Gain, Mutual Information, and Document Frequency thresholding (Yang and Pedersen, 1997)), and χ^2 outperformed.

Table 2: Abbreviation list of the feature set

ID	Description
WORD	Word-level features.
WN-WORD	Word-level features in which only words that present in WordNet are used.
WN-MFS	Sense-level features based on the most frequent sense algorithm.
WN-S-LESK	Sense-level features based on the Simplified Lesk algorithm.
S.SENSE	WordNet semantic label (or WordNet supersense) features.
SENTI	Three sentiment features including <i>posscore</i> , <i>negscore</i> , and <i>neuscore</i> .

5 Experiment and Analysis

We conducted extensive experiments to investigate the impact of different lexical-semantic

features on the APC task. All the feature abbreviations we use are listed in Table 2.

5.1 Experiment Settings

We compared four pipelines based on different lexical-semantic feature settings. In the first and simplest pipeline, the documents are segmented into words used as features. We further refer to this setup as **WORD**. The subsequent feature selection and classification, specified below, is the same for all pipelines. In the second processing pipeline, the documents are segmented to words, and the words are further annotated with their part-of-speech and lemma. With these annotations, we can look them up in WordNet. Only those words, which are present in WordNet, are then used as bag-of-words features. This intermediate step reveals which changes in performance can be attributed to the lexicon coverage as opposed to the WSD quality. We refer to this setup as **WN-WORD**. The third processing pipeline is similar to the previous one, but after the **WN-WORD** lookup step performed, in addition, the WordNet based WSD is employed to extract sense-level features. For each of the words present in WordNet, the resulting sense and its WordNet semantic label (s_SENSE) are both used as two features. There are two possible configurations in the third pipeline, which differ in the WSD algorithm used (see subsection 4.1). We experimented with the most frequent sense baseline (denoted further as $WN-MFS-S_SENSE$) and Simplified Lesk algorithm ($WN-S-LESK-S_SENSE$). Differently from the third pipeline, in the fourth pipeline, for each sense, we calculate three sentiment scores (positive, negative, neutral) by applying SentiWordNet and add them as three extra features. We refer to this setup as $WN-S-MFS-S_SENSE-SENTI$ and $WN-S-LESK-S_SENSE-SENTI$ for the Most Frequent Sense and the Simplified Lesk algorithm correspondingly. All results from the above four different pipelines are shown in Figure 2 and Figure 3. More discussions are present in the following subsections.

5.2 Experimental Result Demonstration

As shown in Figure 2 and 3, though the APC performance of different configurations varies on different datasets, we have some interest-

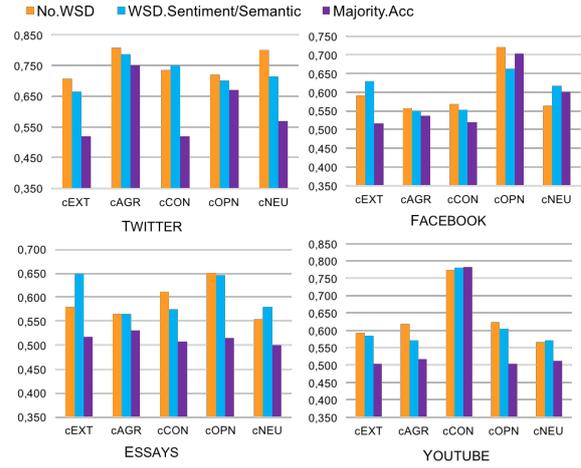


Figure 2: A comparison between not-using WSD (i.e., No.WSD) versus using WSD in a combination with sentiment/semantic features (i.e., WSD.Sentiment/Semantic) in the four datasets. The majority accuracy (i.e., Majority.Acc) is the accuracy when we predict all test instances to a major class.

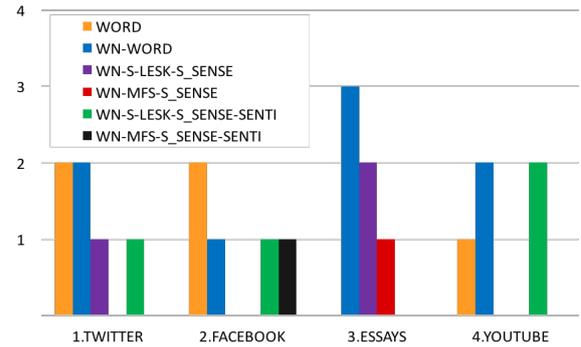


Figure 3: The overall number of times that each feature setting achieves the best performance in the four datasets.

ing observations. For example, for predicting conscientiousness, openness and agreeableness personality traits, using the WSD algorithm always decreases the performance across all datasets, while the prediction performance on extraversion and neuroticism improves 75% cases. The restriction to WordNet-only words is helpful in $10/24 \approx 41\%$ of the cases, especially on ESSAYS dataset. It is noteworthy that the $S-LESK$ related settings (i.e., $S-LESK-S_SENSE$ and $S-LESK-S_SENSE-SENTI$) perform better than MFS related settings (i.e., $MFS-S_SENSE$ and $MFS-S_SENSE-SENTI$).

5.3 Experimental Result Analysis

For the classification results, we have the following two observations: a) The restriction to WordNet words (WN-WORD vs. WORD) helps the most datasets (3 out of 4 datasets) for predicting openness and agreeableness. b) The positive effects of SENTI features on predicting neuroticism (2 out of 4 datasets). Detailed analysis are presented in the following paragraphs.

Impact of word feature (WORD)

We observe that in the all-words approach, there are many pronouns in the top-ranked features. The pronouns are later removed when filtering for WordNet words only. The experimental results show that removing these high-ranked features (e.g., pronouns, particles, and punctuation) increases the accuracy on ESSAYS dataset in all cases, while for other three datasets the feature impact varies based on different data. One possible explanation is that the essays are written in a more thoughtful manner, focused on the inner thoughts. They may, therefore, carry more personality-related information in the content words than the social media data, where the interjection and smileys are more revealing than the topic under discussion. Restriction to WordNet words only thus helps in the essays to better represent the document.

Impact of sentiment feature (SENTI)

In the *WSD-S.SENSE-SENTI* setup, a better result is achieved on cNEU label since *neuroticism* people tend to use more emotional words (Pennebaker and King, 1999).

Comparison with the state-of-the-art results

Table 3: Performance in comparison with the state-of-the-art results on the FACEBOOK dataset.

Trait	Majumder et al. (2017)	Ours (Majority.Acc)
cOPN	62.68	72.10 (70.40)
cCON	57.30	56.80 (52.00)
cEXT	58.09	62.10 (38.40)
cAGR	56.71	55.80 (53.60)
cNEU	59.38	61.70 (39.60)
Avg	58.83	58.64 (50.80)

Given our purpose is not about competing for performance but rather exploring the effectiveness of the general lexical-resources in APC. However, in Table 3, we draw a comparison with the recent best results of Majumder et al. (2017) to show that we get very competitive results on the FACEBOOK dataset. This is a very fair comparison since Majumder et al. used exactly the same evaluation settings as ours. It is worth to mention that, Majumder et al. (2017) used complex neural network models while we used the simple SVM model without tuning parameters. For other datasets, it is difficult to show a fair comparison since previous works (e.g., Farnadi et al. (2016)) regard the APC task as a linear regression problem instead of classification.

5.4 Discussion on Different Pipeline Settings

Figure 3 shows the ratio of the number of times each feature setting achieves the best performance over other pipelines in each dataset. In the picture, we can see the WN-WORD setting works well most of the time across four datasets. Therefore, the restriction to WordNet words is a low-cost and effective process to improve personality prediction.

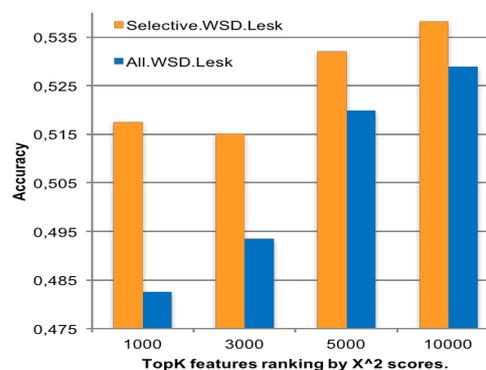


Figure 4: A test on cEXT personal trait of ESSAYS dataset to compare between Selective.WSD and All.WSD.

Impact of WSD on APC

We found that the WSD does not generally lead to an improvement in classification results except arbitrary dataset-specific differences, which can be largely attributed to the lemmatization and POS tagging. However, in contrary to previous beliefs (Sanderson, 1994; Gonzalo et al., 1998), the performance

WORD	χ^2	WN-WORD	χ^2
love	.012	love	.026
boyfriend	.008	music	.010
'd	.008	sleep	.009
me	.007	assignment	.009
so	.006	proud	.008
people	.006	boyfriend	.007
much	.005	worry	.007
we	.005	people	.007
thinks	.005	awkward	.007
WN-MFS	χ^2	WN-S-LESK	χ^2
love _{1v}	.016	love _{1v}	.017
music _{1n}	.009	assignment _{1n}	.009
guy _{1n}	.009	sleep _{1v}	.008
good _{1a}	.009	street _{4n}	.007
proud _{1a}	.008	love _{1n}	.006
assignment _{1n}	.008	sleep _{1n}	.006
boyfriend _{1n}	.008	music _{1n}	.005
real _{1a}	.006	good _{6a}	.005
sleep _{1v}	.006	proud _{3a}	.004

Table 4: The highest ranked features for Extraversion on the ESSAYS dataset, averaged across the 10 cross-validation folds, using the χ^2 feature selection.

of the WSD algorithms is not the major issue for stagnating performance. Rather, it is the reduction of the representative scope of bag-of-words (since function words are not present in the lexicon) and the reduction of the impact of multi-POS words (since those are assigned different senses), which leads to a lower ranking of otherwise highly predictive features. For example, in table 4, in the WN-WORD setup, the word *worry* is ranked to predict *extraversion* with $\chi^2 = .007$, while the sense *worry_{1v}* is ranked to predict *introversion*, i.e., the opposite class of *extraversion*, with $\chi^2 = -.004$. Furthermore, as pointed out in (Gale et al., 1992), if a polysemous word appears two or more times in a discourse, it is likely that all the occurrences will share the same coarse-grained sense. A fine-grained WSD might be therefore counter-productive. However, while the effect of WSD itself in a BoW setup is marginal, we observe that the WSD quality is rather high. This implies that the assigned senses can be reliably used to query additional information about the word meaning (and relations to other words) from the lexical-semantic resources.

Improved impact of WSD

In a more complex setting of WSD, we can partially resolve the issue mentioned above by (1) applying the *Selective.WSD* method and (2) combining WSD with semantic and/or sentiment information. Firstly, in Figure 4, we showed that the *Selective.WSD* method works better than the normal WSD method in selecting sense-level features for the APC. Especially, when we increase the number of topK features, the performance will drop. The reason for this difference was discussed in subsection 4.1. Secondly, we performed various experiments to show the benefit of combining WSD with semantic and sentiment features. Figure 2 indicates the differences between using WSD with semantic and/or sentiment features versus not-using WSD. Briefly, the combination of WSD with semantic and/or sentiment information works better in two cases of less-noise UGC data including ESSAYS and FACEBOOK on cEXT and cNEU personal trait. Our analysis shows that this is because cEXT and cNEU people use more pronouns and emotional words than other personal traits.

6 Conclusion

This paper presents extensive experiments to explore the lexical-semantic resources on APC. Especially, WSD is combined with semantic and sentiment information to pose an improved performance in APC. In summary, we draw the following major conclusions. Firstly, using a dictionary (e.g., WordNet, WiktionaryEN) to remove noise-features often works well in most datasets. Secondly, applying WSD alone, in general, does not work in APC, especially on not-well-written UGC data. However, our proposed *Selective.WSD* works better than a basic WSD. Thirdly, applying WSD combining with semantic and/or sentiment features improve the performance for specific personal traits (i.e., cNEU, cEXT). Moreover, no personality specific resources are required in our method.

Acknowledgments

This work has been supported by the German Research Foundation under grant No. GU 798/14-1 and by Umeå University on federated database research.

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.
- Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. 2012. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference, WebSci '12*, pages 24–32.
- J.I. Biel, O. Aran, and D. Gatica-Perez. 2011. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM)*, pages 446–449.
- P.T. Costa and R.R. McCrae. 2008. The revised neo personality inventory (neo-pi-r). *SAGE Handb. Pers. Theory Assess.*, pages 179–198.
- G. Farnadi, S. Zoghbi, M. Moens, and M. De Cock. 2013. Recognising personality traits using facebook status updates. pages 14–18.
- Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine Cock. 2016. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, pages 109–142.
- Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1805–1816.
- William Gale, Kenneth Ward Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*, pages 415–439.
- William L. Gardner, Brian J. Reithel, Claudia C. Coglisier, Fred O. Walumbwa, and Richard T. Foley. 2012. Matching personality and organizational culture. *Management Communication Quarterly*, 26(4):585–622.
- J. Golbeck, C. Robles, and K. Turner. 2011. Predicting personality with social media. In *Proc. of the 2011 annual conference extended abstracts on Humam factors in computing systems*, pages 253–262.
- F. Verdejo Gonzalo, I. Chugur, and J. Cigarín. 1998. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 38–44.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590.
- Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, pages 197–204.
- F. Iacobelli and A. Culotta. 2013. Too neurotic, not too friendly: structured personality classification on textual data. In *Proceedings of the Workshop on Computational Personality Recognition*, pages 19–22.
- Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, pages 568–577.
- Athanasios Kehagias, Vassilios Petridis, Vassilis G. Kaburlasos, and Pavlina Fragkou. 2003. A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, pages 227–247.
- François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, pages 457–500.
- N. Majumder, S. Poria, A. Gelbukh, and E. Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, pages 74–79.
- Tristan Miller, Nicolai Erbs, Hans-Peter Zorn, Torsten Zesch, and Iryna Gurevych. 2013. DKPro WSD: A generalized UIMA-based framework for word sense disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 37–42.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM Vol. 38*, pages 39–41.
- Saif Mohammad, Xiaodan Zhu, and Joel Martin, 2014. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, chapter Semantic Role Labeling of Emotions in Tweets, pages 32–41.

- Alessandro Moschitti and Roberto Basili. 2004. *Complex Linguistic Features for Text Classification: A Comprehensive Study*, pages 181–196.
- Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15*, pages 15–22.
- J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, pages 1296–1312.
- J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2007. Linguistic inquiry and word count: Liwc [computer software].
- Marten Postma, Ruben Izquierdo Bevia, and Piek Vossen. 2016. More is not always better: balancing sense distributions for all-words word sense disambiguation. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3496–3506.
- T. Rose, M. Stevenson, and M. Whitehead. 2002. The reuters corpus volume 1 from yesterday's news to tomorrow's language resources. *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC*, pages 827–832.
- Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94*, pages 142–151.
- Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the pan/clef 2015 evaluation lab. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF'15)*, pages 518–538.
- D.J. Stillwell and M. Kosinski. 2015. mypersonality project. <http://mypersonality.org/>.
- Laura Uba, 2003. *Asian Americans: Personality Patterns, Identity, and Mental Health*. Psychology, Guilford Press.
- P. Vossen, G. Rigau, I. Alegria, E. Agirre, D. Farwell, and M. Fuentes. 2006. Meaningful results for information retrieval in the meaning project. In *Proceedings of the 3rd Global WordNet Conference*, pages 22–26.
- Xuan-Son Vu and Seong-Bae Park. 2014. Construction of vietnamese sentiwordnet by using vietnamese dictionary. *Proceedings of the 40th Conference of the Korea Information Processing Society*, pages 745–748.
- Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, pages 6–10.
- Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97*, pages 412–420.

Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish

Bolette S. Pedersen¹, Manex Agirrezabal², Sanni Nimb³, Sussi Olsen⁴, Ida Rørmann⁵

University of Copenhagen^{1,2,4,5} & The Danish Society for Language and Literature³

Njalsgade 136, DK-2300 Copenhagen S^{1,2,4,5}, Christians Brygge 1, DK-1219³

bspedersen@hum.ku.dk, manex.agirrezabal@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, idaroermannol-
sen@gmail.com

Abstract

Our aim is to develop principled methods for sense clustering which can make existing lexical resources practically useful in NLP – not too fine-grained to be operational and yet fine-grained enough to be worth the trouble. Where traditional dictionaries have a highly structured sense inventory typically describing the vocabulary by means of main- and subsenses, wordnets are generally fine-grained and *unstructured*. We present a series of clustering and annotation experiments with 10 of the most polysemous nouns in Danish. We combine the structured information of a traditional Danish dictionary with the ontological types found in the Danish wordnet, DanNet. This constellation enables us to automatically cluster senses in a principled way and improve inter-annotator agreement and wsd performance.

1 Lexical resources and word sense disambiguation (WSD)

Dealing with finegrained lexical sense inventories in NLP is a challenging task. Selecting the correct sense in a specific context is incredibly hard when word meaning is richly described with subtle and detailed sense distinctions as found in most wordnets and lexica.

To this end, coarse-grained word-sense disambiguation has become a well-established discipline over the years. One way to obtain a coarse-grained sense inventory is to cluster existing inventories either manually or automatically (Pedersen et al. 1998, Lapata & Brew 2004, Alvez et al. 2008, Izquierdo et al. 2009, McCarthy et al. 2016).

In recent years, also so-called supersense tagging has become popular where WordNet's *first beginners*¹ are applied as a cross-lingual sense inventory. In recent experiments on Danish cor-

pora we achieved state of the art results in both annotator agreement and automatic supersense tagging (Alonso et al. 2015 and 2015b, Pedersen et al. 2016). Nevertheless, our experiments also demonstrated that the inventory was not particularly well suited for our purpose. First of all, the inventory proved *too* coarse in a considerable number of cases (see Alonso et al. 2016 for a discussion), and secondly, the set did not facilitate annotations across part-of-speech as in the case of de-verbal nouns resulting in unbalanced annotations between nouns and verbs.

In the present work, we pursue a slightly different path by returning to the monolingually and corpus-defined sense inventory of our monolingual lexical resources, the Danish wordnet, DanNet, and The Danish Dictionary (Den Danske Ordbog, DDO) on the basis of which DanNet was originally compiled (Pedersen et al. 2009). Our aim is to further examine the potential of a principled method for sense clustering to be performed automatically on existing fully-fledged sense inventories. The basic idea is to combine the structured information of a traditional Danish dictionary with the ontological types found in the Danish wordnet, DanNet, and to develop clustering methods on this basis.

For our lexical sample study, we select 10 of the most polysemous nouns in Danish; we study how the senses are organized in DDO and DanNet and how they can be automatically clustered following two different principles: one allowing for clusters only within the same main sense, and one where also clustering of main senses are allowed except for the cases of homographs. For both sense inventories we perform manual annotation and word sense disambiguation using the LibLINEAR package and compare the results.

¹ Cf. <https://wordnet.princeton.edu/man/lexnames.5WN.html>

2 Sense organization in DDO and Dan-Net

2.1 Senses in DDO

Senses in DDO are according to normal convention organized in main- and subsenses as depicted in figure 1 for the lemma *vold* ('violence'):

vold¹ substantiv, fælleskøn

Vis overblik

BØJNING -en
UDTALE [vɑʌ]
OPRINDELSE fra nedertysk *wal*, oprindelig lånt fra latin *vallum* 'forskansning, pallsadeværk', afledt af *vallus* 'pæl'

Betydninger

- handling eller adfærd som indebærer brug af fysisk magt beregnet på at beskadige, sære eller dræbe nogen
 - 1.a JURA angreb på en anden persons legeme
 - SYNONYMER legemskrænkelser | legemsbeskadigelse | SE OGSÅ voldtægt
 - GRAMMATIK vold mod NOGEN/NOGET
 - EKSEMPLER brutal vold | meningsløs vold | fysisk vold | rå vold | stigende vold | politisk vold | trusler om vold | bruge vold | øve/begå vold | anvendelse af vold | krig og vold
 - Mange mennesker er parate til at anvende vold, når de kommer ud for et problem eller en konflikt [skoleb-rel.92](#)
 - 1.b handling eller adfærd der udgør et overgreb mod et andet menneskes natur og integritet
 - BESLÆGTEDE ORD psykisk vold
 - Passiv psykisk vold foreligger, hvis barnets forældre ikke er i stand til at stimulere barnet og give barnet den nødvendige omsorg og kærlighed [DenSocLinje1992](#)
 - 1.c OVERFØRT overgreb der krænker en rettighed, kultur, tradition el.lign.
 - SE OGSÅ gøre/øve vold mod/på
 - Hun lavede te til dem alle tre, vaskede op, mens Gnags overdøvede deres larmende diskussioner om familiens vold mod børns fantasi [TiBryld84](#)
 - 1.d brug af fysisk kraft eller anstrengelse rettet mod en ting
 - SYNONYMER magt
 - EKSEMPLER med vold
 - Han åbner brevet med vold og læser det hurtigt igennem [SvHolm87](#)
- kontrol eller herredømme som en stærk person eller magt har over nogen
 - GRAMMATIK i NOGEN s/NOGETS vold
 - I 25 timer var han i rockernes vold [BT1991](#)
 - når hadet greb hende, var hun helt i sine følelsers vold [faqb-litt.84a](#)

Figure 1: Main- and subsenses in DDO of *vold* (violence, rampart, bank ..) in its violence sense.

In cases of homography where two lemmas take the same form without sharing etymology, two separate entries are established; in this case also an entry for the lemma *vold* in the sense of 'rampart' (Figure 2).

vold² substantiv, fælleskøn

BØJNING -en, -e, -ene
UDTALE [vɑʌ]
OPRINDELSE fra nedertysk *wal*, oprindelig lånt fra latin *vallum* 'forskansning, pallsadeværk', afledt af *vallus* 'pæl'

Betydninger

- aflang forhøjning af opdyngnet jord, ofte forstærket med sten eller murværk og brugt som del af et fæstningsanlæg omkring en by eller en borg + bruges i vore dage ofte som rekreativt område [SPRØGBRUG](#) især historisk
 - Rundt om i Danmark ligger der flere hundrede kilometer af gamle volde fra jernalderen og vikingetiden [skoleb-hist.92c](#)
 - i bestemt form ofte om en middelalderlig bykernes afgrænsning 1852-67 frigaves arealet uden for voldene, hvorved Nørrebro, Vesterbro og Østerbro fik bymæssig bebyggelse [Fakta1988](#)
 - 1.a aflang forhøjning af opdyngnet jord, sten eller andet materiale, anvendt til ikke-militære formål, fx naturbeskyttelse eller opdæmning
 - BESLÆGTEDE ORD
 - Allerede før Kristi fødsel havde de japanske marker fået deres karakteristiske form, små jordstykker omgivet af volde og forsynet med vand fra mange kanaler [læreb-hist.88b](#)
 - 1.b OVERFØRT værn imod noget ubehageligt, uønsket el.lign.
 - Viveca prøvede at opbygge en uoverstigelig vold imellem dem, som kun en uimodståelig kærlighed kunne nedbryde [SvAMad89](#)

Figure 2: Main- and subsenses in DDO of *vold* (violence, rampart, bank ..) in its 'rampart' sense.

The overall principle for organizing senses within the same lemma follows Cruse (2000) by identifying different kinds of relations between main and subsenses:

- *Auto-hyponymy*: narrowed meaning with same hypernym, as in *to drink alcohol* as a subsense to *to drink*
- *Auto-superordination*: extended meaning with same hypernym as in *man* (male) vs *man* (person)
- *Auto-meronymy*: a part instead of the whole as in *door* meaning a piece of wood, metal or the like in contrast to *door* in the broader opening sense (as in *the door was made of wood* vs. *he closed the door*).
- *Auto-holonymy*: a whole instead of the part as in *body* meaning the whole body in contrast to *body* in the sense of the torso only.
- *Figurative*: sense where only part of the meaning (often its function) is derived from the core sense but used in a figurative/metaphorical context as in *window* in the sense *a window to the world*.

However, also the frequency of the senses (annotated in a set of randomly selected concordance lines (100-200 examples) from a balanced corpus of 40 mill. tokens (DDO Corpus (Norling-Christensen & Asmussen 1998)) was taken into consideration, as well as the communicative effect of the structure. The overall goal was to compile an 'easy to read' printed dictionary, es-

pecially by avoiding very deep sense structures. These two aspects considered, the relational principles defining subsenses to a particular main sense were not always followed. While figurative senses are typically described as subsenses to their main sense, frequent subsenses with a *non-figurative* relation (i.e. one the 4 ‘auto’-relations above) to the main sense were in fact in several cases described as an additional main sense instead of a subsense.

One example is the verb *æde* of which the first main sense describes the eating act of animals, whereas the second describes the eating act of humans, although the second is semantically derived from the first and therefore ought to be described as a subsense.

In other words, the semantic relatedness between word senses which we are looking for in order to be able to cluster senses in a principled way, is not always completely well reflected in the structure of the DDO entry. This inconsistency in structure – which is well-argued and also to our knowledge normal practice in lexicography – indicates why reuse of existing lexical resources in NLP is not just a straight-forward task. It also indicates that more than one experiment should preferably be performed; one where clusters are only established within main senses, and one where clustering also takes place across main senses (see Section 3).

2.2 Senses in DanNet

Senses in DanNet are organized in terms of synsets as in standard in wordnets (Fellbaum 1998). Each synset is assigned an ontological type based on EuroWordNets' top ontology, cf. Vossen 1999).

In contrast to the structure of a conventional dictionary where senses are typically organized in main and subsenses, the synsets that constitute the wordnet all have equal status. Further, each synset is inter-related to other synsets via semantic relations as shown in Figure 3.

slag 7

(lang) vid beklædningsgenstand som ikke har ærmer, ...

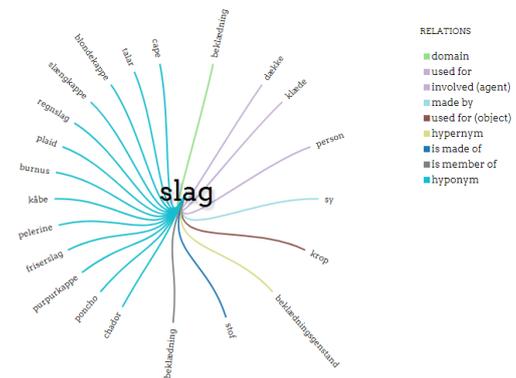


Figure 3: *Slag* in DanNet in its 'cape' sense and corresponding semantic relations

All synsets in DanNet are further assigned a complex ontological type following The EuroWordNet top-ontology (Vossen 1999) as depicted below in Figure 4 and 5.

Origin	Natural	Living	Plant	Human	Creature	Animal
Form	Substance	Solid	Liquid	Gas	Object	
Composition	Part	Group				
Function	Vehicle	Representation	MoneyRepresentation	LanguageRepresentation	ImageRepresentation	
	Software	Place	Occupation	Instrument	Garment	Furniture
	Covering	Container	Comestible	Building		

Fig. 4: Ontological assignments to 1st Order Entities (cf. Vossen 1999:139)

SituationType
Dynamic
BoundedEvent
UnboundedEvent
Static
Property
Relation
SituationComponent
Cause
Agentive
Phenomenal
Stimulating
Communication
Condition
Existence
Experience
Location
Manner
Mental
Modal
Physical
Possession
Purpose
Quantity
Social
Time
Usage

Fig. 5: The EuroWordNet Top Ontology for 2nd and 3rd Order Entities cf. (Vossen et al. 1999:139)

Since our aim is to establish principled methods for sense clustering, it should be noted that the distinction between word senses is in several cases more fine-grained in DDO than the distinction between synsets in DanNet. This means that sometimes senses of the same word in DDO are in fact already members of the same synset in DanNet. These clusters were based on an idiosyncratic lexicographic judgment at the time of compilation of each synset but goes well in line with the more principled approach to sense clustering established here.

3 Establishment of clusters

Following the line of the discussion in Section 2, it does not seem appropriate just to collapse all DDO subsenses with its main sense; this would leave all metaphorical senses (which are indeed very frequent in our corpus) very poorly represented. We combine the information types from both resources: The DDO and DanNet and to this end, we perform three annotation experiments:

- Experiment 1 ('regular') where all main and subsenses are maintained.
- Experiment 2 ('clustered') where subsenses are clustered if they are of the same ontological type, and

- Experiment 3 ('clustered reduced') where also main senses are clustered if they are of the same ontological type.

Even if the ontology enables groupings of synsets which are ontologically similar (for instance artifact/part of artifact artifact/group of artifacts, person/groups of persons), we have in these experiments adopted a rather conservative approach and only clustered senses with the exact same ontological type.

Often a narrowed or an extended sense will have the same ontological type, in other cases a similar one. In contrast, figurative senses are typically of a completely different ontological type and are preserved with this method.

	Ex. 1 regu- lar	Ex. 2 clustered	Ex. 3 clustered reduced
<i>Selskab</i> (company, party, asso- ciation)	10	6	5
<i>Plads</i> (room, space, square, post)	13	9	6
<i>Slag</i> (battle, stroke, cape)	17	11	10
<i>Skud</i> (shot, shoot, dosis)	12	12	11
<i>Skade</i> (harm. injury, magpie, skate)	6	5	4
<i>Kort</i> (card, map)	10	4	3
<i>Vold</i> (vio- lence, bank)	9	7	5
<i>Hul</i> (hole, gap)	14	11	8
<i>Blik</i> (look, glace, tin)	7	6	4
<i>Model</i> (model, pattern, design)	8	7	6

Table 1: Number of sense clusters in ex. 1- 3 excluding idiomatic expressions which do not cluster

4 Corpus and annotation

The texts selected for annotation have been extracted from the 45 million words CLARIN Reference Corpus (Asmussen 2012). This corpus comprises a wide variety of text types and domains: blog, chat, forum, magazine, Parliament debates (written down by professionals), and newswire, of which the latter constitutes 48 % of the entire corpus. In line with the Senseval approach (www.senseval.org), the number of annotated sentences for each noun varies according to the number of DDO senses of the noun ($100 + 15 \cdot \text{no. of senses}$), resulting in 177 to 535 sentences per noun.

It turned out that the otherwise very frequent nouns that we selected are not very frequent in social media texts, and since it is important for the project to have all text types including social media represented in the annotated data, all sentences from this text type that contained the noun in question were extracted from the corpus. Still to reach the specified number of sentences for each noun, we ended up with a majority of sentences from newswire texts.

For the annotation task we used the tool WebAnno (Yimam et al., 2013), which facilitates calculation of the inter-annotator quality and adjudication of the annotated files. For each occurrence of the word to be annotated, the annotators select a sense from the list of clustered senses in a drop down menu, see fig. 6.

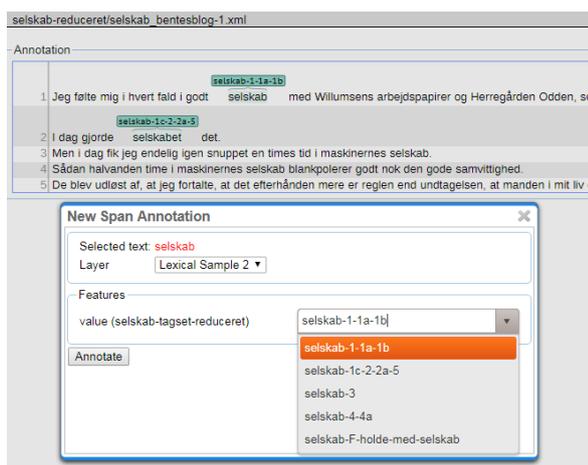


Fig 6: WebAnno annotation of *selskab* (company, party, association ..).

4.1 Annotation results

All sentences have been doubly annotated by advanced students and researchers and around 2% of the examples have been curated. The re-

sults from the three annotation experiments can be seen in Figure 7.

We apply Krippendorffs α (cf. Krippendorffs 2011) which calculates chance corrected agreement coefficients, i.e. sets off the fact that it is easier to agree on few tags than on many. Values range from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. It is customary to require $\alpha \geq .80$ in most annotations tasks, however, for sense annotation where more tentative conclusions are still acceptable, we consider $\alpha \geq .67$ reasonable and useful. With this measure, as can be seen, only experiment 3 achieves 'acceptable' intercoder agreement for all words².

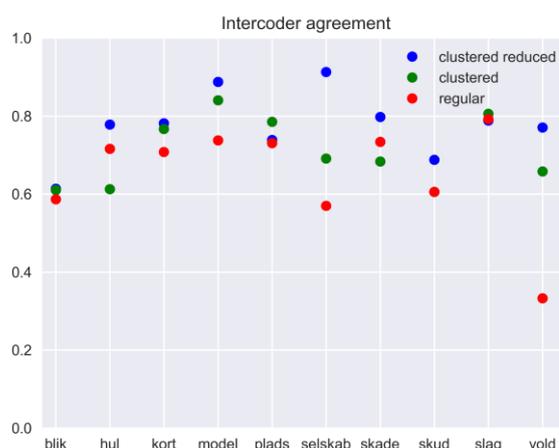


Fig. 7: Intercoder agreement (IA) (Krippendorffs α) in experiment 1-3

When curating 2% of the annotated material, we observed three kinds of discrepancies among annotators:

- *Plain errors*: Diverging annotations due to wrong pos tags or because the annotator had erroneously skipped a word, for instance in cases with more than one lexical occurrence per sentence.
- *Incomplete or unclear tag set*: Diverging annotations in cases where a new/unconventional sense of the word was not covered by the tag set, or where the lexical description of a tag was unclear or blurred.
- *Underspecified examples*: Diverging annotations where the precise word sense

² It should be noted that we are here dealing with some of the most complex and polysemous words in Danish; i.e. agreement measures will most presumably differ for the rest of vocabulary.

could not be deduced from the isolated example (most divergences).

The annotators report that the annotations tasks are generally hard and that they are often in doubt, in particular when annotating with the full sense inventory where the distinctions are often very subtle. In contrast, they report that the generated clusters are somewhat more intuitive for them to work with, a fact which is reflected in an increased annotator agreement for the clustered senses, and also an increased agreement from experiment 2 to experiment 3.

One example is *selskab* (company, association, party) where groups of people doing things together can be more or less temporary resulting in different senses in the fine-grained experiment – but in only one cluster in the cluster experiments; a fact which increased agreement quite a lot. Further, where some clusters at first sight seem awkward, they often prove to ease annotation substantially. An example is *plads* which with its 'space' sense as a physical space/room/area is clustered with the 'square' sense as an urban, open area, square or field. Even though there are slightly different associations with these two senses it proves quite convenient to think of them as part of the same 'physical' cluster. Another noteworthy issue is the associations that we make regarding the digital universe, as in *plads på harddisken* (disc space) or *plads på skrivebordet* (space on the (computer) desktop). Are these examples abstract or concrete? Inter-coder disagreement proves that annotators are in doubt.

In some cases, annotators report that clusters are really too coarse in experiment three, as exemplified with *kort* (card, map ..) where two very different kinds of artifacts are clustered (playing cards and maps) because they are of the same ontological type: Image Representation.

In a few cases, however, the ontologically based cluster separations seem to play a minor role. The ontological types of *fysisk skade* (physical injury/damage) and *psykisk skade* (psychological injury/damage) differ, where a psychological injury is more abstract and non-physical. But is this distinction really crucial? One can argue that the association of being injured, in either one of these ways, is more relevant to the context than whether the damage is physical or not, a fact which is demonstrated by quite a lot of underspecified corpus examples leading to

disagreement among annotators because they had to choose one or the other.

Finally, the annotators meet a dilemma when dealing with metaphors. In the metaphor '*et skud i bøssen*' (one shot left), expressing one's only chance, the word *skud* is not the actual bullet, but rather the figurative sense of a chance. It is important to have a consensus of whether to stay inside the metaphorical picture and annotate within it, or whether to annotate with the actual intention. We chose consensus regarding the former solution, but still these cases lead to disagreement a number of times.

5 Word sense disambiguation using the LibLINEAR package

We also perform an experiment to see how empirical methods can perform in such hard tasks. The task is to disambiguate some specific words in a sentence (lexical sample task), and to see if there is any significant improvement of the prediction accuracies, when using clustered word senses.

The features that we use include a bag of lemmas of the whole sentence. We also include the next and previous four lemmas. These last elements are devised to disambiguate idiomatic expressions whose structure is mostly fixed.

As currently the data includes information from several annotators, training and evaluating Machine Learning classifiers is not straightforward. The main problem is the evaluation of a model. If two or more annotators have tagged a word in a sentence with diverging sense cluster tags, we consider it correct if an ML classifier classifies that instance as one of those sense clusters (either of them). This corresponds well to the fact that most divergences are caused by underspecified corpus examples. For learning, if two different annotators have tagged an instance, we consider it to be two different instances, resulting in some cases where we can have two instances with the same attributes, but with different outputs.

As the amount of data is limited, we decided to perform a 5-Fold Cross-Validation to check if the classifiers work sufficiently. We train a Linear Support Vector Machine for its robustness when used with a high number of features.

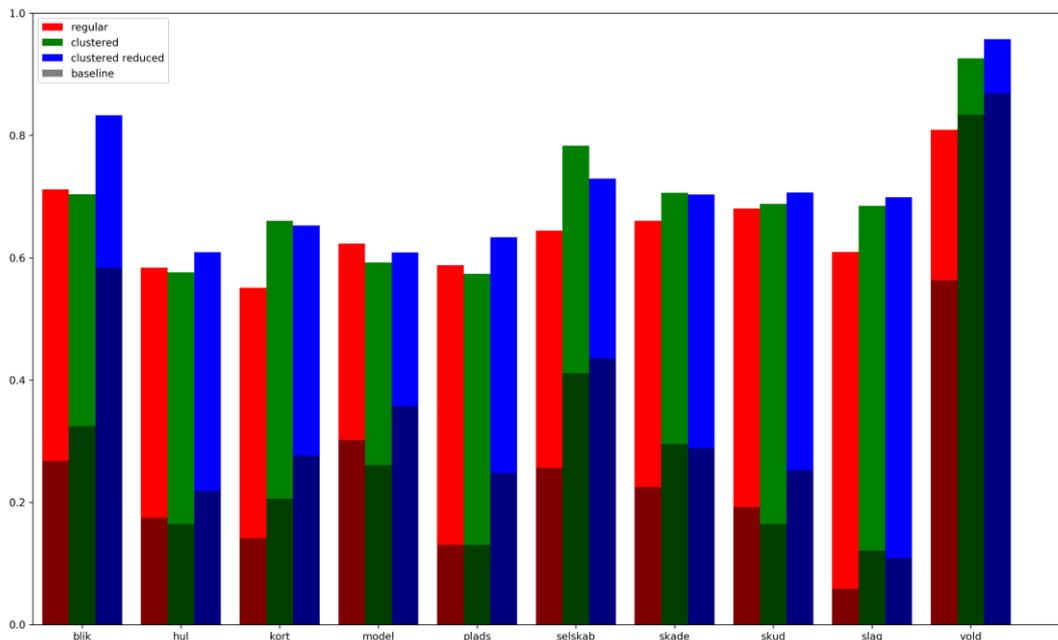


Fig. 8: Accuracies of the three experiments (regular, clustered, reduced clusters) compared to a baseline.

The toolkit that we employ is the well-known LibLINEAR package³ (Fan et al. 2008), included in the module *scikit-learn* (Pedregosa et al. 2011) from Python.

Accuracies of the word disambiguation tasks with the three types of sense inventories compared to a baseline are provided in Figure 8. On average, reduced clusters can be seen to outperform the experiments with the more fine-grained sense inventories.

6 Concluding Remarks

In this paper we have examined how we can cluster noun senses in a principled way based on dictionary and wordnet information in combination (main and sub-senses versus ontological typing). We have dealt with some of the hardest and most polysemous nouns in Danish. We have further examined how systematically clustered noun senses influence inter-annotator agreement and automatic word sense disambiguation in a positive way, resulting in our last experiment (reduced clusters) in a sense inventory which seems actually manageable and well-functioning for both the annotators and the automatic disambiguation system. How our method will apply to verbs and adjectives is still an open question; for these word classes other information types than ontological typing may be more crucial.

It would also be interesting in future work to study how principled clustered based on lexicons and wordnets as presented in this paper compare to the word profiles that appear with word embeddings and sense induction methods.

Finally, only little space has however been left to discuss to which extent the meaning distinctions that are established by our clustering methods are actually relevant. Relevance depends on our purpose and on the kind of language technology service we are aiming at, where translation generally demands a high degree of detail, information search quite less, and question answering maybe something in between. In future work we would like to include relevance criteria as a more dominant feature encompassing also elements such as sense frequency and predominance information of senses; information which is however not directly accessible for Danish at the current stage.

Acknowledgements

Thanks to Anna Braasch, Sara Lee Naldal, and Ida Hauerberg Wolthers for assisting with the sense annotations.

³ <https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

References

- Alvez, Javier, Jordi Atserias, Jordi Carrera, Salvadoar Climent, Egoitz Laparra, Antoni Oliver, German Rigau (2008). Complete and consistent annotation of wordnet using the top concept ontology. *LREC Proceedings* 2008.
- Asmussen, J. (2012). CLARIN-Referencekorpus. Sprogteknologisk Workshop October 31, 2012, University of Copenhagen. <http://cst.ku.dk/Workshop311012/sprogtekno2012.pdf>
- Cruse, D.A (2000). *Meaning in Language*. Oxford: Oxford University Press.
- DDO = *Den Danske Ordbog*. (E. Hjorth et al). 2003-2005. Det Danske Sprog- og Litteraturselskab & Gyldendal, Copenhagen.
- Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug), 1871-1874.
- Fellbaum, Christiane (ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.
- Izquierdo, Rubén, Armando Suárez, and German Rigau. (2009). An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* pp 389-397. The Association for Computational Linguistics.
- Krippendorff, K. (2011). Agreement and Information in the Reliability of Coding. In: *Communication Methods and Measures* 5 (2) pp: 93-112.
- Lapata, Mirella and Chris Brew (2004). Verb Class Disambiguation Using Informative Priors. *Computational Linguistics*, 30(1): 45-73.
- Martínez Alonso, Héctor; Anders Johannsen; Sussi Olsen; Sanni Nimb; Nicolai Hartvig Sørensen; Anna Braasch; Anders Søgaard; Bolette Sandford Pedersen. (2015). Supersense tagging for Danish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015, Linköping Electronic Conference Proceedings #109*, ACL Anthology, Linköping University Electronic Press, Sweden.
- Martínez Alonso, Héctor; Barbara Plank; Anders Johannsen; Anders Søgaard. 2015b. Active learning for sense annotation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015, Linköping Electronic Conference Proceedings #109*, ACL Anthology, Linköping University Electronic Press, Sweden.
- Martínez Alonso, Héctor; Anders Johannsen; Sanni Nimb; Sussi Olsen; Bolette Sandford Pedersen. 2016. An empirically grounded expansion of the supersense inventory. In *Proceedings of Global Wordnet Conference 2016*.
- McCarthy, Diana, Marianna Apidianaki & Katrin Erk (2016). Word Sense Clustering and Clusterability. In: *Computational Linguistics*, Vol. 42, no. 2.
- Norling-Christensen, Ole & Jørg Asmussen: The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8 1998, 223–242
- Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen. 2009. DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation, Computational Linguistics Series*, pp.269-299.
- Pedersen, Bolette Sandford; Braasch, Anna; Johannsen, Anders Trærup; Martínez Alonso, Héctor; Nimb, Sanni; Olsen, Sussi; Søgaard, Anders; Sørensen, Nicolai. 2016. The SemDaX Corpus - sense annotations with scalable sense inventories. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*. Portorož, Slovenia.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
- Peters, Wim, Yvonne Peters & Piek Vossen (1998.). Automatic sense clustering in EuroWordNet. In: *First International Conference on Language Resources & Evaluation 1998*, Granada, Spain.
- Vossen, P (ed). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.
- Yimam, S.M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *Proceedings of ACL-2013*, demo session, Sofia, Bulgaria.

WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation

Tomasz Naskręt¹, Agnieszka Dziob¹, Maciej Piasecki¹, Chakaveh Saedi², António Branco²

¹G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland

²NLX-Natural Language and Speech Group, Department of Informatics
University of Lisbon, Faculty of Sciences, Portugal

{maciej.piasecki, agnieszka.dziob, tomasz.naskret}@pwr.edu.pl

{chakaveh.saedi, Antonio.Branco}@di.fc.ul.pt

Abstract

The paper presents a new re-built and expanded, version 2.0 of WordnetLoom – an open wordnet editor. It facilitates work on a multilingual system of wordnets, is based on efficient software architecture of thin client, and offers more flexibility in enriching wordnet representation. This new version is built on the experience collected during the use of the previous one for more than 10 years of plWordNet development. We discuss its extensions motivated by the collected experience. A special focus is given to the development of a variant for the needs of *MultiWordnet of Portuguese*, which is based on a very different wordnet development model.

1 Introduction

A wordnet is a complex graph of several types of nodes (e.g. lexical units¹, synsets) and edges (e.g. lexical relations, synset relations). Initially Princeton WordNet development was based on manual editing of text files storing wordnet representation (Fellbaum, 1998). Such an approach was error prone and the files edited manually required a lot of error verification and maintenance. At the beginning of the plWordNet project in the year 2005, we developed a wordnet editing system, called WordnetLoom in order to avoid problems with manual editing of wordnet representation. It was based on a database and Graphical User Interface (GUI), and separated users from the internal representation of the wordnet. As plWordNet was developed by a team of linguists, it was important to provide distributed access to the system. WordnetLoom has been constructed in a way providing support for the corpus-based wordnet de-

velopment method used for plWordNet (Maziarz et al., 2013); i.e. enabling close association between editors' decisions and language data, the use of substitution tests and application of semi-automatic methods as tools for editors. A unique feature of WordnetLoom is the possibility to simultaneously browse and edit wordnet graphs directly on the screen. Nevertheless, WordnetLoom was based on a quite inefficient thick client model, as well as it had restricted expressiveness of the applied wordnet representation and limited possibilities to adapt UI to the format extensions. Moreover, WordnetLoom was initially designed to support a monolingual wordnet. It was successfully used for editing plWordNet onto Princeton WordNet mapping, but the simultaneous presentation and editing of the two wordnets was due to a trick: introduction of additional 'English' PoS.

Our goal is to present a new re-built and expanded, version of WordnetLoom 2.0 facilitating work on a multilingual system of wordnets, based on an efficient software architecture of a thin client, and offering more flexibility in enriching wordnet representation. This new version originates from the experience collected during the use of the previous one that has clearly motivated the extensions. We will also discuss its applications and variants, with a special focus on the *MultiWordnet of Portuguese*.

2 Related Works

The first popular wordnet editor was probably *VisDic* (Horák and Smrž, 2004). In *VisDic* the relation definitions were still written in text windows, but an XML based format was utilised. *VisDic* was a monolithic application directly working on XML files, contrary to its descendant *DEBVisDic* (Horák et al., 2006) – a client-server, lexical database editor, based on a general platform for dictionaries called *DEB* (Horak et al.,

¹A triple: lemma, Part of Speech, sense id.

2008). DEBVisDic reimplemented and extended the functionality of VisDic, and offered also more flexibility in adapting XML representation structures. Data presentation was limited and there was no means for visual editing the relation structure. Several other wordnet editors also do not provide elaborated visualisation for wordnet structures, e.g. *Hydra* Rizov (2014) or *OMWEdit* (Morgado da Costa and Bond, 2015).

A web-based system *sloWTool* (Fišer and Novak, 2011, Fišer and Sagot, 2015) offers good UI and visual wordnet browsing and editing. However, presentation is always limited to a small fragment of the wordnet graph (up to two links distance) and there is no means for neither viewing larger parts, nor comparing different parts.

Visualisation of wordnet graphs in most tools follows a radial pattern: a synset in focus is presented in the middle and all links, irrespectively of their types are placed radially around the central element, e.g. *sloWTool* or *WordTies* (Pedersen et al., 2012). *GernEdit* (Henrich and Hinrichs, 2010) offers visualisation of the wordnet structure in the range selected by the user, but it is hierarchical and focused mainly on hypernymy. Moreover the visual presentation does not allow for direct editing of the structures. *WordnetLoom* introduced elaborated presentation of the relation graph and direct visual editing (Piasecki et al., 2013). As it is an open tool, it was used as a basis for the solution presented in this paper.

3 Basic Assumptions

WordnetLoom 1.0 has been used for plWordNet development since 2005 and proved to be a generally useful system. Thus, although software architecture has been reconstructed, the main philosophy of the system was preserved.

In order to avoid errors in the representation format, all editing actions should be done only via GUI client application and the results are stored in the central database. The XML-based format is secondary in relation to the database. WordnetLoom supports distributed group work by a group of linguists on the central database.

plWordNet construction has been following corpus-based wordnet development paradigm. Each iteration starts with the extraction of the most frequent lemmas from a large corpus together with the automated extraction of their semantic description, e.g. as a measure of semantic similarity. New

lemmas are divided into *packages* on the basis of similarity-based clustering. The packages are assigned to linguists as work assignments and presented in WordnetLoom.

Substitution tests² are an intrinsic part of the relation definitions. Test templates are kept together with the relation definitions in the database. Before every editing decision is made, a test for a relation considered by the linguist is presented in a pop-up window and instantiated with the lemmas from the two synsets to be linked.

A wordnet is a network of lexico-semantic relation, and a graph is the basic means for both browsing and editing the wordnet structure. A network of synsets linked by synset relations is visually presented on the screen as a graph. The user can freely browse the network by clicking on synsets and unfolding as many levels of relations as needed, see Fig. 2. Every link can be added or removed directly on the graph presentation. This facilitates better comprehension of the wordnet structure, shorter connection between the editing intention and the resulting change in the wordnet structure, as well as a better understanding of the consequences of the intended and/or performed action to the wordnet structure beyond the local connections of the edited synset.

The same system and the same presentation means should also support the construction of the mappings between wordnets. Thus wordnets for different languages should be presented simultaneously on the screen as graphs that are connected by inter-lingual relations which are also visually presented on the screen. The editing of the mapping is performed in a way similar to monolingual editing by linking synsets or deleting links selected on the screen with the mouse.

Every wordnet includes also elements of the description that are not relations but attributes, e.g.: glosses, usage examples, and different attributes, e.g. stylistic register, sentiment polarity etc. As this kind of information is getting richer with the subsequent versions of plWordNet, we need also to introduce different perspectives on wordnet, not only graph-based, but also more dictionary-oriented. It is not also possible to fit everything into one single screen graph-presentation – the graph would be too cluttered. Attributes for a synset in focus are presented in side panels. Word-

² Each consists of one or more test sentences with slots for the tested lemmas.

netLoom offers three main perspectives on data: the *perspective of lexical units, visualisation and synsets*. The perspective of lexical units presents the wordnet as a dictionary. The searching is focused on lexical units (henceforth, LUs) and their relations, for a selected LU all synsets which it belongs to are listed. In addition the complete description of its attributes and lexical relations is shown. The synset perspective is organised in similar way, but around synsets as basic elements, and the visualisation perspective presents visually wordnet as a network of synsets. For a synset in focus its LUs are presented in the side panels together with their lexical relations.

4 Graph-based Presentation

A wordnet is intrinsically a graph. Lexical meanings are described by subgraphs of lexico-semantic relations. Thus a visual presentation of the wordnet graph should be a basis for a wordnet editing system.

From a formal point of view, there are not many restrictions on the shape of the wordnet graphs. However, the semantics of the relations reveals two basic groups of wordnet relations: relations expressing some aspects of hierarchy (e.g. hypernymy/hyponymy, type/instance) and other relations (e.g. holo/meronymy). The former defines some levels: synsets located at the upper levels are more general, those on the lower – more specific. The latter group does not show any preference concerning the location of elements belonging to one link (a graph arc) on the screen.

In many systems, a wordnet graph is visualised in way following the radial scheme, i.e. for a synset in focus its nearest neighbours are presented around it in equal distance, e.g. (Fišer and Novak, 2011, Pedersen et al., 2012) or the system tries to cover equally the whole area of the screen. In both cases, the important characteristic features of the hierarchical relations are lost together with the information about the hypernymic paths and top synsets which is crucial for the wordnet editors. The wordnet graph cannot be also presented as a tree, because, firstly, the majority of its relations do not form a tree, and secondly, truly hierarchical relations would be visually lost in such a presentation with a significant loss the information for the editors. In order to avoid drawbacks of both basic presentation paradigms, an unique combination of the radial and tree-like presenta-

tion was proposed for WordnetLoom. Structure relations are presented along the vertical dimension, while other relations are presented radially around synsets, but in a way limited to horizontal zone of limited height centred on a given synset (i.e. only two sectors are used for radial presentation for each synset). The proposed visualisation scheme is illustrated in Fig. 1.

In Fig. 1, the octagonals represent synsets, *P 2.3* and *E 3.1* labels – wordnets, navy blue triangles can be clicked to unfold hidden branches, red to fold those shown. If a very large number of links for a synset and presentation direction (top/down, left/right), exceeds a threshold, then the rest is hidden in the green circle symbol and can be ‘taken out’ by user clicking it. The threshold, categorisation of particular relation types as vertical or horizontal, as well as link labels and colours used are defined in the WordnetLoom set-up file.

Division of relations into synset and lexical relations is orthogonal to the previous one. Moreover, lexical relations are linked directly to LUs as graph nodes. In order to visualise lexical relations and synset relations on the same screen, it would be necessary to present two inter-connected graphs, in fact, namely, the graphs of synsets and LUs. What is worse, a synset can be connected to a number of LUs on average. Thus, it would be too much information for one screen to present both graphs in the same time. Such a design of the screen was evaluated by linguists as too much cluttered to be useful. Thus, only synset graph is visually presented, and for a synset in focus its LUs are presented in the middle-right panel, see Fig. 1, and the relations of the selected LU are textually presented in the bottom-right panel.

The largest synsets can include even more than 20 LUs, but the average size is much smaller, e.g. less than 2 in plWordNet. However, the initial tests of the visualisation showed that when the number of the presented synsets on the screen approaches 10, it starts to be perceived as cluttered, when all synset members are visible inside the synset symbols. A kind of dynamic adaptation of the number of synset members presented would be an unnecessary complication (it depends also on synset sizes). So, finally, only one synset member, the first LU from a synset, is presented as its representative, the rest is presented in the middle-right panel. Its different sub-panels give access to the attributes of the given synset. For a LU selected

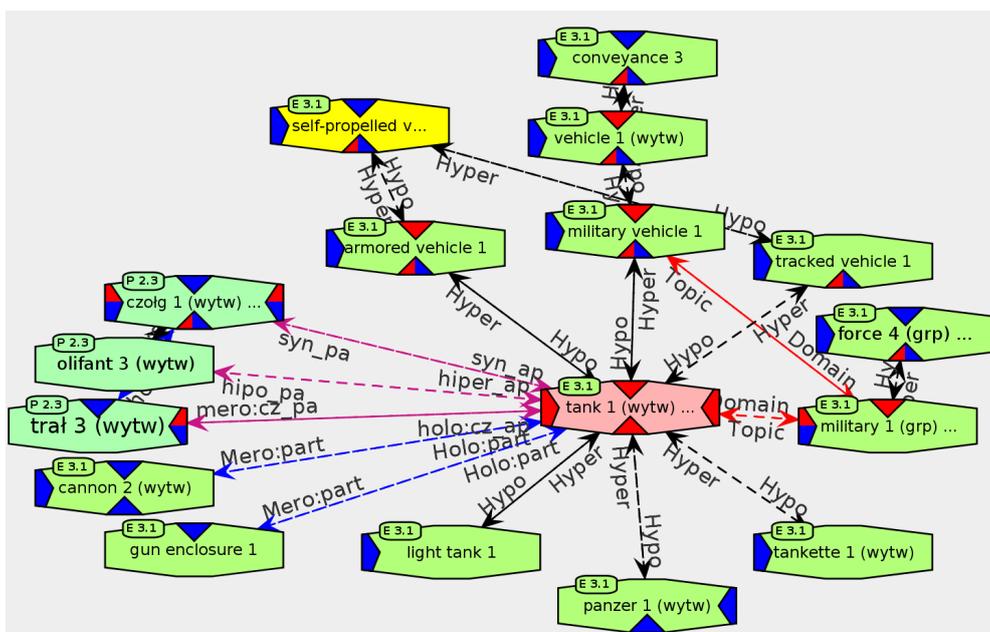


Figure 1: An example of the visualisation scheme.

among the synset members, its list of the lexical relations is shown in the bottom-right panel. Attributes of the synset LUs can be inspected with the context menu (right click).

Users can dynamically open several visualisation panels, e.g. to present search results, compare structures, create links between distant sub-trees or synsets from two wordnets etc.

5 Architecture

Contrary to several other systems, e.g. *DEBVis-Dic*, WordnetLoom 1.0 was written in Java, and the same approach was followed in the version 2.0. Since Java was quite stable, WordnetLoom 1.0 could be easily installed by non-technical users (by simply unpacking files including the *jar* file). The Java-based solution was free of the problems related to the changing versions of web-browsers (like plug-ins do have), and Java provided more flexibility in the implementation in contrast to the script languages used in the Web.

The construction of WordnetLoom 1.0 was initiated in 2005 as a client-server application with a direct connection to the database. The applied trigger mechanism allowed to encapsulate the whole system with a change control mechanism on selected database tables. WordnetLoom 1.0 supported the Polish language only. The lack of dynamic dictionaries made it difficult to expand and every change in the database required redistribu-

tion of the application. At the same time, several supporting tools based on the central wordnet database such as the monitoring system and statistics, API REST, mobile application and web application were created.

In order to adapt the system to new functionalities and other wordnets, we changed its architecture and enhanced the role of the central module which is shared between peripheral applications to increase maintainability of the whole system. In the new three-layer architecture, presented in Fig. 3 an additional, intermediate layer – a *service layer* – was introduced. This layer encompasses now the entire business logic code responsible for CRUD³ operations and validation. The trigger mechanism has been replaced by the *Hibernate Envars*⁴ module allowing the easier undoing of changes. The new schema migration mechanism has been introduced and the application has been secured by mechanisms provided with the *Wildfly Server*. The database schema itself has been rebuilt to be similar to an UBY-LMF structure (Gurevych et al., 2012), new tables have been added, allowing for dynamic construction of dictionaries and a localization mechanism.

WordnetLoom's thick client was transformed into a *thin client* model where all business logic

³The CRUD cycle describes the elemental functions of a persistent database. CRUD stands for Create, Read, Update and Delete

⁴<http://hibernate.org/orm/envars>

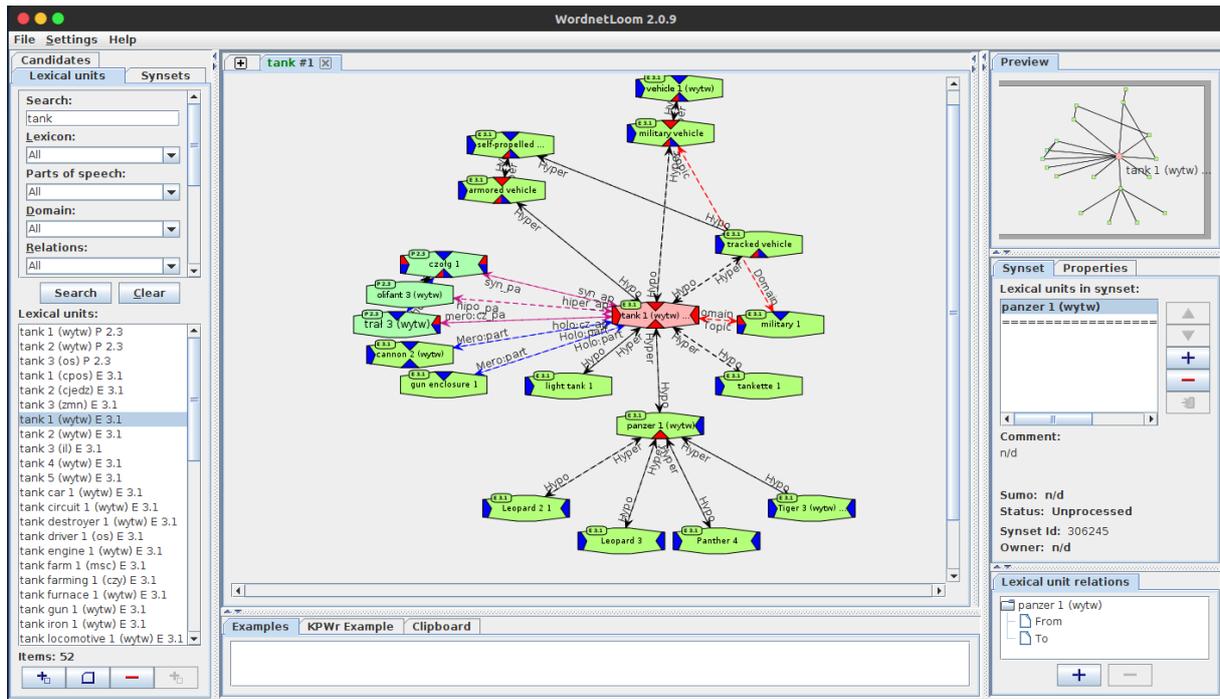


Figure 2: An example of the general layout of the visualisation perspective of WordnetLoom 2.0.

was transferred to the EJB⁵ service module. Some of the UI perspectives for data presentation in tabular manner were removed from the client, while the graphical visualization perspective became the main workspace with all functionality accessible, especially the searching facilities have been expanded in this perspective to fully encompass both synsets and LUs⁶. The language of UI can be changed in any moment. The user can also choose which lexicons, mostly wordnets, they want to work with in the given moment. In the previous version the user was working with all lexicons at the same time. The constantly increasing number of relations, both within individual lexicons and across them obscured the graph visualization window. In order to make work more efficient, a functionality to hide selected lexicons was introduced.

The key business logic module has been implemented as an *EJB module* so that it can be a common unitary element. It offers API access to the data layer and, at the same time, has a common extensible validation module aimed at preventing the establishment of wrong relations and thus forcing the correctness of data input. The key tables are

⁵ EJB is a server-side software component that encapsulates business logic of an application. The EJB specification is a subset of the Java EE specification

⁶ In the visualisation perspective of WordnetLoom 1.0 searching was possible only for LUs and lexical relations.

audited and, in addition, a special table contains a register of all operations carried out. A very important feature of our new version of the system is the fact that each element belongs to a certain lexicon, that gives the possibility to expand with new collections of elements.

The server is based on *MySQL 5.7* database management system⁷ and *Wildfly 10.1.0*⁸. For the construction of the system we have used also the following frameworks: *Java EE 7* enterprise edition platform, *JPA 2.+*⁹ (*Hibernate 5+*¹⁰), *EJB 3.1*¹¹, *JAX-RS*¹², *JSF*¹³ (*PrimeFaces*¹⁴), *Java Swing*¹⁵ and *JUNG 2*¹⁶.

⁷<https://www.mysql.com>

⁸Used as EJB web container which provide a runtime environment for web related components, e.g. computer security, Java servlet lifecycle management, transaction processing, and other web services. <http://wildfly.org>

⁹Java Persistence API. Part of Java EE 7 Specification

¹⁰<http://hibernate.org>

¹¹ <http://download.oracle.com/otndocs/jcp/ejb-3.1-pfd-oth-JSpec>

¹² JAX-RS: Java API for RESTful Web Services is a Java programming specification that provides support in creating web services according to the Representational State Transfer (REST) architectural pattern. Part of Java EE 7 Specification

¹³ JavaServer Faces (JSF) is a Java specification for building component-based user interfaces for web applications

¹⁴ <https://www.primefaces.org>

¹⁵ GUI widget toolkit for Java provides API for building user interfaces

¹⁶JUNG — the Java Universal Network/Graph Framework, <http://jung.sourceforge.net>

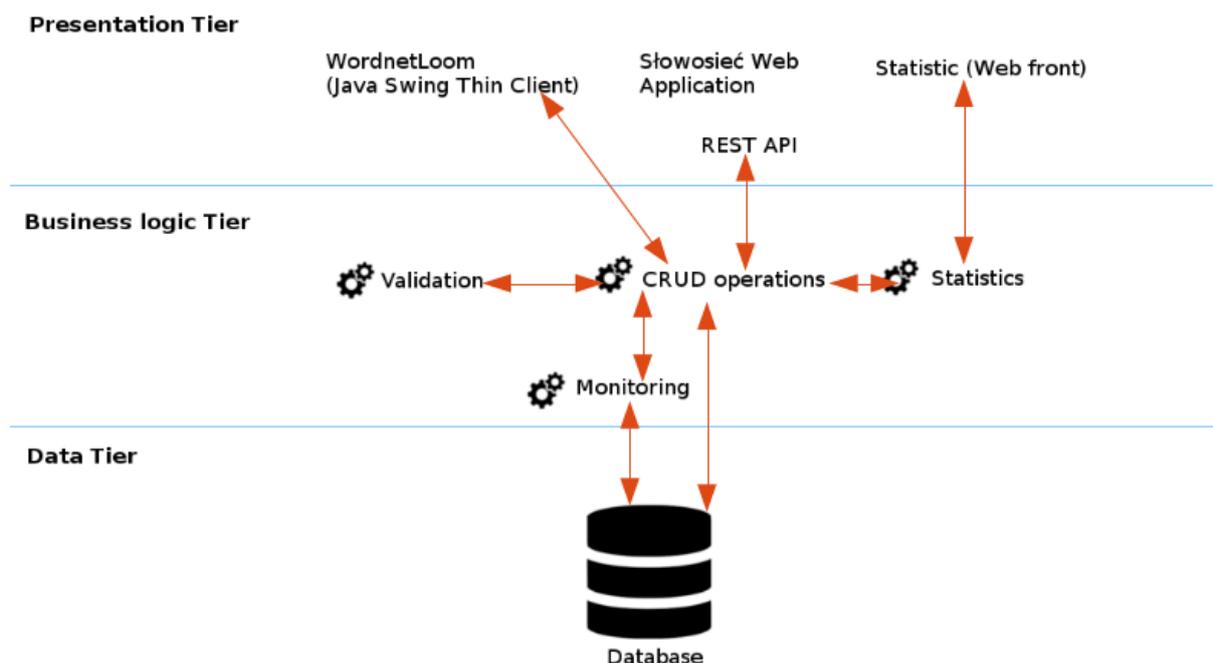


Figure 3: WordnetLoom 2.0 architecture.

6 Extensions and Applications

The architecture of the version 2.0 has been significantly improved in comparison to the previous one, but WordnetLoom has been used for more than 10 years for plWordNet editing (resulting in $\approx 200k$ lemmas, $\approx 300k$ LUs and $\approx 200k$ interlingual mappings processed), as well the new version has become a basis for system's adaptations to other wordnets, e.g. a Portuguese wordnet.

6.1 plWordNet Development

As inter-lingual relations are synset relations, but between synsets in different languages, subgraphs for plWordNet and Princeton WordNet should be presented on the same screen. In the new version we added possibility to work on any number of wordnets for any number of languages. Thus it became necessary to introduce labels representing wordnets (defined in set-up) that are attached to synset symbols. Moreover, searching can be limited to elements of a specified wordnet.

Many improvements requested by users were introduced. In the visualisation perspective, in the bottom-right panel of lexical relations double click on the target of relation, a LU, opens a new graph panel with the synset which this LU belongs to.

Every LU and synset is described by additional, meta-attribute of status with the following values:

not processed (default value), *error*, *verified*, *new*, *partially processed* and *added sense*. Editors can also provide comments to the status, especially important for *error* and *partially processed* statuses, as an explanation of the error, or missing actions, respectively. The status *not processed* marks the material introduced earlier, while *new* signals newly added element especially requiring verification. According to the plWordNet work procedure editors are assigned packages of lemmas, cf Sec. 4, and are obliged to identify and add all LUs for each lemma. However, during their work they may need introduction of LUs for other lemmas than assigned to them, e.g., to add a relation link describing one of the assigned lemmas. In such cases a linguist marks the introduced new LUs and synsets with the *added sense* status that means that some other senses of the same lemma may be lacking. The system of statuses is defined in the database, can be further expanded and supports the management of the linguistic team.

In WordnetLoom 1.0, verb aspect was implicitly expressed by the aspectual relations. In order to facilitate searching and diagnostic procedures, aspect attribute has been added to verbs. Search function was also expanded to cover all attributes, e.g., synset identifiers that are automatically assigned and are not manually edited, but visible in the results of WSD. The search results

can be downloaded in CSV format useful for co-ordinators and plWordNet users.

Diagnostics was also improved by adding PoS tags to variables in substitution tests in the relation definitions stored in the database¹⁷. This PoS specification allows for automated controlling of the correctness of the links that are considered to be added, but also already present in the database.

The introduced easier expansion of the database and UI allows for adding new types of lexicographic files and annotation with semantic domains. The former facilitates wordnet editing (e.g. the extension includes verb classes used in plWordNet), while the latter supports applications. The domains are based on WordNet Domains (Bentivogli et al., 2004), but we plan to manually modify and expand this classification.

6.2 Portuguese Wordnet

As WordnetLoom is getting consolidated, it can be used to help the construction of wordnets other than just plWordNet. This is what is happening with the MultiWordnet of Portuguese, a quality wordnet for Portuguese (Branco et al., 2009).

This Portuguese wordnet is a project started in 2004 as a branch of Multi-WordNet (Pianta et al., 2002), which until now gathered seven different languages (English, Hebrew, Italian, Latin, Portuguese, Romanian and Spanish), and was one of the first consistent initiatives pursuing the goal of establishing a multilingual wordnet that remains open for further languages. The wordnets in these languages, were transitively aligned with each other by resorting to its alignment to Princeton WordNet, whose format all are following, and thus having English as the pivot language.

This pilot application of WordnetLoom to a different wordnet is providing an important testbed to assess its generality, to find aspects where it can be enhanced, and also to check its technical fitness. For instance, there have been a number of usability enhancements whose need emerged by having new users effectively using this application under different conditions and for a different language, thus stretching its usability requirements. A number of technical improvements have been also motivated in this context of extending the cooperative usage of WordnetLoom to further users.

The outcome of this process and key lessons

¹⁷ A dedicated window for editing the definitions is accessible only for the co-ordinators of the linguistic team.

learned with it are reported in this section.

6.2.1 Enhancing WordNet Content

When creating a quality WordNet for a given language, differences among its language variants should be taken into account and be duly recorded. The differences to be registered can be just superficial: the same word may have different spellings in different variants. Or they may be more substantial: a given concept may be expressed by the same words in different variants, or different variants may resort to different words.

Portuguese is the third European language in number of speakers worldwide. It is the official or national language of several countries and territories in four continents, including Portugal and Brazil. While all speakers of Portuguese can easily communicate, this language have a number of variants. In this context, the Portuguese wordnet has synsets that includes words that belong to only one variant. A word in a synset that belongs to all language variants receive no special marking. A word that belongs to one variant but not to others should be registered as expressing that concept *in that variant* (in addition to being included in that synset). Currently, the Portuguese WordNet covers both the European (spoken in Portugal) and American (spoken in Brazil) variant.

This need resulted in a contribution to enhance wordnet's content with which WordnetLoom can cope. There is now a new field by means of which word forms can be associated to one or other variant, or to none, in which that indicates that a word form is common to all variants.

Portuguese WordNet includes the mapping of synsets into concepts in SUMO ontology (Niles and Pease, 2001). A new field in the WordnetLoom database was introduced in order to represent this type of information, that can be also useful for plWordNet for which its mapping to SUMO was stored so far as a separate resource.

6.2.2 Enhancing Lexicographers Work

The quality Portuguese WordNet is being constructed under the semi-automatic methodology of MultiWordnet. After a first projection of tentative synsets and their relations obtained on the basis of Princeton WordNet and bilingual dictionaries, these synsets are adjusted and confirmed by human lexicographers.

In the initial version of WordnetLoom which the Portuguese WordNet started being edited with,

there were just a few search options, namely by word or POS. As the lexicographic labour was proceeding, we realized that it would be faster and easier, if it would be possible to keep track of synsets and senses that have been already checked before, to not check them again, wasting useless effort by the lexicographers. This could be done if there was an identifier for a sense or a synset status, indicating whether it had been checked.

As we discussed in Sec. 6.1, this need resulted in another contribution to enhance the versatility of WordnetLoom to support lexicographers work. In its current version, the users are provided with additional search options based on these statuses, so that they can retrieve only synsets that are yet to be checked or synsets whose edition are finalised.

6.2.3 Enhancing Format Compatibility

There is a main difference between the format of Princeton WordNet and the wordnet designed and developed for plWordNet. The latter is sense-based while the former is synset-based. This creates the need for new information (i.e. data-types and data-relations) in the database. Some instances are “sense relations” and “sense to synset connections”. WordnetLoom was originally designed to be compatible with the Polish wordnet. Hence, before it could be employed, the data of the Portuguese WordNet – in Princeton WordNet format – had to be migrated to the plWordNet format. A converter¹⁸ from the Princeton WordNet format to the WordnetLoom format was developed by the Portuguese team. It can now be reused to convert any wordnet in a format compatible, or convertible to the Princeton WordNet format (a *de facto* standard), into the WordnetLoom format, thus greatly enlarging the number of possible wordnets that now can be uploaded into and edited by WordnetLoom.

This step was rather challenging and demanding as there are substantial differences in the organisation of both representations, although facilitated by higher expressiveness of the plWordNet format (e.g. it allows for assigning a set of attributes to both: synsets and lexical units).

The fact that WordnetLoom is under continuous improvement is a positive aspect as teams can ask for changes according to their needs. These changes might be kept as useful suggestions for the final version of WordnetLoom or could be kept

¹⁸ the link temporarily anonymized for submissions

local for that specific team.

6.2.4 Technical Enhancements

One very important step in developing any system is its testing and debugging. The work on the Portuguese wordnet is part of the former, with the reporting to the central development team about the issues encountered while working with WordnetLoom, thus being contributing to its technical enhancement.

Three examples of more salient issues that were reported, and that were then solved, are indicated here. (1) Problems with multiple senses of a word. This problem occurred for ambiguous words where one of their senses already existed in WordnetLoom database. When adding a new sense, the UI raised a warning about repetitive entry even though it was actually the same word but in a new synset. (2) Some dis-functionality in the UI. There were cases that the buttons did not function correctly or clicking them caused exceptions that forced to restart the client. (3) Difficulties with setting up the server and client. Problems can be categorized into (i) incompatibility of Java versions and Java basic set-ups; (ii) local settings for both the server and each of the clients; and (iii) issues with running Java-Web-Start. The first two of these types of problems are already solved and the resolution of the third category is under progress.

7 Conclusions and Further Works

WordnetLoom incorporates more than 10 years of experience in the development of a very large wordnet by many linguists on daily basis and this rich experience has become a good basis for the development of new version improved with respect to both: technology and functionality. The system is open¹⁹. Its most unique feature is direct work on the visually presented wordnet graph, as well as support for simultaneous editing and inter-linking of many wordnets (editors see the multilingual structures they are going to map).

WordnetLoom adaptation to the needs of the Portuguese Wordnet developed according to completely different method than plWordNet showed system’s potential, and paved way for its adaptations to other resources and tasks. We plan to integrate both variants and continue collaborative development of the system.

¹⁹ <https://github.com/CLARIN-PL/WordnetLoom>

Acknowledgment

Work partially supported by the Polish Ministry of Education and Science, Project CLARIN-PL, and the Portuguese Ministry of Higher Education, Science and technology, by the Infrastructure for the Science and Technology of the Portuguese Language (CLARIN Língua Portuguesa), and by the ANI/3279/2016 grant.

References

- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *COLING 2004 Workshop on “Multilingual Linguistic Resources”*, Geneva, Switzerland, August 28, pages 101–108. ACL, 2004. URL <http://wdomains.fbk.eu/publications/Coling-04-ws-WDH.pdf>.
- António Branco, Francisco Costa, Eduardo Ferreira, Pedro Martins, Filipe Nunes, Joao Silva, and Sara Silveira. Lx-center: a center of online linguistic services. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 5–8. Association for Computational Linguistics, 2009.
- Nicoletta Calzolari et al., editor. *Proc. Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 2012. European Language Resources Association. ISBN 978-2-9517408-7-7.
- Christiane Fellbaum, editor. *WordNet – An Electronic Lexical Database*. The MIT Press, 1998.
- Darja Fišer and Benoît Sagot. Constructing a poor man’s wordnet in a resource-rich world. *Language Resource and Evaluation*, 49(3): 601–635, September 2015. ISSN 1574-020X. doi: 10.1007/s10579-015-9295-6. URL <http://dx.doi.org/10.1007/s10579-015-9295-6>.
- Darja Fišer and Jernej Novak. Visualizing sloWNet. In *Proceedings of eLex*, pages 76–82, 2011. URL <http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-9.pdf>.
- I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, Ch.M. Meyer, and Ch. Wirth. UBY – a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of EACL 2012*. ACL, 2012.
- Verena Henrich and Erhard Hinrichs. GernEđiT – the GermaNet editing tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odiĵk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.
- Ales Horak, Adam Rambousek, and Piek Vossen. A distributed database system for developing ontological and lexical resources in harmony. In *9th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Berlin, 2008. Springer.
- Aleř Horáĵ and Pavel Smrř. New features of wordnet editor VisDic. *Romanian Journal of Information Science and Technology*, 7(1–2): 201–213, 2004.
- Aleř Horáĵ, Karel Pala, Adam Rambousek, and Martin Povolný. DEBVisDic — first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International WordNet Conference — GWC 2006*, pages 325–328. Masaryk University, 2006.
- Hitoshi Isahara and Kyoko Kanzaki, editors. *Advances in Natural Language Processing: Proc. 8th International Conference on NLP, JapTAL*, volume 7614 of *Lecture Notes in Artificial Intelligence*, 2012. Springer-Verlag.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA. URL <http://aclweb.org/anthology/R13-1058>. **ACL Anthology**.
- Luís Morgado da Costa and Francis Bond. Omwedit - the integrated open multilingual wordnet editing system. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 73–78. Association for Computational

Linguistics and The Asian Federation of Natural Language Processing, 2015. doi: 10.3115/v1/P15-4013. URL <http://aclanthology.coli.uni-saarland.de/pdf/P/P15/P15-4013.pdf>.

Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM, 2001.

B.S. Pedersen, L. Borin, M. Forsberg, K. Lindén, H. Orav, and E. Rognvaldsson. Linking and validating nordic and baltic wordnets – a multilingual action in META-NORD. In *Proceedings of 6th International Global Wordnet Conference*, pages 254–260., Matsue, Japan., 2012.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Multiwordnet: developing an aligned multilingual database. 1st gwc. *India, January*, 2002.

Maciej Piasecki, Michał Marcińczuk, Radosław Ramocki, and Marek Maziarz. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232, 2013. doi: 10.1504/IJDM.2013.055861.

Borislav Rizov. Hydra: A software system for wordnet. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Seventh Global Wordnet Conference*, pages 142–147, Tartu, Estonia, 2014. URL <http://www.aclweb.org/anthology/W14-0119>.

Translation Equivalence and Synonymy: Preserving the Synsets in Cross-lingual Wordnets

Oi Yee Kwong

Department of Translation
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

The Princeton WordNet for English was founded on the synonymy relation, and multilingual wordnets are primarily developed by creating equivalent synsets in the respective languages. The process would often rely on translation equivalents obtained from existing bilingual dictionaries. This paper discusses some observations from the Chinese Open Wordnet, especially from the adjective subnet, to illuminate potential blind spots of the approach which may lead to the formation of non-synsets in the new wordnet. With cross-linguistic differences duly taken into account, alternative representations of cross-lingual lexical relations are proposed to better capture the language-specific properties. It is also suggested that such cross-lingual representation encompassing the cognitive as well as linguistic aspects of meaning is beneficial for a lexical resource to be used by both humans and computers.

1 Introduction

The development of multilingual wordnets has been accomplished mostly by starting with the Princeton WordNet for English (Fellbaum, 1998b) and supplying translation equivalents from another language to individual concepts represented by the synsets. When conceptual gaps are identified, they may be handled by the addition or omission of synsets in the new wordnet. While the approach has the merit of good coverage, reliance on translation equivalents may be at the expense of forming non-synsets in the target language wordnet, for which great caution has to be exerted. Past experience from building multilingual wordnets has observed various difficulties, mostly arising from cross-linguistic differences in

lexicalisation, conceptual space and sense distinction (e.g. Vossen, 1998). This paper discusses further observations from the Chinese Open Wordnet (Wang and Bond, 2013), which added new translations from authoritative bilingual dictionaries as a means to increase coverage, to show that translation equivalents need to be very carefully screened to avoid some potential and easily overlooked pitfalls. While a good coverage is appreciated, especially with a view to use the wordnets in a variety of computational and human language applications, it is suggested that alternative representations including additional relational pointers be used to accommodate cross-linguistic differences without disturbing the basic infrastructure of WordNet, in particular its basic definition of synsets.

The rest of this paper is organised as follows: Section 2 reviews the theoretical basis of the Princeton WordNet (PWN) and the construction of the Chinese Open Wordnet (COW). Attention will be focused on adjectives. Section 3 presents some observations from COW in terms of its resulting synsets in the adjective subnet. Section 4 discusses the cross-lingual aspects and proposes alternative ways for representing the lexical semantic relations, followed by a conclusion in Section 5.

2 WordNet Infrastructure

2.1 Synsets as the Building Blocks

The original PWN started as a psycholinguistic project for testing the scalability of relational lexical semantics, where concepts are supposed to be linked by specific relations. Its resulting large lexical database turned out to be well received and popularly used by computational linguists. Concepts are expressed or lexically represented by sets of synonyms (synsets) within individual word classes, and are connected by a variety of relational pointers. This essentially results in four sub-

nets, for nouns, verbs, adjectives and adverbs, respectively (Fellbaum, 1998a).

It is therefore well-known that the basic building blocks of the original PWN are the “synsets”, which are unordered sets of words that “denote the same concept and are interchangeable in many contexts”, and the main relation in WordNet is synonymy¹. PWN defines word senses by means of synsets. Given the mutual substitutability that holds among members in a synset, membership of a lexical item in a certain synset indicates a particular sense of the word.

2.2 The Adjective Database

Although PWN has four subnets, it is obvious that the noun database and verb database have been the most discussed and utilised (for PWN and multilingual wordnets alike), not only because they contain a larger number of synsets, but perhaps also for the more clearly defined relations established in them. For example, the hypernymy/hyponymy relation for nouns and the troponymy relation for verbs are typical. The adjective database, on the other hand, appears to receive far less attention.

According to Fellbaum et al. (1993), WordNet contains descriptive adjectives and relational adjectives. Descriptive adjectives ascribe a value of an attribute to a noun, such as “heavy” as a value for “weight”, indicated in the database by the attribute pointer. The descriptive adjective synsets are not hierarchically ordered as nouns, and apart from the basic semantic relation, antonymy, the semantics of adjectives is more naturally perceived as an N-dimensional space. Adjectives similar in meaning may not all have antonyms, and the similarity pointer is used to mark this phenomenon. Not all gradable attributes have most gradation lexicalised. As remarked by Fellbaum et al. (1993), “It would not be difficult to represent ordered relations by labeled pointers between synsets, but it was estimated that not more than 2% of the more than 2,500 adjective clusters could be organized in that way. Since the conceptually important relation of gradation does not play a central role in the organization of adjectives, it has not been coded in WordNet.” In fact, adjectives are considered very polysemous and of limited usefulness in conveying information, and they are not even included in EuroWordNet (Fellbaum, 1998b). But whether this phenomenon is equally

insignificant for other languages and its exclusion will not affect the construction of wordnets in those languages may require further thought, and will be discussed in the following sections. It is also noted that “adjectives expressing evaluations (good/bad, desirable/undesirable) can modify almost any noun; those expressing activity (active/passive, fast/slow) or potency (strong/weak, brave/cowardly) also have wide ranges of applicability”, which is also a key point to consider when multilingual wordnets are built.

2.3 Wordnets with Translation Equivalents

Since the inception of the EuroWordNet project (Vossen, 1998), which aimed at building a multilingual lexical database for several European languages in the form of PWN, subsequent development of wordnets in other languages has often similarly followed one of the two approaches: the Merge Model or the Expand Model. With the Merge Model, vocabulary selection and synsets are developed separately and locally, followed by generating equivalence relations to PWN. The Expand Model, on the other hand, starts with PWN vocabulary and synsets, and translates the synsets using bilingual dictionaries into equivalent synsets in the other languages.

There have been various attempts for Chinese wordnet (e.g. Huang et al., 2004; Huang et al., 2010; Wang and Bond, 2013; Xu et al., 2008). They primarily relied on some ways to identify translation equivalents, including automatic means and human verification (e.g. Huang et al., 2004). Some limited the number of translation equivalents to be included for a synset (e.g. Huang et al., 2004), while others (e.g. Wang and Bond, 2013) intentionally added more entries.

The Chinese Open Wordnet (COW), in particular, followed the Expand Model and started with the core synsets in PWN (Boyd-Graber et al., 2006), and formulated detailed guidelines to build a better Chinese wordnet. According to Wang and Bond (2013), among the 4,960 core synsets, adjectives occupy only 13.8% of the total. In building the COW, Chinese translations for the core synsets were first obtained by merging existing data from the Southeast University Chinese Wordnet (Xu et al., 2008) and the Open Multilingual Wordnet linked with lemmas extracted from the English Wiktionary (Bond and Foster, 2013). The resulting translations were checked manually, with dele-

¹<http://wordnet.princeton.edu/>

tions and amendments as necessary, while new translations found from authoritative bilingual dictionaries were added. The lexical semantic relations were also checked with a random sample from the database (Wang and Bond, 2013).

The manual checking was intended to ensure that the Chinese translations match the English synsets in terms of meanings and parts of speech. Cross-linguistic differences have been recognised all along, especially with respect to lexicalisation, where a specific lexicalised concept in English may not find an equivalent lexicalised form in Chinese, and in such cases a phrase or definition will be used for representing the concept in the Chinese wordnet. Wang and Bond (2013) have also identified a range of situations for which discrepancy within synsets may be found. Where conceptual meaning is concerned, there are cases where two languages may have similar basic conceptual meanings that differ in severity and usage scope. Where affiliated meaning is concerned, words may differ in their affection, genre, and time. Strictly speaking, such cases should be ruled out from the synsets, although a looser standard was adopted for COW, which keeps them to ensure higher coverage but admittedly lower accuracy.

2.4 Potential Blind Spots

In addition to the above known facts, translation equivalents have yet to be more cautiously handled to avoid other potential problems, especially with respect to any incompatibility with the basic WordNet structure. For example, consider the following PWN synset with its correspondence in COW:

01586342-a

nice (pleasant or pleasing or agreeable in nature or appearance)

体贴(的), 合意(的), 美好(的), 和蔼(的), 友好(的), 令人愉快(的), 令人快乐(的), 讨人喜欢(的)

The English synset has only one lexical item, which is not really a problem itself. The tricky part is the “generalness” of this concept, as expressed by the word “nice”, in terms of its meaning and usage contexts. As hinted by its gloss, this sense of “nice” can mean “pleasant” or “pleasing” or “agreeable”, and such good quality can apply to the “nature” or “appearance”

of something. In other words, almost anything can be described as “nice”, to mean something good in general without specifying any particular attributes and qualifying how good it is. So strictly speaking, and to be as general as it is, the Chinese equivalent 好 *hǎo* would suffice, and all the items listed above are in a certain sense “over-translation”, as they are only conceptually equivalent under certain contexts. For example, 和蔼 *hé’ǎi* can only describe a person, and 美好 *méihǎo* for something inanimate and often more abstract. Meanwhile, 和蔼 *hé’ǎi* is also among the set of words in another adjective sense corresponding to a synset for “kind”, as follows:

01372049-a

kind (having or showing a tender and considerate and helpful nature; used especially of persons and their behavior)

体谅(的), 体贴(的), 善良(的), 仁慈(的), 和善(的), 宽厚(的), 友善(的), 好心(的), 好心肠(的), 亲切(的), 温和(的), 和蔼(的), 宽宏大量(的), 友好(的), 乐于助人(的)

Similarly, strictly speaking this sense of “kind” is also quite encompassing, and its fuzziness may be more equivalently represented by 仁慈 *réncí* and 好心 *hǎoxīn*, while leaving others like 友善 *yǒushàn* for “friendly”, 乐于助人 *lèyúzhùrén* for “helpful”, and 体贴 *tǐtiē* for “considerate”.

Given the co-existence of the same lexical items like 和蔼 *hé’ǎi* in correspondence to two synsets relating to “nice” and “kind” separately in PWN, whereas the conceptual distinction in PWN has not considered the two senses synonymous², and there is no obvious evidence for multiple senses for 和蔼 *hé’ǎi* according to most dictionaries, it is questionable to treat it as a translation equivalent for the two PWN senses. On the other hand, despite the vague definition for synonymy (as defined by substitutability in a given context), it is readily realised that the criterion is not met for the above examples. No dictionary seems to consider 和蔼 *hé’ǎi* and 体贴 *tǐtiē*, for instance, synonymous in any case as they refer to different qualities of a person. In other words, the set of Chinese words can no longer be qualified as a “synset” as originally

²The specific sense of “kind” is not linked to the specific sense of “nice” in PWN via the see-also and similar-to connections. The sense distinction is thus different from other resources, such as the Roget’s Thesaurus, where “nice” and “kind” co-exist in group 884 for their sense of “amiable”.

defined for the WordNet structure. Moreover, to a certain extent, the conceptual meaning is mingled with specific contextual usage. Thus, when we refer to someone being nice (as in “he is very nice”), it is only as much as saying 他这个人很好 *tā zhège rén hěn hǎo*. Only with more specific context or additional information given could one decide on the way in which he is nice, such as being easy to get along with, very helpful, very generous, or others.

Complete equivalents are generally rare (Svensen, 1993), especially for distant language pairs like English and Chinese, except for very domain-specific concepts and terminologies. The difference in lexicalisation of concepts is also an issue. Since other wordnets are centered on PWN, the lexicalisation in English is taken as a default, which may lead to the use of longer expressions in a synset in other languages. This brings up two issues in constructing wordnets in other languages. One is the seriousness of the problem with respect to different parts of speech. Given the references available for nouns and verbs, and the fuzziness and subjectivity involved in adjectives, we expect that the problem is more pronounced among adjectives. Second, when the coverage of the meanings by the translation equivalents is at the expense of violating the requirements for synsets, are there better ways to handle such cases? In the following sections, we analyse the situation with reference to COW, and discuss possible alternatives for representing the lexical semantics therein.

3 Synsets in COW

The Chinese Open Wordnet (COW)³ consists of 42,312 synsets (Nouns 65.9%, Verbs 12.2%, Adjectives 20.2%, Adverbs 1.7%) with 80,009 lexical items (Nouns 57.9%, Verbs 16.7%, Adjectives 22.9%, Adverbs 2.5%). The following discussion covers the three major word classes, namely nouns, verbs and adjectives, with focus on adjectives, and adverbs are excluded.

3.1 Synset Size and Polysemy

In terms of synset sizes, as measured by the number of items in a synset, the largest range was observed for nouns, from 1 to 39 items in a synset, followed by adjectives and verbs, from 1 to 15 and from 1 to 13 respectively. As shown in

³Downloaded from <http://compling.hss.ntu.edu.sg/omw/>

Figure 1, noun synsets tend to be of smaller sizes than adjective synsets, and there are relatively even more larger synsets for verbs. Many of the extreme examples in the noun database have to do with biological nomenclature, as when a certain plant species is known by many formal and informal names in Chinese, as well as culture-specific items which lack one-to-one correspondences, such as:

12896307-n

black nightshade, common nightshade, poison-berry, poisonberry, Solanum nigrum (Eurasian herb naturalized in America having white flowers and poisonous hairy foliage and bearing black berries that are sometimes poisonous but sometimes edible)

老鸦酸浆草, 乌归菜, 野葡萄, 酸浆草, 救儿草, 黑姑娘, 天泡果, 地戎草, 七粒扣, 山海椒, 黑茄, 野茄子, 天泡草, 地泡子, 天天茄, 天茄子, 野辣角, 野海椒, 后红子, 天茄苗儿, 老鸦眼睛草, 水茄, 水苦菜, 野伞子, 天茄菜, 山辣椒, 狗钮子, 苦葵, 苦菜, 野茄菜, 飞天龙, 龙葵, 耳坠菜, 乌疗草, 野辣椒

09823502-n

aunt, auntie, aunty (the sister of your father or mother; the wife of your uncle)

妯, 姑母, 伯母, 姑姑, 老大妈, 阿姨, 妯母, 叔母, 姑妈, 舅母, 姑, 姨妈, 姨, 舅妈, 婶子, 婶婶, 姨母, 婶母

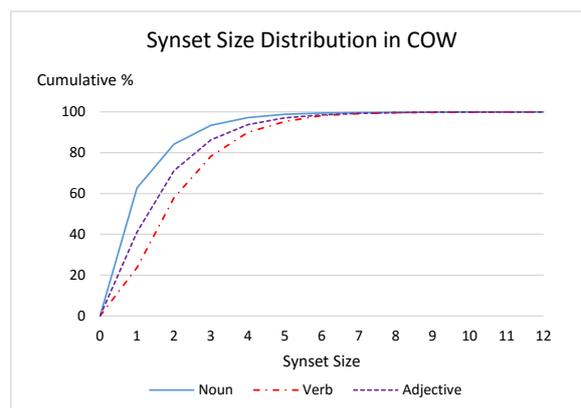


Figure 1: Synset Size Distribution for Various Word Classes in COW

The above two examples actually reveal two very different scenarios. Although we probably need a biologist or an expert in herbal medicine

to verify the many renditions for the very same plant species, as far as they are valid names, they can certainly be considered synonymous. But the second case corresponds to an obvious difference in sense distinction as a consequence of cultural difference. While “aunt” can refer to one of the many female relatives as indicated in the gloss, the Chinese words are not all interchangeable because each of them only refers to one type of the female relatives, e.g. 姑母 *gūmǔ* and 姑姑 *gūgu* for “the sister of one’s father” (further distinguished as the elder and younger sister respectively in some dialects), 舅母 *jiùmǔ* and 舅妈 *jiùmā* for “the wife of the brother of one’s mother”, etc. In other words, although they can be considered translation equivalents for “aunt” in a given context, they are definitely not synonyms.

The issue is also quite different from what can be observed from the adjective database and verb database. The large synsets in them do not really contain multiple renditions for the same conceptual meaning as in the noun examples above, but more often reflect the polysemy contained by the concepts as represented by the English synsets which results in translational differences in Chinese, such as:

01256332-a

hot (extended meanings; especially of psychological heat; marked by intensity or vehemence especially of passion or enthusiasm)

流行(的), 热切(的), 激烈(的), 热门(的), 才发行(的), 急躁(的), 销路好(的), 刚出版(的), 轰动一时(的), 最新(的), 紧缺(的), 激动(的), 狂热(的), 热烈的(的), 时新(的)

01215137-v

arrest, pick up, nail, apprehend, nab, collar, cop (take into custody)

捕捉, 捉到, 捕获, 逮捕, 拘留, 拘押, 拘捕, 抓住, 抓获, 当场逮捕, 擒获, 逮住

The adjective example is another typical one, like those mentioned in an earlier section, which apparently violates the requirements for synsets. It is least likely that one would equate 急躁 *jízào* (impatient) with 流行 *liúxíng* (popular), although the examples given in the English synset include a whole lot of extended usage of “hot” as in “a hot temper”, “a hot topic”, “a hot new book”, “a hot love affair”, and “a hot argument”, while the

encompassing “hot” has to be rendered according to its subtle sense difference according to the noun it modifies. Thus the “hotness” associated with “temper” is not the same “hotness” associated with “topic” in Chinese, which are therefore non-synonyms.

As for the verb example, the English synset obviously refers to “arrest by police”. Nevertheless, the Chinese expressions like 捕捉 *bǔzhuō* (catch) may be too general while those like 当场逮捕 *dāngchǎng dài bǔ* (arrest on the scene) are seemingly over-specific. Issues with the verb synsets are no less complex than those pertaining to adjectives, and will not be pursued further in the current discussion. However, the verbal synset above can also illustrate a logical issue. It is not appropriate to find 逮捕 *dài bǔ* (arrest) and 当场逮捕 *dāngchǎng dài bǔ* (arrest on the scene) in the same synset, not only because the latter is a more specific meaning than the former, but also the latter is a phrasal expression (with modifier and verb) which cannot logically mean the same thing as the simple lexical verb.

3.2 Adjectives and Non-synsets

We selected 200 top-sized adjective synsets from COW and examined the synonymy therein. It turns out that at most 27 out of the 200 synsets do not contain phrasal members (in addition to lexicalised items)⁴. While this does not necessarily mean that over 85% of the English adjectives in these synsets do not have lexicalised translation equivalents in Chinese, it at least shows that bilingual dictionaries may tend to provide translated definitions or paraphrase instead of or in addition to translation equivalents. Although this is an unavoidable practice in bilingual lexicography (Atkins and Rundell, 2008), its compatibility with WordNet structure is questionable. It is thus worth to reconsider their representation in the resource, adhering to the lexicalisation criterion on the one hand (e.g. Huang et al., 2010) and expanding the overall coverage on the other (e.g. Wang and Bond, 2013).

The lexicalisation issue aside, it was observed from the selected data that non-synsets often result from one or more of the following situations:

⁴Some common and fixed four-character expressions are considered single words, e.g. 无忧无虑 *wúyōuwúlǜ* (care-free), while those containing an obvious combination of two or more words are considered phrasal expressions, e.g. 轻松愉快 *qīngsōng yúkuài* (relaxed and happy).

1. Different sense distinctions

The difference in the division of semantic space and granularity of sense distinction is particularly salient with the more “general” adjectives already illustrated above. But even for the less “general” adjectives, the broadened coverage may not always match the sense granularity indicated in PWN, especially as PWN is known for its possibly over-fine-grained senses. For example, “civilised” belongs to two synsets in PWN, and here are their parallel Chinese synsets:

00411886-a

civilized, civilised (having a high state of culture and development both social and technological)

文明化(的), 有礼貌(的), 有教养(的), 开化(的), 文明(的), 文雅(的)

01947741-a

cultured, polite, civilized, civilised, cultivated, genteel (marked by refinement in taste and manners)

文雅(的), 有礼貌(的), 优雅(的), 有教养(的), 有礼(的), 文明(的), 有先进文化(的), 有修养(的)

The two senses of “civilised” are quite distinct, such that the first refers to a general high state of development in a collective sense and the second specifically relates to more personal and individual behaviour. But the Chinese synsets overlap considerably, especially when 有礼貌 *yǒulǐmào* (polite), 有教养 *yǒujiàoyǎng* (cultivated) and 文雅 *wényǎ* (elegant) are more relevant to the second sense than the first.

2. Over-interpretation of concepts

In addition to the examples like “hot” and “kind” discussed above, over-interpreting a concept may lead to obscure results as in:

02328659-a

docile (willing to be taught or led or supervised or directed)

易管教(的), 驯服(的), 易教育(的), 易驾驭(的), 可教导(的), 容易教(的), 听话(的), 驯良(的), 愿学习(的), 易训练(的), 温顺(的), 顺从(的), 易控制(的)

While lexicalised items like 驯服 *xúnfú* and 温顺 *wēnshùn* may already satisfactorily represent the concept in Chinese, the others like 易管教 *yì guǎnjiào* (easy to teach) and 易驾驭 *yì jiàoyù* (easy to control) may still be acceptable except that they are phrasal expressions. However, 愿学习 *yuàn xuéxí* (willing to learn) seems to have over-interpreted in the sense that “willing to learn” may not necessarily mean “willing to be taught / well-behaved / easy to control”.

3. Multiple facets of concepts

Relating less to sense granularity but more to individual context of usage, some adjectives may highlight different facets of a certain quality when modifying different things. For example:

02964782-a

Chinese (of or pertaining to China or its peoples or cultures)

中国文化(的), 汉, 华, 中文(的), 中国人(的), 汉语(的), 中国话(的), 中国(的), 中

As clearly indicated by its gloss, the adjective “Chinese” in this synset pertains to various aspects relating to China, while the Chinese synset, although reflecting these many potential facets, does not really contain synonyms, as 中国人 *zhōngguó rén* (Chinese people) and 中国话 *zhōngguó huà* (Chinese language) are both included.

4. Related but subtly different words

This situation is not simply a one-to-many correspondence, but there are more subtly defined Chinese lexical items which may only be coarsely represented by the same set of synonymous English words. For example:

00372111-a

brown, brownish, dark-brown, chocolate-brown (of a color similar to that of wood or earth)

咖啡色(的), 呈褐色(的), 黑褐色(的), 茶褐色(的), 棕色(的), 褐色(的)

Strictly speaking the Chinese words correspond to different hues and intensities of “brownness”, which are more specific than the English synset.

5. Contradictory connotation

Logically, lexical items or expressions with opposite connotations cannot be synonyms as they are not mutually substitutable in all contexts. For example:

00438909-a

sharp, shrewd, astute (marked by practical hardheaded intelligence)

狡黠(的), 锐利(的), 精明(的), 狡猾(的), 机敏(的), 诡计多端(的), 锋利(的)

The English items are somewhat neutral or even positive, which are more or less equivalently represented by 精明 *jīngmíng* and 机敏 *jīmǐn*, but 狡黠 *jiǎoxiá*, 狡猾 *jiǎohuá* and 诡计多端 *guǐjìduōduān* are obviously derogatory.

4 Handling Extra-synset Information

While it is intrinsically more difficult to define the synsets and concepts represented by adjectives due to their polysemy, even in PWN, the adjective database also reveals important conceptual and lexical gaps across languages. Multilingual wordnets, in this regard, would provide useful resources for language learning and translation, by humans and machines alike. It has been shown from the above discussion that apart from paying attention to cultural and linguistic differences across languages, building wordnets in other languages based on translation equivalents from bilingual dictionaries does not necessarily result in equivalent and valid synsets. This issue is a salient one, especially for languages with very different morphological properties and word formation mechanisms from English. For instance, while new words can easily be formed by inflectional and derivational morphology in English, the meaning carried by the additional morphemes may often be straightforwardly rendered with an extra word in Chinese, such as *un-X* to 不X (e.g. unhappy 不快乐 *bù kuàilè*) and *X-able* to 可X (e.g. respectable 可尊敬 *kě zūnjìng*)⁵.

Realising the importance and potential use of the multiple forms and renditions of a given meaning in Chinese, or other languages which are similarly distant from English, it would therefore be

⁵Sometimes disyllabic words as a more lexicalised form are available, e.g. 不快 *bùkuài* or 不乐 *bùlè* for “unhappy” and 可敬 *kějìng* for “respectable”, although they might be considered leaning toward classical Chinese.

value-adding to accommodate them in wordnets in some way. But the thesis in the current discussion is that the basic structure of synsets foundational to PWN should be maintained in multilingual wordnets. The following proposals are thus made to ensure that synsets are preserved as much as possible in target language wordnets while enabling language-specific properties and useful information to be captured:

1. An equivalent synset to a PWN synset should preferably contain only lexicalised items in the target language, unless no lexicalised translation equivalent is available. It is easy to get too far and result in over-interpretation with phrasal or clausal expressions. For example, synset **01251128-a** *cold* (having a low or inadequate temperature or feeling a sensation of coldness or having been made cold by e.g. ice or refrigeration) could be represented with 冰 *bīng*, 冻 *dòng*, 冷 *lěng*, 寒 *hán*, and perhaps the near-synonymous disyllabic words 冰冻 *bīngdòng*, 冰冷 *bīnglěng*, and 寒冷 *hánlěng*. The expressions above the lexical level, such as 气温低 *qìwēndī*, 温度不足 *wēndù bùzú* and 温度没有达到要求 *wēndù méiyǒu dá dào yāoqiú*, which are actually parallel to the gloss, should better be excluded from the synset.
2. The other non-lexicalised expressions which nevertheless convey the meaning close enough to the sense of the original synset, including but not limited to the examples above, could be stored in a separate class in a language-specific structure, instead of the core wordnet structure or the Inter-Lingual-Index. These separate and language-specific classes can be linked to the base concepts in WordNet with an *extension* pointer.
3. For very general adjectives, or those that are highly polysemous depending on the nouns being modified, similarly general equivalents, if available, should be included in the corresponding synset. The collocation-specific equivalents (that is, possible words actually used in the target language when the adjective is used to modify a particular noun) are different facets or even senses of the general adjective, and should therefore be captured at yet another subsuming level. This could be done in one of the two

ways. If PWN does not have a synset corresponding to a specific meaning of the general adjective, an extra synset can be introduced in the target language wordnet, with a *sub-level* pointer from the general adjective synset to the relevant senses as distinguished in the target language. Meanwhile, if there are existing adjective synsets corresponding to the specific adjectives in PWN, they could be linked as in PWN by relational pointers like *similar_to*. For example, synset **02569558-a** *sagacious, perspicacious, sapient* (acutely insightful and wise) could correspond to a Chinese synset with 睿智 *ruìzhì* with a pointer to the more general adjective synset like **02569130-a** *wise* (having or prompted by wisdom or discernment), while synset **00438909-a** *sharp, shrewd, astute* (marked by practical hardheaded intelligence) as discussed above, revised as 精明 *jīngmíng*, 机敏 *jīmǐn*, can point to synset **00438707-a** *smart* (showing mental alertness and calculation and resourcefulness). The two more general adjectives (wise and smart) can correspond to the more general Chinese adjectives like 聪明 *cōngmíng* and 聪颖 *cōngyǐng*.

4. In fact, very similar words like “clever”, “wise”, “smart”, “intelligent”, “sharp”, “sagacious”, “canny”, and many others, are not easy to distinguish in a clear manner. Subtle differences are also found among the many similar words in Chinese such as 聪明 *cōngmíng*, 聪颖 *cōngyǐng*, 聪敏 *cōngmǐn*, 机智 *jīzhì*, 睿智 *ruìzhì*, 英明 *yīngmíng*, 精明 *jīngmíng*, 明智 *míngzhì*, etc. It is nevertheless obvious, and perhaps intuitive to the native speakers, that 聪明 *cōngmíng* describes cleverness in a most general sense, and others describe a more specific aspect of cleverness, such as being mentally quick (e.g. 机智 *jīzhì*) or able to make wise decisions (e.g. 英明 *yīngmíng*). It is thus linguistically unsatisfactory to merge all these items into a particular synset. On the one hand, they may not be equally synonymous with one another as they tend to be used for a particular aspect of intelligence, depending on the usage context. On the other hand, the appearance of the same item in too many synsets may defeat the purpose of defining senses as such,

giving a distorted picture of sense distinction and polysemy. In this regard, the *pertainym* relation in PWN could be utilised in a target language wordnet for connecting adjective synsets with noun synsets to enhance the cross-POS relations in wordnets in addition to the morphosemantic links, like the synset with 英明 *yīngmíng* can pertain to both “human” and “decision”.

5. To ensure logical validity, words with contradictory connotation should be avoided in a synset. Similarly, phrasal expressions should be prudently handled as the same concept should not really correspond to both one lexical item and another form of it qualified by a degree adverb or so. For example, “very drunk” cannot be at the same time 喝醉 *hēzùi* and 烂醉 *lànzùi*, as the former only means “drunk after drinking” while the latter indicates how seriously one is drunk. Similarly, 贫困 *pínkùn* (impoverished) and 极度贫困 *jídù pínkùn* (extremely impoverished) cannot mean the same thing at the same time. The item which most matches the concept represented by the synset will suffice.

5 Conclusion

This paper has thus raised the issue of preserving the synonymy relation holding in synsets as the basic building blocks for wordnets in other languages, while taking advantage of the translation equivalents from other lexical resources as a starting point. Examples from Chinese were highlighted to illustrate how cross-linguistic differences especially in morphology and word formation may result in non-synsets in the process of building wordnet in a target language. It has been shown that the adjective database is particularly prone to the problem, especially for the relatively “general” concepts expressed by adjectives which can be used to describe many different entities and qualify a wide range of properties. To avoid non-synsets, it is thus suggested that partial equivalence be handled in a target wordnet by connecting the context-dependent equivalents to the basic synset with extra relational pointers. Although the alternative representation may not make any significant difference as far as the coverage and actual usage of the resource is concerned, it is nevertheless fundamentally important to keep the theoretical foundation intact.

Acknowledgements

The work described in this paper was partially supported by grants from the Faculty of Arts of the Chinese University of Hong Kong (Project No. 4051094) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 14616317).

References

- B.T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1352–1362, Sofia, Bulgaria.
- Jordan Boyd-Graber, Christiane Fellbaum, Daniel Oserson, and Robert Schapire. 2006. Adding dense, weighted, connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*, Jeju, Korea.
- Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Adjectives in WordNet. In George A. Miller, editor, *Five Papers on WordNet*. <http://wordnetcode.princeton.edu/5papers.pdf>.
- Christiane Fellbaum. 1998a. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Christiane Fellbaum. 1998b. A semantic network of English: The mother of all WordNets. *Computers and the Humanities*, 32(2/3):209–220.
- Chu-Ren Huang, Ru-Yng Chang, and Shiang-Bin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1553–1556.
- Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese WordNet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23.
- Bo Svensen. 1993. *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford University Press.
- Piek Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32(2/3):73–89.
- Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan.
- Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In John Domingue and Chutiporn Anuntariya, editors, *The Semantic Web: 3rd Asian Semantic Web Conference*, volume 5367, pages 302–314. Springer.

Lexical Perspective on Wordnet to Wordnet Mapping

Ewa Rudnicka,[♡] *Francis Bond*,[♣]
Lukasz Grabowski,[♣] *Maciej Piasecki*[♡] and *Tadeusz Piotrowski*[◇]

[♡]Wrocław University of Technology

[♣]Nanyang Technological University, Singapore

[♣]University of Opole

[◇]University of Wrocław

{ewa.rudnicka,maciej.piasecki}@pwr.edu.pl, bond@ieee.org, lukasz@uni.opole.pl, tadeusz.piotrowski@uwr.edu.pl

Abstract

The paper presents a feature-based model of equivalence targeted at (manual) sense linking between Princeton WordNet and plWordNet. The model incorporates insights from lexicographic and translation theories on bilingual equivalence and draws on the results of earlier synset-level mapping of nouns between Princeton WordNet and plWordNet. It takes into account all basic aspects of language such as form, meaning and function and supplements them with (parallel) corpus frequency and translatability. Three types of equivalence are distinguished, namely strong, regular and weak depending on the conformity with the proposed features. The presented solutions are language-neutral and they can be easily applied to language pairs other than Polish and English. Sense-level mapping is a more fine-grained mapping than the existing synset mappings and is thus of great potential to human and machine translation.

1 Introduction

Currently, bi- and multilingual wordnets are most commonly inter-linked on the synset level, (e.g., Bond and Foster, 2013). Synsets can be composed of one or more lexical units (lemma-PoS-synset triples, also called senses; henceforth, LUs), so such inter-wordnet links may be of three types: *1-to-1* sense link (between two synsets each built of a single LU); *1-to-many* sense link (between two synsets, one built of a single LU, the other of more than one); and *many-to-many* sense link (between two multiple-LU synsets). The (large) majority of inter-linked wordnets use one simple equivalence relation to connect their synsets (ef-

fectively synonymy). If, due to substantial differences between languages, such a link cannot be introduced, sometimes *artificial* synsets are created to provide equivalents (e.g., Bentivogli and Pianta, 2004; Lindén and Carlson, 2010). When we consider 1-to-many and many-to-many sense links, the question arises whether the correspondence between all their component LUs is of the same strength. Basic principles of language economy state that within one language there should not exist two different forms that share identical function and meaning, so there have to be slight differences between component LUs of a given synset, and even larger differences between the LUs from two synsets representing two different languages (even if those synsets are linked by I-synonymy). Existing research on inter-wordnet mapping between plWordNet (Maziarz et al., 2016) and Princeton WordNet (Fellbaum, 1998), especially 1-to-many and many-to-many sense links, has shown the potential for creating stronger links between some LUs from a given pair of synsets (Rudnicka et al., 2016). To give an example, in the pair of synsets: $\{\text{złoto}_{n:3}, \text{Au}_{n:1}\}^{PL}$ I-syn $\{\text{gold}_{n:3}, \text{Au}_{n:1}, \text{atomic number } 79_{n:1}\}^{EN}$ — $\text{złoto}_{n:3}^{PL}$ and $\text{gold}_{n:3}^{EN}$ and $\text{Au}_{n:1}^{PL}$ and $\text{Au}_{n:1}^{EN}$ seem the best-fitted equivalents due to the agreement not only in sense, but also in register. The words from the first pair belong to the general register, while the ones from the second pair are from the specialist register. Bi- and multilingual wordnets are used by translators who would certainly appreciate such a more detailed mapping.

2 Background

Equivalence is a popular concept used in, among others, translation studies and bilingual lexicography – see Rudnicka et al. (2017b) for a more detailed discussion, also regarding typologies of

equivalence). The concept has many faces depending on which features of language or texts researchers focus on. For example, one may find binary oppositions such as, for instance, natural and directional equivalence, semantic and pragmatic equivalence, or full and partial equivalence, e.g. Pym (2007); Svensen (2009). When studying recent approaches to equivalence developed in the field of bilingual lexicography, one may also find a distinction between cognitive and translational equivalence (Adamska-Sałaciak, 2010; Heja, 2016). Cognitive equivalents are typically general ones; they first come to the mind of a language user (even without any context) and when it comes to translation they may fit many contexts. Translational equivalents, which may be extracted from corpus data, may be less obvious and sometimes they may slightly differ in their basic meaning; however, they may fit more specific contexts. In Rudnicka et al. (2017a), we analysed basic equivalence types from translation and lexicographic literature and verified their relevance for synset-level wordnet mapping. We assumed that LUs in the linked pWLN-PWN synsets can be treated as bilingual dictionary data. We checked if pairs of LUs might be treated as cognitive and translational equivalents depending on their frequency of use as equivalents in translation in a particular co-text and context. We put forward an initial proposal of sense-level mapping designed to cross-cut through cognitive and translational equivalence. In this paper we present an extended and verified version of our initial proposal with carefully defined equivalence features, equivalence types and a sense-level linking procedure supported by a number of examples. At this point, it is important to note that the term equivalence has been also used in the context of wordnets; more precisely, it was first used in the wordnet world to name a set of inter-lingual relations holding between synsets in the EuroWordNet project (Vossen, 2002, p.:38). Inter-lingual synonymy was defined as a simple equivalence relation “which only holds if there is 1-to-1 mapping between synsets”. The remaining types of inter-lingual relations were called Complex Equivalence relations and allowed to obtain between one-to-many and many-to-many synset pairs. Since many of EuroWordNet wordnets relied on translation approaches, many of the senses can be translational equivalents. In designing a strategy for

mapping synsets between Princeton WordNet and pWordNet, Rudnicka et al. (2012) built on this proposal in the set of I-relations.

Currently, the overall size of pWordNet amounts to 217,426 synsets, 282,749 senses (lexical units) and 190,555 lemmas and these numbers are constantly growing. The synset mapping between pWordNet and Princeton WordNet encompasses 230,185 links, with 43,740 inter-lingual synonymy links. The number of Polish synsets with at least one inter-lingual relation is 177,634 (only I-synonymy is unique to a synset pair). The majority of I-links form noun links, 122,811 instances covering about 92% of Polish noun synsets (132,380 in total), the next are adjective links: 45,282 instances, covering 96% (46,721 in total), and last come adverb links with 9,541 instances, covering 84% of Polish adverb synsets (11,256 in total). At the present stage there are no inter-lingual links between verbal synsets (27,069), but we are working on the mapping procedure for them. Looking from the Princeton WordNet direction, we have mapped 80% of noun synsets (72,621), 43% of adjective synsets (7,905), and 47% of adverb synsets (1,737).

Since nouns are the most stable semantic category, we have decided to make them the starting point for our procedure for sense mapping. Other categories may require category-specific treatment which is outside the scope of the present paper.

3 Equivalence Features

In this section we discuss a set of features that will determine the strength of equivalence holding between (particular) LUs from the mapped Polish and English synsets. Each feature will be followed by a short definition and examples. First, we will look at *formal features*, such as grammatical category, number, countability and gender. Next, we will delve into *semantic* and *pragmatic* ones, such as sense, lexicalisation (of concepts), register, collocations, co-text and context. Finally, we will consider translatability based on dictionary listing and translation equivalences extracted from the Polish-English parallel corpus *Paralela*¹ (Peżik, 2016).

¹ <http://paralela.clarin-pl.eu>

3.1 Formal features

The first, basic, formal feature is identity in *grammatical category* between source and target LUs. Since sense-level mapping will be based on the results of an earlier synset-level mapping, this feature will be treated as ‘given’. The inter-lingual relations that will be taken into consideration include I-synonymy, I-partial synonymy, I-hyponymy and I-hypernymy, all of which hold between the same part of speech synsets. Our focus will be the relations between nouns.

The more interesting formal features are *number* and *countability*. For regular, countable nouns, agreement in these features is usually also given, because both in plWordNet and in Princeton WordNet lemmas appear in singular form. Still, some cases of ‘mixed’ Princeton WordNet-synsets were already tracked e.g. *{dumpling_{n:1}, dumplings_{n:1}}* (Rudnicka et al., 2012, 2016). Such mixed synsets currently serve as inter-lingual hypernyms for both singular and plural Polish synsets e.g. *{pieróg_{n:2}, pierog_{n:1}}* ‘dumpling’ or *{pierogi ruskie_{n:1}}* ‘Russian dumplings’. Still, sense level mapping will allow to resolve such inconsistencies in the synset built-up. In regular cases, the agreement in number will always be observed in the mapping.

A different case are pluralia and singularia tantum that have regular countable nouns as equivalents in another language such as, for instance, *{drzwi_{n:1}}* ‘door^{pl}’ I-syn *{door_{n:1}}*, *{grabie_{n:1}}* ‘rake^{pl}’ I-syn *{rake_{n:1}}*, *{centrala_{n:2}}* I-syn *{headquarters_{n:1}}*, or *{stajnia Augiasza_{n:1}}* ‘Augeas’ stable’ I-syn *{Augean stables_{n:1}}*. A re-analysis of their relation structures and glosses shows very close meaning correspondence and leads to the conclusion that the difference in number is only a difference in grammaticalisation of the same concept. A similar case are regular nouns mapped to mass or group nouns, such as *{grzmot_{n:1}}* ‘thunder^{sg}’ I-syn *{thunder_{n:2}}* or *{błyskawica_{n:1}}* ‘lightning^{sg}’ I-partial-syn *{lightning_{n:2}}*. There are also cases of pluralia tantum mapped to uncountable nouns e.g. *{wagary_{n:1}}* ‘truancy^{pl}’ I-syn *{truancy_{n:1}, hooky_{n:1}}*. On the basis of the above examples, we want to argue that identity in number and countability is an important criterion only in the case of regular, countable nouns. Cases of singularia and pluralia tantum should be dealt with on an individual basis. The features may gain more importance in the case

of 1-to-many and many-to-many sense pairs e.g. *{odwiedziny_{n:1}, wizyta_{n:1}}* I-syn *{visit_{n:1}}*.

The last formal feature is *gender*. One of the typical differences between a morphologically synthetic language (Polish) and an analytical one (English) is the degree of gender lexicalisation. Gender is systematically lexicalised in Polish, marked by derivational suffixes e.g. *nauczyciel* ‘teacher’ and *nauczycielka* ‘female teacher’, while it is much less lexicalised in English — it is sometimes signalled by derivational suffixes e.g. *emperor* – *empress*, sometimes by different, derivationally unrelated words e.g. *mare* – *stallion*. We suggest to constrain sense links with gender identity between LUs only in cases where both languages lexicalise the distinction, while in the remaining, contrasting cases mark the equivalence as slightly weaker than in former ones. Such a proposal is motivated by the fact that we consider information about natural gender to be an additional meaning component. Thus, we get very close correspondence between LUs in the following synset pairs: *{ogier_{n:1}}* I-syn *{stallion_{n:1}, entire_{n:1}}* and *{klacz_{n:1}}* I-syn *{mare_{n:1}, female horse_{n:1}}*, while just close correspondence between the pairs *{nauczyciel_{n:1}}* and *{nauczycielka_{n:1}}* I-hypo to *{teacher_{n:1}}*.

3.2 Semantic features

As already alluded in the previous section, the key denominator for LU mapping will be the correspondence in sense. By definition, the component LUs of a given synset do share the same (basic) meaning (Fellbaum, 1998). Still, in such a model, some more subtle meaning distinctions may not be captured, such as shades of meaning going beyond Leibniz’s (1704) truth-conditional understanding of synonymy. Other factors that determine meaning are similarities and differences in lexicalisation of concepts, register, style, typical co-texts and contexts. They need not be of importance in some language processing tasks, but are always important for a translator. Therefore, the proposed sense-level mapping aims to go beyond the existing synset level mapping in the granularity and specificity of links. Currently, the I-synonymy link between synsets signals their correspondence in sense based mainly on their synset relation network (and partly on glosses and examples of use that come with synsets in Princeton WordNet and with LUs in plWordNet). In LU

mapping, we would like to re-analyse the existing inter-lingual synset links, and wherever possible, establish sense links of a stronger character. We see the potential for stronger sense links especially in the case of 1-to-many and many-to-many sense pairs. For these purposes, we will need to consult external resources such as mono- and bilingual dictionaries, encyclopaedia, and mono- and parallel corpora.

An example of 1-to-many sense pair is the Polish synset {*narzeczone*_{n:1}} ‘fiancee’ linked via I-synonymy to the English synset {*fiancee*_{n:1}, *bride-to-be*_{n:1}}. The Polish gloss can be translated as “a woman who obliged herself to marry a concrete man (her fiance), made him such a promise”, while the English one is just “a woman who is engaged to be married”. Having consulted a couple of monolingual English dictionaries Cobuild (2012); CALD (2013); LDCE (2014), we find that *fiancée* is defined as “the woman that a man is engaged to/going to marry”, while *bride-to-be* as “a woman who is going to be married soon”. Clearly, there is an additional meaning component in the case of *bride-to-be*, namely *soon*, not included in the general synset gloss. The synset gloss corresponds more closely to the dictionary definitions of *fiancée* and to the Polish gloss of *narzeczone*_{n:1}. Therefore, there is a stronger link between lexical units *narzeczone*_{n:1} and *fiancée*_{n:1} than between *narzeczone*_{n:1} and *bride-to-be*_{n:1}.

An important factor influencing equivalence between LUs of the two languages are similarities and differences in lexicalisation of the same concepts. These will be judged by comparing the denotations of bilingual pairs of LUs. An example is the Polish word *zabytek*_{n:2} ‘historic monument’ with the gloss: “stary budynek, przedmiot” ‘an old building, artefact’ which denotes anything of historic value no matter of its size. There is no direct equivalent of this word in English. One has to use a different noun depending on the size of an object e.g. *historic monument*, *building*, *site*, *landmark*. The Princeton WordNet synset with the closest meaning is {*monument*_{n:2}} with the following gloss: “an important site that is marked and preserved as public property”, an instance hyponym {*Stonenhenge*_{pn:1}} and a hyponym {*market cross*_{n:1}}. The two synsets {*zabytek*_{n:2}} and {*monument*_{n:2}} are linked by I-partial synonymy. In some contexts *monument*_{n:2} will be the best translation of *zabytek*_{n:2}, yet their overall mean-

ing correspondence is partial.

Another area to look for more meaning specification is *register*. More precisely, registers are marked only for very few Princeton WordNet synsets by means of the *Domain Usage* relation, of which a couple of specifiers are of interest to us, namely {*archaism*_{n:1}}, {*colloquialism*_{n:1}}, {*disparagement*_{n:1}}, {*ethnic slur*_{n:1}}, {*formality*_{n:3}}, {*vulgarism*_{n:1}} and {*slang*_{n:2}}. In plWordNet registers are marked for lexical units and the following ones are distinguished: *general*, *official*, *specialist*, *literary*, *colloquial*, *common*, *vulgar*, *obsolete*, *regional*, *slang/argot* and *non-normative*. There are some cases of correspondence in register systems between English and Polish e.g. {*big fish*_{n:1}, ...} linked by Domain Usage relation to {*colloquialism*_{n:1}} and via I-partial synonymy relation to the Polish synset {*gruba ryba*_{n:1} ‘big fish’, *ważniak*_{n:2} ‘VIP’} with both its LUs marked for the colloquial register. However, such simple cases are rare. Both in Princeton WordNet and in plWordNet, LUs of different registers can co-occur in the same synset. However, in the latter only LUs of *compatible* register can be grouped in one synset or linked by some relation, e.g. hypernymy. A set of rules was defined for this purpose in plWordNet (Maziarz et al., 2014), while this aspect is largely unconstrained in Princeton WordNet. General, specialist, literary, and official registers can co-occur in one synset; the same holds for general and colloquial ones (provided that that specialist, literary and official are not found in the same synset). Colloquial, common and vulgar can also come together. On the other hand, regional, obsolete, slang/argot and non-normative always come on their own. An example is the Polish synset {*okulary*_{n:1} ‘glasses’: *general* register, *patrzalki*_{n:1}, *szkła*_{n:1} ‘specs’: *colloquial* register, *binokle*_{n:2} ‘eyeglasses’: *colloquial* register}. *okulary*_{n:1}’s gloss is translated to “an optical device built of a pair of lens and a frame enabling fitting the lens in front of the eyes most often by ear arms, usually used to correct sight acuity, weakened by an illness, injury or age).. It is linked by I-synonymy relation to the English synset {*spectacles*_{n:1}, *specs*_{n:1}, *eyeglasses*_{n:1}, *glasses*_{n:1}} “(plural) optical instrument consisting of a frame that holds a pair of lenses for correcting defective vision”. There is no information about register for the Princeton WordNet synset. Still, when we look

up its component LUs in English dictionaries we find that *spectacles* is classified as either formal or old-fashioned, *specs* as informal and *eyeglasses* as North American. That suggests a strong link between *okulary*_{n:1} with *glasses*_{n:1} (both of a general register), and possibly also with *eyeglasses*_{n:1} (though maybe by a slightly weaker link), while *patrzalki*_{n:1} and *szkła*_{n:1} with *specs*_{n:1} (all of an informal or colloquial register). In fact, the Polish word *binokle*_{n:2} marked with a colloquial register also has an old-fashioned flavour, which makes it a good equivalent for the English *spectacles*_{n:1}.

An important means for disambiguating sense are *collocations*, *co-text* (co-occurring words and text fragments) and *context* (type of situation, speaker, target audience, purpose of communication, style etc.). Words with the same meaning that appear in similar language environments in two languages tend to be equivalents of each other. It can be illustrated by LUs from the following pair of synsets: {*centrala*_{n:2}} linked via I-synonymy to {*headquarters*_{n:1}, *office*_{n:1}, *main office*_{n:1}, *home office*_{n:2}, *home base*_{n:2}}. The pair of LUs *centrala*_{n:2} – literary a noun LU derived from the adjective *centralny* ‘central’– and *headquarters*_{n:1} gets 40 hits in the *Paralela* corpus and a couple of concordances illustrating the use of these two equivalents in their co-text can be distinguished, e.g.:

- *Jesienią 2007 r. duńska centrala firmy Arriva poszukiwała ponad 400 kierowców autobusów...*
‘In the autumn of 2007, Arriva’s Danish headquarters were looking for more than 400 bus drivers...’
- *Ponieważ jej europejska centrala znajduje się w Irlandii, ...*
‘As their European headquarters is located in Ireland, ...’
- *Do pierwszego sprawozdania centrala wydała krótki komentarz,...* ‘Headquarters commented briefly on the first report, ...’

Other LUs in the English synset (*central office*_{n:1}, *main office*_{n:1}, *home office*_{n:2}, and *home base*_{n:2}) either do not appear in a pair with *centrala* or are quite rare.

3.3 Translatability

We have already seen in the previous section that dictionaries and corpora are indispensable re-

sources in determining many features of equivalence, because they provide different types of information that may be missing in wordnets (e.g. register, collocations or typical co-text or contexts). In the process of construction of contemporary bilingual dictionaries a lot of emphasis is put on the translatability of the provided equivalents (e.g., Zgusta, 1971), with better translation equivalences listed first. Therefore, we would like to suggest that *dictionary listing* be treated as one of the indicators of the strength of equivalence between LUs. The main Polish-English/English-Polish dictionaries to be consulted will be PWN-Oxford (2007), Collins-YDP (1997) and Słownik-Kościszewski (2014). An issue that immediately emerges here is directionality of translation. It is known that not all equivalents work equally well both ways, that is from L1 to L2 and from L2 to L1. It can be verified by the so-called back-translation, also using dictionaries. In the extreme case it there is not always an equivalent provided for a headword when you try to back translate.

Translation theorists distinguish between natural equivalence and directional equivalence. According to Pym (2007), natural equivalence describes the correspondence between words, expressions or text chunks on all dimensions of meaning. It typically concerns terminology (e.g. {*duck*_{n:1}} I-syn {*kaczka*_{n:1}} ‘duck’, both belonging to the semantic domain *animal*), prefabricated chunks of texts and specialized uses of words (e.g. *whereas* – *zważywszy, że* as found in certain legal texts), so it seems to exist prior to translation. On the other hand, directional equivalence refers to situations when translators actively search for equivalents of source words in the target language (often in cases of lexical or cultural gaps), so it is by definition uni-directional or one-way. An example is the Polish synset {*stachanowiec*_{n:1}, *przodownik pracy*_{n:1}} whose gloss translates to ‘in the Eastern Block countries: a person competing for a title of a most efficient worker’. It is linked via I-hyponymy to the English synset {*toiler*_{n:1}} gloss: “one who works strenuously”. As shown by the gloss, *stachanowiec* is a typical cultural gap; the concept is specific to Eastern Block countries. Its I-hypernym, *toiler* can serve as a translational equivalent from Polish to English, yet back-translation does not work in this case (cf Techland-Dictionary (2006): *toiler* – *człowiek ciężkiej pracy* ‘a man of hard work’.)

4 Equivalence types

Relying on equivalence features described in the previous section, we will define three equivalence types of a variable strength: *strong*, *regular* and *weak* (implied). The categorisation to a given type will be based on values of features a bilingual pair of LUs will agree in. The types will be later reflected in three kinds of links between LUs.

Some features will be agreed across all types, while some other feature will differ. Summing up the discussion in Section 3.1, there will always be an agreement in grammatical category (only noun-to-noun pairs are taken into consideration) and in most cases in number, countability and gender. Instances of pluralia and singularia tantum as well as count-to-mass mappings will be dealt on an individual basis – the agreement will not always have to hold. Cases of lexicalised natural gender in Polish will be treated in a similar way.

4.1 Strong equivalence

By its very name, the strong equivalence will be the strongest type of link. It will require identity in sense, similarity in lexicalisation of concepts, compatibility in register, a shared set of typical co-texts, dictionary listing (preferably as the first equivalent), bidirectionality (but not uniqueness) of translation and, preferably, frequent parallel corpora hits. The most suitable candidates for such strong correspondence are LUs from one element (LU) synsets linked via I-synonymy synset relation. A couple of examples are given below (for their full descriptions see Sections 3.1 and 3.2):

- *drzwi*_{n:1} I-syn *door*_{n:1}

- *grzmot*_{n:1} I-syn *thunder*_{n:2}

All strong because of identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

The second group of examples to consider are one-to-many sense pairs of synsets linked via I-synonymy. It is likely that there will be at least one pair of LUs that will meet the strong equivalence criteria. Below we present instances of such pairs of LUs (for their full descriptions see Sections 3.1 and 3.2):

- *narzeczona*_{n:1} I-syn *fiancee*_{n:1}

- *centrala*_{n:2} I-syn *headquarters*_{n:1}

- *gruba ryba*_{n:1} I-partial-syn *big fish*_{n:1}

All strong because of identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

The last group of synsets to look at are many-to-many sense pairs, among which we are likely to find pairs of LUs that can function as strong equivalents of each other. These are illustrated below (for their full description see Section 1 and 3.2):

- *złoto*_{n:3}^{PL} I-syn *gold*_{n:3}^{EN}

- *okulary*_{n:1}^{PL} I-syn *glasses*_{n:3}^{EN}

For all, identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

4.2 Regular equivalence

The regular equivalence will be a slightly weaker type of link than the strong one, but it will still signal clear correspondence in a number of features. It will require large similarity in sense, compatibility in register, dictionary listing, bidirectionality of translation, a similar set of typical co-texts and, preferably, some parallel corpora hits. It will allow for some differences in lexicalisation of concepts. Examples of regular equivalence links from one-to-many sense pairs are given below (for their full descriptions see Section 3.2):

- *zabytek*_{n:1} I-partial-syn *monument*_{n:2}

Lexical gap (on the English side)

- *narzeczona*_{n:1} I-syn *bride-to-be*_{n:1}

Additional (temporal) sense specification on the English side; few parallel corpora hits

- *centrala*_{n:2} I-syn *central office*_{n:1}

Few parallel corpora hits for this pair

Instances of regular equivalence can also be found within many-to-many sense pairs. Below we illustrate them with instances of Polish grammaticalised gender (for their full description see Section 3.1) :

- *nauczyciel*_{n:1} I-syn *teacher*_{n:1}

- *nauczycielka*_{n:1} I-hypo *teacher*_{n:1}

Examples of Polish grammaticalised gender

4.3 Weak equivalence

Since translatability can be achieved by very different means, we would like to point out that in certain contexts even LUs from pairs that do not meet all the criteria for strong or regular equivalence can function as translational equivalents. We will call such type of equivalence weak (or implied) equivalence. It will be postulated for pairs of LUs from plWordNet and Princeton WordNet synsets linked by I-synonymy, I-partial synonymy and I-hypernymy that do not meet the criteria for strong or regular equivalence, and can be automatically derived from the synset-level links. Often these will be instances of culture specific concepts absent from the second language (cultural gaps) and linked via I-hyponymy relation. An example of such weak equivalence link is given below. It obtains for both component LUs of the Polish synset given below (for its full description see Section 3.2.):

- {*stachanowiec*_{n:1}, *przodownik pracy*_{n:1}} I-hypo {*toiler*_{n:1}}
Polish culture specific term, with no direct equivalent

We expect that, except for instances of lexical gaps and gender lexicalisation where bidirectionality of translation does hold, the majority of I-hyponymy and I-hypernymy synset links will be unidirectional in terms of translation and thus pairs of their component LUs will be treated as weak equivalents.

5 Linking procedure

Having defined the equivalence features and types, below we put forward a linking procedure for lexicographers. In the procedure we lead lexicographers from simpler to more complex features and from wordnet data to dictionary and corpora data. We believe that there is no need for a lexicographer to verify each feature separately, but that they can be analysed in groups or pairs on the basis of the data provided by a specific resource.

We will illustrate the linking procedure on the example of the pair of synsets {*centrala*_{n:2}} linked via I-synonymy to {*headquarters*_{n:1}, *central office*_{n:1}, *main office*_{n:1}, *home office*_{n:2}, *home base*_{n:2}}. Formal features that is number, countability and gender should be verified first. Gender is not relevant here, since we do not deal with an animate noun. On the other hand, we have

an instance of a pluralia tantum in the English synset: *headquarters*_{n:1}. The remaining lexical units are regular countable nouns. Next, we move to semantic (and partly pragmatic) features starting from the data provided in wordnets that is relations, glosses, qualifiers and examples. The key relations are hypernyms and hyponyms, as well as their I-synonyms or I-hypernyms. The Polish synset {*centrala*_{n:2}} has {*ośrodek*_{n:2}, ...} - 'center' as its hypernym, which is an I-synonym of the English {*centre*_{n:4}, ...}. It is glossed as: "siedziba centrali, główny ośrodek czegoś" - 'the headquarters' seat, main centre of something'. It has general register and the usage example is the following: "Pożar centrali mleczarskiej w miejscowości obok było widać z daleka." - 'The fire in the dairy center in the nearby place could be seen from the distance.' The English synset {*headquarters*_{n:1}, ...} has {*office*_{n:1}, *business office*_{n:1}} as its hypernym. It is attributed with the following gloss and example: "(usually plural) the office that serves as the administrative center of an enterprise; "many companies have their headquarters in New York." There is no information about the register provided.

Next, in order to gather still more information about semantics and pragmatics as well as translatability of pairs of particular LUs, lexicographers are asked to consult external resources such as dictionaries and encyclopedias as well as a Polish-English parallel corpus *Paralela*. Looking up *centrala* in a couple of Polish-English dictionaries (see ...), we find that its most frequent equivalents are *headquarters*, *head office* and *central office*. Interestingly, *head office* does not appear in Princeton WordNet at all. Looking up *headquarters* in English-Polish dictionaries, we obtain *centrala* and *siedziba główna* (the latter term appearing in the gloss of the Polish synset); checking *central office*, we get *siedziba główna* and *centrala*. In the next step, we check the frequency of the pairs *centrala* – *headquarters* and *centrala* – *central office* in the *Paralela* corpus and we learn that the pair *centrala* and *headquarters* gets 40 hits, while *centrala* – *central office* gets only 3 hits. In the last step, we analyse the most frequent contexts of occurrence of *centrala* – *headquarters* and we get a couple of typical shared contexts and collocations (examples given in Section 3.2.) On the basis of the whole discussed data, we want to argue that the lexical units *centrala*_{n:2} -

*headquarters*_{n:1} form a pair of strong equivalents, *centrala*_{n:2} - *central office*_{n:1} are regular equivalents, while *centrala*_{n:2} - *main office*_{n:1}, *home office*_{n:2}, *home base*_{n:2} should be treated as weak equivalents.

6 Conclusions

The strategy for sense-level mapping between Princeton WordNet and plWordNet nouns put forward in this paper is a new initiative in the wordnet world. It offers a possibility for fine-grained mapping that is of great potential especially for human and machine translation. It is illustrated with examples from the Polish-English language pair, but the set of features described in this paper are language-neutral and they can be easily extended to wordnets of other languages of the Indo-European family. As for (non)-Indo-European language pairs, it is necessary to analyse whether the two languages share all the features that will be taken into account. Also, the strategy may be extended to other grammatical categories such as adjectives and adverbs, which are already partially mapped on the synset level, and, eventually, to verbs after some mapping between verb synsets is accomplished. It may well be that additional features will need to be introduced while some of the ones proposed for nouns might be dismissed as irrelevant.

The proposed strategy is designed for manual mapping, but we plan to develop an automatic system of prompts that will support lexicographers' work. The new system will be an extension of an earlier system of automatic prompts for mapping of noun synsets and based on a modification of the Relaxation Labelling algorithm of Daudé et al. (1999) joined with lemma-pair checking and filtering by a large Polish-English cascade dictionary Kędzia et al. (2013) and translation probabilities from bilingual corpora.

As regards future avenues, this study may be continued in a number of possible ways. Firstly, the strategy of sense-level mapping described in this paper should be further tested on a structured and balanced sample of concrete and abstract nouns representing the whole variety of semantic domains (lexicographers' files). We plan to extract the lists of Polish-English lexical unit pairs from the Polish-English pairs of synsets linked by I-synonymy, I-partial synonymy and I-hyponymy (both Polish-English and English-Polish). The

reason for that is that pairs linked by these relations are most likely to yield strong and regular equivalents. We will (proportionally) explore all three possible types of pairing, that is 1-to-1 sense match, 1-to-many sense match and many-to-many sense match.

Secondly, in order to pinpoint any translation tendencies, the next step should be to calculate translation probabilities for pairs of equivalents, preferably in both directions, extracted from parallel corpora (e.g. *Paralela*). This would enable the verification of the degree to which sense-level mapping is reflected in translated texts found in a parallel corpus. Obviously enough, translation probabilities should be interpreted with caution given the limitations of any parallel corpus used (its size, structure, representativeness, balance, scope of annotation, etc.). At this point, it is also important to note that searching through parallel corpora is problematic when one deals with polysemous lexical units. The lack of word-sense disambiguation (or, in other words, semantic tagging of bilingual corpus data) means that when we consult a parallel corpus, we search for language forms rather than senses; that is why translation probabilities should be calculated in a way reflecting polysemy of lexical units. All this should enable one to further test, verify and improve the linking procedure proposed in this paper, which can be useful for anyone interested in applying it for sense-level mapping of wordnets representing languages other than Polish and English.

Acknowledgment

This work was supported by the National Science Centre in Poland under the agreement No. UMO-2015-/18/M/HS2/00100.

References

- Arleta Adamska-Sałaciak. 2010. Examining equivalence. *International Journal of Lexicography*, 23(4):387–409.
- Luisa Bentivogli and Emanuele Pianta. 2004. Extending wordnet with syntagmatic information. In *Proceedings of the Second Global WordNet Conference*, pages 47–53. Brno, Czech Republic.
- Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.
- CALD. 2013. *Cambridge Advanced Learner’s Dictionary*. Cambridge University Press, Cambridge, fourth edition.
- Cobuild. 2012. *Collins Cobuild Advanced Dictionary of English*. Heinle ELT, Boston, 7th edition.
- Collins-YDP. 1997. *Multimedialny słownik angielsko-polski i polsko-angielski Collins*. Polska Oficyna Wydawnicza, Warszawa. Version 3.03 (computer program).
- Jordi Daudé, Lluís Padró, and German Rigau. 1999. Mapping multilingual hierarchies using relaxation labeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 12–19.
- Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.
- Eniko Heja. 2016. Revisiting translational equivalence: Contributions from data-driven bilingual lexicography. *International Journal of Lexicography*.
- Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies*. To appear.
- LDCE. 2014. *Longman Dictionary of Contemporary English*. Pearson Education, Harlow, 6th edition.
- Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNordica*, 17.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plWordNet as the cornerstone of a toolkit of lexico-semantic resources. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312. University of Tartu Press, Tartu, Estonia. URL <http://aclweb.org/anthology/W14-0142>.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL. URL <http://aclweb.org/anthology/C/C16/>.
- PWN-Oxford. 2007. *Wielki słownik angielsko-polski PWN-Oxford*. Wydawnictwo Naukowe PWN S.A. and Oxford University Press, Warszawa.
- Anthony Pym. 2007. Natural and directional equivalence in theories of translation”. *Target*, 19(2):271–294.
- Piotr Pęzik. 2016. Exploring phraseological equivalence with paralela. In Ewa Gruszczyńska and Agnieszka Leńko-Szymańska, editors, *Polish-Language Parallel Corpora*, pages 67–81. Instytut Lingwistyki Stosowanej UW, Warszawa.
- Ewa Rudnicka, Francis Bond, Łukasz Grabowski, Maciej Piasecki, and Tadeusz Piotrowski. 2017a. Towards equivalence links between senses in plWordNet and Princeton WordNet. *Lodz Papers in Pragmatics*, 13(1):3–24.
- Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proc. COLING 2012, posters*, pages 1039–1048.
- Ewa Rudnicka, Maciej Piasecki, Tadeusz Piotrowski, Łukasz Grabowski, and Francis Bond. 2017b. Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies / Études cognitives*, 17. In print.
- Ewa Rudnicka, Wojciech Witkowski, and Łukasz Grabowski. 2016. Towards a methodology for

filtering out gaps and mismatches across word-nets: the case of noun synsets in plWordNet and Princeton WordNet. In B. Barbu Mititelu, C. Forascu, Ch. Fellbaum, and P. Vossen, editors, *Proceedings of the 8th International Global WordNet Conference 2016*, pages 344–351. Global WordNet Association, Bucharest, Romania. URL <http://gwc2016.racai.ro/proceedings.pdf>.

Bo Svensen. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press., Cambridge.

Słownik-Kościuszkowski. 2014. *Nowy słownik Fundacji Kościuszkowskiej polsko-angielski i angielsko-polski*. TAIWPN Universitas, Kraków.

Techland-Dictionary. 2006. *Wielki słownik angielsko-polski, polsko-angielski*. Techland. Version 1.0.1 (computer program).

Piek Vossen. 2002. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam.

Ladislav Zgusta. 1971. *Manual of Lexicography*, volume 39 of *Series Maior*. Janua Linguarum, The Hague. Pari.

ReferenceNet: a semantic-pragmatic network for capturing reference relations

Piek Vossen, Marten Postma, Filip Ilievski

VU University Amsterdam, Netherlands

{piek.vossen, m.c.postma, f.ilievski}@vu.nl

Abstract

In this paper, we present ReferenceNet: a semantic-pragmatic network of reference relations between synsets. Synonyms are assumed to be exchangeable in similar contexts and also word embeddings are based on sharing of local contexts represented as vectors. Co-referring words, however, tend to occur in the same topical context but in different local contexts. In addition, they may express different concepts related through topical coherence, and through author framing and perspective. In this paper, we describe how reference relations can be added to WordNet and how they can be acquired. We evaluate two methods of extracting event coreference relations using WordNet relations against a manual annotation of 38 documents within the same topical domain of gun violence. We conclude that precision is reasonable but recall is lower because the WordNet hierarchy does not sufficiently capture the required coherence and perspective relations.

1 Introduction

Synonyms from the same synset (Fellbaum, 1998) are assumed to be exchangeable in contexts. Similarly, word embeddings are based on sharing of contexts represented as vectors (Mikolov et al., 2013; Baroni et al., 2014). Both synsets and word embeddings capture some variation in language, but they do not fully capture variation in reference and coreference. Reference relations are different in that they cross local (sentence) contexts. We typically tell stories in discourse in which entities or events play different roles and reflect different phases in relation to the same incident (the topic of the story). Furthermore, authors may frame these entities and events differently either within the same story or across different stories. We can thus consider a story as a larger topical context within which co-referring expressions occur in different local contexts. Each local context of a co-referring expression may represent a different concept. The set of local contexts within a topical context is therefore expected to express not only similarity, but also topical coherence and author framing and perspective.

The next two examples show two fragments from two news articles that make reference to the same incident (topical coherence) in which a man shot several people in a bar in Pittsburgh. The first fragment is published shortly after the incident when the suspect has not yet been identified. The second fragment is published later after the suspect was identified, found guilty and sentenced to prison (changing perspective).

Investigators continue to look for suspects after one person was killed and four others were injured when gunfire erupted overnight at a bar in Homewood Several witnesses , [...] They believe the gunman was not searched by the four security guards who left the business before police arrived .

Man Gets 15 - 30 Years For Deadly Shooting At Homewood Bar . PITTSBURGH (KDKA) A man has pleaded guilty in a 2014 shooting that left four men injured and one dead . Cornell Poindexter , 30 , appeared in court Monday and pleaded guilty to one count of 3rd degree murder , four counts of aggravated assault and one count of person not to possess a firearm According to our partners at The Pittsburgh Post - Gazette , 23-year - old Corey Clark was originally accused of being the gunman , but those charges were dropped .

The following words and expressions are used to make reference to the incident or parts of the incident: *killed, injured, gunfire erupted* (first fragment) and *deadly shooting, shooting, left injured and dead, murder, aggravated assault* (second fragment). The references to the shooter are made through *suspects, gunman* and through *man, Cornell Poindexter, person not to possess a firearm, 23-year - old Corey Clark* and *gunman* respectively. References to the events differ across the text due to the legal view, e.g. *murder*, whereas the entity references differ due to having more knowledge on the identity of the suspects and the fact that one suspect turned out to be innocent and another was convicted. Making reference is more than similarity of meaning, as it is also governed by pragmatic principles related to information sharing, relevance, salience, and framing. In the different sentences of a discourse, we tend to tell different things about the same referents. These sentences thus represent different local contexts, which are connected through the topical context of the story that is told. From a language understanding and generation

perspective, WordNet synsets and word embeddings are not expected to provide sufficient information to predict usage of one expression over the other, or to infer from the referential usage of expressions what is the semantic implicature (coherence or framing).

We therefore propose to add a layer to WordNet, that captures variation in making reference within a topical context across different synsets or word embeddings that represent local contexts. In this paper, we describe how these relations can be acquired as a **ReferenceNet**. The relations in a ReferenceNet exceed the notion of synonymy and partially also hyponymy and capture a broader range of roles, perspectives, and also different phases of processes. Referential relations can not only help detecting coreference and coherence relations, but also help distinguishing roles from rigid types which is important for further ontologisation of semantic networks, and capturing different ways of framing the same things.

This paper is structured as follows. In section 2, we discuss related work and present the motivation for adding referential relations to WordNet. In section 3, we define the model for expressing these relations. We present two approaches to acquire these relations in sections 4 and 5. Section 6 describes the evaluation data created and section 7 contains the evaluation results. Finally, we conclude and discuss future work in section 8.

2 Related work and motivation

Reference and identity have been discussed extensively in the philosophical literature (Quine and Van, 1960; Kripke, 1972; Putnam, 1973; Frege, 1892; Rast and others, 2007; Wittgenstein, 2010). The linguistic field of lexical pragmatics (Levinson, 1983; Matsumoto, 1995; Blutner, 1998; Weigand, 1998) tries to explain variation in reference as a function of pragmatic principles such as the Gricean maxims (Grice et al., 1975): be maximally informative but no more informative than necessary. Variation of form is partly explained through pragmatic licensing: the least complex form that yields the most salient implicatures is preferred among all forms that can potentially yield these implications. Such principles may predict how we make reference to real-world situations using certain words and expressions and not others, given the shared knowledge we have about these situations.

The way we make reference is however not only determined by efficiency, salience and information sharing, but also by the framing of referents by the author. FrameNet (Baker et al., 2003) is a large resource that describes different ways in which situations can be framed. Frames and frame elements in FrameNet are very specific and typically model the specific realisation of lexical units in texts. It is not clear how to generalise over the specific frames (what do they share or have in common) nor to derive from the database which combinations of frames can be expected within specific

topical contexts.

We believe it is worthwhile to investigate empirically the actual referential relations that occur within topical contexts at a large scale, as well as to describe the observed lexical variation according to both pragmatic principles of quality and efficiency and framing principles. We therefore propose a ReferenceNet as a data structure that captures the observed coherence and framing relations between WordNet synsets. ReferenceNet therefore extends WordNet with a new orthogonal relation, which is less strict and limited than FrameNet, and more specific than for instance WordNet Domains (Strapparava et al., 2004). We argue that such data can be potentially very valuable, as it enables our community: 1. to investigate the semantic-pragmatic implications of making reference 2. to learn about the contextual roles and perspectives that govern the usage of these words and expressions, and 3. to improve the detection of these relations by coreference systems.

3 The ReferenceNet model

We define a ReferenceNet as a collection of **ReferenceSets**. A ReferenceSet consists of:

1. the **words and expressions** that have been used to make reference to the same individual in a topical context
2. the list of different **synsets** associated with these words and expressions in this context
3. the **type of topical context** in which the reference relation was observed

As synsets represent concepts, the variety of synsets reflects the range of things or denotation that is captured in a single ReferenceSet. As this range is not ontologically defined, it reflects the *typical* ways in which we frame and conceptualize individuals in topical situations. Typically, these synsets cannot be disjoint (mutually exclusive): they should either belong to the same hypernym chain (being more or less specific), or should be orthogonal according to formal ontological criteria (Guarino, 1999). A ReferenceSet may consist of one or more synsets and the same synset may participate in more than one reference set, thus constituting a ‘many-to-many’ relation. In addition to the synset of the expression, we also need to record the actual form or synonym from the synset that was used to make reference.¹ As the constraints for making reference with different expressions and different concepts are mildly ontological, it make sense to register the referential usage of expressions and synsets using counters.

¹Note that in case of proper names, we abstract from the proper name to the most specific WordNet synset or entity type of which the entity is an instance. When building ReferenceSets from large text collections it makes sense to leave out the proper name references, as we would otherwise include all people’s names in the ReferenceNet.

Finally, ReferenceSets include an attribute to mark the type of topical context within which referential variation is observed. The topical context underlies the coherence relations within a discourse. Moreover, it explains the variation in making reference to the same entities and events either through shifting roles, phases, and aspects, or through framing by the author. The topical context allows us to abstract from references to individual entities and events, by generalising the observations to the surface forms and synsets. For example, the same person may be referenced during *school*, *family*, *leisure*, or at *work*. It makes little sense to combine all the references to the same person in a single ReferenceSet. Instead, we gather reference to individuals across all different incidents within the same type of topical context. This captures our general ways of framing persons and events within these topics and according to some topical schema. ReferenceSets thus will reflect which synonyms from which synsets are used how frequently to make reference within the same topical context.

Figure 1 shows two examples of a ReferenceSet for the two texts in the introduction that report on the same incident and thus the same topical context of *gun-violence*. We see separate ReferenceSets for the *shooter* and for the *shooting*. Each ReferenceSet consists of a list of *synset-ref* elements.² The *synset-ref* element has attributes for the CILI identifier *iid* (Bond et al., 2016; Vossen et al., 2016b), the language specific WordNet synset, and the *corefcount* attribute to express how often this entity was mentioned in the text. Each *synset-ref* contains a list of *surface-form* elements with the surface form and its observed token frequency of making reference.

We can see that the words span different synsets and also different parts-of-speech tags. The first ReferenceSet exhibits the perspective of the *shooter* and the *suspect* before the trial. We abstracted from the actual names of the people through a separate element and counter *proper-name*. The second ReferenceSet shows different granularities of the event: the overall *incident*, the *shooting*, *hitting* and the *outcome*, and it shows the legal judgments: *murder*, *assault*. This illustrates that the reference relations are often orthogonal to hypernym relations.

ReferenceSets as in Figure 1 can be derived from collections of texts in which coreference relations are resolved across documents making reference to the same incident, involving the same entities and events. ReferenceSet can then be formed by aggregation across incidents of the same topic type, based on sufficient overlap between surface forms and synsets of incidents. We discuss methodologies for building a ReferenceNet in detail in the next section.

²At the end of each *synset-ref element* we list the corresponding WordNet synonyms as a comment.

4 Methodologies for building a ReferenceNet

Semantic parsing aims at generating a representation of entities and events from their mentions throughout this text. It relies on a broad range of NLP techniques such as tokenization, parsing, named-entity recognition and linking, and semantic role labeling. Coreference modules often operate on top of the output of such modules. Words and phrases that make reference to the same individual or event are coreferential. If different documents report on the same entities, these would ideally result in cross-document coreference. Applying coreference modules to large collections of texts potentially gives us the different ways in which people make reference to the same entities and events in the world. If for all these referential expressions, we would also know the WordNet synsets, we can abstract from coreferential mentions to their synsets and derive ReferenceSets for the semantic *types* of referents. This requires Word Sense Disambiguation (WSD) to run in addition to establishing coreference relations.

The feasibility of this approach depends on the quality of all the underlying modules (among which WSD) as well as the quality of the coreference modules. A distinction can be made between nominal/entity and event coreference, as they are defined and approached differently by different research groups. As we are primarily interested in the topical coherence underlying texts in this paper, we focus in this paper on event coreference and leave nominal or entity coreference for future work. We discuss two methods for obtaining event ReferenceNet data from text collections using semantic parsing: 1) text-to-data and 2) data-to-text.

Text-to-data involves semantic text parsing without knowing the referents a priori and without knowing which texts report on the same incident. It therefore relies on high-quality cross-document event coreference resolution and it is computationally very expensive as it requires comparing all event mentions with each other (within and across documents). Automatic event coreference is a difficult task (Hovy et al., 2013) and made little progress over the years. To compare different approaches on the ECB+ dataset (Cybulska and Vossen, 2014), Yang et al. (2015) reimplemented state-of-the-art algorithms proposed by Bejan and Harabagiu (2010) and Chen and Ji (2009), as well as their own approach. They report 58.7 CoNLL-F1 (Luo et al., 2014) on ECB+ for their own approach, compared to 53.6 CoNLL-F1 for (Bejan and Harabagiu, 2010) and 55.2 CoNLL-F1 for (Chen and Ji, 2009). They obtained their results however only after boosting event detection from an original 65F to 95F by training a separate event detection system on part of the ECB+ data. Without such nearly perfect event detection, their results are much lower. All three approaches are clustering approaches over the dataset using event mentions as input. Likewise, they can only recover coreference relations between mentions that match local structural

```

1 <ReferenceSet topic="gun-violence">
2   <synset-ref corefcount="2" wid="pwn30:eng-10287213-n" iid="i90357"> <!-- gunman, gun -->
3     <surface-form "tokencount="2">gunman</surface-form>
4   </synset-ref>
5   <synset-ref corefcount="2" wid="pwn30:eng-10152083-n" iid="i91182"> <!-- man, adult.male -->
6     <surface-form "tokencount="2">man</surface-form>
7   </synset-ref>
8   <synset-ref corefcount="1" wid="pwn30:eng-10681383-n" iid="i93471"> <!-- suspect -->
9     <surface-form "tokencount="1">suspect</surface-form>
10  </synset-ref>
11  <synset-ref corefcount="3" wid="pwn30:eng-00007846-n" iid="i35562"> <!-- person, individual, someone, somebody, mortal, soul -->
12    <surface-form "tokencount="1">person</surface-form>
13    <proper-name "tokencount="2"/>
14  </synset-ref>
15 </ReferenceSet>
16
17 <ReferenceSet topic="gun-violence">
18   <synset-ref corefcount="1" wid="pwn30:eng-00355365-v" iid="i23513"> <!-- kill -->
19     <surface-form "tokencount="1">kill</surface-form>
20   </synset-ref>
21   <synset-ref corefcount="2" wid="pwn30:eng-00260470-v" iid="i23019"> <!-- hurt, injure -->
22     <surface-form "tokencount="2">injure</surface-form>
23   </synset-ref>
24   <synset-ref corefcount="2" wid="pwn30:eng-00225150-n" iid="i36591"> <!-- shooting -->
25     <surface-form "tokencount="2">shooting</surface-form>
26   </synset-ref>
27   <synset-ref corefcount="1" wid="pwn30:eng-00095280-a" iid="i500"> <!-- dead -->
28     <surface-form "tokencount="1">dead</surface-form>
29   </synset-ref>
30   <synset-ref corefcount="1" wid="pwn30:eng-00045888-s" iid="i233"> <!-- deadly -->
31     <surface-form "tokencount="1">deadly</surface-form>
32   </synset-ref>
33   <synset-ref corefcount="1" wid="pwn30:eng-00123783-n" iid="i36562"> <!-- gunfire, gunshot -->
34     <surface-form "tokencount="1">gunfire</surface-form>
35   </synset-ref>
36   <synset-ref corefcount="1" wid="pwn30:eng-00220522-n" iid="i36562"> <!-- murder, slaying, execution -->
37     <surface-form "tokencount="1">murder</surface-form>
38   </synset-ref>
39   <synset-ref corefcount="1" wid="pwn30:eng-00767826-n" iid="i39445"> <!-- assault -->
40     <surface-form "tokencount="1">assault</surface-form>
41   </synset-ref>
42 </ReferenceSet>

```

Figure 1: ReferenceSets for the text fragments referencing the shooter and the event

features, hence exhibit limited variation. Another approach implemented by Vossen and Cybulska (2016), logically matches semantic representations of the *action* mentions, the participants, the time, and the place. Assuming again near-perfect event detection, this approach results in a CoNLL-F1 score of 67.13. For comparison, a baseline system that applies a one-lemma-one-referent heuristics already scores 53.4 CoNLL-F1. As argued in (Cybulska and Vossen, 2014), the ECB+ dataset is very limited in terms of referential variation and within each topic there are only two potential referents to choose between. Concluding, we observe that event coreference systems perform poorly, especially with respect to recall. Applying these corpora to large collections of texts is not likely to give us reliable referential data to derive a ReferenceNet and will not capture sufficient variation. However, the advantage of this approach is that it can be applied to any collection of text.

The **data-to-text** method starts from structured data in which the referents are predefined and searches for texts that make reference to this data, so-called reference texts. Structured event data paired with reference texts appear to exist and are publicly available: GunViolenceArchive (GVA),³ FireIncidentRe-

ports (FR),⁴ Railwaysarchive (RA),⁵ Gun Violence Database (GVDB),⁶ ASN incident database,⁷ ASN Wikibase.⁸ These resources register event incidents with rich properties such as participants, location, and incident time, and often even provide pointers to one or more reference texts. The number of events and documents is usually high, i.e. there are ~ 9 k incidents in RA, and ~ 30 k incidents in GVA.

In the *data-to-text* approach, we convert the structured data from such archives to what we call a *microworld*. A microworld is an RDF⁹ representation of the referents related to a specific event (e.g. human calamities or economic events) but no more than that. *Reference texts* are then news, blogs, and Wikipedia pages that report on this data. Given the a-priori pairing of microworlds with reference text, we can apply the simple *one-mention-one-referent* principle to obtain reference relations for event mentions for free with a relatively high confidence. By increasingly mixing microworlds and reference texts, we approximate the complexity of reference relations in reality across large

⁴<https://www.firerescue1.com/incident-reports/>

⁵<http://www.railwaysarchive.co.uk/eventlisting.php>

⁶<http://gun-violence.org/>

⁷<https://aviation-safety.net/database/>

⁸<https://aviation-safety.net/wikibase/>

⁹We use the Simple Event Model (SEM-RDF) to represent events (Van Hage et al., 2011)

³<http://gunviolencearchive.org/reports/>

volumes of text. By collecting news from different sources on the same or similar events, we approximate true variation in making reference from different perspectives. For example, we can not only take news from different sources with different stances but also vary the time between the event date and the publication date to get articles with different historical perspective. Furthermore, the fact that the data on events from which we start has been created from the perspective of general human interest (e.g. gun violence incident reports) avoids discussion on what establishes an event in text, as we consider only those mentions that directly refer to the reported incident or salient subevents of these incidents.

Although this method may result in more precise reference relations as there is little ambiguity for paired microworlds and reference texts, its downside is the dependency on the availability of the structured data coupled with such reference texts. While for certain types of events such as calamities, sports, and business there may be sufficient data, for others people are less inclined to register events for longer periods. Alternatively, structured event data can also be obtained from DBpedia (Knuth et al., 2015; Elbassuoni et al., 2010), Wikidata (Vrandečić and Krötzsch, 2014), and YAGO2 (Hoffart et al., 2013). As these databases are often linked to Wikipedia articles, references in these articles can be used to find reference texts. Note that we only need the structured data to reconstruct a minimal representation of the referents and we do not need the full representation of the event. Another downside of this approach is that the granularity of the incident is more coarse than the granularity at which the events are reported in the associated news articles. To illustrate this, the GVA collection provides a summary on the incident outcome, whereas the corresponding news documents report on the process that led to this outcome: *firing, hitting, killing, getting injured, dying, etc.*

5 The NewsReader event coreference system

We used the NewsReader system (Vossen et al., 2016a) to simulate both a text-to-data and data-to-text process. In both cases, we apply generic semantic parsing to articles, obtaining representations of entities, events, and roles. The output is represented in the Natural Language Annotation Format (NAF) (Fokkens et al., 2014). Coreference for events within a single NAF file is based on a number of steps described in (Vossen and Cybulska, 2016):¹⁰

1. all predicates from the semantic role layer in NAF are considered as event mentions;
2. we collect all mentions with the same lemma of an SRL predicate throughout the text and consider them to be coreferential;

¹⁰Speech acts and so-called grammatical verbs (aspect, auxiliaries) are excluded from this process.

3. we take the output of WSD for each mention to obtain the best scoring synsets above a threshold. From these synsets, we obtain the highest scoring synsets across all mentions as the most *dominant synsets* for the lemma in the document;
4. we create a coreference set from all the lemma mentions with their dominant senses;
5. all lemma-based coreference sets are compared with each other (cross-lemma) by applying WordNet similarity to the dominant senses across lemma sets
 - (a) if their similarity exceeds a preset threshold, we merge the coreference sets across the lemmas aggregating the dominant synsets. In addition, we include the lowest-common-subsumer synset that was responsible for the similarity match.
 - (b) if below the threshold, we keep the sets distinct
6. we iterate over the reference sets until there are no changes

For WSD, NewsReader uses the UKB system (Agirre and Soroa, 2009), as well as the supervised It-Makes-Sense system of Zhong and Ng (2010). The output of both systems is used to vote for the most dominant synsets associated with a mention of a predicate. For WordNet similarity, NewsReader uses the WordNet distance measure proposed by Leacock and Chodorow (1998).¹¹ To be able to capture similarity across nouns and verbs, we extended the WordNet hyponym relations with morphological relations of the type *event* across noun and verb synsets, obtained from the Princeton WordNet website.¹² Below we show two examples of event coreference sets in NAF obtained from two text fragments on the same incident, where the similarity threshold was set to 1.0 and the dominant sense threshold was set to the 80% best-scoring synsets in WSD.

Curry Bryson , the father of the 11-year - old who police say shot and killed a 3-year - old , appeared in court today for a hearing Barney says it is not the charges against him that have torn his client apart . It is the fact Bryson 's 11-year - old son is accused of shooting and killing 3-year - old Elijah Walker .

```

1 <coref id="coevent13" type="event">
2   <span> <target id="t4"/> </span> <!--shooting-->
3   <span> <target id="t35"/> </span><!--shot-->
4   <span> <target id="t104"/> </span><!--torn-->

```

¹¹We used the implementation in <https://github.com/cltl/WordnetTools> which allows us to include cross-part-of-speech relations

¹²<http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls>

```

5 <exReferences>
6 <exRef conf="1.38" ref="eng-30-02055267-v" source="lcs"/>
7 <exRef conf="0.85" ref="eng-30-01134781-v" source="dom"/>
8 <exRef conf="0.70" ref="eng-30-01597286-v" source="dom"/>
9 <exRef conf="0.74" ref="eng-30-01002740-v" source="dom"/>
10 <exRef conf="0.75" ref="eng-30-02061495-v" source="dom"/>
11 <exRef conf="1.0" ref="eng-30-02484570-v" source="dom"/>
12 <exRef conf="0.72" ref="eng-30-01003249-v" source="dom"/>
13 <exRef conf="0.70" ref="eng-30-02055267-v" source="dom"/>
14 <exRef conf="0.90" ref="eng-30-01137138-v" source="dom"/>
15 </exReferences>
16 </coref>

```

An 11-year - old Detroit boy has been charged with manslaughter in the fatal shooting of 3-year - old Elijah Walker

```

1 <coref id="coevent28" type="event">
2 <span><target id="t148"/></span><!--shooting-->
3 <exReferences>
4 <exRef conf="0.83" ref="eng-30-00225150-n" source="dom"/>
5 <exRef conf="1.0" ref="eng-30-00122661-n" source="dom"/>
6 </exReferences>
7 </coref>

```

In the first fragment, the software lumped together verbal mentions of *shooting*, *shot*, and *torn*. The first two share the same lemma, while they were matched with *torn* through the lowest-common-subsumer (source="lcs") synset `eng3002055267v:buck;charge;shoot_down;shoot;tear`. The similarity score was 1.38. Setting the similarity threshold to 1.5 would prevent merging these mentions. In the second fragment, there is only one mention of the noun *shooting*. We can see that across the documents the verbal and nominal senses will not match on the basis of just the synset identifiers. However, they may still be merged through the form *shooting* or using cross-part-of-speech similarity. From all the mentions, we obtain the most dominant synsets (source="dom") associated by the WSD system according to the threshold setting. The lowest-common-subsumer and the dominant synsets form the basis to compare event coreference sets across documents.

In order to match reference sets across documents, NewsReader first converts NAF representations to SEM-RDF, in which each coreference set represents a unique instance of an event (represented by a unique URI). Each event instance is described with the semantic information associated from all mentions throughout the document. However for the cross-document comparison reported here, we have chosen to match coreference sets only in terms of the overlap of WordNet synsets and surface forms, thus ignoring participants, roles, and temporal relations. The proportion of overlap across instances of events can be set through a parameter. In our experiment, 5% of the synsets or surface forms (in case a lemma has no synsets) need to match for merging instances across different documents.

To simulate the text-to-data approach, *all* the RDF representations of events are compared across *all* the documents. In order to simulate a data-to-text approach, we applied the above cross-document strategy in such a way that events are only compared when the reference texts report on the same incident according to

the structured data. This means that *shootings* in documents reporting on different incidents are never compared and cannot constitute coreference relations.

6 Evaluation data

To evaluate both these methodologies, we manually annotated 38 news articles associated with 20 incidents from the GVA data set. The articles were grouped by the incident on which they report together with the structured data on the incidents, e.g. which people got injured or died. We used an annotation schema that differentiates events at different levels of granularity and with respect to the most salient implication derived from the event mention:

incident The incident as a whole is referred to, corresponding to an entry in the structured database.

firing a gun The event of operating a gun without implying somebody got hit.

hit Somebody got hit as a result of shooting without implying death or injury.

miss A gun was used but the bullet missed a person.

injure Somebody got injured as a result of being hit.

die Somebody died as a result of being hit.

For each mention of these events, the annotator creates a unique instance identifier based on the incident, the assigned event type, and the affected victims. When annotating events in documents reporting on the same incident, identity results from same type and victims assigned to mentions whereas non-identity results from a difference in type and/or victim. Documents that report on different incidents never result in identity regardless of the type or victims annotated. Shooting the same person in different incidents is not the same, as well as shooting a different person in the same incident.

The annotation resulted in 138 event instances and 874 mentions in 38 documents. In total, 77 different lemmas were used to make reference to these events. Given these annotations, we can abstract from the instances and group lemmas that make reference to the same *type* of event. Table 1 shows the ReferenceSets derived from the manual annotation for the types of event. Note that the total number of mentions and lemmas is higher because the same word, e.g. *shooting* may occur in multiple reference sets.

Table 1 reveals the large variation based on just 38 documents. We also observe that the event implications follow from very different expressions. For example, *death* can be concluded forward from *fatal shot* or backward from *autopsy*. Especially words making reference to the complete incident show a lot of variation, reflecting different judgments and appraisals.

Table 1: ReferenceSets at the event type level, derived from manual annotation for 20 incidents on gun-violence

Event type	Nr. Variants	Nr. Mentions	ReferenceSets
incident	27	229	accident:39, incident:34, it:34, this:17, murder:15, hunting:14, reckless:14, tragedy:9, happen:8, felony:7, manslaughter:5, what:5, homicide:4, shooting:4, assault:3, case:2, endanger:2, endangerment:2, that:2, violence:2, 's:1, crime:1, event:1, go:1, mistake:1, on:1, situation:1
fire	21	148	shooting:48, fire:25, discharge:16, go:12, shot:9, pull:7, gunman:6, gun:5, gunshot:4, firing:3, shoot:2, turn:2, accidental:1, act:1, action:1, at:1, handle:1, it:1, return:1, shootout:1, shotgun:1,
hit	11	196	shot:131, discharge:17, shooting:17, strike:16, hit:4, blast:3, victim:3, striking:2, gunshot:1, into:1, turn:1
injure	16	73	wound:36, surgery:13, treat:5, injure:3, stable:3, injurious:2, send:2, bodily:1, critical:1, hit:1, hospitalize:1, hurt:1, injury:1, put:1, stabilize:1, unresponsive:1
die	16	246	death:60, die:52, dead:45, kill:34, fatal:13, lose:9, fatally:7, loss:7, autopsy:6, body:4, take:3, homicide:2, claim:1, deadly:1, life:1, murder:1
Total	114	1043	

7 Evaluation results

We automatically generated ReferenceSets from the 38 annotated documents using the NewsReader pipeline. We used standard settings for dominant-senses (80% top-scoring senses) and similarity (similarity of 2 or higher). Following the methodologies described in section 4, we processed the data twice:

1. **without-i:** comparing all events across all 38 documents, without considering the document-to-incident links from the structured data. This corresponds to the traditional cross-document text-to-data approach.
2. **with-i:** comparing only events across documents if these documents report on the same incident. This method is enriched with data-to-text knowledge.

. In both settings, we only compare event mentions detected by the system and we exclude knowledge about participants, location, and time expressions. We expect *without-i* to lead to more drift in the coreference sets as it will match mentions of events across all documents without the microworld and reference text association. In table 2, we show the coverage results for both, where we make a distinction between the proportion of gold mentions detected and the proportion of gold lemmas. Lemma recall (r) and precision (p) is calculated by comparing the set of lemmas detected by the system to the set of lemmas in the gold annotation. For the mentions evaluation, we compared the frequencies of the lemmas in the texts.

Table 2: Mention and lemma coverage evaluation (r=recall, p=precision, f=harmonic mean) of the NewsReader system output with (with-i) and without (without-i) incident association

	Mentions (874 gold)		Lemmas (77 gold)	
	with-i	without-i	with-i	without-i
r	20.25%	18.19%	49.35%	49.35%
p	59.80%	35.57%	62.30%	46.34%
f	30.26%	24.07%	55.07%	47.80%

We see that *with-i* (incident pairing) performs better in terms of mention recall (+2), precision (+14) and f-score (+6) than *without-i*. For lemma coverage, the recall is the same, but the incident-aware version *with-i* has much higher precision (+16). Overall recall is significantly lower than precision for both methods.

The precision of the data-to-text approach with incident pairing is reasonable (around 60%), though not very high. This can be improved by using better WSD and/or by making the cross-document matching more strict. In the current setting only 5% of the synsets or phrases need to match across documents.

In table 3, we show per event type the ReferenceSets generated by the systems that matched at least one lemma from the gold annotation (the matching lemmas are in bold). We can make a number of observations from these data. First of all, we see that automatic ReferenceSets are more fine-grained than gold sets. This is mainly due to the fact that we use WSD and WordNet similarity to group event mentions in coreference sets. The WordNet synsets and hypernyms do not cover the diverse relations that we annotated for the incidents. Having more relations would merge together reference sets. Furthermore, we see that *with-i* obtains more ReferenceSets, but also more precise ReferenceSets, in comparison to *without-i*.¹³ This is to be expected because *with-i* is not allowed to create ReferenceSets across incidents. Finally, we see that multiwords are not considered by NewsReader, which leads to semantic drift for words such as *pull* (the trigger), *take* (a life).

The recall for both methods is really low: around 20% for mentions and 50% for lemmas. Error analysis on the missed recall shows that most of these are not detected as predicates by the semantic role labeler: pronouns (*it, this, what*), adjectives (*fatal, fatally, reckless, injurious, accidental, deadly*), and nouns (*dead, incident, surgery, felony, autopsy*). Predicate detection is based on the Mate tool (Johansson and Nugues, 2008),

¹³The only exception are the ReferenceSets that include *take*, where the incident pairing generated 4 ReferenceSets and included more wrong mentions than without incident pairing.

Table 3: ReferenceSets at the event type level, derived from automatic annotation for 20 incidents on gun-violence

Type	Reference set with-i	Reference set without-i
textbfaccident:3 incident	accident:3 act:1 action:1 case:3 crime:1 happen:14 fact:1 fact:1 happen:1 hunting:1 manslaughter:1 murder:1 shootout:1 tragedy:1 victim:3	call:9 make:4 name:2 act:1 action:1 holler:1 case:3 crime:1 happen:14 occur:2 fact:1 hunt:2 hunter:1 hunting:1 manslaughter:1 murder:1 shootout:1 tragedy:1 victim:3
fire gun	discharge:3 fire:5 gun:1 gunman:1 address:1 deal:1 handle:1 speak:1 pull:4 return:3 turn:3 use:2 mother:1	fire:5 discharge:3 release:3 complete:2 gun:1 gunman:1 pull:4 force:1 return:3 turn:3 grow:2 raise:2 mother:6 use:2 bill:1
hit/fire gun	shoot:5 shooting:4 shot:2 shoot:26 shot:7 shooting:4 hit:3 charge:1 shoot:2 charge:1	shoot:23 shot:5 charge:3 hit:3 shooting:2
hit	hit:3 shoot:3 strike:2	strike:2
injury	send:7 post:4 message:1 message:1 send:1 treat:2 wound:3 hurt:3	send:6 carry:5 post:5 letter:1 message:1 transport:1 handling:3 treat:2 deal:1 handle:1 manage:1 wound:3 hurt:3 back:2 suffer:2 support:2
die	death:7 die:4 die:9 death:7 run:1 kill:12 house:2 live:2 life:1 life:8 live:5 house:2 lose:4 loss:1 put:4 place:1 place:1 put:1 say:52 take:21 involve:10 need:9 come:8 get:8 tell:3 ask:2 bring:2 carry:2 want:2 conduct:1 involve:10 come:8 get:8 take:4 need:3 bring:2 want:2 carry:2 take:2 ask:2 take:2 need:1	die:9 death:5 run:5 kill:12 family:13 life:7 home:5 live:4 house:1 lose:4 loss:1 put:4 place:2 set:2 lay:1 say:146 tell:17 take:14 involve:7 need:6 ask:5 conduct:3 state- ment:2 want:2 bring:1

which is trained on PropBank (Kingsbury and Palmer, 2002), and NomBank (Meyers et al., 2004). Improving the recall for lexical coverage therefore primarily requires improving the coverage of these resources.

8 Conclusions and future work

In this paper, we present ReferenceNet: a network of referential relations between synsets that is complementary to WordNet and word embeddings. ReferenceNet consists of ReferenceSets that group synsets and words that make reference to similar entities and events within similar topical contexts. Typically, ReferenceSets reflect different local contexts and perspectives within a shared topical context as opposed to synsets and word embeddings that capture similar local contexts. We described two methods to derive ReferenceSets from textual data. We evaluated the approaches against a manually annotated data set. We concluded that precision is reasonable, whereas recall is low, mainly due to poor recall of predicates. We also observed that coreference relations are missed because WordNet does not sufficiently capture coherence and

perspective relations, resulting in smaller ReferenceSets. The evaluation further showed that ReferenceSets created with a data-to-text approach have higher recall and precision. In future work, we want to capture more referential variation. Event coreference can be improved using other coherence measures, especially when comparing coreference sets across documents. The fact that WSD already restricts the association of concepts by part-of-speech limits the matching in the current system. We will also extend to other types of events and contexts. Finally, entity coreference can be included by exploiting semantic matches of noun phrases and semantic roles.

Acknowledgements

The work presented in this paper was funded by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines". We also thank the reviewers for their constructive comments.

References

- Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.
- C. F. Baker, C. J. Fillmore, and B. Cronin. 2003. The structure of the framenet database.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.
- Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422. Association for Computational Linguistics.
- Reinhard Blutner. 1998. Lexical pragmatics. *Journal of semantics*, 15(2):115–162.
- Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*.
- Z. Chen and H. Ji. 2009. Graph-based event coreference resolution. pages 54–57.
- Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552.
- Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, and Gerhard Weikum. 2010. Searching rdf graphs with sparql and keywords. *IEEE Data Eng. Bull.*, 33(1):16–24.
- Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–16.
- G. Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophischen Kritik*, 100:25–50.
- H. P. Grice, P. Cole, and J. Morgan. 1975. Logic and conversation. pages 41–58.
- Nicola Guarino. 1999. The role of identity conditions in ontology design.
- J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28??61.
- E. Hovy, T. Mitamura, F. Verdejo, J. Araki, and A. Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity.
- Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187. Association for Computational Linguistics.
- P. Kingsbury and M. Palmer. 2002. From treebank to propbank. pages 1989–1993.
- Magnus Knuth, Jens Lehmann, Dimitris Kontokostas, Thomas Steiner, and Harald Sack. 2015. The dbpedia events dataset. In *International Semantic Web Conference (Posters & Demos)*.
- Saul A Kripke. 1972. Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.
- Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.
- S. Levinson. 1983. *Pragmatics*. Cambridge University Press.
- A. Xiaoliang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June.
- Yo Matsumoto. 1995. The conversational condition on horn scales. *Linguistics and philosophy*, 18(1):21–60.
- Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Hilary Putnam. 1973. Meaning and reference. *The journal of philosophy*, 70(19):699–711.
- W Quine and O Van. 1960. Word and object: An inquiry into the linguistic mechanisms of objective reference.
- Erich H Rast et al. 2007. *Reference and indexicality*. Logos-Verlag.

- Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.
- Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136.
- Piek Vossen and Agata Cybulska. 2016. Identity and granularity of events in text.
- Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016a. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.
- Piek Vossen, Francis Bond, and J McCrae. 2016b. Toward a truly multilingual global wordnet grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Edda Weigand. 1998. *Contrastive lexical semantics*, volume 171. John Benjamins Publishing.
- Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.
- Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *arXiv preprint arXiv:1504.05929*.
- Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.

Wordnet-based Evaluation of Large Distributional Models for Polish

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, Paweł Kędzia

G4.19 Research Group, Computational Intelligence Department

Wrocław University of Science and Technology, Wrocław, Poland

{maciej.piasecki,arkadiusz.janz}@pwr.edu.pl

Abstract

The paper presents construction of large scale test datasets for word embeddings on the basis of a very large wordnet. They were next applied for evaluation of word embedding models and used to assess and compare the usefulness of different word embeddings extracted from a very large corpus of Polish. We analysed also and compared several publicly available models described in literature. In addition, several large word embeddings models built on the basis of a very large Polish corpus are presented.

1 Introduction

Distributional Semantics (DS) is focused on describing semantic associations between words on the basis of their distributional patterns in texts by applying statistical methods. DS methods are used to extract different kinds of the *Measures of Semantic Relatedness* (MSR) from corpora. An MSR can cover the whole range of semantic relations from topic or domain based till lexico-semantic relations. For many applications it is desirable to obtain an MSR which is close to a *Measure of Semantic Similarity* (MSS), i.e. a measure which assigns the highest values to words associated by linguistic lexico-semantic relations. Recently, word embeddings have become one of the best tools of DS. However, word embeddings, e.g. (Mikolov et al., 2013), are based on predicting a word occurrence in a context (mostly a sequence) of other words. This aspect of co-occurrence prediction in a local context can influence an MSR built on the basis of word embeddings. An MSS can be an important source of knowledge supporting wordnet development, e.g. (Piasecki et al., 2009). However, the question is how to evaluate to which extent the given MSR resembles an proper MSS? Experiments with the participation of humans are laborious, costly and the datasets created as a result are of limited size. It is hard to construct an evaluation by application in a way revealing the properties of a potential MSS.

A large wordnet is built on knowledge originating from humans. It includes directly the knowledge about lexico-semantic relations and offers an opportunity to build large scale, realistic tests. Our goal is to construct large scale test datasets for word embeddings on the basis of a large wordnet, apply them for evaluation of word embedding models and next to analyse and com-

pare the usefulness of different word embeddings extracted from a very large corpus of Polish. Finally, we want to publishing word embedding models of known properties built on the basis of a very large corpus of Polish.

2 Related Works

MSR evaluation methods can be roughly divided into intrinsic and extrinsic. The former are based on the direct evaluation of the MSR properties, e.g. by assessment by humans or comparison with a gold standard. The latter is based on applying an MSR as knowledge source in some NLP application.

Typical datasets used in the intrinsic evaluation are small, e.g. (Rubenstein and Goodenough, 1965), WS-353 (Finkelstein et al., 2002) and most of the all 10 data sets discussed in (Faruqui and Dyer, 2014), where only two of them include ≈ 2000 and ≈ 3000 word pairs. They were used in many tests, in fact overused. Small sizes of these datasets make performing proper evaluation more difficult, e.g. because of the lack of the common partitioning into training, tuning and testing parts.

Datasets for MSR evaluation are often collected during experiments based on testing human judgement in reaction to some prompting signal, which is close to reaction to a stimuli, e.g. (Auguste et al., 2017) measured the correlation between the reaction times in the context of priming with ranking based on word embeddings. However, this is slightly different situation than analysis of lexical meanings during language utterance interpretation, especially a textual utterance. MSR is extracted from a text corpus, and it is more natural to evaluate it against language resources. Moreover, (Faruqui et al., 2016) noticed that the distinction between similarity and relatedness is not well defined and consistently expressed in most popular test datasets.

(Schnabel et al., 2015) evaluated systematically different DS models, but finally all tests were based on data collected during crowdsourcing experiments using Amazon Turk. (Jastrzebski et al., 2017) performed “evaluation focused on data efficiency” with respect to 4 categories, namely: “Similarity, Analogy, Sentence and Single word”. In the case of similarity, which is most interesting for us, they used only well known data sets for English. For each type of dataset different combinations of preprocessing and classification algorithms were applied.

It is worth to notice, that the cost of preparing larger datasets for another language than English is quite sub-

stantial. This is one of the reasons that it is hard to find such approaches for other languages, with notable exceptions e.g. (Hartmann et al., 2017) for Portuguese. In our case we want to explore the possibility of constructing of large test datasets on the basis of an already existing wordnet. As the primary application we focused on is support for wordnet development, so comparison with data collected in experiments with humans is not necessarily the best solution for us.

3 Wordnet-based Evaluation

In many approaches a wordnet was used to generate a wordnet-based measure of semantic similarity that was next used to assess the correlation between it and an MSR, e.g. (Lin, 1998). It was assumed that similarity rankings generated by the two measures should be similar. However, there are many wordnet-based similarity measures of different properties and some of them depend on additional knowledge like information about the frequency of word senses. Thus, the result of the comparison can be different depending on the wordnet-based similarity measure applied and in all cases is not straightforward in interpretation. We want to follow a different approach and to explore two methods that are free of these problems.

3.1 Synonymy tests

(Freitag et al., 2005) proposed a wordnet-based synonymy test (WBST) in which for a *question word* x an n -tuple is automatically generated:

$\mathbf{D} = \langle d_1, \dots, d_n \rangle$, such that one the elements: d_i is the correct *answer*, i.e. it is synonymous with x and belongs to the same synset as x , and all other $d_j \neq d_i$ are *detractors*, i.e. false answers that are not synonymous with x . Elements of \mathbf{D} and the position of the correct answer are randomly selected. MSR is tested by using its values in selecting a possible answer for the problem word x .

In the case of some wordnets, including pWordNet, many synsets are singletons and include only one word. Thus they would be excluded from the test, and this could bias the evaluation result.

To prevent this, in *Hypernymy-expanded WBST* (HWBST) (Piasecki et al., 2009) answers for singleton synsets are selected from their hypernym synset, and in the same time these hypernyms are excluded from possible detractors. For a large wordnet, WBST and HWBST can include many thousands of (question – answer) pairs enabling very intensive testing of an MSR and partitioning the set in many different ways, e.g. test vs train, frequent vs infrequent or according to the domains of words.

Because detractors in WBST and HWBST are selected completely randomly, the majority of them come from those parts of the wordnet that are very remote in relation to the question word. Thus these types of tests are relatively easy to be solved on the basis of an MSR. In order to make the test harder we need

to select detractors in such a way that words from synsets semantically similar to the question words have a higher probability of being selected than words from the synsets of small similarity. This version of the test is called *Extended WBST* (EWBST) (Piasecki et al., 2009). EWBST consists of pairs $\langle x_l, \mathbf{D}_l \rangle$, where x_l is a question word and $\mathbf{D}_l = \langle d_1, \dots, d_n \rangle$ is a sequence of possible answers such that d_i is the correct answer, i.e. a synonym or hypernym of x_l , as in HWBST, while the rest of $d_j \in \mathbf{D}_l \wedge d_j \neq d_i$ are selected randomly from the whole wordnet but with the probability correlated to the *wordnet-based similarity measure* (WSM) between d_j and x_l . Any WSM can be used to generate EWBST, but in the experiments presented in this work, we use a simple measure (1) proposed in (Agirre and Edmonds, 2006) based on the normalised length of a shortest path in the wordnet graph. It can be computed without knowing the frequency of senses:

$$WSM(w_1, w_2) = -\log \frac{\text{path}(w_1, w_2)}{2D_m} \quad (1)$$

In (1), w_1 and w_2 are words, $\text{path}(w_1, w_2)$ is the shortest path in the extended hypernymy graph between two synsets including, respectively: w_1 and w_2 , and D_m is the mean depth of the extended hypernymy graph. While (Agirre and Edmonds, 2006) used normalized path distance, in the recent version of pWordNet many synsets are far away from the root. This effectively flattens the probability distribution to the point where it is no different than uniform random sampling as per HWBST. Using average depth D_a instead reflects better relations contained in pWordNet and promotes synsets closer to the question word. However, this modification results also in negative values of WSM, so they had to be capped off at 0:

$$WSM_a(w_1, w_2) = \max\left(-\log \frac{\text{path}(w_1, w_2)}{2D_a}, 0\right) \quad (2)$$

This reduces probability of choosing a detractor with distance $2D_a$ or greater to 0, so the tests become more difficult due to the elimination of trivial detractors unrelated to the question. The idea of EWBST is to make detractors more similar to the correct answer and more difficult to be properly distinguished from the correct answer on the basis of MSR values.

The graph was built from hypernymy relations and type/instance relations. In addition, as pWordNet hypernymy is not a single-rooted structure, we added to the graph several SUMO concepts (Pease, 2011) as top level nodes on the basis of the mapping of pWordNet hypernymy root synsets onto SUMO concepts.

3.2 Cut-off rendering tests

WBST-family tests illustrate the ability of an MSR to distinguish between words whose senses are located in different parts of the wordnet graph, while EWBST gives also insights into the sensitivity to small local differences. However, WBST-family tests concentrate on

synonymy and hypernymy, as these two relations are mostly used in selection of the correct answers and detractors. Nevertheless, from a good MSS we can also expect an ability to express other types of lexico-semantic relations. This can be measured with the help of a simple *Wordnet-based Cut-off Rendering* test (WBCR). In WBCR for each question word x a bag-of-words of words is generated in which they come from:

- the synset S_x of x
- and synsets S_i connected directly and also indirectly to S_x by selected wordnet relations.

S_x and S_i are indirectly connected, if there is a path in the graph of wordnet relations such that it consists of a proper sequence of wordnet relations. Depending on the type of relations allowed for direct and indirect connections, as well as the assumed patterns for the paths and their maximal length, we can define different types of bags-of-words. Next, the evaluated MSR is used to reconstruct the extracted bag-of-words:

1. for the problem word x a ranking list of the words most related to x on the basis of the MSR values is generated; such a list will be called the *k-nearest neighbours* list (henceforth *k-NNL*) of x .
2. for the assumed k , the top k words from the list are collected as a reconstructed bag-of-words,
3. the reconstructed bag-of-words for x is compared with the wordnet-based bag-of-words, and precision, recall and F-measure are calculated.

This simple test is meaningful only for large, comprehensive wordnets or wordnets describing well some selected domains. However, WBCR has very simple interpretation and can be easily tuned to different subsets or domains of words and senses.

4 Experiments

During experiments, we built several word embeddings models from the largest corpus of Polish available. Next we evaluated them in several tests based on plWordNet 3.1 (i.e. the most contemporary version) and compared with other word embedding models for Polish extracted from smaller corpora and published in the web.

4.1 Corpora and preprocessing

As a basis for the experiments we selected plWordNet 3.1 – a very large wordnet of Polish including $\approx 190,500$ different words, described by $\approx 282,500$ senses, more than 217,000 synsets and more than 750,000 relation links. plWordNet has been built by corpus-based wordnet developed method (Maziarz et al., 2013) and expresses very good coverage of words in large corpora (Maziarz et al., 2016).

We calculated our word embeddings model on the basis of plWordNet Corpus 10.0 (plWNC) of Polish,

which includes more than 4 billion words¹. It is also probably the largest corpus of Polish built in a controlled way and was used during the plWordNet development.

plWNC was used in the experiments in two versions of preprocessing:

plWNC-lem the corpus was first morphosyntactically tagged and lemmatised with the help of WCRFT2 tagger (Radziszewski, 2013; Radziszewski and Warzocha, 2014); strings: “lemma:grammatical class” were in the input to *word2vec* (Mikolov et al., 2013).

plWNC-multi in the morpho-syntactically tagged plWNC Proper Names and multiword expressions described in plWordNet 3.1 were merged to single tokens.

plWNC-multi was prepared with the help of *Liner2* tool (Marcinićzuk et al., 2013) for recognition and classification of PNs. plWordNet 3.1 includes almost 60,000 Polish MWEs represented as lexical units and described by lexicalised morpho-syntactic constraints that allow for their efficient and accurate recognition in tagged texts (Kurc et al., 2012). We represent Proper Names (one and multiword, including many common words) and multiword expressions as single tokens in *plWNC-multi* in order to block the interpretation of their components as individual words. Components of PNs and MWEs can have very specific meanings (e.g. in non-compositional MWEs) that can influence the resulting word embeddings.

Corpora created from the Polish Wikipedia data alone (of $\approx 600M$ words) were used in two experiments reported in the literature. We evaluated these published word embedding models against our tests, too, see Sec. 5

4.2 Word embedding models tested

For the generation of word2vec models *Gensim* library was used (Řehůřek and Sojka, 2010). On the basis of the set of 6 parameters, we selected during pre-experiments 9 different types of models to be evaluated experimentally, i.e. the following combinations:

1. vector size: 100, 300 and 1000,
2. algorithm type: *Skip-gram*, *CBOW ns* (with negative subsampling) and *CBOW hs* (with hierarchical softmax).

¹It consists of IPI PAN Corpus (Przepiórkowski, 2004), the first annotated corpus of Polish, National Corpus of Polish (Przepiórkowski et al., 2012), Polish Wikipedia (from 2016), *Rzeczpospolita* Corpus (Weiss, 2008) – corpus of electronic editions of a Polish newspaper from the years 1993-2003, supplemented with text acquired from the Web – only text with small percentage of words unknown to a very comprehensive morphological analyser Morfeusz 2.0 (Woliński, 2014) were included; duplicates were automatically eliminated from the merged corpus.

Vector size	Min freq.	Model	WBST	HWBST	EWBST
1000	1000	w2w- <i>plWNC-multi-skipg-ns</i>	92.43	89.00	63.97
		w2w- <i>plWNC-multi-cbow-hs</i>	91.54	89.34	63.21
		w2w- <i>plWNC-multi-cbow-ns</i>	91.68	89.31	62.99
	200	w2w- <i>plWNC-multi-skipg-ns</i>	92.52	89.80	62.51
		w2w- <i>plWNC-multi-cbow-hs</i>	92.71	90.11	60.94
		w2w- <i>plWNC-multi-cbow-ns</i>	92.58	90.11	60.97
	30	w2w- <i>plWNC-multi-skipg-ns</i>	90.43	88.84	58.92
		w2w- <i>plWNC-multi-cbow-hs</i>	92.56	90.05	57.35
		w2w- <i>plWNC-multi-cbow-ns</i>	92.51	90.07	57.30
300	1000	w2w- <i>plWNC-multi-skipg-ns</i>	90.81	88.24	62.50
		w2w- <i>plWNC-multi-cbow-hs</i>	90.32	88.12	61.00
		w2w- <i>plWNC-multi-cbow-ns</i>	90.70	88.49	62.13
	200	w2w- <i>plWNC-multi-skipg-ns</i>	91.81	89.36	61.24
		w2w- <i>plWNC-multi-cbow-hs</i>	91.46	89.29	59.45
		w2w- <i>plWNC-multi-cbow-ns</i>	91.11	89.50	60.76
	30	w2w- <i>plWNC-multi-skipg-ns</i>	90.99	89.43	58.25
		w2w- <i>plWNC-multi-cbow-hs</i>	91.36	89.41	55.97
		w2w- <i>plWNC-multi-cbow-ns</i>	91.35	89.79	57.50
100	1000	w2w- <i>plWNC-multi-skipg-ns</i>	88.84	86.01	59.42
		w2w- <i>plWNC-multi-cbow-hs</i>	87.71	86.14	58.26
		w2w- <i>plWNC-multi-cbow-ns</i>	88.14	86.71	59.34
	200	w2w- <i>plWNC-multi-skipg-ns</i>	89.78	87.53	58.52
		w2w- <i>plWNC-multi-cbow-hs</i>	88.97	87.33	56.75
		w2w- <i>plWNC-multi-cbow-ns</i>	89.05	87.57	58.12
	30	w2w- <i>plWNC-multi-skipg-ns</i>	89.79	88.21	55.99
		w2w- <i>plWNC-multi-cbow-hs</i>	89.44	87.62	53.52
		w2w- <i>plWNC-multi-cbow-ns</i>	89.63	88.13	55.27
	1000	pl-embeddings-cbow	71.63	69.36	43.71
		pl-embeddings-skip	76.30	74.54	47.16
		fastText.wiki.pl	80.01	78.17	52.42
	200	pl-embeddings-cbow	71.79	69.46	42.31
		pl-embeddings-skip	76.89	74.65	45.53
		fastText.wiki.pl	80.11	79.16	51.40
	30	pl-embeddings-cbow	71.49	70.35	41.85
		pl-embeddings-skip	77.41	75.69	45.28
		fastText.wiki.pl	81.44	80.27	51.39

Table 1: WBST-like tests generated from noun in plWordNet 3.1 and applied to word embedding models extracted from *plWNC-multi*.

Thus, we tested: Skip-gram 100, Skip-gram 300, Skip-gram 1000, CBOW ns 100, CBOW ns 300, CBOW ns 1000, CBOW hs 100, CBOW hs 300 and CBOW hs 1000. In all models the minimal frequency of tokens (i.e. tagged lemmas and/or PN and MWE tokens) was set to ≥ 8 (min_count=8). Pre-trained models are readily available²

(Rogalski and Szczepaniak, 2016) first preprocessed a text corpus based on the Polish Wikipedia³ by changing the text to lower case, numbers were divided into separate digits, and some non-text elements were deleted. Next two word embedding models were constructed: CBOW and Skip-gram models with negative sampling and the vector size: 300. The extracted models are publicly available in the internet⁴ and following the original names they will be called in the experiments, respectively: *pl-embeddings-cbow* and *pl-embeddings-skip*.

²<https://clarin-pl.eu/dspace/handle/11321/442>

³<https://pl.wikipedia.org>

⁴http://publications.it.p.lodz.pl/2016/word_embeddings/

(Bojanowski et al., 2016) built Skip-gram models⁵ using *fastText* technique with the vector size 300 for many languages on the basis of Wikipedia data. For the extraction of the models a novel method in which “each word is represented as a bag of character n-grams”, cf (Bojanowski et al., 2016), was applied. It was designed for languages with richer inflection and was meant to better deal with a large number of word forms in such languages. Their model will be simply called *fastText.wiki.pl* in the experiments.

The Polish language has a very rich morphology, which is why we also decided to examine *fastText* models, but *plWNC* 10.0 corpus was used for training. All of our *fastText* models were trained with the Skip-gram architecture and the vector of size 300. We tested the Skip-gram 300 model with minimal word frequencies of 5, 20 and 50. These models will be named according to given schema *fastText.plWNC* in our experiments.

Another set of models was introduced in (Mykowiecka et al., 2017). For our experiments

⁵<https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>

Model	<i>VS</i>	Score	Model	<i>VS</i>	Score
w2w- <i>plWNC-lem-cbow-hs</i>	100	39.29	w2w- <i>plWNC-lem-cbow-ns</i>	300	55.61
w2w- <i>plWNC-multi-cbow-hs</i>	100	40.82	w2w- <i>plWNC-multi-cbow-ns</i>	300	57.14
w2w- <i>plWNC-lem-cbow-hs</i>	300	48.47	w2w- <i>plWNC-lem-skipg</i>	100	45.92
w2w- <i>plWNC-multi-cbow-hs</i>	300	48.98	w2w- <i>plWNC-multi-skipg</i>	100	48.98
w2w- <i>plWNC-lem-cbow-ns</i>	100	47.96	w2w- <i>plWNC-lem-skipg</i>	300	60.20
w2w- <i>plWNC-multi-cbow-ns</i>	100	47.96	w2w- <i>plWNC-multi-skipg</i>	300	59.18
ft- <i>plWNC-multi-skipg-mC5</i>	300	50.75	ft- <i>plWNC-lem-skipg-mC5</i>	300	50.75
ft- <i>plWNC-multi-skipg-mC20</i>	300	53.30	ft- <i>plWNC-lem-skipg-mC20</i>	300	54.23
ft- <i>plWNC-multi-skipg-mC50</i>	300	50.75	ft- <i>plWNC-lem-skipg-mC50</i>	300	59.28
<i>nep-lemmas-all-100-cbow-hs</i>	100	48.72	<i>nep-forms-all-100-cbow-hs</i>	100	28.18
<i>nep-lemmas-all-100-cbow-ns</i>	100	46.67	<i>nep-forms-all-100-cbow-ns</i>	100	35.00
<i>nep-lemmas-all-100-skipg-hs</i>	100	44.10	<i>nep-forms-all-100-skipg-hs</i>	100	34.55
<i>nep-lemmas-all-100-skipg-ns</i>	100	44.10	<i>nep-forms-all-100-skipg-ns</i>	100	39.55
<i>nep-lemmas-all-300-cbow-hs</i>	300	55.38	<i>nep-forms-all-300-cbow-hs</i>	300	35.91
<i>nep-lemmas-all-300-cbow-ns</i>	300	57.95	<i>nep-forms-all-300-cbow-ns</i>	300	43.18
<i>nep-lemmas-all-300-skipg-hs</i>	300	56.92	<i>nep-forms-all-300-skipg-hs</i>	300	43.64
<i>nep-lemmas-all-300-skipg-ns</i>	300	54.36	<i>nep-forms-all-300-skipg-ns</i>	300	46.82
<i>nep-lemmas-restricted-100-cbow-hs</i>	100	49.74	<i>nep-forms-restricted-100-cbow-hs</i>	100	32.27
<i>nep-lemmas-restricted-100-cbow-ns</i>	100	47.69	<i>nep-forms-restricted-100-cbow-ns</i>	100	39.55
<i>nep-lemmas-restricted-100-skipg-hs</i>	100	43.59	<i>nep-forms-restricted-100-skipg-hs</i>	100	36.82
<i>nep-lemmas-restricted-100-skipg-ns</i>	100	45.13	<i>nep-forms-restricted-100-skipg-ns</i>	100	40.00
<i>nep-lemmas-restricted-300-cbow-hs</i>	300	52.82	<i>nep-forms-restricted-300-cbow-hs</i>	300	40.00
<i>nep-lemmas-restricted-300-cbow-ns</i>	300	59.49	<i>nep-forms-restricted-300-cbow-ns</i>	300	43.64
<i>nep-lemmas-restricted-300-skipg-hs</i>	300	54.87	<i>nep-forms-restricted-300-skipg-hs</i>	300	42.73
<i>nep-lemmas-restricted-300-skipg-ns</i>	300	54.87	<i>nep-forms-restricted-300-skipg-ns</i>	300	47.27

Table 2: Analogy tests from (Mykowiecka et al., 2017) applied to the different word embeddings models, where k is 10, all results in (%).

we selected the models trained with Skip-gram and CBOW architectures and the vector size of 100 and 300. These pre-trained models were generated on National Corpus of Polish. Due to the anticipated problems with the morpho-syntactic tagging, (Mykowiecka et al., 2017) utilised two versions of the corpus: full, further called ‘*nep-lemmas*’ or ‘*nep-forms*’ and “restricted data sets [...] which only included tokens classified as nouns, adjectives, adverbs, verb forms, and abbreviations, which constitute 19 parts of speech (POS) out of the 34 foreseen in” NCP.

4.3 Tests

4.3.1 Wordnet-based Synonymy Tests

All three types of tests, namely: WBST, HWBST and EWBST were generated on the basis of the noun part of plWordNet 3.1 in three versions corresponding to the minimal frequency of words in plWNC 10.0: 30, 200 and 1000, i.e. in a given test all question, answer and detractor words had to express the predefined minimal frequency in the corpus. However, still the generated tests are very large e.g. EWBST(min. 1000) includes 19,996 question – answers pairs, HWBST (min. 30) includes 48,263 pairs, WSBT, and WBST(min. 1000) includes 9,100 pairs – the smallest set because singleton synsets are omitted. All tests are open and accessible⁶.

4.3.2 Wordnet-based Cut-off Rendering tests

As in the case of the WBST-like tests, the cut-off tests were generated on the basis of nouns in plWordNet 3.1

⁶<https://clarin-pl.eu/dspace/handle/11321/446>

and in three main versions with respect to the minimal frequency of nouns in plWNC 10.0: 30, 200 and 1000 (the numbers of bag of words are smaller than the number of pairs in WBST-like tests but similarly large).

The wordnet context of a problem word x , which was represented as a bag of words was defined in three different ways:

Cnt – all words linked to x by direct relation links, i.e. from synsets linked directly to the synset of x and also by direct lexical relations to one of the x senses; it also includes synonyms of x .

CntH – **Cnt** expanded with all indirect hyponyms and hypernyms of x up to the hypernymy and hyponymy paths of the maximal length 3.

CntHC – **CntH** expanded with all $k = m + n$ cousins of x with $k = 3$, i.e. words from synsets accessible from the synsets of x by hyper/hyponymy paths of up to m hypernymy and n hyponymy links.

Thus, **Cnt** measures the ability of an MSR to find words in very close relations (e.g. as a potential tool supporting description of x senses), **CntH** illustrates the use of the MSR as a tool supporting construction of hyper/hyponymy structures, and **CntHC** characterises, e.g., a possibility of using the given MSR for identifying small wordnet subgraphs which lemma senses belong to. All cut-off tests were applied to the k -best neighbours lists with $k \in \{10, 100\}$ generated for nouns from plWordNet.

Cut-off Precision								
Model	k NN	Min. f.	10			100		
			Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w- <i>plWNC-multi-cbow-hs</i>	1000	1000	13.42	15.12	35.67	3.31	4.29	17.04
w2w- <i>plWNC-multi-cbow-ns</i>	1000	1000	13.62	15.16	34.25	3.30	4.22	15.96
w2w- <i>plWNC-multi-skipg</i>	1000	1000	12.35	13.47	28.07	2.66	3.18	10.12
ft- <i>plWNC-multi-skipg</i>	1000	1000	8.74	9.24	15.72	2.59	3.00	8.14
w2w- <i>plWNC-lem-cbow-hs</i>	1000	1000	12.86	14.26	33.38	3.11	3.93	15.75
w2w- <i>plWNC-lem-cbow-ns</i>	1000	1000	9.65	10.58	25.40	2.17	2.60	9.71
w2w- <i>plWNC-lem-skipg</i>	1000	1000	11.61	12.61	27.15	2.47	2.92	9.82
ft- <i>plWNC-lem-skipg</i>	1000	1000	7.39	7.72	13.31	2.25	2.54	7.25
w2w- <i>plWNC-multi-cbow-hs</i>	200	200	11.54	12.94	32.91	2.70	3.47	15.48
w2w- <i>plWNC-multi-cbow-ns</i>	200	200	11.17	12.34	30.92	2.61	3.29	14.41
w2w- <i>plWNC-multi-skipg</i>	200	200	10.37	11.23	25.06	2.15	2.55	9.20
ft- <i>plWNC-multi-skipg</i>	200	200	8.42	8.84	16.09	2.21	2.54	7.95
w2w- <i>plWNC-lem-cbow-hs</i>	200	200	10.50	11.57	30.01	2.46	3.07	14.21
w2w- <i>plWNC-lem-cbow-ns</i>	200	200	8.20	8.94	23.26	1.79	2.12	9.08
w2w- <i>plWNC-lem-skipg</i>	200	200	9.64	10.42	24.03	1.99	2.33	8.84
ft- <i>plWNC-lem-skipg</i>	200	200	7.05	7.32	12.98	1.93	2.16	6.87
Cut-off Recall								
Model	k NN	Min. f.	10			100		
			Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w- <i>plWNC-multi-cbow-hs</i>	1000	1000	10.33	7.10	3.42	20.83	15.69	8.61
w2w- <i>plWNC-multi-cbow-ns</i>	1000	1000	10.09	6.84	3.24	20.27	14.84	8.16
w2w- <i>plWNC-multi-skipg</i>	1000	1000	9.24	6.26	2.91	17.22	12.20	6.26
ft- <i>plWNC-multi-skipg</i>	1000	1000	7.33	4.87	2.18	17.54	12.22	5.80
w2w- <i>plWNC-lem-cbow-hs</i>	1000	1000	8.74	6.05	2.85	17.67	13.03	7.03
w2w- <i>plWNC-lem-cbow-ns</i>	1000	1000	6.71	4.61	2.18	13.20	9.46	4.99
w2w- <i>plWNC-lem-skipg</i>	1000	1000	8.19	5.64	2.60	15.12	10.82	5.41
ft- <i>plWNC-lem-skipg</i>	1000	1000	5.92	4.04	1.82	14.88	10.40	4.85
w2w- <i>plWNC-multi-cbow-hs</i>	200	200	10.76	7.40	3.89	20.90	15.75	9.42
w2w- <i>plWNC-multi-cbow-ns</i>	200	200	9.82	6.64	3.54	19.53	14.24	8.76
w2w- <i>plWNC-multi-skipg</i>	200	200	9.18	6.22	3.19	16.65	11.76	6.71
ft- <i>plWNC-multi-skipg</i>	200	200	8.56	5.70	2.84	18.40	12.81	6.92
w2w- <i>plWNC-lem-cbow-hs</i>	200	200	8.45	5.91	3.19	16.89	12.48	7.65
w2w- <i>plWNC-lem-cbow-ns</i>	200	200	6.86	4.78	2.60	13.20	9.52	5.69
w2w- <i>plWNC-lem-skipg</i>	200	200	8.04	5.59	2.91	14.53	10.44	5.93
ft- <i>plWNC-lem-skipg</i>	200	200	6.98	4.83	2.49	15.63	11.04	5.90
F measure								
Model	k NN	Min. f.	10			100		
			Cnt	CntH	CntHC	Cnt	CntH	CntHC
w2w- <i>plWNC-multi-cbow-hs</i>	1000	1000	11.67	9.66	6.23	5.72	6.74	11.44
w2w- <i>plWNC-multi-cbow-ns</i>	1000	1000	11.59	9.42	5.92	5.68	6.57	10.80
w2w- <i>plWNC-multi-skipg</i>	1000	1000	10.57	8.55	5.27	4.61	5.05	7.73
ft- <i>plWNC-multi-skipg</i>	1000	1000	7.97	6.38	3.83	4.51	4.82	6.77
w2w- <i>plWNC-lem-cbow-hs</i>	1000	1000	10.41	8.49	5.25	5.29	6.04	9.72
w2w- <i>plWNC-lem-cbow-ns</i>	1000	1000	7.91	6.42	4.02	3.73	4.08	6.59
w2w- <i>plWNC-lem-skipg</i>	1000	1000	9.60	7.80	4.75	4.24	4.60	6.98
ft- <i>plWNC-lem-skipg</i>	1000	1000	6.57	5.30	3.20	3.90	4.09	5.81
w2w- <i>plWNC-multi-cbow-hs</i>	200	200	11.13	9.42	6.96	4.78	5.68	11.71
w2w- <i>plWNC-multi-cbow-ns</i>	200	200	10.45	8.63	6.35	4.60	5.35	10.89
w2w- <i>plWNC-multi-skipg</i>	200	200	9.74	8.01	5.66	3.81	4.19	7.76
ft- <i>plWNC-multi-skipg</i>	200	200	8.49	6.93	4.83	3.94	4.23	7.40
w2w- <i>plWNC-lem-cbow-hs</i>	200	200	9.37	7.82	5.77	4.30	4.93	9.95
w2w- <i>plWNC-lem-cbow-ns</i>	200	200	7.47	6.23	4.68	3.15	3.47	7.00
w2w- <i>plWNC-lem-skipg</i>	200	200	8.76	7.28	5.20	3.49	3.81	7.10
ft- <i>plWNC-lem-skipg</i>	200	200	7.02	5.82	4.17	3.43	3.62	6.35

Table 3: Wordnet-based Cut-off Rendering tests for nouns in p1WordNet 3.0 applied to word embedding models extracted from *plWNC-multi* (vec. size=300), where kNN is the length of the k -NN lists, all results in (%).

Cut-off Precision							
k NN		10			100		
Model	Min. freq.	Cnt	CntH	CntHC	Cnt	CntH	CntHC
pl-embeddings-cbow	1000	4.97	6.10	20.98	1.37	1.94	10.51
pl-embeddings-skipg	1000	4.19	4.91	15.32	1.27	1.66	7.58
fastText.wiki.pl	1000	4.03	4.24	7.24	1.52	1.78	6.04
pl-embeddings-cbow	200	3.92	4.81	17.77	1.08	1.52	8.78
pl-embeddings-skipg	200	3.42	4.05	13.58	1.02	1.34	6.65
fastText.wiki.pl	200	3.90	4.07	7.33	1.31	1.51	5.76
pl-embeddings-cbow	30	3.28	4.03	15.56	0.90	1.27	7.67
pl-embeddings-skipg	30	2.99	3.55	12.56	0.88	1.16	6.14
fastText.wiki.pl	30	3.72	3.87	7.41	1.15	1.33	5.49
Cut-off Recall							
k NN		10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
pl-embeddings-cbow	1000	3.07	2.27	1.13	7.44	6.02	3.35
pl-embeddings-skip	1000	2.79	1.98	0.96	7.24	5.57	2.91
fastText.wiki.pl	1000	3.13	2.12	0.96	9.52	6.62	3.12
pl-embeddings-cbow	200	2.79	2.08	1.12	6.81	5.55	3.29
pl-embeddings-skipg	200	2.68	1.95	1.02	6.90	5.43	3.04
fastText.wiki.pl	200	3.74	2.55	1.26	9.86	6.89	3.62
pl-embeddings-cbow	30	2.55	1.90	1.05	6.29	5.10	3.13
pl-embeddings-skipg	30	2.64	1.93	1.04	6.75	5.34	3.09
fastText.wiki.pl	30	4.07	2.78	1.44	9.79	6.87	3.85
F measure							
k NN		10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
pl-embeddings-cbow	1000	3.79	3.31	2.15	2.32	2.94	5.08
pl-embeddings-skipg	1000	3.35	2.82	1.80	2.15	2.56	4.20
fastText.wiki.pl	1000	3.52	2.83	1.70	2.63	2.81	4.12
pl-embeddings-cbow	200	3.26	2.90	2.11	1.86	2.38	4.79
pl-embeddings-skipg	200	3.01	2.63	1.89	1.77	2.15	4.17
fastText.wiki.pl	200	3.82	3.14	2.16	2.31	2.48	4.45
pl-embeddings-cbow	30	2.87	2.58	1.97	1.58	2.03	4.44
pl-embeddings-skipg	30	2.80	2.50	1.93	1.55	1.90	4.11
fastText.wiki.pl	30	3.89	3.24	2.41	2.06	2.22	4.53

Table 4: Wordnet-based Cut-off Rendering tests for nouns in plWordNet 3.0 applied for word embedding models extracted from the Polish Wikipedia, where kNN is the length of the k -NN lists, all results in (%).

4.3.3 Analogy tests

One of the most popular techniques for word embeddings is to test their ability of reflecting word analogies, e.g. applied also in (Mykowiecka et al., 2017) for testing word embeddings for Polish. Analogy consists of 2 pairs of words, the relation between the first pair being similar to second pair. For example, we can say that the relation between *winter* and *snow* is analogous to *autumn* and *rain*, the relation being the typical weather in given season. Another common example is often used to showcase analogy is *man-woman:king-queen*, with the relation of male-female counterparts.

For word embeddings, the relation between words is simply the difference between their vectors and therefore the analogy can be written as $\vec{a} - \vec{b} = \vec{c} - \vec{d}$, where \vec{a} , \vec{b} , \vec{c} , \vec{d} are embedding vectors for words.

For the purpose of testing, the above equation is transformed into $(\vec{b} + \vec{c}) - \vec{a} = \vec{d}$. The left hand side is evaluated by the means of vector arithmetic and k vectors most similar to the result are found. If one of the vectors is \vec{d} , then the model is said to pass the analogy test. If one of the words in the analogy is not present in the model then the single analogy is omitted

as it cannot be evaluated. We used analogy tests from (Mykowiecka et al., 2017) with a kind permission and help of the authors. Only the semantic part of 196 test items was applied.

5 Results

The results of the tests in Tab. 1 illustrate well the differences in the difficulty of the tests: WBST is the simplest one, EWBST the hardest. The difference between EWBST and the two other tests is striking in all experiments. The difficulty of EWBST can be tuned by changing the wordnet-base similarity measure it is based on and the dependency between the similarity measure and the probability distribution of the selection of detractor words.

Skip-gram model is better than CBOW according to WBST and EWBST in most of the cases while in the other cases the differences are small. Only in HWBST CBOW-ns achieved higher result that can be attributed to a kind of generalisation introduced by the inclusion of hypernyms into correct answers. Also among the models from the literature, models based on Skip-gram

Cut-off Precision							
k NN		10			100		
Model	Min. freq.	Cnt	CntH	CntHC	Cnt	CntH	CntHC
<i>nep</i> -lemmas-all-cbow-hs	1000	11.71	13.27	32.67	2.88	3.73	15.51
<i>nep</i> -lemmas-all-cbow-ns	1000	12.15	13.51	31.57	2.95	3.74	14.61
<i>nep</i> -lemmas-all-skipg-hs	1000	10.40	11.52	27.19	2.51	3.11	12.00
<i>nep</i> -lemmas-all-skipg-ns	1000	10.00	10.92	23.15	2.17	2.57	8.42
<i>nep</i> -lemmas-all-cbow-hs	200	9.56	10.79	28.55	2.28	2.93	13.42
<i>nep</i> -lemmas-all-cbow-ns	200	9.70	10.72	27.38	2.30	2.88	12.72
<i>nep</i> -lemmas-all-skipg-hs	200	8.67	9.60	24.67	2.03	2.51	10.85
<i>nep</i> -lemmas-all-skipg-ns	200	7.98	8.69	19.66	1.70	1.99	7.25
Cut-off Recall							
k NN		10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
<i>nep</i> -lemmas-all-cbow-hs	1000	8.16	5.81	2.75	16.73	12.88	6.81
<i>nep</i> -lemmas-all-cbow-ns	1000	8.10	5.59	2.60	16.66	12.36	6.54
<i>nep</i> -lemmas-all-skipg-hs	1000	7.39	5.11	2.37	15.16	11.23	5.82
<i>nep</i> -lemmas-all-skipg-ns	1000	7.26	5.02	2.30	13.79	9.88	4.90
<i>nep</i> -lemmas-all-cbow-hs	200	8.01	5.77	3.07	16.16	12.52	7.28
<i>nep</i> -lemmas-all-cbow-ns	200	7.64	5.34	2.83	15.62	11.62	6.94
<i>nep</i> -lemmas-all-skipg-hs	200	7.45	5.25	2.76	15.04	11.30	6.50
<i>nep</i> -lemmas-all-skipg-ns	200	6.73	4.71	2.41	12.71	9.17	5.11
F measure							
k NN		10			100		
		Cnt	CntH	CntHC	Cnt	CntH	CntHC
<i>nep</i> -lemmas-all-cbow-hs	1000	9.62	8.08	5.07	4.91	5.78	9.46
<i>nep</i> -lemmas-all-cbow-ns	1000	9.72	7.91	4.80	5.01	5.74	9.04
<i>nep</i> -lemmas-all-skipg-hs	1000	8.64	7.08	4.36	4.30	4.88	7.84
<i>nep</i> -lemmas-all-skipg-ns	1000	8.42	6.88	4.18	3.75	4.07	6.20
<i>nep</i> -lemmas-all-cbow-hs	200	8.71	7.52	5.54	4.00	4.75	9.44
<i>nep</i> -lemmas-all-cbow-ns	200	8.55	7.13	5.12	4.01	4.61	8.98
<i>nep</i> -lemmas-all-skipg-hs	200	8.01	6.79	4.96	3.58	4.11	8.13
<i>nep</i> -lemmas-all-skipg-ns	200	7.30	6.11	4.29	3.00	3.28	6.00

Table 5: Wordnet-based Cut-off Rendering tests for nouns in plWordNet 3.0 and applied for word embedding models from (Mykowiecka et al., 2017), where kNN is the length of the k -NN lists, all results in (%).

scheme, including *fastText.wiki.pl* (which is a Skip-gram model too) express higher results. This is especially visible in the case of the more difficult EWBST test. The *wiki.pl* was superior among the models built only on the data from Wikipedia, i.e. several times smaller than plWNC 10.0. However, all models built on much smaller corpus produced much worse results. We tested also models based on *plWNC-lem* version of the large corpus and all models were slightly but significantly worse in the WBST-family of tests.

Contrary to the synonymy tests, in the case of WBCR evaluations of the models generated from *plWNC-multi* presented in Tab. 3, we can notice that CBOW models are superior in all cases in comparison to Skip-gram models. It means that Skip-gram models are better in describing differences between word meanings, while CBOW enable broader exploration of potential lexico-semantic relations. However, relatively good precision signals that instances of lexico-semantic relations receive higher values. Definitely the results of the test are negatively biased by lacking relation instances in plWordNet. This kind of tests and evaluations can be used also as a diagnostic tool to spot these subdomains in a wordnet that are potentially not well enough described by relation links. We can ob-

serve also that the application of hierarchical softmax consistently produces better results in all frequency ranges. However, hierarchical softmax should result in better estimation of the representation.

We evaluated also word embedding models extracted from *plWNC-lem*, i.e. a version without folding PNs and MWEs into single tokens. WBCR tests showed lower performance due to the lack of MWE description. We also plan to apply tests not including MWEs in the future in order to investigate the effect of folding more precisely. Thus, models based on *plWNC-multi* offer a unique opportunity of obtaining good distributional description of PNs and MWEs.

Quite surprisingly, we can observe in Tab. 4 that models built on a smaller corpus of Wikipedia behave in a slightly different way in WBCR tests for less frequent words than those constructed on a very large corpus. In Tab. 4 Skip-gram models express higher recall, *fastText* Skip-gram with sub-word representation have much higher recall for words with lower frequency. However, this can be an effect of different preprocessing and filtering of the data. Nevertheless, all results obtained on the Polish Wikipedia are worse than those in Tab. 3 generated from a very large corpus (including also the Wikipedia data). It means that for WBCR

tests, that cover a wider spectrum of relations, larger data used result in the improvement of the model.

Finally, in analogy tests, see Table 2, Skip-gram model built on the very large corpus *p/WNC* is still the best one, like in EWBST, but the difference to models constructed on much smaller NCP is minimal. However, the analogy tests of (Mykowiecka et al., 2017) include mostly general and frequent words. Moreover, the differences are small only for models based on the restricted version of NCP, i.e. focused on the words included in the tests. Potential influence of the corpus preprocessing and filtering on distinguishing relations and lexical meanings is worth further investigation.

6 Conclusions

We showed that a large comprehensive wordnet can be successfully used as a basis for two different types of MSR evaluation methods, namely the family of Wordnet-based Synonymy Tests and Wordnet-based Cut-off Rendering tests. In both types of tests very large datasets can be generated allowing for very intensive testing and high statistical significance of the test results. The datasets are enough large to conveniently partitioned according to the frequency criteria of semantic criteria. In fact the datasets and tests are based on human decisions expressed in the wordnet structure. Both tests describe the ability of an MSR to be used as a basis for developing a lexico-semantic language resource.

WBST-family tests focus on the ability of an MSR to distinguish between different lexical meanings, while WBCR is sensitive more to representation of different types of wordnet relations by an MSR. As a result both types of tests are quite complementary. Moreover, by changing the similarity and context definitions in EWBST we can obtain tests of different difficulty.

In the further work, we develop a wordnet-based test that has properties of contextual tests, e.g. which is similar to Stanford contextual word similarity dataset (SCWS) (Huang et al., 2012).

We will also expand the presented evaluation to the dataset covering all four PoS, namely nouns, adjectives, verbs and adverbs.

The constructed word embedding models and evaluation datasets have been published on open licences under the link: <https://clarin-pl.eu/dspace/handle/11321/446>.

Acknowledgments

Co-financed by the Polish Ministry of Education and Science, CLARIN-PL Project.

References

- Eneko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.
- Jeremy Auguste, Arnaud Rey, and Benoit Favre. 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector-Space Representations for NLP*, pages 21–26, Copenhagen, Denmark, September. ACL.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proc. of ACL: System Demo*.
- Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 30–35, Berlin, Germany, August. ACL.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1).
- Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Nathan Hartmann, Erick R. Fonseca, Christopher Shulby, Marcos Vinícius Treviso, Jessica Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR*, abs/1708.06025.
- Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*.
- Stanislaw Jastrzebski, Damian Lesniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *CoRR*, abs/1702.02170.
- Roman Kurc, Maciej Piasecki, and Bartosz Broda. 2012. Constraint based description of polish multiword expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2408–2413, Istanbul, Turkey, may. European Language Resources Association (ELRA).

- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *International Conference On Computational Linguistics (COLING'98). Proceedings of the 17th International Conference on Computational Linguistics*, volume 2, pages 768–774. ACL.
- Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2 – a customizable framework for proper names recognition for Polish. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, pages 231–253.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In N. Calzolari, Y. Matsumoto, and R. Prasad, editors, *Proc. of COLING 2016, 26th Inter. Conf. on Computational Linguistics*, pages 2259–2268. ACL.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2017. Testing word embeddings for Polish. *Cognitive Studies | Études cognitives*, 17:1–19.
- Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.
- Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences.
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN.
- Adam Radziszewski and Radosław Warzocha. 2014. WCRFT2. CLARIN-PL digital repository, <http://hdl.handle.net/11321/36>.
- Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.
- Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- Marek Rogalski and Piotr S. Szczepaniak. 2016. Word embeddings for the polish language. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *15th International Conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016, Proceedings*, volume 9692 of *LNAI, Artificial Intelligence and Soft Computing*, pages 126–135. Springer.
- Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of ACM*, 8(10).
- Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September. ACL.
- Dawid Weiss. 2008. Korpus Rzeczpospolitej [Corpus of text from the online edition of “Rzeczpospolita”]. <http://www.cs.put.poznan.pl/dweiss/rzeczpospolita>.
- Marcin Woliński. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavik, Iceland. ELRA.

Distant Supervision for Relation Extraction with Multi-sense Word Embedding

Sangha Nam, Kijong Han, Eun-kyung Kim and Key-Sun Choi

School of Computing, KAIST

Daejeon, Republic of Korea.

{nam.sangha, han0ah, kekeeo, kschoi}@kaist.ac.kr

Abstract

Distant supervision can automatically generate labeled data between a large-scale corpus and a knowledge base without utilizing human efforts. Therefore, many studies have used the distant supervision approach in relation extraction tasks. However, existing studies have a disadvantage in that they do not reflect the homograph in the word embedding used as an input of the relation extraction model. Thus, it can be seen that the relation extraction model learns without grasping the meaning of the word accurately. In this paper, we propose a relation extraction model with multi-sense word embedding. We learn multi-sense word embedding using a word sense disambiguation module. In addition, we use convolutional neural network and piecewise max pooling convolutional neural network relation extraction models that efficiently grasp key features in sentences. To evaluate the performance of the proposed model, two additional methods of word embedding were learned and compared. Accordingly, our method showed the highest performance among them.

1 Introduction

Relation extraction refers to the task of extracting the relation between two entities in a sentence. For example, a relation extraction system extracts ‘*Founder(Facebook, Mark Zuckerberg)*’ from the sentence “*Mark Zuckerberg is the founder of Facebook*”. In recent years, the importance of knowledge bases has emerged, and studies for constructing large-scale knowledge bases such as DBpedia, YAGO, and Wikidata are actively underway. Furthermore, the research on extracting knowledge from web-scale corpus is also underway. However, since many studies use machine learning to design a relation extraction system, there is a high-cost problem in generating a large amount of supervised training data. To solve this problem, the distant supervision assumption is introduced in this paper (Mintz *et al.*, 2009). The dis-

tant supervision assumption means, “*If two entities are linked with a certain relation in the knowledge base and there is a collected sentence that contains both entities from the corpus, then the collected sentences may describe the certain relation between the two entities.*” Figure 1 is an example of automatically collected labeled data using the distant supervision assumption.



Figure 1. Example of labeled data collection based on distant supervision

The distant supervision method is relatively efficient in that it automatically generates training/labeled data between a large corpus and a large knowledge base, but the veracity of the labeled data is sometimes ambiguous. As shown in Figure 1, among the collected sentences that contain both ‘*Facebook*’ and ‘*Mark Zuckerberg*’, the first sentence means that Mark Zuckerberg is a founder of Facebook, but the second sentence does not. Various studies (Riedel *et al.*, 2010; Hoffmann *et al.*, 2011; Surdeanu *et al.*, 2012) have been introduced to solve this problem. However, they use traditional natural language processing (NLP) features such as part of speech (POS) tagging and dependency tree, so the errors occurring in NLP tools propagate to the relation extraction system. Therefore, these papers (Kim, 2014; Zeng *et al.*, 2014) proposed a relation extraction system that used word embedding and deep neural network (DNN) approaches without the above NLP features, and showed improved performance than previous studies. Especially, the piecewise max pooling convolution neural network (PCNN) model introduced in (Zeng *et al.*, 2015) transforms the convolution neural network (CNN) model into a form more suitable for relation extraction task.

However, these studies have a disadvantage in not reflecting the sense of words in word embedding. For example, the word ‘bow’ could be divided into various meanings such as ‘baU – greeting’ and ‘boU – archer’s weapon’. Therefore, if a relation extraction model is learned with lexical ambiguity, it may result in not properly reflecting the characteristics of the homograph. Thus, it is necessary to apply multi-sense word embedding to the relation extraction model. However, to the best of our knowledge, there are no studies applying multi-sense word embedding to relation extraction models.

In this paper, we introduce a distant supervision relation extraction model with multi-sense word embedding. We use two relation extraction models, CNN proposed in (Kim, 2014) and PCNN proposed in (Zeng *et al.*, 2015). To learn the multi-sense word embedding, we use the results of the word sense disambiguation (WSD) module and Skip-gram algorithm. To demonstrate the superiority of our method, we compared the relation extraction performances of two other word embedding models. The first is the most common word-token-based word embedding, and the second is the morpheme-based word embedding. In chapter 4, we present the experimental results of learning and evaluation of these models based on Korean Wikipedia and K-Box, which extended knowledge base on Korean DBpedia.

2 Related Work

2.1 Skip-gram Model

Word embedding is a way of expressing words in real-valued vectors, and expresses the meaning of a word on the vector space. Thus, it is easy to grasp the semantic similarity between words by a simple vector operation, and therefore, it is widely used in various NLP fields. The skip-gram model, which is type of word embedding learning method, learns by predicting words that appear around the

target word. The skip-gram model proceeds to maximize the following objective function.

$$J(\theta) = \sum_{(w_t, c_t) \in D^+} \sum_{c \in c_t} \log P(D = 1 | v(w_t), v(c)) + \sum_{(w_t, c_t) \in D^-} \sum_{c' \in c_t'} \log P(D = 0 | v(w_t), v(c'))$$

w_t is a target word and c_t stands for the word actually appearing around w_t in the corpus, and c_t' are randomly selected words that do not appear around w_t . That is, the learning is performed in such a manner as to maximize the probability of predicting words actually appearing around a target word and the probability of not predicting words that did not actually appear.

2.2 PCNN Relation Extraction Model

CNN is a deep neural network that shows excellent performance in image classification and sentiment classification. One of the features and advantages of CNN is that it efficiently finds key features in input data. Accordingly, the authors in (Kim, 2014; Zeng *et al.*, 2014) proposed a relation extraction model using CNN. In (Zeng *et al.*, 2014), the authors suggest the position embedding concept and adding it to the input vector of their CNN relation extraction model, and then the performance is improved. Position embedding is the embedding of the relative distance between two entity and non-entity words in a sentence as an n-dimensional vector. For example, as shown in Figure 3, the word ‘co-founder’ is three words away from the ‘Mark Zuckerberg’ entity and two words away from the ‘Facebook’ entity. This relative distance is embedded into an n-dimensional vector to create the position embedding, and the value is used as part of the input vector of model learning.



Figure 3. Example of the relative distance of position embedding

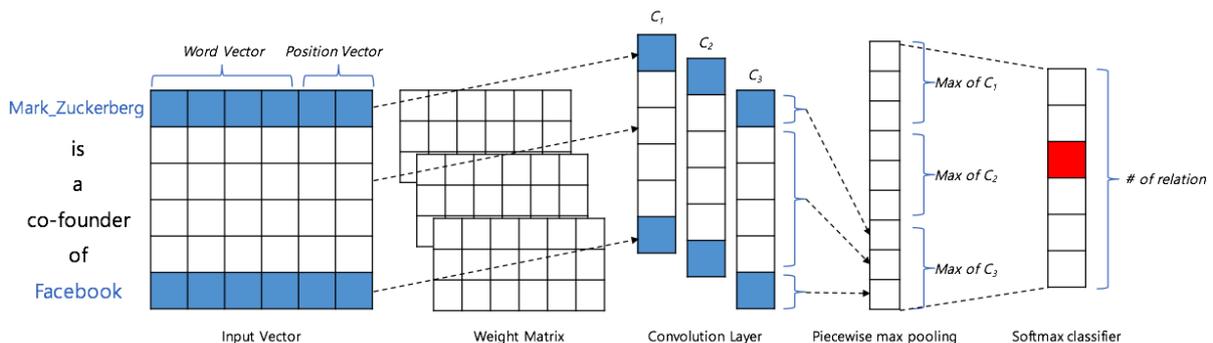


Figure 2. Architecture of PCNN

PCNN is an extended CNN model proposed in (Zeng *et al.*, 2015). The structure of PCNN is shown in Figure 2. The entire structure is made up of input vectors, three convolution layers, piecewise max-pooling layer, and softmax output layer. The input vector consists of a word vector and a position vector. The major difference is that extends the single max-pooling layer to the piecewise max-pooling layer. In CNN, max pooling is the method of extracting the largest value, i.e., the most important feature, in the output matrix of the convolution layer. However, it is difficult to grasp the key features required for relation extraction by selecting only one maximum value among the convolution layer result values in the single max-pooling layer. To solve these weaknesses, PCNN proposed a piecewise max-pooling layer by dividing the single max-pooling layer into three. Since the sentence used in relation extraction always contains two entities, it is possible to divide the sentence into three subunits based on two entities, and then the maximum value is extracted for each subunit in the piecewise max-pooling layer.

3 Methodology

In this paper, we propose a relation extraction model using multi-sense word embedding. We use CNN and PCNN for the relation extraction model, and generate multi-sense word embedding using the WSD module.

The structure of our relation extraction system is as shown in Figure 4, and it consists largely of the word embedding and distant supervision relation extraction model. First, we take the corpus as input and perform WSD module. Next, entity-padding tokenization is performed as described in Section 3.1. Next, the multi-sense word embedding is learned by the skip-gram algorithm, so that the tokens with the sense number have their own embedding vectors. In this way, the same form of lexical token has different embedding vectors based on the sense number.

Distant supervision is performed between Knowledge Base and Corpus, and the collected labeled data are word sense disambiguated and tokenized in the same manner. Then this data is divided into two groups—one for learning and the other for evaluation.

3.1 Multi-sense Word Embedding

In general, word embedding is a method of dividing the input corpus into word tokens and then mapping tokens with similar meaning onto similar vector spaces. In English, a token is usually generated in word units. However, Korean language is not as good as English when the word embedding is generated on a word token due to the plurality of elements constituting a word such as postposition, ending, and suffix. Therefore, when learning word embedding in Korean, a token is formed by a stem unit, and a POS tag is sometimes used as a constituent element of the token. The advantage of using a POS tag in learning of word embedding is that it can be divided into whether the same lexical word is used as a verb or as a noun. For example, in Korean, the word ‘*Ga-Ji*’ can be used as a noun to mean ‘*branch*’ or as a verb to mean ‘*get*’. Moreover, as an example in English, the word ‘*wind*’ can be used as ‘*movement of air*’ for nouns and ‘*twist*’ for verbs. Therefore, when learning word embedding, it is effective to use POS tags together to construct a token because some ambiguity could be resolved.

However, there is a problem in that common word embedding does not reflect the actual meaning of words, which is the same in Korean as well as English. For example, the word ‘*apple*’ is used both as a fruit and as a company. As mentioned earlier in Chapter 2, word embedding is based on what the surrounding words appear to be. The words around ‘*apple-fruit*’ and the words around ‘*apple-company*’ are definitely different, but all these words appear around the word ‘*apple*’, so the word ‘*apple*’ has only one n-dimensional real-valued vector that cannot distinguish between

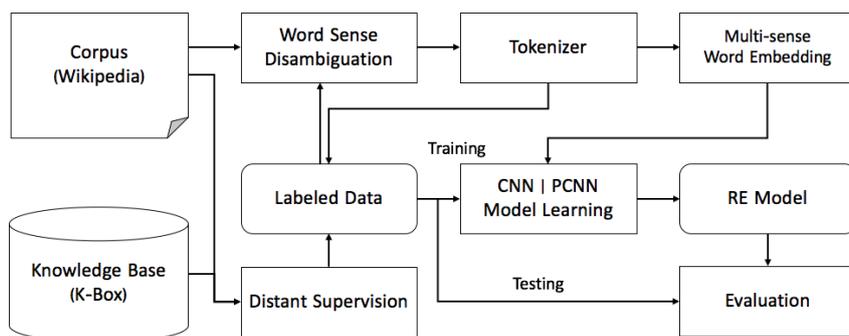


Figure 4. Architecture of relation extraction system with multi-sense word embedding

‘apple-fruit’ and ‘apple-company’. Thus, the triangle inequality problem (Neelakantan *et al.*, 2015) can occur.

$$\text{distance}(a, c) \leq \text{distance}(a, b) + \text{distance}(b, c)$$

For example, there is a problem that the distance between ‘(a) pollen – (c) refinery’ is smaller than the sum of the distances between ‘(a) pollen – (b) plant’ and ‘(c) refinery – (b) plant’. In other words, the similarity between the two words ‘pollen’ and ‘refinery’ is closer to the actual semantic distance centered on the homonym of ‘plant’. To solve this problem, several papers (Neelakantan *et al.*, 2015; Rothe and Schütze, 2015) have been published that learn word embedding by the actual meaning of words using a method is called multi-sense word embedding.

We learn multi-sense word embedding using a WSD module to distinguish the meaning of words in advance. Our WSD module is based on the unsupervised learning approach and uses the Markov Random Field (MRF) algorithm which resolves the ambiguity based on the semantic category of CoreNet (Choi *et al.*, 2004). In MRF, the node is composed of common noun, verb, and adjective, and the edge between the nodes is set as long as the distance is only one on the dependency path, in a similar way to this paper (Chaplot *et al.*, 2015).

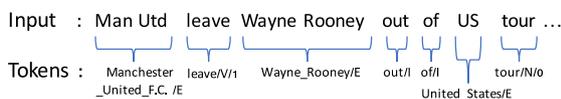


Figure 5. Example of Tokenization

The tokenization example for the input sentence is shown in Figure 5. The second word ‘leave’ is tokenized with a POS tag and a sense number. In addition, to make a word embedding suitable for relation extraction, the multiword entity was grouped into one token. As shown in Figure 5, ‘Man Utd’ and ‘Wayne Rooney’, a multiword entity, was bundled into a single token, and solved the entity disambiguation problem. Even if an entity consists of several words, learning to have a single word embedding value is proper for designing a word embedding and relation extraction model. We use personal entity tags in Wikipedia for entity linking as shown in Figure 6.

Facebook is an American [for-profit corporation](#) and an online [social media](#) and [social networking service](#) based in [Menlo Park, California](#). The Facebook website was launched on February 4, 2004, by [Mark Zuckerberg](#),

Figure 6. Example of Multiword Entity in Wikipedia

These blue entities, such as ‘for-profit corporation’, ‘social media’, and ‘social networking service’, are hand-tagged by Wikipedia content writers, so they are very accurate.

3.2 Relation Extraction Model

We use CNN and PCNN relation extraction models. The input representation consists of a 100-dimensional word vector and a 10-dimensional position vector. Three convolution layers were constructed and the weight matrix size was 3×110 , and the stride is one. CNN model is implemented as a single max-pooling layer and PCNN model is implemented as a piecewise max-pooling layer. The softmax layer is sized according to the relation number of the classification.

4 Evaluation

4.1 Data

For the experiment, we used 6,941,760 sentences of Korean Wikipedia (2017. 07. 01) and K-box. K-Box is a knowledge base that extends triple to Korean DBpedia, and the added triple is a conversion of Korean local property into ontological property. For example, the conversion of a Korean property such as ‘prop-ko:chul-saeng-ji’ into a ‘dbo:birthPlace’. The mapping table is created manually by three human experts. Through distant supervision, 358,464 labeled data were collected on 451 properties in all, but many properties were long tail problems with a small amount of collected data. In the multi-class classifier model, since learning does not proceed properly if there are few data per class, we used total 200,323 labeled data of 68 properties based on the number of collected data, which is 1000 or more per class.

4.2 Evaluation Results

To demonstrate the excellence of our proposed method, three types of word embedding have been learned. The first is learning by tokenization in word unit (Word), the second is tokenization by morpheme unit and POS tag (+POS), and the third is tokenization by morpheme unit, POS tag, and word sense (++WSD). All of these types of learning proceeded with the same parameters; 100-dimension, 5 window sizes, 1 minimum word count.

As given in Table 1, the result of multi-sense word embedding clusters the words in a sense-specific manner. In addition, since we apply multiword entity embedding, we can see that the multiword entity is learned by one embedding vector, and the similar words are also meaningful.

Token	Word	Similar Words
+POS	Si-Jang	invest, distribution, profit, export, assets, conglomerate, sales, import, industry, price
	Sa-Gwa	ask, apology, sorry, condolences, pass, envelope, report, complain, explanation, comment
++WSD	Si-Jang - Market	industry, business, competitiveness, small businesses, enterprise, investment, antioxidant, finance
	Si-Jang - Mayor	superintendent of education, self-government director, Park Soonja, The 5 th Local Elections in Korea
	Sa-Gwa - Apology	apology, pass, accusation, sorry, morning star, :’(
	Sa-Gwa - Apple	fruit, pea, chestnut, apricot, walnut, grape, nut products, poison ivy
	Entity	UN

Table 1: Similar words of ‘Si-Jang’ and ‘Sa-Gwa’ by word embedding. ‘Si-Jang’ is a Korean word, and it is mainly used for market or mayor. ‘Sa-Gwa’ is also a Korean word, and it is mainly used for apology or apple. All of the similar words are written by translating Korean words into English.

We perform the held-out evaluation of the relation extraction model using the multi-sense word embedding. Held-out evaluation is a method for measuring precision, recall, F1-score by dividing the collected data in half, and one is used for learning and the other for evaluation. The evaluation results are shown in Table 2. To verify the effectiveness of our method, we used three different embedding models, as mentioned above, as inputs of the CNN/PCNN models, and measured the performance. The hyper-parameters settings for two models are as follows; both models were set to ReLU activation and 1 drop-out, but CNN use Adadelta optimizer and PCNN use Adam optimizer.

Owing to the evaluation, both models showed better performance of using morpheme embedding (+POS) than word embedding (Word), and the performance of sense embedding (++WSD) is also improved than morpheme embedding

(+POS). Additionally, the performance of the CNN model was higher than that of the PCNN model because, in Korean, the position of two entities in the sentence often appears at the top of the sentence and the two entities are often placed consecutively.

Model	Embedding	Precision	Recall	F1-score
CNN	Word	0.5537	0.3506	0.4275
	+POS	0.5315	0.4279	0.4739
	++WSD	0.5921	0.5039	0.5443
PCNN	WSD	0.457	0.3251	0.3799
	+POS	0.4555	0.3472	0.394
	++WSD	0.4529	0.3713	0.4081

Table 2: Performance of relation extraction model by word embedding

5 Conclusion

In this paper, we propose a method for improving the performance of a distant supervision relation extraction model using multi-sense word embedding, and experimentally evaluated two relation extraction models based on CNN and PCNN. In addition, we used entity-padding word embedding, which bundles multi-word entity into a single token, when generating word embedding. Accordingly, it was confirmed that the multi-sense word embedding improves the performance of the relation extraction model.

In the future, we plan to apply the convolutional RNN model, which is a combined model of CNN and recurrent neural network (RNN), to the relation extraction task. We will also study the method of removal of error data, which is one of the problems when collecting distant supervised training data.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

- Chaplot, D. S., Bhattacharyya, P., & Paranjape, A. 2015. *Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser*. In Proceedings of AACL, pages 2217-2223.
- Choi, K. S., Bae, H. S., Kang, W., Lee, J., Kim, E., Kim, H., ... & Shin, H. 2004. *Korean-Chinese-Japanese*

- Multilingual Wordnet with Shared Semantic Hierarchy*. In Proceedings of LREC, pages 1131-1134.
- Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pages 541-550.
- Kim, Y. 2014. *Convolutional neural networks for sentence classification*. In Proceedings of EMNLP.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013. *Distributed representations of words and phrases and their compositionality*. In Advances in neural information processing system, pages 3111-3119.
- Mintz, M., Bills, S., Snow, R., & Jurafsky, D. 2009. *Distant supervision for relation extraction without labeled data*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, pages 1003-1011.
- Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. 2015. *Efficient non-parametric estimation of multiple embeddings per word in vector space*. arXiv preprint arXiv:1504.06654.
- Riedel, S., Yao, L., & McCallum, A. 2010. *Modeling relations and their mentions without labeled text*. Machine learning and knowledge discovery in databases, pages 148-163.
- Rothe, S., & Schütze, H. 2015. *Autoextend: Extending word embeddings to embeddings for synsets and lexemes*. arXiv preprint arXiv:1507.01127.
- Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. 2012. *Multi-instance multi-label learning for relation extraction*. In Proceedings of EMNLP, Association for Computational Linguistics, pages 455-465.
- Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. 2014. *Relation Classification via Convolutional Deep Neural Network*. In Proceedings of COLING, pages 2335-2344.
- Zeng, D., Liu, K., Chen, Y., & Zhao, J. 2015. *Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks*. In Proceedings of EMNLP, pages 1753-1762.

Cross-Lingual and Supervised Learning Approach for Indonesian Word Sense Disambiguation Task

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo,
Ruli Manurung, and Mirna Adriani

Faculty of Computer Science, Universitas Indonesia

Depok 16424, West Java, Indonesia

rahmad.mahendra@cs.ui.ac.id, {heninggar, haryo97}@gmail.com

Abstract

Ambiguity is a problem we frequently face in Natural Language Processing. Word Sense Disambiguation (WSD) is a task to determine the correct sense of an ambiguous word. However, research in WSD for Indonesian is still rare to find. The availability of English-Indonesian parallel corpora and WordNet for both languages can be used as training data for WSD by applying Cross-Lingual WSD method. This training data is used as an input to build a model using supervised machine learning algorithms. Our research also examines the use of Word Embedding features to build the WSD model.

1 Introduction

One of the biggest challenges in Natural Language Processing (NLP) is ambiguity. Ambiguity exists when there are many alternatives of linguistic structures that can be composed for an input language. Some words can have more than one meaning (word sense). For example the word “kali” in Indonesian can possess two senses, i.e. river and frequency (as described in Table 1)

Word Sense Disambiguation (WSD) is a task to determine the correct sense of a polysemous word. Even though it becomes a fundamental task in NLP, research on WSD for Indonesian language has not attracted many interests. To our knowledge, the only published work was Uliniansyah and Ishizaki (2005). Uliniansyah and Ishizaki applied the corpus-based approach using Naive Bayes as the classifier. The training data was collected from news websites and manually annotated. The words in training data were processed using the morphological analysis to obtain lemma. The features being used were some words around the target word (including the words before and

after the target word), the nearest verb from the target word, the transitive verb around the target word, and the document context. Unfortunately, neither the model nor the corpus from this research is made publicly available.

This paper reports our study on WSD task for Indonesian using the combination of the cross-lingual and supervised learning approach. Training data is automatically acquired using Cross-Lingual WSD (CLWSD) approach by utilizing WordNet and parallel corpus. Then, the monolingual WSD model is built from training data and it is used to assign the correct sense to any previously unseen word in a new context.

2 Related Work

WSD task is undertaken in two main steps, namely listing all possible senses for a word and determining the right sense given its context (Ide and Véronis, 1998). To list possible senses, we can use dictionaries, knowledge resources (e.g. thesaurus, WordNet), and transfer directory (e.g. translation from other language). To determine the right sense, we can use the information from the context where the word is used, and also external knowledge resource such as dictionary or encyclopedia.

Among various approaches to WSD, supervised learning approach is the most successful one to date. The supervised WSD uses machine learning techniques for inducing a classifier from sense-annotated data sets. Training data used to learn the classifier contains a set of examples in which each occurrence of an ambiguous word has been annotated with the correct sense according to existing sense inventory. (Navigli, 2009)

Despite of its success, the supervised learning approach has a drawback of requiring manually sense-tagged data. Manually labeling data for training set is costly and time-consuming. As an alternative, the sense labeling can be done automatically by utilizing existing resources. Cross

Sentence in Indonesian	English translation	Meaning of “kali”
Saya makan dua kali pagi ini	I ate twice this morning	frequency
Rumah saya di dekat kali	My house is near the river	river

Table 1: Word Ambiguity Example in Bahasa Indonesia

lingual approach is able to disambiguate word sense based on the evidence from the translation information. The rationale behind this approach is that a different sense of a word typically has different translations in other languages. The plausible translations of a word in context restrict the number of its possible senses. Cross-Lingual Word Sense Disambiguation (CLWSD) aims to automatically disambiguate a text in one language by exploiting its differences to other language(s) in a parallel corpus.

Before being a dedicated task in SemEval-2013 (Lefever and Hoste, 2010), CLWSD has been explored in several works. Brown et.al. (1991) proposed an unsupervised approach for WSD. The word alignment was performed on a parallel corpus, and then the most appropriate translation was determined for a target word based on a set of contextual features.

Ide et.al (2002) conducted an experiment using translation equivalents derived from parallel corpus to determine the sense distinctions that can be used for automatic sense-tagging and other disambiguation tasks. They found that sense distinctions derived from cross-lingual information are at least as reliable as those made by human annotators. In their study on seven languages (English, Romanian, Slovene, Czech, Bulgarian, Estonian, and Hungarian), Ide et.al exploited EuroWordNet as a knowledge source.

Sense intersection, an approach described in Gliozzo et al. (2005) and Bonansinga and Bond (2016), inspires CLWSD process in our study. Gliozzo et.al. proposed an unsupervised WSD technique to automatically acquire sense tagged data that exploited the polysemic differential between two languages using aligned corpora and multilingual lexical databases. An aligned multilingual lexical resource (e.g. MultiWordNet) allowed them to disambiguate aligned words in both languages by simply intersecting their senses. Bond and Bonansinga (2015) then applied the sense intersection approach in multilingual settings. Bonansinga and Bond (2016) considered four languages, e.g. English, Italian, Romanian,

and Japanese in their experiment to reduce more ambiguity.

For the supervised learning approach to WSD tasks, the common features are the surrounding words of target word, POS tags of the surrounding words, and local collocation. Current studies (Taghipour and Ng, 2015) (Iacobacci et al., 2016) examined the potential use of Word Embedding as a feature for the sense classification task. Iacobacci et.al. (2016) described four different strategies to use Word Embedding, e.g. concatenation, average, fractional decay, and exponential decay.

3 CLWSD for Building Indonesian WSD Training Data

We utilize CLWSD using parallel corpus and WordNet to acquire WSD training data. The model is then learned from the training data to disambiguate the word sense in testing data. Our CLWSD approach is illustrated in Figure 1.

The input for CLWSD process is English-Indonesian parallel corpus. The corpus used in our experiment is Identic++, which is Identic corpus (Larasati, 2012) that has been extended by adding the instances of English-Indonesia parallel sentences from movie subtitles. In addition to the parallel corpus, we harnessed lexical database, namely Princeton WordNet (Miller, 1995) and WordNet Bahasa (Noor et al., 2011).

CLWSD process consists of several steps:

1. Align the words in the parallel corpus using GIZA++ (Och and Ney, 2003) to obtain translation pairs.
2. Assign the sense label to the words by using sense-ID from WordNet. English words are labeled with the sense-ID from Princeton WordNet, and Indonesian words are labeled with the sense-ID from Indonesian WordNet. There may be more than one possible sense-ID for a single word. To disambiguate the word sense, we find the intersection between English and Indonesian sense inventory of the words in the translation pairs. Since our

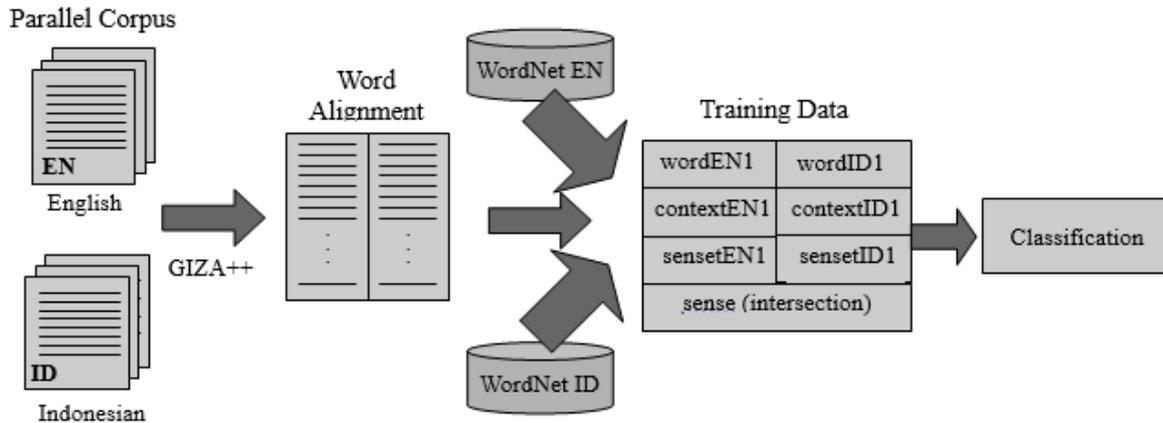


Figure 1: CLWSD Methodology

study aims to obtain the training data with high precision, we consider only the word pair instances that have exactly one intersected sense-ID.

3. Extract the content words surrounding each word to be disambiguated in the corpus.

An example is given to illustrate the process. Consider the following entry of the parallel corpus.

“EN: She reads page 50 of that book”.
 “ID: Dia membaca halaman 50 dari buku itu”.

That sentence pair (along with the rests in the corpus) is processed with GIZA++. Word “page” is aligned with “halaman”. The sense labels for the words “page” and “halaman” are listed below.

page → **06256697**, 11220149, 10391416, 10391248, 10391086
 halaman → 00193486, 00227165, 00754560, **06256697**

Among many senses corresponding to each English and Indonesian words, there is one intersected sense. Therefore, the word “halaman” in the sentence “dia sedang membaca halaman 50 dari buku itu” is labelled with the sense-ID 06256697. Moreover, the content words in this sentence include “dia”, “sedang”, “membaca”, “dari”, “buku”, and “itu”.

Word	Translation	Number of Instances
alam	nature	251
	universe	225
atas	above	546
	top	441
kayu	timber	51
	wooden	49
anggur	wine	272
	grape	21
perdana	prime	302
	premier	22
dasar	primary	225
	underlying	11
All samples		2,416

Table 2: Sample Words for Monolingual WSD

We have retrieved 352,816 pairs of aligned words between Indonesian and English from the Identific++ corpus. Among of them, 4,237 Indonesian words are polysemous. The rest of words may have only one sense (not ambiguous) or no corresponding sense found in WordNet towards them. Finally, 752 different words can be disambiguated using sense intersection approach.

4 Supervised Learning for Indonesian Language WSD

The sense-tagged words acquired in CLWSD process are used to train classifier. The classifier induced the model for Indonesian WSD. For evaluation of the supervised learning approach, we performed monolingual lexical sample WSD task. We tested sample of 2,416 sentences that contain

Word	Baseline	NB	MLP	RF	SVM	XGB
alam	36.41	62.45	94.54	94.75	95.17	96.01
atas	39.41	69.79	71.95	71.16	71.69	72.21
kayu	34.45	69.98	66.52	71.98	70.71	73.98
anggur	89.38	89.17	91.81	89.86	90.92	89.40
perdana	90.03	89.12	93.77	91.23	91.64	91.04
dasar	93.06	92.42	95.90	93.06	94.29	94.29
Average	63.79	78.82	85.75	85.34	85.74	86.16

Table 3: F1 Score of Baseline vs Machine Learning Models using BoW Features

one of 6 target words. Each of these target words has two possible senses. Sample Indonesian words for monolingual WSD experiment are listed in Table 2.

We ran the experiment in 10-fold cross validation setting. We built the model using five different supervised machine learning algorithms, namely Naive Bayes, Multi Layer Perceptron (MLP), Random Forest (Breiman, 2001), Support Vector Machine (SVM) (Boser et al., 1992), and XGBoost (Chen and Guestrin, 2016). For the baseline evaluation, we assign the most frequent sense label to each instance.

Using the content words as bag-of-words (BoW) representation, any machine learning models tested in our experiment outperformed the baseline evaluation. All machine learning models but Naive Bayes obtained the average F-1 score >85%. XGBoost model achieved the best average F-1 score, that is 86.16%. On other hand, MLP performed better compared to other models to disambiguate the words with imbalanced sense label distribution (e.g. “anggur”, “perdana”, and “dasar”). Complete evaluation of the baseline and machine learning methods is presented in Table 3.

4.1 POS and Word Embedding as Features

A word in the different parts of speech (POS) has the different sense. A word used in the different senses is likely to have the different set of POSs around it. So, the POS information of content words can be potential cue to determine the word sense.

To obtain the POS feature, we used Indonesian POSTag model from Rashel et.al (2014). In general, incorporating the POS into the bag-of-words features improve WSD performance in our experiment. Average F-1 scores of SVM and MLP models increase, but there is a slight decrease in F-1

score of XGBoost model.

Beside that, we conducted other experiments using the Word Embedding features. We transformed each word in the sentence into continuous-space vector representation using skip gram model pre-trained by Word2Vec (Mikolov et al., 2013). We considered two different strategies to incorporate the Word Embedding in monolingual Indonesian WSD task. First, the vectors of the content words are concatenated into a larger vector that has a size equal to the aggregated dimension of all the individual embeddings (**concat**). Second, the vectors of the content words are summed up and the resultant vector is divided by number of content words (**avg**).

Sense classification using the Word Embedding features produced promising result. MLP and XGBoost model that make use of the Word Embedding on the basis of average strategy reach the F-1 score respectively 86.80% and 86.34%. These scores are higher than the best result achieved by same models using the traditional bag-of-words only. The experimental result related to use of the features in lexical sample WSD task is reported in Table 4.

4.2 Effect of Stemming and Stopword Removal to BoW Features

Stemming is a common technique used in information retrieval to eliminate the morphology variations to obtain the basic form of a word. On the other hand, stopword removal is the process of removing common words that are often used in many sentences, e.g. “and”, “or”, “is”. We had a hypothesis that the stemming and stopword removal can affect the WSD system performance. Stemming is used in order that the words with different morphological forms can be counted as the same content words. In addition, stopwords removal is used to prevent the matrix representing

Word	BoW			BoW+POS			WE (concat)			WE (avg)		
	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB
alam	95.17	94.54	96.01	95.60	94.13	96.22	89.10	88.26	85.53	87.42	89.94	92.87
atas	71.69	71.95	72.21	73.29	74.72	74.38	63.72	67.38	68.37	67.20	69.81	70.10
kayu	70.71	66.52	73.97	73.72	67.91	68.36	71.25	78.21	74.23	82.12	82.16	77.21
anggur	90.92	91.81	89.40	90.42	92.06	88.68	87.72	89.01	89.79	91.84	92.23	91.06
perdana	91.64	93.77	91.03	93.54	93.54	92.36	90.11	89.04	90.83	89.95	91.03	91.12
dasar	94.28	95.90	94.28	93.06	92.85	94.98	93.77	92.64	92.85	95.64	95.65	95.67
Average	85.74	85.75	86.16	86.61	85.87	85.83	82.61	84.09	83.60	85.70	86.80	86.34

Table 4: WSD Experiment Using POS and Embedding Features

Word	BoW			stem			no stopword			stem & no stopword		
	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB	SVM	MLP	XGB
alam	95.17	94.54	96.01	95.80	93.49	95.80	95.38	92.85	96.22	96.01	93.48	96.43
atas	71.69	71.95	72.21	71.00	74.18	72.06	69.31	70.77	63.89	70.82	72.09	67.22
kayu	70.71	66.52	73.98	78.79	70.96	67.97	75.59	61.46	70.45	82.89	65.89	70.96
anggur	90.92	91.81	89.40	91.41	90.17	89.60	90.26	91.56	88.87	90.60	91.31	89.38
perdana	91.64	93.77	91.04	93.01	93.03	91.04	92.80	94.43	92.21	92.80	93.01	91.64
dasar	94.29	95.90	94.29	95.00	96.25	94.59	93.06	94.90	94.07	94.04	93.06	94.67
Average	85.74	85.75	86.16	87.50	86.35	85.18	86.07	84.33	84.29	87.86	84.81	85.05

Table 5: Effect of Stemming and Stopword Removal to WSD Model

content words becomes too sparse, as well as to remove unimportant words from the content words.

We used the Indonesian stemmer (Adriani et al., 2007) to derive the stem of content word, while stopword removal was conducted using dictionary of Indonesian stopwords (Tala, 2003).

The effect of stemming in this study is increasing the F1-score (for SVM and MLP model). The initial F1-score of SVM model using the bag-of-words feature is 85.74% and after the stemming the F1-score becomes 87.50%. The words, that were previously considered different because of the morphological variations, are counted as the same words after the stemming, so two sentences that were considered unlike now become similar. On the other hand, the effect of stopwords removal is not as good as stemming. MLP and XGBoost models have decreased the F1-scores when the stopwords are excluded from the bag-of-words feature. We argue that the stopwords list may still contain the words that are discriminative enough to explain the context of the sentence.

5 Summary

In our study, CLWSD has been implemented to provide the training data and then the model based on the training data is built by the classifier to perform monolingual Indonesian WSD. We took

advantage of existing of the parallel corpus and WordNet to obtain the sense-labeled words by a cross lingual approach. We retrieved all possible senses for the translation pairs and then found the intersection between senses from English and Indonesian language. The data acquired by CLWSD process is released at <https://github.com/rmahendra/Indonesian-WSD>. We ran several experiments on monolingual WSD task and concluded that any supervised machine learning model outperforms the baseline method. Moreover, we found that the use of embedding vector can produce better F1-score of sense classification than the use of the traditional bag-of-words features.

The study still has rooms for improvement. We need to test our methodology in larger corpus and involve more target words for experiment. Detail evaluation of CLWSD to produce Indonesian training data can be more explored. On the other hand, it is interesting to check how sensitive our proposed approach works when considering the semantic difference between senses.

Acknowledgments

The authors gratefully acknowledge the support of the PITTA UI Grant Contract No. 410/UN2.R3.1/HKP.05.00/2017

References

- Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi, and Hugh E. Williams. 2007. Stemming indonesian: A confix-stripping approach. *Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33, December.
- Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proceedings of the Eighth Global WordNet Conference*, pages 44–49.
- Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In: Second Italian Conference on Computational Linguistics CLiC-it.
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, pages 144–152, New York, NY, USA. ACM.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, October.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91*, pages 264–270, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, pages 785–794, New York, NY, USA. ACM.
- Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Crossing parallel corpora and multilingual lexical databases for wsd. In *In: Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, vol 3406, CICLing 2005*, pages 242–245, Berlin, Heidelberg. Springer.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.
- Nancy Ide and Jean Véronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):2–40.
- Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8, WSD '02*, pages 61–66, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Septina Dian Larasati. 2012. Identical corpus: Morphologically enriched indonesian english parallel corpus. In *Proceedings of LREC*.
- Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13*, pages 3111–3119, USA. Curran Associates Inc.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Communication of ACM*, 38(11):39–41.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10:1–10:69, February.
- Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an indonesian rule-based part-of-speech tagger. In *Proceedings of 2014 International Conference on Asian Language Processing (IALP)*. IEEE.
- Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 314–323.
- Fadillah Z Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia. Master's thesis, University of Amsterdam, the Netherlands.
- Mohammad Teduh Uliniansyaht and Shun Ishizaki. 2005. A word sense disambiguation system using modified naive bayesian algorithms for indonesian language. *Natural Language Processing*, 12(1):33–50.

Recognition of Hyponymy and Meronymy Relations in Word Embeddings for Polish

Gabriela Czachor, Maciej Piasecki, Arkadiusz Janz

Department of Computational Intelligence, Wrocław University of Science and Technology,
{maciej.piasecki, arkadiusz.janz}@pwr.edu.pl

Abstract

Word embeddings were used for the extraction of hyponymy relation in several approaches, but also it was recently shown that they should not work, in fact. In our work we verified both claims using a very large wordnet of Polish as a gold standard for lexico-semantic relations and word embeddings extracted from a very large corpus of Polish. We showed that a hyponymy extraction method based on linear regression classifiers trained on clusters of vectors can be successfully applied on large scale. We presented also a possible explanation for contradictory findings in the literature. Moreover, in order to show the feasibility of the method we extended it to the recognition of meronymy.

1 Introduction

A very large wordnet, e.g. plWordNet (Maziarz et al., 2014) describes many lexico-semantic relations, linking lexical units¹ (or word senses) by thousands of relation instances. However, even in a very large wordnet some relation instances can be omitted and typically wordnets are very biased towards only a few relations, e.g. hypernymy/hyponymy for nouns, with much smaller coverage for the other. Measures of semantic relatedness constructed on the basis of word embeddings (Mikolov et al., 2013b) are known to express many different lexico-semantic relations, e.g. on a list of the k most related words to a word x we can typically find words associated with x by different relations. However, word embeddings are very heterogeneous with respect to the types of semantic associations expressed, also including syntactic and pragmatic relations. What is worse, word embeddings have problems with representing different senses of a word (typically only a few most frequent ones can be spotted on the lists) and with proper representation of less frequent words (even words with frequency 100–200 per 1G can be erroneously

¹Lexical units here are triples: lemma, Part of Speech and sense id.

described, not mentioning those < 100). So, the question is whether we can successfully recognise among the associations suggested by word embeddings those that correspond to lexico-semantic relations, i.e. whether we can interpret word embeddings in a meaningful way for humans. The works presented for English are contradictory even in the case of *hypernymy* – the relatively simplest relation: from successful extraction (Fu et al., 2014) till denial of the feasibility of such a method (Levy et al., 2015).

We want to re-approach this intriguing issues, first checking the contradictory points of the view on large corpora and comprehensive wordnet for Polish, second by expanding this research with one more relation, a more difficult one, namely *meronymy*. This is a part of a broader work on the automated extraction of lexico-semantic relations that are under-represented in wordnets, e.g. in order to improve wordnet-based WSD.

2 Related Works

ClassHyp system of (Piasecki et al., 2008) used a measure of semantic relatedness based on Distributional Semantics together with several statistical knowledge sources extracted from corpora (e.g. describing specificity of features for a noun) to build a classifier in a supervised Machine Learning process. The classifier was trained to recognise selected wordnet relations among word pairs.

(Fu et al., 2014) assumed that as a hyponym extends features of its hypernym (i.e. shares features with its hypernym and adds more specific ones), the hyponym’s and hypernym’s word embedding vectors should be related in some characteristic way, i.e. we can find some aspect of semantic inclusion when comparing both vectors. They proposed to use offsets between embedding vectors as representation of the projection (or “mapping”) of a hyponym on its hypernym. Offsets were simply calculated by subtracting vectors representing a hypernym y and a hyponym x : $\mathbf{y} - \mathbf{x}$. (Fu et al., 2014) observed that hypernymy relation can vary beyond one uniform projection. As a result, they proposed to cluster the offset vectors for training pairs into a number of groups and next to train

a separate classifier based on linear regression for each group. Training examples were taken from a large Chinese thesaurus, but quite shallow and with coarse grained sense distinction. It included 5 level hierarchy with the fourth level including non-lexicalised concepts. Vector offsets for direct and indirect hypernymy pairs were clustered into separate groups, but nevertheless the indirect pairs represented quite close relations (max. length 3). The number of clusters was established experimentally on a separate development set. A test example was classified as a positive, if it received a positive decision from at least one classifier.

(Levy et al., 2015) analysed several methods for the extraction of relations that can express different forms of lexical inference, e.g. hypernymy, entailment or causation. They tested four different ways for representing pairs of words by feature vectors based on word embeddings, namely: *concatenation*, *subtraction* (called *difference*) and representations by single vectors of one of the words. Several different tests proposed in literature were used. In most of the cases supervised methods based on two-vector representation were only slightly better than the single vector representation of the more general word from the give pair. (Levy et al., 2015) proposed also an evaluation experiment in which negative pairs were artificially built from words included in the positive pairs. By using a SVM classifier they showed that the correlation between match error and recall (positive) is close to the perfect correlation in a series of experiments. As a result, (Levy et al., 2015) concluded that the supervised classifiers proposed in literature, including (Fu et al., 2014), are learning whether “ y is a prototypical hypernym (i.e. a category) regardless of x , rather than learning a concrete relation between x and y .” They called this potential effect “lexical memorizing”. They also claimed that “contextual features might lack the necessary information to deduce how one word relates to another”. However, it is worth to notice that the reported results for supervised approaches based on two vectors were in fact in the most cases slightly but significantly better than single-vector results, and (Levy et al., 2015) did not apply the original approach of (Fu et al., 2014) in their key tests (sic!). Moreover, all evaluations were done only for English and for quite limited test data.

However, there are also many other works that report on successful extraction of hypernymy from contextual features, e.g. (Shwartz et al., 2016).

3 Search for Relations in Word Embeddings

The method proposed by (Fu et al., 2014) intuitively seems to be correct: elements of the word embeddings are derived from the occurrence con-

texts and correspond to the semantic features of words, while the similarity of features of two words correspond to the amount of overlapping in the values. This inspired us to revisit the method of (Fu et al., 2014) in a new setting and confront it once again with the objections of (Levy et al., 2015).

3.1 Corpora and Vector-based Representation

As a gold standard for lexico-semantic relations we used plWordNet – a very large wordnet of Polish (Maziarz et al., 2016). It is substantially bigger than Princeton WordNet (Fellbaum, 1998), and was constructed from scratch using a corpus-based wordnet development method. As a result plWordNet has much better coverage of words in large corpora than other wordnets including Princeton WordNet. As a source of text data, we utilised plWordNet Corpus (henceforth plWNC) which includes ≈ 4 billion words and combines all publicly available Polish corpora, and a very large number of Polish texts collected from the Web².

Using *word2vec* tool (Mikolov et al., 2013a) we built embedding vectors as representations for all words from plWNC with the frequency ≥ 8 (min_count=8). We tested several different settings of *word2vec*, see (?), and finally, we selected the Skip-gram model and two vector sizes: 100 and 300, as best performing during the wordnet-based evaluation.

3.2 Classifiers based on Clusters

Following (Fu et al., 2014), we represent a hyponymy instance (link): $\langle x, y \rangle$, where x – a hyponym, and y – a hypernym, are lemmas belonging to two separate synsets by the difference of two word embedding vectors: $\mathbf{x} - \mathbf{y}$. It is also assumed that comparison of the difference vectors should reveal a *projection* which corresponds to the feature sharing pattern that is characteristic for hyponymy. This hypernymy projection, reducing the hyponym specific features, can be expressed by a linear projection of the vector \mathbf{x} , i.e. $\Phi\mathbf{x}$, on a vector \mathbf{y}' . However, both vectors can represent other semantic aspects beyond the feature sharing (e.g. polysemy, differences in contexts of occurrences etc.) and the set of additional features introduced by a hyponym. So the difference can

²It consists of IPI PAN Corpus (Przepiórkowski, 2004), the first annotated corpus of Polish, National Corpus of Polish (Przepiórkowski et al., 2012), Polish Wikipedia (from 2016), *Rzeczpospolita* Corpus (Weiss, 2008) – a corpus of electronic editions of a Polish newspaper from the years 1993-2003, supplemented with text acquired from the Web – only text with small percentage of words unknown to a very comprehensive morphological analyser Morfeusz 2.0 (Woliński, 2014) were included; duplicates were automatically eliminated from the merged corpus.

be biased beyond the capabilities of a representation by a single common hypernymy projection. In order to obtain a more regular picture, difference vectors for the training hyponymy instances are automatically clustered and for each cluster a different classifier is trained. The *k-means* algorithm was used for clustering and for each cluster a separated classifier was trained by the linear regression method. In a similar way, negative examples of non-hyponymic pairs constructed on the basis of p1WordNet are clustered and negative classifiers are built. A test difference vector for a pair $\langle x, y \rangle$ is classified as representing hyponymy if it is positively classified by at least one of the created classifiers.

Semantic representation based on word embeddings has several limitations, e.g.:

1. the whole model can be biased by the particular selection of texts,
2. senses of polysemous words are merged together, i.e. represented by a single vector,
3. and the representation of less frequent words and senses can be blurred by the statistical noise.

This problem was not explicitly and well enough treated in both contradictory works, namely: (Fu et al., 2014) and (Levy et al., 2015).

In order to decrease the potential bias, the point 1, we used as large and diversified corpus as possible. To understand the influence of uneven representation of different senses, the point 2, we divided experiments into two groups of: monosemous words only, and all words. We used also a large number of words in the experiments. To avoid noise caused by low frequency of data, we took into account in all experiments only words with more than 1,000 occurrences³.

(Fu et al., 2014) tested their method for several different number of clusters achieving the best results with larger numbers. We established the value of this parameter by automated optimisation on a *development* subset. In each experiment the data were randomly divided into three subsets: *training*, *testing* and *development* in the ratio 6:2:2.

As noun hypernymy in p1WordNet forms quite deep hierarchical structures (in some cases beyond 20 levels). Thus, testing indirect hyponyms with longer hypernymy paths could also make the analysis of the results more difficult. That is why we limited positive examples only to direct hyponymic pairs. As negative ones, we created word pairs that do not overlap with direct and indirect hyponymic pairs up to the distance of three links.

³A heuristic threshold applied in many works and which seems to heuristically demarcate the area of robust representations, obviously on average

4 Experiments

Following the suggestion of (Levy et al., 2015), we constructed data sets for experiments in two ways:

1. *random* division into subsets,
2. and *lexical train/test splits* rule proposed by them.

In (2) the division is random but positive cases in the test set (i.e. true hyponymy instances) cannot include hypernyms occurring in the training set. Moreover the negative cases are also constructed in a tricky way explained below in Eq 1–3, where T^+ is a set of word pairs belonging to the given relation:

$$T_x^+ = \{x \mid (x, y) \in T^+\} \quad (1)$$

$$T_y^+ = \{y \mid (x, y) \in T^+\} \quad (2)$$

$$S = (T_x^+ \times T_y^+) \setminus T^+ \quad (3)$$

As a result S contains false relation instances, but constructed from words included in the positive examples, also hypernyms that are suspected to be the signal recognised by classifiers in the data. This type of division is meant to prevent training a classifier which recognises prototypical hypernyms instead of hyponymy relation.

Results of experiments on recognition of the hyponymy relation are presented in Tab. 1. Each experiment was performed in $k = 10$ fold cross-validation setting. Due to the limited space, only average results from the 10 folds are presented in Tab. 1. In the following experiments we have analysed:

Hypo-Mono – hyponymy recognition for monosemous words: 6,000 hyponymy pairs including only monosemous words as positive examples, 6,000 negative examples; the two variants of the generation of negative examples were applied: *random* and *lexical split*; the size of the embedding vectors was 100.

Hypo-Poly – 20,000 hyponymy pairs including polysemous words; 20,000 negative examples were selected using the assumed two methods; the vector size was 100.

Hypo-Mono300 – as in Hypo-Mono but the vector size is 300 in order to check a more fine-grained description, only lexical split method was used for the generation of the negative examples, i.e. the more difficult one.

Hypo-Poly300 – as above, but 20,000 hyponymy pairs including polysemous words were used, 20,000 negative pairs were selected by the lexical split, the vector size was: 300.

Mero-Poly – 7,900 meronymy pairs (only the main subtype *part of*), 8,000 pairs of words that are not connected in *plWordNet* at all or are connected by paths longer than 3 links were selected as negative examples by the lexical split method, the vector size was: 100.

5 Results

For pairs of experiments performed using two different ways of the selection of negative examples, as well as for two different sizes of the vectors we checked the statistical significance of the differences.

First we tested if the results obtained in different folds come from the normal distribution by applying Shapiro-Wilk test, e.g. for **Hypo-Mono** we obtained p value of 0.8082 for the random selection results and 0.8648 for the lexical split series, so with the confidence level of 0.05 we cannot reject the null hypothesis that the results come from a normal distribution. Having normality confirmed, we applied *t-Student* test to the differences between results, e.g. in the case of **Hypo-Mono** p value is 0.4583 and with the confidence level 0.05 we cannot reject the null hypothesis of the lack of a difference between both series of results.

In the case of **Hypo-Poly** the fold results do not come from a normal distributions according to Shapiro-Wilk test, so we applied *Mann-Whitney U test* and the obtained p value of 0.008931 shows the lack of statistical significance of the differences. In a similar way we checked that the differences between results for different vector sizes are significant, namely **Hypo-Mono** vs **Hypo-Mono300** and **Hypo-Poly** vs **Hypo-Poly300**. We did not analysed differences between the results for monosemous and polysemous words, but these differences are very visible.

5.1 Hyponymy Recognition

In Tab. 1, we can observe that in all experiments very good results in the recognition of hyponymy relation were achieved. As (Levy et al., 2015) expected, the lexical split selection of negative samples caused the decrease of the results. However, the observed differences are small ≈ 1 in the value of percents for monosemous and ≈ 2 for polysemous, while e.g. (Shwartz et al., 2016) reported the difference of ≈ 20 . Moreover, these differences are not statistically significant. It means that recognition methods trained on other hypernyms that those in the test set are still working, properly recognise hyponymy instances and are not simply deviating to prototype recognition as suggested by (Levy et al., 2015). Moreover, the small difference between the random and lexical split selections can be also attributed to the imperfection of the linear projection based on a lim-

ited number of clusters, that is less precisely tuned for hypernyms coming from different subbranches of *plWordNet*. In all cases recall is higher than precision, but in applications, e.g. in wordnet development, this is a required property, as we do not want to loose potential hyponymy instances. Significantly lower results that were obtained for longer embedding vectors of 300 elements, especially for **Hypo-Mono300** are surprising. This can be caused by insufficient number of training examples, as in the case of **Hypo-Poly300** the results are higher when using a larger training set.

In order to test a potential influence of the training data size on the hypernymy recognition we performed a series of experiments on randomly selected subsets of **Hypo-Mono** with the increasing subset size. The sequence of results is presented in Tab. 2. The trends observed in them are illustrated in Fig. 1. We can observe that in Accuracy, Precision and F-measure values are increasing with the increasing size of the data, and it is difficult to definitely say whether this process saturates with the size 6,000. It is quite surprising that Recall is decreasing. However it quickly goes high, so the later small decrease can result from a better ability of the model to separate positive and negative cases. On the basis of this experiment we can conclude that larger volume of data improves the performance of this type of a classifier.

The substantial discrepancy of our findings with the claimed inability to train supervised recognition on the basis of word embedding vectors observed in (Levy et al., 2015) can be also caused by the choice of different classification methods: so far we followed the work of (Fu et al., 2014) and we combined unsupervised clustering with the construction of supervised classifiers based simply on linear regression, while (Levy et al., 2015) used only SVM algorithm. To complete the picture we also repeated for all experiments the error analysis proposed by (Levy et al., 2015), e.g. for **Hypo-Mono** it is presented in Fig. 2.

In Fig. 2 the ratio of the matching error (see Tab. 1), a kind of ‘negative’ recall, and the positive recall for different folds is presented. If a classifier recognises not relation instances, but hypernyms as prototypes, than it reacts in a similar way to both negative and positive examples as the negative ones prepared by lexical split include hypernyms from the training data. (Levy et al., 2015) showed that this ratio for different experiments is perfectly set on the diagonal. In our case all values are far way from the diagonal.

We also used the training and testing data prepared according to the lexical split from **Hypo-Mono** and a SVM-based classifier. Many experiments were performed with different settings of the classifier (kernels: linear, polynomial and ra-

Experiment	Acc	P	R	F	Err	Type	Vec. Size
Hypo-Mono	85.22%	78.91%	96.27%	86.72%	27.91%	Rnd	100
<i>std. dev.</i>	0.64%	1.00%	0.65%	0.65%	1.92%	Rnd	100
Hypo-Mono	84.98%	78.90%	95.18%	86.27%	28.05%	Lex. split	100
<i>std. dev.</i>	0.61%	1.59%	0.79%	0.91%	2.22%	Lex. split	100
Hypo-Poly	78.94%	74.35%	88.35%	80.74%	31.63%	Rnd	100
<i>std. dev.</i>	0.65%	0.41%	1.70%	0.79%	1.78%	Rnd	100
Hypo-Poly	77.23%	73.83%	84.66%	78.85%	30.54%	Lex. split	100
<i>std. dev.</i>	0.79%	1.40%	2.39%	1.04%	2.25%	Lex. split	100
Hypo-Mono300	73.31%	65.16%	98.20%	78.32%	–	Lex. split	300
<i>std. dev.</i>	1.11%	1.82%	0.39%	1.31%	–	Lex. split	300
Hypo-Poly300	82.54%	84.51%	94.72%	89.32%	–	Lex. split	300
<i>std. dev.</i>	1.01%	1.11%	0.69%	0.73%	–	Lex. split	300
Mero-Poly300	79.95%	74.66%	90.43%	81.77%	–	Lex. split	100
<i>std. dev.</i>	1.05%	1.71%	1.38%	0.99%	–	Lex. split	100

Table 1: Supervised recognition of lexico-semantic relations on the basis word embedding vectors, where *Acc* is the percentage of correct decisions, *P* – positive precision, *R* – positive recall, *F* – F-measure from *P* and *R*, *Err* – the match error, $2FP/(TN + FP)$, a ‘reversed’ recall, *Type* – the selection method for negative examples and *Vec. size* – the size of the embeddings vectors. All results are average from the 10 folds cross validation. In *std. dev.* standard deviation calculated for 10 results is provided. In the case of similar experiments only the differences between **Hypo-Mono** vs **Hypo-Mono300** and **Hypo-Poly** vs **Hypo-Poly300** are statistically significant.

Dataset Size	Accuracy	Precision	Recall	F-measure
1000	61.84%	60.92%	82.28%	68.60%
1500	61.04%	61.00%	81.07%	65.93%
2000	66.90%	61.54%	98.95%	75.84%
2500	66.41%	60.26%	98.39%	74.68%
3000	71.81%	63.52%	98.30%	77.11%
3500	74.55%	66.69%	97.71%	79.25%
4500	77.81%	71.09%	96.16%	81.73%
5000	78.43%	70.97%	96.72%	81.85%
5500	80.15%	72.78%	96.41%	82.94%
6000	80.73%	73.10%	96.26%	83.07%

Table 2: Average values (from 10 folds) for different evaluation measures with respect to the size of the training-testing dataset selected from **Hypo-Mono**

dial, cost $C \in \{1, 10, 100, 1000\}$, example influence $\gamma \in \{0.001, 0.0001\}$) and 10-fold cross validation. The results were compared in the ratio analysis presented in Fig. 3. There is varied distribution of the ratio values in contrast to the univocally bad situation reported in (Levy et al., 2015). However, many of the points are located on the diagonal or close to it. This suggests that the pessimistic conclusions of (Levy et al., 2015) maybe limited only to some settings of the SVM classifier, when it was applied to the relation recognition in the word embedding vectors.

5.2 Analysis of Clusters

In order to get insight into the work of the classifiers, we have also examined the structure of clusters built on the basis of vector differences. We

were analysing if one cluster corresponds to one specific, possibly domain-dependent realisation of the hyponymy relation. As it was very unclear how this could be assessed automatically – all clustered pairs represented hyponymy – we performed manual inspection of selected clusters built on the basis of **Hypo-Mono** dataset.

We used K-means algorithm and we set the number of clusters to be equal to the number of different top most hypernyms in the dataset. We could observe that created clusters include very often pairs of the same hypernym, see examples presented in Tab. 3. This seems to suggest that clusters do not represent different realisations of hyponymy, contrary to the assumption of (Fu et al., 2014), but rather group difference vectors according to the more general lemmas, i.e. their most

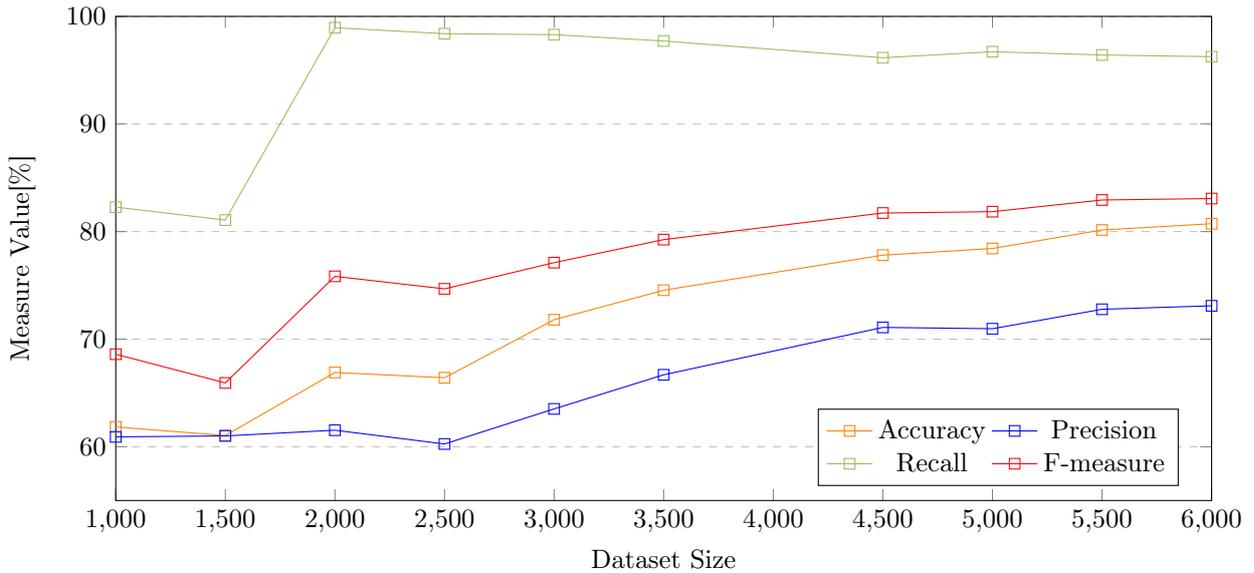


Figure 1: Average values (from 10 folds) for different evaluation measures with respect to the size of the training-testing dataset selected from **Hypo-Mono**.

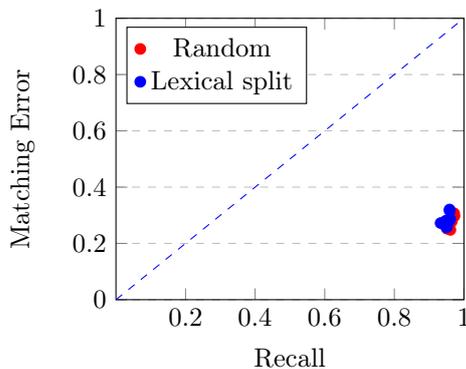


Figure 2: Ratio between the matching error and recall for different folds in **Hypo-mono** experiment and for the two methods of the selection of negative samples.

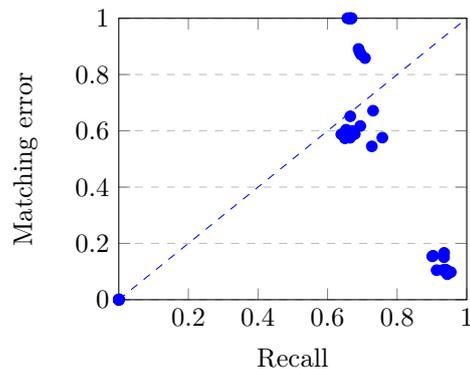


Figure 3: Ratio between the matching error and recall for supervised recognition of hyponymy by using different configurations of SVM

prominent, but quite specific features, corresponding somehow to potential prototypes. However, this tendency was not a general rule confirmed by all inspected cases and our manual analysis was too selective to formulate strong conclusions.

In addition, we applied hierarchical agglomerative clustering for monosemous hyponymic pairs, direct and indirect (from **Hypo-Mono**), including several very different hypernyms: *mebel* ‘a piece of furniture’, *szafa* ‘a wardrobe’, *fotel* ‘a chair’, *zwierzę* ‘an animal’, *kot* ‘a cat’ and *pies* ‘a dog’. We used cosine measure for clustering. In the results we could observe that pairs related to animal and furniture were linked together only in later stages of clustering (on higher levels of the tree). Pairs of the same hypernym were merged on earlier stages of clustering. Also pairs of hypernyms from the

same category, e.g. cat and dog were merged quite early into common clusters. However, we could also observe that literal and metaphorical senses of the lemma *zwierzę* ‘an animal’ were initially separated into different subtrees.

Next we added to the set a small subset of negative examples constructed by exchanging hyponyms in the pairs. In the resulting cluster hierarchy, the negative pairs first were merged together and only after this their subbranches were linked with other clusters. It seems that vectors of true hyponymic pairs were distinguished from the negative ones. The structure of the cluster tree was dominated by hypernyms, but still very different hypernyms dominated separate subtrees. So, we can provisionally conclude that the information expressed in the difference vectors is mixture of the information about domains and prototypes. This

Hypernym	Hyponym	Cluster ID
wyziew ‘vapour’	spaliny ‘engine exhausts’	973
usługa ‘service’	przewóz ‘transport’	973
usługa ‘service’	fryzjerstwo ‘hairdressing’	973
usługa ‘service’	outsourcing ‘outsourcing’	973
usługa ‘service’	usługa powszechna ‘common service’	973
usługa ‘service’	usługa telekomunikacyjna ‘telecommunication service’	973
usługa ‘service’	produkt bankowy ‘bank product’	973
nudziarz ‘bore’	szywniak ‘staffed shirt’	973
dysputa ‘≈debate’	polemika ‘polemic’	1101
dostojnik ‘high official’	podsekretarz ‘undersecretary’	1101
dostojnik ‘high official’	wiceminister ‘vice-minister’	1101
dygnitarz ‘dignitary’	wiceminister ‘vice-minister’ 1	1101
oficjel ‘high-up’	wiceminister ‘vice-minister’	1101
dostojnik ‘high official’	wicepremier ‘deputy prime minister’	1101
dygnitarz ‘dignitary’	wicepremier ‘deputy prime minister’	1101
oficjel ‘high-up’	wicepremier ‘deputy prime minister’	1101
dezaprobata ‘disapproval’	wotum nieufności ‘vote of censure’	1101

Table 3: Examples of clusters of lemma pairs constructed by k-Means algorithms from difference vectors.

is one more suggestion that the test applied (Levy et al., 2015) does not necessarily work in a general setting, but the specific setting of SVM.

5.3 Meronymy Recognition

It is worth to emphasise that we achieved a very good result for meronymy, see Tab. 1 by applying exactly the same method of (Fu et al., 2014) as for hyponymy recognition, i.e. linear regression classifiers trained on clusters of difference vectors. It was helpful that we concentrated on the *part of* subtype of the meronymy, i.e. probably, the most prototypical subtype. However, meronymy is usually more difficult relation to be extracted. Its recognition in the word embeddings vectors cannot be explained by sharing a prototype, as it is a more complex relation, and in this particular experiment we were using the lexical split technique, too. The obtained results are much lower, but the experiment was performed on monosemous and polysemous lemma pairs together, i.e. in the more difficult setting. Also in this case recall is higher than precision and probably use of a larger amount of data could improve this.

6 Conclusions

The claim of (Levy et al., 2015) that supervised classifiers trained on combinations of word embeddings vectors are learning in fact that one of the words is a prototypical hypernym, instead recognising the pair as an instance of the hyponymy relation seemed to be well motivated. However, it was contradictory with intuition and many results reported in literature. One of them, namely (Fu et al., 2014), presented good results, but tested on a limited scale of only 412 words. In our work

we verified both claims using a very large wordnet of Polish (developed in a more linguistically oriented way and closer to the language data in corpora) as a gold standard for lexico-semantic relations and word embeddings extracted from a very large, automatically pre-processed corpus of Polish. We showed that the method proposed by (Fu et al., 2014) can be successfully applied to the extraction of the hyponymy relations. In series of carefully conducted and evaluated experiments we verified negatively the objections of (Levy et al., 2015). This contradiction can be due to different languages and datasets used, but also to the fact that they concentrated their attention on the use of SVM classifiers only, while we showed that in some settings SVM classifier can produce much worse results for this particular task.

In addition we applied successfully the method of (Fu et al., 2014) also to the recognition of meronymy achieving very good results tested on a very large data sample prepared manually. We plan to expand this approach to other relations, e.g. lexico-semantic relations manifested derivationally that are quite numerous in Polish. We aim at building a semi-automated system for improving the density of relations in a wordnet. It will be also very valuable to continue the research on types of classifiers and experimental settings that make extraction methods of this types successful.

Acknowledgments

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL.

References

- Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *ACL (1)*, pages 1199–1209.
- Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan. 2015. Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*, pages 970–976.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plwordnet as the cornerstone of a toolkit of lexico-semantic resources. In *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312.
- Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Pawel Kedzia. 2016. plwordnet 3.0 - a comprehensive lexical-semantic resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.
- Maciej Piasecki, Michał Marcińczuk and Stanisław Szpakowicz, and Bartosz Broda. 2008. Classification-based filtering of semantic relatedness in hypernymy extraction. In Bengt Nordström and Arne Ranta, editors, *Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Sweden, August 25-27, 2008, Proceedings*, volume 5221 of *LNCS*, pages 393–404. Springer.
- Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences.
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN.
- Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany, August. Association for Computational Linguistics.
- Dawid Weiss. 2008. Korpus Rzeczpospolitej [Corpus of text from the online edition of “Rzeczpospolita”]. <http://www.cs.put.poznan.pl/dweiss/rzeczpospolita>.
- Marcin Woliński. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

Simple Embedding-Based Word Sense Disambiguation

Dieke Oele

Rijksuniversiteit Groningen,
Groningen,
d.oele@rug.nl

Gertjan van Noord

Rijksuniversiteit Groningen,
Groningen,
g.j.m.van.noord@rug.nl

Abstract

We present a simple knowledge-based WSD method that uses word and sense embeddings to compute the similarity between the gloss of a sense and the context of the word. Our method is inspired by the Lesk algorithm as it exploits both the context of the words and the definitions of the senses. It only requires large unlabeled corpora and a sense inventory such as WordNet, and therefore does not rely on annotated data. We explore whether additional extensions to Lesk are compatible with our method. The results of our experiments show that by lexically extending the amount of words in the gloss and context, although it works well for other implementations of Lesk, harms our method. Using a lexical selection method on the context words, on the other hand, improves it. The combination of our method with lexical selection enables our method to outperform state-of the art knowledge-based systems.

1 Introduction

The quest of automatically finding the correct meaning of a word in context, also known as Word Sense Disambiguation (WSD), is an important topic in Natural Language Processing (NLP). WSD systems that are based on supervised learning methods gain best results (Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli and Lapata, 2007; Navigli, 2009; Zhong and Ng, 2010). However, they require a large amount of manually annotated data for training. Also, even if such a supervised system obtains good results in a certain domain, it is not readily portable to other domains (Escudero et al., 2000).

As an alternative to supervised systems,

knowledge-based systems do not require manually tagged data and have proven to be applicable to new domains (Agirre et al., 2009). An example of such a system is the Lesk algorithm (Lesk, 1986) that exploits the idea that the overlap between the definition of a word and the definitions of the words in its context can provide information about its meaning. It only requires two types of information: a set of dictionary entries with definitions (hereafter referred to as glosses) for each possible word meaning, and the context in which the word occurs. A popular variant of the algorithm is the “simplified” Lesk algorithm (Kilgarriff and Rosenzweig, 2000), which disambiguates one word at a time by comparing each of its glosses to the context in which the word is found. This variant avoids the combinatorial explosion of word sense combinations the original version suffers from when trying to disambiguate multiple words in a text.

A problem with the aforementioned method, however, is that, when a gloss is matched against the context of a word, in most cases the lexical overlap is very small. As a solution to this problem, we use a WSD-method that, instead of counting the number of words that overlap, takes embeddings as input to compute the similarity between the gloss of a sense and the context of the word. Although our method works well on its own, its simplicity allows us to explore whether other extensions to the Lesk algorithm that have proven to be successful can improve it further.

As both the Lesk algorithm and our extension rely on the definition of the words and the words that surround it, it is interesting to see whether adapting both sources of information would improve either of them. In this light, there are two possibilities: expansion or reduction. For the first option, the existing words of the context and glosses can be expanded with additional words that have similar meanings. For example, Miller

et al. (2012) use a distributional thesaurus, that is computed from a large parsed corpus to lexically expand the context and sense information. They show that, using these expanded context and glosses, improves two variants of Lesk. When reducing the amount of words in either the context or the target words’ sense, methods are required that prohibit the loss of informative words. Vasilescu et al. (2004) shows that a pre-selection of words in the context of the target word improves Simplified Lesk. In this paper we describe experiments where both methods are used in combination with our method that is based on word- and sense embeddings.

2 Related work

In the past few years, much progress has been made on learning word embeddings from unlabeled data that represent the meanings of words as contextual feature vectors. A major advantage of these word embeddings is that they exhibit certain algebraic relations and can, therefore, be used for meaningful semantic operations such as computing word similarity (Turney, 2006), and capturing lexical relationships (Mikolov et al., 2013b).

A disadvantage of word embeddings is that they assign a single embedding to each word, thus ignoring the possibility that words may have more than one meaning. This problem can be addressed by associating each word with a number of sense-specific embeddings. For this, several methods have been proposed in recent work. For example, in Reisinger and Mooney (2010) and Huang et al. (2012), a fixed number of senses is learned for each word that has multiple meanings by first clustering the contexts of each token, and subsequently relabeling each word token with the clustered sense before learning embeddings.

Although such sense embedding methods have demonstrated good performance, they use automatically induced senses. They are, therefore, not readily applicable for applications that rely on WordNet-based senses, such as machine translation and information retrieval and extraction systems (see Morato et al. (2004) for examples of such systems). Recently, features based on sense-specific embeddings learned using a combination of large corpora and a sense inventory have been shown to achieve state-of-the-art results for supervised WSD Rothe and Schütze (2015; Jauhar et al. (2015; Taghipour and Ng (2015).

Our system makes use of a combination of sense embeddings, context embeddings, and gloss embeddings. Similar approaches have been proposed by Chen et al. (2014) and Pelevina et al. (2016). The main difference to our approach is that they automatically induce sense embeddings and find the best sense by comparing them to context embeddings, while we add gloss embeddings for better performance. Inkpen and Hirst (2003) apply gloss- and context vectors to the disambiguation of near-synonyms in dictionary entries. Also Basile et al. (2014) use a distributional approach to representing definitions and the context of the target word. They create semantic vectors for glosses and contexts to compute similarity of the gloss and the context of a target word, while we also compute the similarity of a sense and its context directly using sense embeddings.

3 Lesk++

Our WSD algorithm takes sentences as input and outputs a preferred sense for each polysemous word. Given a sentence $w_1 \dots w_i$ of i words, we retrieve a set of word senses from the sense inventory for each word w . Then, for each sense s of each word w , we consider the similarity of its lexeme (the combination of a word and one of its senses (Rothe and Schütze, 2015)) with the context and the similarity of the gloss with the context.

For each potential sense s of word w , the cosine similarity is computed between its gloss vector G_s and its context vector C_w and between the context vector C_w and the lexeme vector $L_{s,w}$. The score of a given word w and sense s is thus defined as follows:

$$\text{Score}(s, w) = \cos(G_s, C_w) + \cos(L_{s,w}, C_w) \quad (1)$$

The sense with the highest score is chosen. When no gloss is found for a given sense, only the second part of the equation is used.

Prior to the disambiguation itself, we sort the words by the number of senses it has, in order that the word with the fewest senses will be considered first. The idea behind this is that words that have fewer senses are easier to disambiguate (Chen et al., 2014). The algorithm relies on the words in the context which may themselves be ambiguous. If words in the context have been disambiguated already, this information can be used for the ambiguous words that follow. We, therefore, use the

resulting sense of each word for the disambiguation of the following words starting with the “easiest” words.

Our method requires lexeme embeddings $L_{s,w}$ for each sense s . For this, we use AutoExtend (Rothe and Schütze, 2015) to create additional embeddings for senses from WordNet on the basis of word embeddings. AutoExtend is an auto-encoder that relies on the relations present in WordNet to learn embeddings for senses and lexemes. To create these embeddings, a neural network containing lexemes and sense layers is built, while the WordNet relations are used to create links between each layer. The advantage of their method is that it is flexible: it can take any set of word embeddings and any lexical database as input and produces embeddings of senses and lexemes, without requiring any extra training data.

For each word w we need a vector for the context C_w , and for each sense s of word w we need a gloss vector G_s . The context vector C_w is defined as the mean of all the content word representations in the sentence: if a word in the context has already been disambiguated, we use the corresponding sense embedding; otherwise, we use the word embedding. For each sense s , we take its gloss as provided in WordNet. In line with Banerjee and Pedersen (2002), we expand this gloss with the glosses of related meanings, excluding antonyms. Similar to the creation of the context vectors, the gloss vector G_s is created by averaging the word embeddings of all the content words in the gloss.

4 Lexical expansion and lexical selection

We use the method of Miller et al. (2012) to expand the glosses and the contexts of the target words before using our adaptation of the Lesk system.¹ For each content word we retrieve the 30 most similar terms from the distributional thesaurus and add them to the context or gloss while occurrences of the target word are removed.

For the selection of context words, we use the lexical chaining technique as applied in Vasilescu et al. (2004) that use the idea of creating lexical chains from Hirst and St-Onge (1998). Lexical chains are sequences of words that are semantically related. Similar to Vasilescu et al. (2004), we use the synonymy and hypernymy relations in

¹We use the distributional thesaurus downloaded from www.lt.informatik.tu-darmstadt.de/de/data/distributional-thesauri.

WordNet in combination with a similarity measure (Jaccard formula (Manning and Schütze, 1999)), to verify whether a context word is a member of such a lexical chain. For both the target word w and each context word c in its context, we retrieve a set of sense definitions of all the synonyms and hypernyms of w according to the WordNet hierarchy. A context word is added to the context if the similarity score for the set of w and the set of c is greater than an experimental threshold.

5 Experiments

We test our method on both Dutch and English data. We build 300-dimensional word embeddings on the Dutch Sonar corpus (Oostdijk et al., 2013) using word2vec CBOW (Mikolov et al., 2013a), and create sense- and lexeme embeddings with AutoExtend. For English, we use the embeddings from Rothe and Schütze (2015)². They lie within the same vector space as the pre-trained word embeddings by Mikolov et al. (2013a)³, trained on part of the Google News dataset, which contains about 100 billion words. This model (similar to the Dutch model) contains 300-dimensional vectors for 3 million words and phrases.

Our sense inventory for Dutch is Cornetto (Vossen et al., 2012) and for English, we use WordNet 1.7.1 (Fellbaum, 1998) as this version matches the AutoExtend embeddings. In Cornetto, 51.0% of the senses have glosses. In the Princeton WordNet, almost all of them do. The DutchSemCor corpus (Vossen et al., 2013b) is used for Dutch evaluation and, for English, we use SemCor (Fellbaum, 1998). A random subset of 5000 manually annotated sentences from each corpus was created. Additionally, we test on the Senseval-2 (SE-2) and Senseval-3 (SE-3) all-words datasets (Snyder and Palmer, 2004; Palmer et al., 2001).

We evaluate our method by comparing it with a random baseline and Simplified Lesk with expanded glosses (SE-Lesk) (Kilgarriff and Rosenzweig, 2000; Banerjee and Pedersen, 2002). We do not compare our system to the initial results of AutoExtend (Rothe and Schütze, 2015) as they tested it in a supervised setup using sense embeddings as features. However, as is customary in WSD evaluation, we do compare our system to the most frequent sense baseline, which is notoriously

²<http://www.cis.lmu.de/sascha/AutoExtend/>

³<https://code.google.com/p/word2vec/>

	DSC	SC	SE-2	SE-3		DSC	SC	SE-2	SE-3
SE-Lesk	28.1%	53.2	52.1%	50.1%	Lesk++	45.9%	55.1%	54.9%	59.3%
+LE	29.6%	56.5%	51.0%	49.3%	+LE	42.5%	47.8%	43.8%	46.2%
+LS	16.0%	40.7%	48.1%	54.3%	+LS	47.3%	67.2%	58.4%	59.4%
+LE,LS	25.2%	40.6%	46.2%	46.0%	+LE,LS	41.0%	66.9%	49.1%	43.5%

Table 1: Results for DutchSemCor (DSC), SemCor (SC), Senseval-2 (SE-2) and Senseval3 (SE-3) for Simplified Extended Lesk (SE-Lesk) and Lesk++. The following columns use lexical selection (LS), lexical extension (LE) and both extension and selection (LE,LS).

difficult to beat due to the highly skewed distribution of word senses (Agirre and Edmonds, 2007). As this baseline relies on manually annotated data, which our system aims to avoid, we consider this baseline to be semi-supervised.

Additionally, we compare our system to a state-of-the-art knowledge-based WSD system, UKB (Agirre and Soroa, 2009), that, similar to our method, does not require any manually tagged data. UKB can be used for graph-based WSD using a pre-existing knowledge base. It applies random walks, e.g. Personalized PageRank, on the Knowledge Base graph to rank the vertices according to the context. We use UKBs Personalized PageRank method word-by-word with WordNet 1.7 and eXtended WordNet for English, as this setup yielded the best results in Agirre and Soroa (2009). For Senseval-2 (SE-2) and Senseval-3 we use the WSD evaluation framework of Raganato et al. (2017), which provides evaluation datasets and output of other knowledge-based WSD systems. From those systems we report on the Extended Lesk version of Basile et al. (2014), (DSM)⁴ which is most similar to our approach.

The manually annotated part of DutchSemCor is balanced *per sense* which means that an equal number of examples for each sense is annotated. It is therefore not a reliable source for computing the most frequent sense. Alternatively, similar to Vossen et al. (2013a), we derive sense frequencies by using the automatically annotated counts in DutchSemCor⁵. The most frequent sense baseline for Dutch is, therefore, lower as compared to the English, where the most frequent sense of a word is fully based on manual annotation.

⁴We use <https://github.com/pippokill/lesk-wsd-dsm> without sense frequency for comparability

⁵In DutchSemCor senses are annotated with an SVM, trained on the manually annotated part of the corpus, see Vossen et al. (2013a) for more details.

6 Results

Table 1 shows the results of both SE-Lesk and our method (Lesk++) with lexically extended (LE) and selected (LS) context and gloss vectors. The use of word and sense embeddings yields overall better results as compared to SE-Lesk. Remarkably, lexical extension, that is very beneficial for SE-Lesk, does serious harm to our method. Selecting words in the context, on the other hand, improves our method and makes SE-Lesk perform worse.

Table 2 shows the results of the best performing combinations, SE-Lesk with lexical extension and Lesk++ with lexical selection, compared to three baselines. Our system, when used in combination with the lexical selection method, performs better than the other purely knowledge-based methods.

	DSC	SC	SE-2	SE-3
Lesk++LS	47.3%	67.2%	58.4%	59.4%
SE-Lesk,LE	29.6%	56.5%	51.0%	49.3%
UKB	38.9%	57.6%	56.0%	51.8%
DSM	-	-	51.2%	42.3%
Random	26.5%	33.6%	39.9%	34.9%
MFS	36.0%	70.9%	65.6%	66.2%

Table 2: Results for Simplified Extended Lesk (SE-Lesk) with lexical extension (LE) and Lesk++ with lexical selection (LS), UKB, DSM, a random and a most frequent sense baseline

7 Discussion

The difference in results for Dutch and English can be explained by the coverage of the datasets. The Cornetto coverage is about 60%, compared to Princeton Wordnet, with an average polysemy of 1.07 for nouns, 1.56 for verbs and 1.05 for adjectives while, for English it is 1.24 for nouns, 2.17 for verbs and 1.40 for adjectives. Also, not all Dutch senses have corresponding glosses while

most of the English ones do. As our method relies greatly on gloss vectors, this could affect its performance.

The different performance of both extensions to SE-Lesk and Lesk++ shows that both algorithms capture different types of information and therefore require a different type of input. As SE-Lesk counts on the direct overlap of words, it depends highly on a larger amount of words. Lesk++ on the other hand, already overcomes this problem and clearly benefits from more “quality” information in the contexts.

In future work we would like to try other vector types such as Melamud et al. (2016) that represents contexts that outperform the context representation of averaged word embeddings. Also, it would be nice to see whether other Knowledge-based sense embeddings, such as the ones from Camacho-Collados et al. (2016), could improve our results.

8 Conclusions

We compared several extensions to the Lesk algorithm with an adaption which uses sense, gloss and context embeddings to compute the similarity of word senses to the context in which the words occur. We try two different methods that could improve ours, one that further extends the information in both the context and the glosses by utilizing Distributional thesauri (Miller et al., 2012) and one that pre-selects context words using the WordNet hierarchy (Vasilescu et al., 2004). Although our approach is a straightforward extension to the Lesk algorithm, it achieves better performance compared to Lesk and a random baseline. When using a selection scheme before creating context vectors, its performance is better than our knowledge based baselines. The main advantage of our method is its simplicity which makes it fast and easy to apply to other languages. It furthermore only requires unlabeled text and the definitions of senses, and does not rely on any manually annotated data, which makes our system an attractive alternative for supervised WSD.

References

- Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition.
- Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 33–41.
- Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. 2009. Knowledge-based WSD on specific domains: Performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1501–1506.
- Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1591–1600.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif. Intell.*, 240:36–64.
- Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1025–1035.
- Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 172–180.
- Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.
- Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press.
- Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882.
- Diana Zaiu Inkpen and Graeme Hirst. 2003. Automatic sense disambiguation of the near-synonyms in a dictionary entry. In *Computational Linguistics and Intelligent Text Processing, 4th International Conference*, pages 258–267.

- Sujay Kumar Jauhar, Chris Dyer, and Eduard H. Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 683–693.
- Adam Kilgarriff and Joseph Rosenzweig. 2000. English senseval: report and results. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, May. European Language Resources Association (ELRA).
- Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86*, pages 24–26, New York, NY, USA. ACM.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.
- Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 746–751.
- Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of COLING 2012*, pages 1781–1796.
- Jorge Morato, Miguel Ngel Marzal, Juan Llorns, and Jos Moreiro. 2004. Wordnet applications. In *Proceeding of the Second Global Wordnet Conference*.
- Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1683–1688.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, February.
- Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman, 2013. *The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch*, pages 219–247. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24.
- Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183.
- Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110. Association for Computational Linguistics.
- Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117.
- Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1793–1803.
- Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, July.
- Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, May–June.
- Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics, Volume 32, Number 3, September 2006*.
- Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings*

of the Fourth International Conference on Language Resources and Evaluation.

- Piek Vossen, Attila Görög, Rubén Izquierdo, and Antal van den Bosch. 2012. Dutchsemcor: Targeting the ideal sense-tagged corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 584–589, may.
- Piek Vossen, Rubén Izquierdo, and Attila Görög. 2013a. Dutchsemcor: in quest of the ideal sense-tagged corpus. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing*, pages 710–718.
- Piek Vossen, Isa Maks, Roxane Segers, Hennie der van Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke, 2013b. *Cornetto: A Combinatorial Lexical Semantic Database for Dutch*, pages 165–184. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 78–83.

Semi-automatic WordNet Linking using Word Embeddings

Kevin Patel[†], Diptesh Kanojia^{†,♣,*}, Pushpak Bhattacharyya[†]

[†]Indian Institute of Technology Bombay, India

[♣]IITB-Monash Research Academy, India

^{*}Monash University, Australia

[†]{kevin.patel, diptesh, pb}@cse.iitb.ac.in

Abstract

Wordnets are rich lexico-semantic resources. Linked wordnets are extensions of wordnets, which link similar concepts in wordnets of different languages. Such resources are extremely useful in many Natural Language Processing (NLP) applications, primarily those based on knowledge-based approaches. In such approaches, these resources are considered as gold standard/oracle. Thus, it is crucial that these resources hold correct information. Thereby, they are created by human experts. However, manual maintenance of such resources is a tedious and costly affair. Thus techniques that can aid the experts are desirable. In this paper, we propose an approach to link wordnets. Given a synset of the source language, the approach returns a ranked list of potential candidate synsets in the target language from which the human expert can choose the correct one(s). Our technique is able to retrieve a winner synset in the top 10 ranked list for 60% of all synsets and 70% of noun synsets.

1 Introduction

Wordnets (Fellbaum, 1998) have been useful in different Natural Language Processing applications such as Word Sense Disambiguation (Tufiş et al., 2004; Sinha et al., 2006), Machine Translation (Knight and Luk, 1994) etc.

Linked Wordnets are extensions of wordnets. In addition to language specific information captured in constituent wordnets, linked wordnets have a notion of an interlingual index, which connects similar concepts in different languages. Such linked wordnets have found their application in machine translation (Hovy, 1998), cross-lingual information retrieval (Gonzalo et al., 1998), etc.

Given the extensive application of wordnets in different NLP applications, maintenance of wordnets involves expert involvement. Such involvement is costly both in terms of time and resources. This is further amplified in case of linked wordnets, where experts need to have knowledge of multiple languages. Thus, techniques that can help reduce the effort needed by experts are desirable.

Recently, deep learning has been extremely successful in a wide array of NLP applications. This is primarily due to the development of word embeddings, which have become a crucial component in modern NLP. They are learned in an unsupervised manner from large amounts of raw corpora. Bengio et al. (2003) were the first to propose neural word embeddings. Many word embedding models have been proposed since then (Collobert and Weston, 2008; Huang et al., 2012; Mikolov et al., 2013c; Levy and Goldberg, 2014). They have been efficiently utilized in many NLP applications: Part of Speech Tagging (Collobert and Weston, 2008), Named Entity Recognition (Collobert and Weston, 2008), Sentence Classification (Kim, 2014), Sentiment Analysis (Liu et al., 2015), Sarcasm Detection (Joshi et al., 2016)

Mikolov et al. (2013a) made a particularly interesting observation about the structure of the embedding space of different languages. They noted that there is a linear mapping between such spaces.

In this paper, we address the following question:

“Can information about the structure of embedding spaces of different languages and the relation among them be used to aid linking of corresponding wordnets?”

We demonstrate that this is true at least in the case of English and Hindi WordNets. We propose an approach to link them using word embeddings. Given a synset of the source language, the approach provides a ranked list of target synsets. This makes the overall linking task easy for human

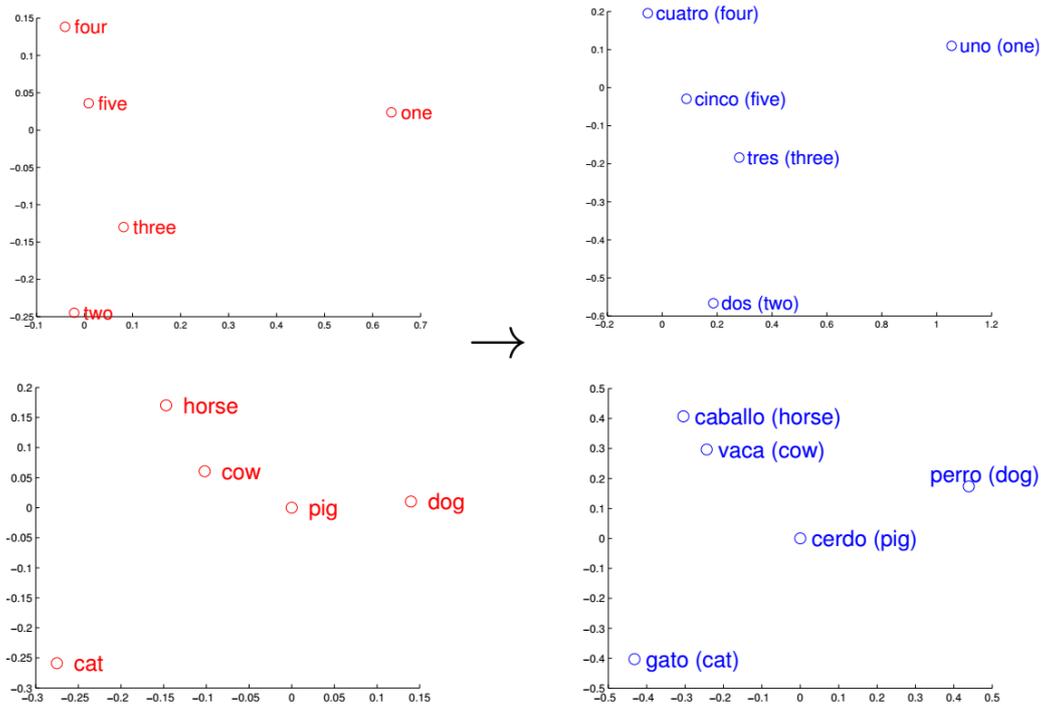


Figure 1: Word embeddings of numbers and animals in English (left) and Spanish (right) (taken from (Mikolov et al., 2013a)).

experts, as they have to choose from a relatively small set of potential candidates. Our evaluation shows that our technique is able to retrieve a winner synset in the top 10 ranked list for 60% and 70% of all synsets and noun synsets respectively.

2 Background and Related Work

Princeton WordNet or the English WordNet was the first wordnet and inspired the development of many other wordnets. EuroWordNet (Vossen and others, 1997) is a linked wordnet comprising of wordnets for European languages, *viz.*, Dutch, Italian, Spanish, German, French, Czech and Estonian. Each of these wordnets is structured in the same way as the Princeton WordNet for English (Miller et al., 1990) - synsets (sets of synonymous words) and semantic relations between them. Each wordnet separately captures a language-specific information. In addition, the wordnets are linked to an Inter-Lingual-Index, which uses Princeton WordNet as a base. This index enables one to go from concepts in one language to similar concepts in any other language. Such features make this resource helpful in cross-lingual NLP applications.

IndoWordNet (Bhattacharyya, 2010) is a linked wordnet comprising of wordnets for major In-

dian languages, *viz.*, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. These wordnets have been created using the expansion approach using Hindi WordNet as a pivot, which is partially linked to English WordNet. Previously, Joshi et al. (2012a) come up with a heuristic based measure where they use bilingual dictionaries to link two wordnets. They combine scores using various heuristics and generate a list of potential candidates for linked synsets.

Singh et al. (2016) discuss a method to improve the current status of Hindi-English linkage and present a generic methodology *i.e.*, manually creating bilingual mappings for concepts which are unavailable in either of the languages or not present as a synset in the target wordnet. Their method is beneficial for culture-specific synsets, or for non-existing concepts; but, it is cost and time inefficient, and requires a lot of manual effort on the part of a lexicographer.

Our approach is mainly geared towards reducing effort on the part of the lexicographers.

3 Problem Statement

Given wordnets of two different languages E and F with sets of synsets $\{s_E^1, s_E^2, \dots, s_E^m\}$ and $\{s_F^1, s_F^2, \dots, s_F^n\}$ respectively, find mappings of the form $\langle s_E^i, s_F^j \rangle$ which are semantically correct.

4 Approach

We adapted the technique of translating words in Mikolov et al. (2013a) to translate synsets (see fig 1). In order to do so, however, we need "synset embeddings". We computed the same by assigning to a synset-id, the average of the "word embeddings" of its synset-members. To the best of our knowledge, this is a first attempt at solving this problem using word embeddings. The following is a detailed description of the technique.

Let E and F be two languages. Let $|E|$ and $|F|$ be the number of synsets in wordnets of E and F respectively. Let s_E^i and s_F^j be the i^{th} and j^{th} synsets of E and F respectively, with $s_E^i = \{e_\alpha^1, e_\alpha^2, \dots, e_\alpha^{m_i}\}$ and $s_F^j = \{f_\beta^1, f_\beta^2, \dots, f_\beta^{n_j}\}$, where e_α^p and f_β^q are words in vocabulary of E and F respectively for $1 \leq p \leq m_i$ and $1 \leq q \leq n_j$, and $1 \leq i \leq |E|$ and $1 \leq j \leq |F|$.

Let $v_{e_\alpha^p}$ be the word embedding corresponding to e_α^p . Then we estimate $v_{s_E^i}$, the embedding for synset s_E^i , as

$$v_{s_E^i} = \frac{1}{m_i} \sum_{p=0}^{m_i} v_{e_\alpha^p} \quad (1)$$

Similarly,

$$v_{s_F^j} = \frac{1}{n_j} \sum_{q=0}^{n_j} v_{f_\beta^q} \quad (2)$$

Given links of the form $\langle s_E^i, s_F^j \rangle$, we learn W such that the error Err

$$Err = \|W.v_{s_E^i} - v_{s_F^j}\|^2 \quad (3)$$

is minimized.

Now, to find a mapping for a new synset s_E^k , one needs to

1. Calculate $v' = W.v_{s_E^k}$
2. Find $v_{s_F^l}$ such that $v_{s_F^l} \cdot v'$ is maximized
3. Create link $\langle s_E^k, s_F^l \rangle$

Our hypothesis is that for a given synset-id, the noise added to its representative embedding by a highly polysemous synset-member will be canceled out, while the actual information content pertaining to that synset-id will be enhanced, due to contribution from other, relatively less polysemous, synset members.

5 Experiments

Datasets

We applied our technique to link Hindi and English Wordnets. We obtained a dataset of mappings between English and Hindi wordnets from the developers of IndoWordNet. These mappings are of the form $\langle hindi_synset_id, english_synset_id, link_type \rangle$, where $link_type \in \{\text{DIRECT}, \text{HYPERNYMY}, \text{etc.}\}$. For this experiment, we focused solely on DIRECT links. There are a total of 6,883 such mappings, the distribution among classes of which is mentioned in table 1

Class	Count
Noun	4757
Adjective	1283
Verb	680
Adverb	143

Table 1: Distribution of available links among various classes

For the English language, we used the pre-trained word embeddings published by Google that were trained on part of Google News Dataset (about 100 billion tokens). These embeddings are of dimension 300, and are created using CBOW model with negative sampling. For the Hindi language, we trained word embeddings on BOJAR HindMonoCorp dataset (Bojar et al., 2014). Mikolov et al. (2013b) suggests that the input embeddings' dimension should be at least 2.5 to 4 times that of the output dimension. But we also wanted to check what happens when they are equal. Therefore, we trained two sets of embeddings, one of dimension 300, and the other of dimension 1200.

Evaluation Metric

We use the $accuracy@n$ measure, i.e the prediction is said to be correct if one out of the top n results returned is correct. This is because $accuracy@1$ is an underestimate of the system's per-

formance, as higher-ranking synonym translations will be counted as mistakes.

	Predicted Label	Accuracy @1	Accuracy @3	Accuracy @5
True label	Prediction1			
	Prediction2			
	Prediction3			
	Prediction4			
	Prediction5			

Figure 2: Accuracy@n: The green colored cells indicate the predictions considered for exact match for a given accuracy@n

6 Results and Discussion

Table 2 shows the overall accuracy@n of the system, for different values of n. We also performed a per word-class evaluation, along with different settings for the embedding dimensions. Table 3 and Table 4 shows the accuracy for different word classes ¹.

Acc@1	Acc@3	Acc@5	Acc@8	Acc@10
0.29	0.45	0.52	0.58	0.60

Table 2: Results for the overall setting: Dimension of English embeddings=300, Dimensions of Hindi embeddings=300

Word Class	Acc@1	Acc@3	Acc@5	Acc@8	Acc@10
Noun	0.35	0.53	0.60	0.65	0.67
Adjective	0.26	0.44	0.50	0.57	0.60
Verb	0.15	0.25	0.29	0.33	0.37
Adverb	0.28	0.51	0.59	0.70	0.73

Table 3: Results for the setting: Dimension of English embeddings=300, Dimensions of Hindi embeddings=300

We observe that except for verbs, the approach performs decently. Here we mention some of the reasons for poor performance, as well as possible methods to address them.

- The approach to create synset embeddings is inadequate. The current averaging approach only takes the synset members into account, while ignoring gloss and examples, which could provide additional information. A potential candidate approach for creating synset embeddings should properly utilize the set of

¹All values reported are the average values obtained from 3-fold cross validation.

Word Class	Acc@1	Acc@3	Acc@5	Acc@8	Acc@10
Noun	0.35	0.52	0.58	0.63	0.66
Adjective	0.12	0.20	0.24	0.30	0.32
Verb	0.17	0.27	0.32	0.35	0.39
Adverb	0.38	0.52	0.65	0.76	0.80

Table 4: Results for the setting: Dimension of English embeddings=300, Dimensions of Hindi embeddings=1200

French synonyms, gloss, example sentences, and synset relations.

- Synset members are often phrases instead of words. Creating phrase embeddings is a different problem altogether.
- Currently, we utilized a word embedding model which gives only one embedding per word. That is one of the reasons for ambiguity. A model which provides one embedding per sense of a word will be a more appropriate.
- The linear transformation approach is incorrect. While (Mikolov et al., 2013a) shows the linear relation between English and Spanish languages, this may not be true for all pairs of languages.
- Perhaps, something is fundamentally missing in word embeddings. Probably presence of only co-occurrence information and lack of other information such as word order, argument frames(for verbs), *etc.* leads to this poor performance.

However, we were unable to find an explanation for the degradation of results of adjectives when using 1200 dimensions for Hindi word embeddings.

7 Conclusion and Future Work

In this paper, we described an approach to link wordnets. It entails creating synset embeddings using the word embeddings of the synset members, and learning a function to map the embedding of a synset from the source language to an embedding in the space of target language, and returning the nearest neighbors as potential candidates for linking. Our evaluation shows that our technique is able to retrieve a winner synset in the top 10 ranked list for 60% and 70% of all synsets and noun synsets, respectively. Although, it did

not achieve significantly good results for other classes, especially verbs. We discussed the possible reasons for poor performance and suggested mechanisms to address the same.

In future, we plan to continue this work, and explore each of the above possible reasons for poor performance, in order to mitigate them. We will also evaluate it in an active learning setting. Eventually, we aim to integrate our work with tools such as the ones created by Joshi et al. (2012b), *etc.* so that our work can be used by lexicographers and researchers alike.

References

- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.
- Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.
- Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindMonoCorp 0.5.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. 1998. Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.
- Eduard Hovy. 1998. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 535–542.
- Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Salil Joshi, Arindam Chatterjee, Arun Karthikeyan Karra, and Pushpak Bhattacharyya. 2012a. Eating your own cooking: automatically linking wordnet synsets of two languages.
- Salil Joshi, Arindam Chatterjee, Karthikeyan Arun Karra, and Pushpak Bhattacharyya. 2012b. Eating your own cooking: Automatically linking wordnet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, pages 239–246. The COLING 2012 Organizing Committee.
- Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, Austin, Texas, November. Association for Computational Linguistics.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.
- Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.
- Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308.
- Pengfei Liu, Shafiq R Joty, and Helen M Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *EMNLP*, pages 1433–1443.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Meghna Singh, Rajita Shukla, Jaya Jha, Laxmi Kashyap, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Mapping it differently: A solution to the linking challenges. In *Eighth Global Wordnet Conference*. GWC 2016.
- Manish Sinha, Mahesh Reddy, and Pushpak Bhattacharyya. 2006. An approach towards construction and application of multilingual indo-wordnet. In *3rd*

Global Wordnet Conference (GWC 06), Jeju Island, Korea.

Dan Tufiș, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1312. Association for Computational Linguistics.

Piek Vossen et al. 1997. Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

Multilingual Wordnet sense Ranking using nearest context

E Umamaheswari Vasanthakumar and Francis Bond

School of Humanities

Nanyang Technological University

Singapore

umavasanth28@gmail.com, bond@ieee.org

Abstract

In this paper, we combine methods to estimate sense rankings from raw text with recent work on word embeddings to provide sense ranking estimates for the entries in the Open Multilingual WordNet (OMW). The existing Word2Vec pre-trained models from Polygot2 are only built for single word entries, we, therefore, re-train them with multiword expressions from the wordnets, so that multiword expressions can also be ranked. Thus this trained model gives embeddings for both single words and multiwords. The resulting lexicon gives a WSD baseline for five languages. The results are evaluated for Semcor sense corpora for 5 languages using Word2Vec and Glove models. The Glove model achieves an average accuracy of 0.47 and Word2Vec achieves 0.31 for languages such as English, Italian, Indonesian, Chinese and Japanese. The experimentation on OMW sense ranking proves that the rank correlation is generally similar to the human ranking. Hence distributional semantics can aid in Wordnet Sense Ranking.

1 Introduction

Most of the existing Word-net sense rankings (Navigli, 2009) use document level statistics to find the prominent sense of the given word. McCarthy and Carroll (2003) showed that predominate senses could be learned from a sufficiently large corpus, and this work has since been extended by various researchers. Words that appear nearest to the given word convey the context/meaning of a word (Lim, 2014;

Liu et al., 2015; Pocostales, 2016; Rong, 2014; Long et al., 2016), and this can be used to estimate the most frequently used senses. This proposed work uses nearest context words to predict the senses and computes the frequency of occurrence of these senses within the corpus. Since most of the existing WSD systems utilize the Most Frequent Sense (MFS) as a baseline, it is important to rank the Wordnet senses in a meaningful way.

Two well-known software packages used to train word embeddings, are Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove model (Pennington et al., 2014). Polyglot (Al-Rfou et al., 2013) is a natural language pipeline that supports many NLP based tasks such as tokenization, Language detection, Named Entity Recognition, Part of Speech Tagging, Sentiment Analysis, Word Embeddings, Morphological analysis and Transliteration for many languages. This work utilizes their Word embeddings. Existing polyglot word embeddings (Al-Rfou et al., 2013) support 137 languages. We have planned to use the word embeddings for the 35 hand-built wordnets currently in OMW (Ruci, 2008; Elkateb et al., 2006; Borin et al., 2013; Pedersen et al., 2009; Simov and Osenova, 2010; Gonzalez-Agirre et al., 2012; Pociello et al., 2011; Wang and Bond, 2013; Huang et al., 2010; Pedersen et al., 2009; Fellbaum, 1998; Stamou et al., 2004; Sagot and Fišer, 2008; Ordan and Wintner, 2007; Mohamed Noor et al., 2011; Isahara et al., 2008; Montazery and Faili, 2010; Lindén and Carlson., 2010; Garabík and Pileckytė, 2013; Vossen and Postma, 2014; Piasecki et al., 2009; de Paiva et al., 2012; Tufiş et al., 2008; Darja et al., 2012; Borin et al., 2013; Thoongsup et al., 2009; Pianta et al., 2002; Oliver et al., 2015; Raffaelli et al., 2008; Toral et al., 2010).

We use corpus based frequencies for five of these languages (English, Chinese, Japanese, Italian and Indonesian) from the NTU Multilingual Corpus (NTU-MC: Tan and Bond, 2013) and use them to evaluate the learned sense rankings. Our major contribution is training and testing on large numbers of multiword expressions, which are often neglected in the word embedding literature. We identify the multi-word expressions found in the hand-built lexicons and train our own model for them using Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove (Pennington et al., 2014).

This paper is structured as follows. Section 2 discusses the related work in Word embedding and its application in WordNet Synset Ranking. Section 3 describes the data, methods, and Section 4 discusses the evaluation of results obtained from word embedding and its effect in WordNet Sense Ranking. Finally, Section 5 concludes with the findings and future plans to improve the results.

2 Related Work

Word embedding techniques have been popular in recent years in Word Sense Disambiguation (*WSD*) research. Similar to this proposed work, (Bhingardive et al., 2015b) computes word embeddings with the help of pretrained Word2Vec(Mikolov et al., 2013; Rong, 2014) and matches with the sense embeddings obtained from the Wordnet features. They have attempted Wordnet sense ranking for Hindi and English. Since the Word2Vec (Mikolov et al., 2013; Rong, 2014) model is based on the words frequency of occurrence in the corpus, finding the nearest context words that occur infrequently in the corpus is difficult.

Panchenko (2016) compares sense embeddings of AdaGram (Bartunov et al., 2015) with BabelNet (Navigli and Ponzetto, 2010) synsets and proved that sense embeddings can be retrieved by automatically learned sense vectors. Sense embeddings for a given target word are identified by finding the similarity between the AdaGram Word embeddings list with the BabelNet Synsets words list. Rothe and Schütze (2015) proposed an approach that takes word embeddings as input and produces synset, lexeme embeddings without retraining

them. They used WordNet lexical resource to improve word embeddings.

Arora et al. (2016) showed that word vectors can capture polysemy and word vectors can be thought of as linear superpositions of each sense vector. They have attempted discourse analysis to find the cluster of sense vectors.

Although the basic idea of word embeddings is not tied to any one languages, the preprocessing steps are language specific. Kang et al. (2016) presented a cross-lingual word embedding for English and Chinese Word Sense Disambiguation (WSD). They have experimented with the performance of WSD using different word embeddings such as Word2Vec (Mikolov et al., 2013; Rong, 2014) and Glove model. Bhingardive et al. (2015a) compared word embeddings obtained from the Word2Vec (Mikolov et al., 2013; Rong, 2014) model and the sense embedding obtained from the WordNet for English and Hindi languages and restricted to Nouns. They used various WordNet features similar to this proposed work to find the predominant sense. Their approach outperforms SemCor baseline for words with the frequency below five.

In this research context words are identified with the help of Polyglot(Al-Rfou et al., 2013) word embeddings.

3 Methodology

In this work, we use word embeddings to find the nearest context of a given word and compare it with the senses obtained from the OMW to find the most frequently used senses. Our aim is to rank the senses obtained from the OMW with the help of the context words and their frequency of occurrence. Initially, we use the pretrained polyglot word embedding model (Al-Rfou et al., 2013) to retrieve the nearest context words and found multiwords are unidentified. Hence in this work, we have trained our own model similar to polyglot for both single and multi-words. Our aim is to train this model for all 35 languages supported by OMW, for this paper we present only the results for the five languages for which we have evaluation data: English, Chinese, Japanese, Italian and Indonesian.

3.1 Corpus Cleaning and preprocessing

We exploit the openly available Polyglot wiki dump corpus (Al-Rfou et al., 2013) for English, Chinese, Japanese, Italian and Indonesian. We chose this as it contains various domains and languages. Before training our own model, the corpus texts are preprocessed by removing symbols, numbers and shortest text. Stop words have been removed with the help of the NLTK toolkit (Bird et al., 2009). However, NLTK does not support stop-words for all languages. Hence we have included stop words of Chinese, Japanese, Indonesian, Italian from publicly available online utilities to NLTK toolkit. For English, Indonesian and Italian we have lemmatized each word of the cleaned text to find their base form. Chinese does not inflect, and Japanese inflections are normally split off by the tokenizer. Hence we have used Mecab to tokenize/lemmatize Japanese texts. After preprocessing the text, each sentence of the corpus is tokenized into single and multiple terms. In order to identify the multiwords from the corpus, we have used the existing Wordnet MWE lexicon ($MWEs$). The terms of each sentence are matched with the existing wordnet $MWEs$ lexicon and if an MWE is found it is rewritten to a single token, with spaces replaced by an underbar “_” symbol. The preprocessed MWE tagged texts are given as input to train our own model. So, for example, a sentence like *I looked five words up* will be preprocessed to `I look_up word`.

3.2 Training Model

Word embeddings for the above five languages have been trained using the Polyglot2 (Al-Rfou et al., 2013) package and *Global Vectors for Word Representation Glove Model* (Pennington et al., 2014). Polyglot2 is a software package that enables building your own language models. It learns the distributed representations of words/word embeddings for the given corpus. GLOVE is another unsupervised learning algorithm used for obtaining vector representations of words. Training is performed by considering global word-word co-occurrence statistics from a corpus and results with the linear substructures of the word vector space. We can build our own word em-

beddings with the help of Polyglot2 and Glove models.

3.3 Predominant Sense Scoring

To find the predominant senses for the given word w , the senses obtained from the OMW are represented as $S_w = S_1, S_2, \dots, S_n$. The neighbouring context obtained from Polyglot2 or Glove is represented as $S_w^N(w, d)$ where N represents the number of neighbouring contexts from word embedding obtained for the senses S_w that can vary from 1 to N , and d represent the distance score between the S_w and S_w^N . $P_s(S_w)$ represents the predominant score of S_w based on the WordNet synset similarity.

$$P_s(S_w) = \log(\text{sum}(S_w^N(w, d)) + M_W^T/T^N W_e) + [H_s(M)/T^N W_e] \quad (1)$$

M_W^T - represents the number of matching terms between the *OMW* synset definitions and example sentences with respect to polyglot word embeddings.

$T^N W_e$ - represents the number of word embeddings obtained from Polyglot2.

After computing the predominant score $P_s(S_w)$ for each word-net entries the semantic similarity between the word embedding with the OMW ontology hierarchy is measured. $H_s(M)$ represents the number of concepts such as Hypernyms and Hyponyms of WordNet Ontology that match with the number of terms obtained in the polyglot word embeddings. The intuition behind is that the words in the word embedding will have similar words that can appear in WordNet hierarchy. For example, the word *party* may refer to a person, organization or an occasion. If it refers to a *person*, the hypernyms are *person* and the hyponyms are *assignee*, *assignor*, *contractor*, *intervenor*. Similarly for organization the hypernyms is *set* and hyponyms are *fatigue_party*, *landing_party*, *party_to_the_action*, *rescue_party*, *each_party*, *stretcher_party*, *war_party* and for considering *occasion* as sense the hypernyms are *affair* and hyponyms are *bash*, *birthday party*, *bunfight*, *ceilidh*, *cocktail_party*, *dance*, *fete*, *house_party*, *jolly*, *tea_party*, *whist_drive*.

When we give *Person* as Input to Polyglot2 (Al-Rfou et al., 2013), we will get the following word embeddings. *person-0.575121*, *contractor-0.628679*, *team-0.619203*, *division-0.682174*, *unit-0.700489*, *government-0.62491*, *strategy-0.725378*, *event-0.692839*, *camp-0.689145* *program-0.688767*. The terms such as *person* and *contractor* matched with the Wordnet hypernyms and hyponyms. Thus *person* sense is the most predominantly used when compared to *organization* and *event* senses since it shares the semantics with WordNet hierarchy. Similarly, we can match with other features of WordNet senses to infer which sense is important.

4 Results and Evaluation

In this section, the word embedding models such as (Glove: Al-Rfou et al., 2013) and (Word2Vec: Pennington et al., 2014) have been evaluated on two different tasks such as word-sense ranking of Wordnet and query expansion for clinical texts, then we present some examples of word embeddings for intuitive comprehension. The word sense ranking and trained word embeddings have been tested for 5 languages English, Chinese, Japanese, Indonesian and Italian languages of Semcor dataset for the words with more than one sense. The Polyglot2 word embedding model have been trained with the Context Window Size as 14, Initial learning rate as 0.025, Hidden Layer size as 32 and minimum word count as 2 (Al-Rfou et al., 2013). Glove word embedding model has been trained with the minimum word count as 2, Vector size as 100, Maximum Iteration as 100 and Context Window size as 14 (Al-Rfou et al., 2013).

We use two metrics to measure the efficiency of the baseline and the proposed word embedding model.

- Accuracy - The fraction of relevant word embeddings among the top 10 word embeddings are measured based on the human-relevant judgment.
- Rank Biased Overlap (RBO) - The rank correlation metrics that measures similarity and dissimilarity between two ranked list.

4.1 Baseline

We have taken two baseline approaches. One based on the corpus frequency based approach and the other based on the Topic model distribution score (LexSemTm). Corpus frequency-based approach ranks the synset based on the frequency of occurrence of the lemma across the corpus whereas the LexSemTm used an unsupervised sense distribution learning method (LexSemTm) (Bennett et al., 2016), that utilizes *HDP-WSI* based sense learning (Lau et al., 2014). In Bennett et al. (2016), the sense distribution of words for each sense is obtained by estimating the maximum likelihood of terms with the topics.

Both the baseline approaches used SemCor Dataset. Here the SemCor Dataset is separated into groups of lemmas with frequency 1-3 (*Group I*), 4-8 (*Group II*), 9-20 (*Group III*) and greater than 21 (*Group IV*) as described by Bennett et al. (2016). In each group, the sense distribution for each lemma is obtained from LexSemTm and the senses are ranked in descending order based on the sense distribution score and similarly for corpus frequency based method the senses are ranked based on the frequency of lemma. Then these results are compared with the proposed work.

4.2 Analysis on Word embedding

Evaluation was carried out on English, Japanese, Chinese, Indonesian and Italian word embedding using Polyglot2 (Word2Vec) and Glove. We found that the *Glove* model gave a better result when compared to the *Polyglot2(Word2Vec)* model. However, existing Word2Vec model Polyglot2¹ can capture the single terms well and to a very lesser degree the Multi-words are handled. In order to test this across domains, we have taken 5,611 unique terms from a clinical corpus and found that existing pre-trained model handles 1,500 terms semantically correct and the remaining 4,111 terms are not handled. The reason is pre-trained polyglot2 Word2Vec model is trained on wiki corpus and unable to scale up to the specific domain. Moreover, it is not trained for Multi-words. Some samples of semantic-based word embedding obtained

¹<https://sites.google.com/site/rmyeid/projects/polyglot>

from the existing model in each language (Polyglot2) are listed below:

List of semantic-based word embedding obtained in each language for *Location* as query term are listed below:

- *Indonesian*

- *lokasi(location)*
:Peta, persimpangan, pelabuhan, fondasi, celah, ruangan, wilayah, potensi, batas, otoritas-(**Map, intersection, harbor, foundation, gap, room, territory, potential, limit, authority**)

- *Italian:*

- *luogo(location)* - Teatro, motivo, periodo, servizio, passato, punto, campo, caso, segno, paese- (**Theater, pattern, period, service, past, point, field, case, sign, country**)

- *English:*

- *Location-site, map, structure, area, direction, building, locality, settlement, line, Bridge*

- *Japanese:*

- *ロケーション (Location)*
: クルージング, デモンストレーション, 個室, バナー, ガレージ, 買い物, バルコニ, ウォーキング, ナビゲーション -(**Cruising, demonstration, private room, banner, garage, shopping, balcony, walking, navigation**)

- *Chinese:*

- *位置 (Location)*
: 方向, 形式, 功能, 部分, 大小, 排列, 材料, 以上, 原本, 描述- (**Direction, Form, Feature, Section, Size, Arrangement, Material, Above, Original, Description**)

Since this proposed work has been trained for both single and multi-word expressions, we have specifically analyzed the embeddings for multi-words and the resultant samples are shown below.

Sample List of Multi-words and Nearest Context Word:

- *Query–English:*

deficit_hyperactivity_disorder:

- *attention, memory, deficit_hyperactivity_disorder, adhd, rigidly, proliferative, splinted, treat_attention, allergic_rhinitis, special*

- *Query–Japanese:*

プリンス _ オヴ _ ウェールズ (Prince of Wales):

- *トレハラーゼ, ろかく, レゼルヴ, フリーア, グローヴス, レインボーカップファイナル, mishnaic, traininfomation, カタリココ*

- *(Trehalase, fighting, reserve, free, Groves, Rainbow Cup Final, mishnaic, traininfomation, Catalina Coco)*

- *Query–Chinese:*

足球 _ 运动员 (soccer player):

- *大 _ 祭台, 阅览, 鑑, 諫, 分内事, 大捷, 新交, 續, 井底*
- *(Large altar, learning clang, remonstrance, sub-ministry, victory, new cross, play, bottom*

- *Query–Indonesian:*

erosi_pantai(erosion):

- *: Mikrokimerisme, gerunggang, membuat_bangkrut, mikkeli, lille, superintendent, thur, cibinuang, operasi_boolean*
- *(Microcimerism, rider, bankruptcy, mikkeli, lille, superintendent, thur, cibinuang, boolean operation)*

- *Query–Italian:*

seconda_guerra_mondiale (Second World War):

- *tisiddu, smetlivyi, pelligra, mortificava, skavronskij, tureaud, preprocessing, telemolise, quetzalctl*
- *(Mixed with other language text)*

Results of semantic based word embedding obtained for each language of Glove are listed below:

- *Seconda guerra mondiale*(Second World War)(Italian):
 - *prima guerra mondiale, scoppio, guerra, conflitto, dopoguerra, militare, bellico, militari, guerra mondiale, sovietica* (WWI, outbreak, war, conflict, war, military, war, Word war, military)
- *jus lemon*(lemon juice)(Indonesian):
 - *Memberikan tenaga, Mengasamkan, operated, menguapkan, boya, memfermentasi, efektif, recoil, mwh, meluapkan.*(provide power, acidity, ooperated, Evaporate, boya, ferment, effective, recoil, mwh, vent)
- Chinese: 参考 资料 (Reference Information):
 - 注释, 脚注, 参考, 迈, 资料 来源, 内部 网络, 注解, 服务 设施, 参见, 出处 (Reference information, Annotations, Footnote, reference, Side, Information source, Internal network, annotation, Service Facilities, See also, Source)
- English: *Treadmill test*:
 - *Stress test, exercise, physiology, suggestion, participate, vigorous, walking, prescription, intensity*

English, Chinese and Italian word embeddings gave better results; whereas for Indonesian documents, the results are often mixed with other language texts, even though we are able to get meaningful word embeddings. We also found that the Japanese text corpus is tagged with minimal multi-word expressions and noisy. The reason is Japanese text has different writing styles that degrade the accuracy of MWE tagging because the MWE lexicon basically includes the standard scripts. Hence we need to fine tune the MWE tagging

Accuracy(Word2Vec)	Accuracy(Glove)
0.35	0.67

Table 1: Accuracy of Word embedding score for medical text(English)

by properly filtering the character-level, word level non-standard noisy text.

The overall accuracy of the Glove model is 0.47 and Word2Vec is 0.31. Since existing polyglot model (Al-Rfou et al., 2013) handles single terms well and the trained glove model (Pennington et al., 2014) handle most of the terms meaningfully, we have planned to merge both the models to handle single and multi-terms word embeddings.

4.3 Scalability

In order to check, the scalability of these models in different domains, We have tested with Singapore Clinical Practical Guidelines documents of Dental, Medical, Nursing, and Pharmacy of 72 documents, available from *Ministry of Health, Singapore* (2016).² There are 124.2 MB in all. The results are shown in Table 1. Again the accuracy of Glove model³ is better when compared to the Word2Vec Polyglot learned model because Glove model computes co-occurrence statistics across the corpus whereas Word2Vec computes co-occurrence statistics within the context window size. The word embedding results also depend on the context window size and minimum frequency count. If we increase both the context window size and minimum frequency count to a certain extent, we can achieve semantically relevant word embeddings. However, the recall will be low.

In order to find the optimum value to maintain precision and recall, we need to run the test with different values for few test samples. The quality and size of the corpus may also impact the results. Since clinical text contains only domain-specific terms which are unambiguous, we are able to achieve meaningful results. Whereas We found difficulty in Wikipedia dump corpus(5 languages) because it contains a lot of noisy

²They are online at https://www.moh.gov.sg/content/moh_web/healthprofessionalsportal/doctors/guidelines/cpg_medical.html.

³<https://nlp.stanford.edu/projects/glove/>

data. Our purpose of this work is to check, how far this distributional semantics can help in Word Sense Ranking and Clinical Information Retrieval.

Another validation on PubMed corpus have also been taken to check the scalability of this work. BioASQ⁴ releases Word2Vec model for PubMed Abstracts of size 3.5GB (uncompressed). Their PubMed word2vec corpus consists of 10,876,004 English abstracts of biomedical articles that are publically available. We have taken a sample of PubMed corpus with 1.3 GB of data for training with our model and achieved average precision for multiword expressions as 0.55 and for single terms 0.72.

4.4 Quality of Ranking

To evaluate the quality of rankings produced by this method, we have compared the human/authors judgment rank (*Approach 1*) *A1* with three approaches such as Word embedding (*Approach 2*) *A2*, Corpus frequency ranking (*Approach 3*) *A3* and LexSemTm approach (*Approach 4*) *A4*. There are basically two well-defined algorithms such as *Spearman's* and *Kendall's tau* (Kumar and Vassilvitskii, 2010) rank correlation have been used to find the statistical difference in ranking. DCG (*Discounted Cumulative Gain*) (Harman, 2011) measures both relevance and ranking, whereas rank correlation helps to find statistically significant difference in order. *Webber et al (2010)* (Webber et al., 2010) proposed a method to compare ranking quality of two methods and addressed the top-relatedness issue. Since this proposed work needs to consider the concordance and discordance of ranked results based on position, We have used this measure to find the correlation between the two ranked lists. The correlation score is measured with *Approach 1 to Approach 2*, *Approach 3* and *Approach 4* for the Semcor dataset. The statistics of test data is shown in Table 5. For example, when we give "gleam" as query, the resulted ranking of *A1*, *A2*, *A3* are shown in Table 4, Table 2, Table 3, respectively. The rank overlapping between Approach 1 to Approach 2 and Approach 3

⁴<http://bioasq.lip6.fr/tools/BioASQword2vec/>

Synsets (gleam)

be shiny, as if wet
 shine brightly, like a star or a light
 appear briefly
 an appearance of reflected light
 a flash of light (especially reflected light)

Table 2: Ranking result of Approach 2 (Proposed)

Synsets (gleam)

a flash of light (especially reflected light)
be shiny, as if wet
 appear briefly
shine brightly, like a star or a light
an appearance of reflected light

Table 3: Ranking result of Approach 3 (Baseline - Corpus Frequency)

are calculated. Here in this example, the baseline (Corpus frequency) ranking (Approach 3) is dissimilar in all positions except the third position, whereas with human judgment (Approach 1) only the 3rd synset is moved to the last position and the remaining ranking is similar to the proposed approach (Approach 2). Hence the Rank correlation for Approach 3 to Approach 1 is 0.52 and Approach 2 to Approach 1 is 0.88. Thus the rank quality depends on how much it is similar to the human judgment.

The results are shown in table 6. Table 7 shows the comparison of the Rank overlapping value of *A1-A2*, *A1-A3* and *A1-A4*. We found that the average correlation between *A1* to *A2* is greater than *A1* to *A3* and *A1* to *A4*. This result provides an additional validation of our model as it demonstrates that the sense ranking can capture the sense preferred by a human. Hence the word embedding score definitely aid in wordnet sense ranking. When we analyze the rare sense words with frequency 1-3 and 4-8, the word embedding and Wordnet feature influence the results by providing most relevant result on the first hit. We have

Synsets (gleam)

be shiny, as if wet
 shine brightly, like a star or a light
 an appearance of reflected light
 a flash of light (especially reflected light)
appear briefly

Table 4: Ranking result of Approach 1 (Human)

Languages	Lemma Count (MWs)	Lemma Count (Single words)
English	2,361	8,187
Chinese	2,067	12,341
Japanese	473	5,289
Italian	262	9,606
Indonesian	1,134	5,178

Table 5: Statistics of Test data

Languages	A1 to A3	A1 to A2
English	0.55	0.75
Chinese	0.62	0.68
Japanese	0.64	0.69
Italian	0.61	0.67
Indonesian	0.44	0.56

Table 6: Average Rank correlation analysis between *A1 to A3* and *A1 to A2*

Groups	Freq	Lemma Count	A1 to A2	A1 to A3	A1 to A4
I	1-3	1896	0.73	0.50	0.57
II	4-8	567	0.82	0.49	0.48
III	9-20	327	0.77	0.46	0.47
IV	>20	124	0.87	0.49	0.48

Table 7: Average Rank correlation analysis

Language	Lemma	First Hit Results
English	contact	a channel for communication between groups
English	intrusion	any entry into an area not previously occupied
English	celebration	a joyful occasion for special festivities to mark some happy event
English	no more	referring to the degree to which a certain quality is present
English	write up	a short account of the news
Japanese	名人 (expert)	a person with special knowledge or ability who performs skillfully
Japanese	召集 (convene)	a group gathered in response to a summons
Japanese	ビル (building)	a structure that has a roof and walls and stands more or less permanently in one place
Chinese	适应 (adopt)	adapt or conform oneself to new or different conditions
Chinese	加入 (join)	a process of increasing by addition (as to a collection or group)
Chinese	修复 (repair)	restore by replacing a part or putting together what is torn or broken
Italian	detenzione(custody)	the state of being imprisoned
Italian	piuma(feather)	the light horny waterproof structure forming the external covering of birds
Italian	esaminare(examine)	look at carefully; study mentally
Indonesian	kehidupan(life)	the period between birth and the present time
Indonesian	barang(goods)	goods carried by a large vehicle
Indonesian	hanya(alone)	without any others being included or involved

Table 8: First Hit Analysis Results

observed that the first hit obtained from each synset ranking found most appropriate when compared to LexSemTm (A4) and OMW Corpus frequency ranking (A3). A sample list of terms and the results of the first hit have been shown in table 8.

5 Conclusion

OMW has over 150 languages with word-nets built automatically, ranging from major languages like German or Korean for which there are no free word-nets, to smaller languages such as Volapuk. For all languages for which Polyglot has data (which is most of them) we will learn rankings and incorporate them into

OMW, so that the lexicon is maximally useful for speakers of as many languages as possible. In future, we planned to extend this work to identifying missing senses by comparing the trained model over the sense-annotated corpus with the existing pre-trained models like polyglot. Since the Glove model is based on co-occurrence context, it gave better results even for a tiny corpus, hence we have planned to extend our model to sentence embedding using Glove model for finding nearest context sentences for a given synset example sentence to further improve our wordnet ranking.

Acknowledgments

This research was supported by the MOE Tier 1 grant *Semi-Automatic Implementation of Clinical Practice Guidelines in Singapore Hospitals* (RG25/13).

References

- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192. Association for Computational Linguistics, Sofia, Bulgaria. URL <http://www.aclweb.org/anthology/W13-3520>.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.
- Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skipgram. *arXiv preprint arXiv:1502.07257*, pages 47–54.
- Andrew Bennett, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. Lexsemntm: a semantic dataset based on all-words unsupervised sense distribution learning. In *ACL (1)*. The Association for Computer Linguistics.
- Sudha Bhingardive, Dharendra Singh, Rudra Murthy, Hanumant Redkar, and Pushpak Bhattacharyya. 2015a. Unsupervised most frequent sense detection using word embeddings. In *DENVER*. Citeseer.
- Sudha Bhingardive, Dharendra Singh, Rudra-murthy V, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya. 2015b. Unsupervised most frequent sense detection using word embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1238–1243.
- Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O’Reilly. (www.nltk.org/book).
- Lars Borin, Markus Forsberg, and Lennart Lönngrén. 2013. Saldo: a touch of yin to WordNet’s yang. *Language Resources and Evaluation*, 47(4):1191–1211. URL [dx.doi.org/10.1007/s10579-013-9233-4](https://doi.org/10.1007/s10579-013-9233-4).
- Fišer Darja, Jernej Novak, and Tomaž. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.
- Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMAP technical report, Escola de Matemática Aplicada, FGV, Brazil.
- Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.
- Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to lithuanian wordnet. In Katarína Gajdošová and Adriána Žáková, editors, *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80. Lüdenscheid: RAM-Verlag. http://korpus.juls.savba.sk/attachments/publications/lithuanian_wordnet_2013.pdf.
- Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.
- Donna Harman. 2011. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119.
- Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure.

- Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).
- Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.
- Hong Jin Kang, Tao Chen, Muthu Kumar Chandrasekaran, and Min-Yen Kan. 2016. A comparison of word embeddings for english and cross-lingual chinese word sense disambiguation. *arXiv preprint arXiv:1611.02956*.
- Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 571–580. ACM, New York, NY, USA. URL <http://doi.acm.org/10.1145/1772690.1772749>.
- Jey Han Lau, Paul Cook, Diana Mccarthy, Ana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.
- Aiden Si Hong Lim. 2014. *Acquiring Predominant Word Senses in Multiple Languages*. Ph.D. thesis, School of Humanities and Social Sciences, Nanyang Technological University.
- Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.
- Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1501–1511.
- Teng Long, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2016. Leveraging lexical resources for learning entity embeddings in multi-relational data. *arXiv preprint arXiv:1605.05416*.
- Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Nurril Hirfana Mohamed Noor, Suerya Sapan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.
- Mortaza Montazery and Hesham Faily. 2010. Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, pages 846–850.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.
- Antoni Oliver, K. Šojat, and M. Srebačić. 2015. Automatic expansion of Croatian wordnet. In *Proceedings of the 29th CALS international conference “Applied Linguistic Research and Methodology”*. Zadar (Croatia).
- Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.
- Alexander Panchenko. 2016. Best of both worlds: Making word sense embeddings in-

- terpretable. In *the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.
- Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.
- Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. URL http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf, (ISBN 978-83-7493-476-3).
- Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.
- Joel Pocostales. 2016. Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. *Proceedings of SemEval*, pages 1298–1302.
- Ida Raffaelli, Božo Bekavac, Željko Agić, and Marko Tadić. 2008. Building Croatian wordnet. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference 2008*, pages 349–359. Szeged.
- Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.
- Sascha Rothe and Hinrich Schütze. 2015. Autoextend: extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.
- Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora. (<http://fjalnet.com/technicalreportalbanet.pdf>).
- Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco.
- Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA), Valletta, Malta. URL <http://www.lrec-conf.org/proceedings/lrec2010/summaries/848.html>.
- Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, page 781–784. Lisbon.
- Liling Tan and Francis Bond. 2013. XLING: Matching query sentence to parallel corpus using topic models for word sense disambiguation. In *International Workshop on Semantic Evaluation (SemEval 2013)*.
- Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokrat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP)*, Suntec, Singapore.
- Antonio Toral, Stefania Bracal, Monica Mona-

- chini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.
- Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceaşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452. Szeged.
- Piek Vossen and Marten Postma. 2014. Open Dutch wordnet. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. (presentation only).
- Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38. URL <http://doi.acm.org/10.1145/1852102.1852106>.

Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton WordNet

Kiril Simov¹, Alexander Popov¹, Iliana Simova², Petya Osenova¹

¹IICT-BAS, Sofia, Bulgaria, {kivs|alex.popov|petya}@bultreebank.org

²Saarland University, Saarbrücken, Germany, ilianas@coli.uni-saarland.de

Abstract

In this paper we present an approach for training verb subatom embeddings. For each verb we learn several embeddings rather than only one. These embeddings include the verb itself as well as embeddings for each grammatical role of this verb. To give an example, for the verb ‘to give’ we learn four embeddings: one for the lemma ‘give’, one for the subject, one for the direct object and one for the indirect object. We have exploited these grammatical role embeddings in order to add new syntagmatic relations to WordNet. The evaluation of the new relations quality has been done extrinsically through the Knowledge-based Word Sense Disambiguation task.

1 Introduction

In this paper we present an approach to extending the knowledge graph, based on Princeton English WordNet (PWN) — (Fellbaum, 1998) — with syntagmatic relations. Our aim is to improve the knowledge-based word sense disambiguation (KWSD). In several papers we showed that adding syntagmatic relations from syntactic and semantic annotated corpora improves the performance of KWSD — (Simov et al., 2015) and (Simov et al., 2016). The main types of syntagmatic relations extracted from these corpora are the ones corresponding to the grammatical roles: verb-subject (**has-subj**), verb-direct object (**has-dobj**) and verb-indirect object (**has-iobj**). Although we managed to extract good sets of new relations, the main problem is that corpora annotated with semantic and syntactic information contain only a fraction of all the possible syntagmatic relations.

The inheritance over the hierarchies of PWN is problematic because the hierarchies of PWN are not monotonic. For that reason, in this paper we use feature learning in low dimensional vectors of real numbers known as *embeddings*.

Word Embeddings play an important role in the new stream of natural language processing applications, providing latent features for lexical items. It is expected that the necessary features are encoded within the embedding space. For example, a verb embedding represents information for its valency frame elements. Unfortunately, we can check this information only indirectly. In the paper we report embeddings on the subatom level¹ that make explicit some of the features related to the semantic selectional restrictions on grammatical roles of words in text. Thus our goal is not to learn an embedding for a verb, but rather embeddings for the participants in the event (or state) denoted by that verb.

Such an explicit embedding of the valency frame elements has many potential applications. In this work we exploit these embeddings for adding new syntagmatic relations to PWN with the aim to improve applications such as KWSD. Evaluation in the paper is performed by automatically extending WordNet with ranked relations within the context of KWSD. We show that adding higher ranked relations improves the performance of KWSD. Further we provide manual inspection and validation of the new relations that also supports the feasibility of our approach. Our approach is similar to the approach of (Paperno et al., 2014) who started with the *lexical function* model where each functional lexical item is represented via $n+1$ tensor if it is an n -ary functor. In order to escape from using tensor

¹By subatom level we mean the arguments of a predicate.

with three and more dimensions they proposed a representation where for each argument a matrix is used. Each matrix determines the incorporation of the corresponding argument semantics into the compositional semantics of the whole phrase.

Our method is also similar to the other popular methods for relation extraction. The main difference is that we do not implement relation embeddings, but rather a general embedding for one of the entities involved in the relation. Also we work with relations that are not present in the knowledge source we extend — PWN in our case. In this way we hope that our method is applicable also to the under-resourced languages.

The structure of the paper is as follows: Section 2 briefly discusses related work. In section 3 we present our motivation to extend WordNet with syntagmatic relations. Section 4 outlines an example of subatom sentential semantics based on the ideas behind Minimal Recursion Semantics. Section 5 describes the mechanism for creating grammatical role embeddings. In section 6 the experiment setup is presented and the results are discussed. The last section concludes the paper.

2 Related Work

The success of KWSD approaches apparently depends on the quality of the knowledge graph – whether the knowledge represented in terms of nodes and relations (arcs) between them is sufficient for the algorithm to pick the correct senses of ambiguous words. Several extensions of the knowledge graph constructed on the basis of WordNet have been proposed and implemented. With respect to the extension of WordNet with syntagmatic information there exist many works such as (Bentivogli and Pianta, 2004) and (Lothar Lemnitzer and Gupta, 2008).

Here we present in more detail only one approach similar to ours. It is described in Agirre and Martinez (2002) and explores the extraction of syntactically supported semantic relations from manually annotated corpora. In this line of research SemCor — (Miller et al., 1993), being a semantically annotated corpus, was processed with the MiniPar dependency parser and the subject-verb and object-verb

relations were consequently extracted. The new relations were represented on several levels: as word-to-class and class-to-class relations. The extracted selectional relations were then added to WordNet and used in the WSD task. The main differences with the approach described here are as follows: we used a bigger set of relations (since it includes also indirect-object-to-verb relations). Apart from that, the new relations reported in this paper are not added as selectional relations, but as semantic relations between the corresponding synsets. This means that the specific syntactic role of the participant is not taken into account, but only the connectedness between the participant and the event is registered in the knowledge graph. Also, in our work we use embeddings as filters, instead of the selectional restrictions approach undertaken in Agirre and Martinez (2002).

In the range of distributional semantics, the representation of word semantics for compositionality was suggested as $n+1$ dimension tensors for n -ary functor words. For example, an adjective is treated as a function over the modified noun. In order to implement this idea in practice, the semantics of the adjective is represented as a matrix which by multiplication with the vector representation of the noun produces the semantics of the noun phrases. Thus, if we assume 300-size vectors for representation of nouns, the adjectives are represented as a 300×300 matrix. For transitive verbs the representation is a $300 \times 300 \times 300$ tensor. This approach is called *lexical function* model by (Paperno et al., 2014). However, it has been criticized because the number of parameters to be learned exponentially increases. In order to solve this problem (Paperno et al., 2014) proposed a representation of functor words as a vector of a vector for the semantics of the word itself and a matrix for each of its arguments — $\langle \vec{a}, \overset{\square}{a}_1, \dots, \overset{\square}{a}_n \rangle$. Each of the matrices corresponds to a function-argument relation, such as subject-verb, noun-modifying adjective, etc. The compositional semantics of a phrase is defined as sum of the vector for the semantic functor and the multiplication of the vectors for the arguments and the corresponding matrix. This approach is called *practical lexical function (plf)* model. (Paperno et

al., 2014) demonstrate the feasibility of *plf* by testing it on several benchmarks that represent different aspects of sentence-level semantics composition. Our main goal is to learn features for the prototypical grammatical roles. In principle, they might be constructed from the *plf* representation, but the derivation of the prototypical roles representation would require an additional mechanism of abstraction. We hope that our representation will facilitate the selectional restrictions of the corresponding predicates. This is not possible to be done directly by the *plf* model. One direction of future research is to combine both approaches. This could be done by learning argument matrices that work in combination with the argument vector and the grammatical role vector. Our intuition is that such matrices could not be attached to a specific lexical unit, but to a class of lexical units.

There is also a huge number of works on extending world knowledge oriented graphs with new relations (see (Minervini et al., 2015), and (Nguyen et al., 2016) among others). The main difference in our case is that we do not learn instances of the required relations from the corpora, but we learn semantic restrictions over the arguments of the relations. The candidate relations are generated from knowledge base itself (WordNet here).

3 WordNet Extensions with New Relations

As mentioned above, in our previous works we extended PWN with syntagmatic relations using semantically annotated corpora such as SemCor. The idea was that if there is a subject-verb syntactic relation in the corpus, and the related verb and noun are manually annotated with synset ids from PWN, we could reliably assume that there is a semantic relation between the noun and the verb synsets in PWN. At a more general level we call this relation **has-participant**. It is directed from the verb to the noun synset. In order to draw a distinction between the different participants in an event (state), we use subrelations named after their grammatical roles: **has-subj**, **has-dobj**, and **has-iobj**.²

²In future work we plan to switch to semantic role names.

Adding a **has-participant** relation between two synsets in WordNet imposes two questions: (1) Does this relation hold for more specific synsets? (2) Does this relation generalize to more general synsets? In our previous research on extending WordNet with new relations from semantically and syntactically annotated corpora — (Simov et al., 2015) and (Simov et al., 2016) — we showed that using inference over the WordNet hierarchy adds new appropriate relations between verb synsets and noun synsets. Especially with respect to the **has-participant** relation, we assume that the relation holds when the noun synset is substituted with a hyponymic synset and that it also holds when the verb synset is substituted with a hypernymic synset. We noticed that in many cases such an inheritance is not correct. For example, if we have “A **doctor** operates a patient”, it does not entail that all doctors can operate. Thus we cannot reliably substitute the synset for ‘doctor’ with each of its hyponymic synsets. It is also true that the verb synset allows many more participants than the instances in the corpus. For example, “A **surgeon** cures a patient” does not imply that only hyponymic synsets are appropriate to substitute ‘surgeon’. Thus, although the extraction from syntactically and semantically annotated corpora is a reliable method for adding syntagmatic relations to WordNet, their generalization to all possible syntagmatic relations is problematic. Another problem is that such manually annotated corpora are relatively small and many verbs and nouns do not appear in them. Thus, we need a new mechanism for selection of appropriate noun synsets for participants of verbs. In this paper we used subatom semantic embeddings for checking which ones are appropriate. Such subatom semantic embeddings for each verbal synset are constructed for the appropriate grammatical roles: subject, direct object and indirect object. Having these embeddings, we rank each noun synset in PWN with respect to the corresponding grammatical role. The closer the noun synset embedding to the grammatical role embedding, the more appropriate is the noun synset as a participant for the corresponding grammatical roles in the selected verbal synset. In the rest of the paper

we present some additional motivation why such subatomic embeddings are useful, how we could train and evaluate them.

4 Minimal Recursion Semantics

An additional piece of motivation for subatom semantic embeddings is the construction of a logical form for a sentence. In many semantic theories the lexical semantics is represented not only by using predicates from first order logic, but by exploring a more complicated schema which would allow access to a more detailed representation of the semantic interpretation. As an illustration of such a kind of semantics we assume Minimal Recursion Semantics (MRS) — (Copestake et al., 2005). An MRS structure is a tuple $\langle GT, R, C \rangle$, where GT is the top handle, R is a bag of EPs (elementary predicates) and C is a bag of handle constraints, such that there is no handle h that outscopes GT . Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). Here is an example of an MRS structure for the sentence “*Every dog chases some white cat.*”

```
<h0, {h1:every(x,h2,h3), h2:dog(x),  
      h4:chase(e, x, y), h5:some(y,h6,h7),  
      h6:white(y), h6:cat(y)}, {}>
```

The top handle is $h0$. The quantifiers are represented as the relations $every(x, y, z)$ and $some(x, y, z)$, where x is the bound variable, y and z are handles determining the restriction and the body of the quantifier. The conjunction of two or more relations is represented by sharing the same handle ($h6$ above). The outscope relation is defined as a transitive closure of the immediate outscope relation between two elementary predications — EP immediately outscopes EP’ iff one of the scopal arguments of EP is the label of EP’. In the example the set of handle constraints is empty, which means that the representation is underspecified with respect to the scope of both quantifiers.

In order to use semantic embeddings over MRS structures we need to determine the interactions of the latent features for each of the

predicate arguments. For example, the features from the embeddings for ‘every’, ‘dog’, and ‘chase’ have to agree on the common argument denoted by the variable ‘ x ’. In order to control this interaction in a better way, we would like for each multiargument predicate to learn an embedding per argument. Thus for the above MRS structure we will need to have embeddings for ‘ x ’, ‘ y ’, ‘ e ’, ‘ $h0$ ’, ... ‘ $h7$ ’. When we have them, we would like also to create an embedding related to the first argument of ‘every’. The argument of ‘dog’ and the second argument of ‘chase’ have to “agree”.

Our long-term goal is to train such subatom embeddings. Here we present an approach for learning such embeddings for grammatical roles. Then we use these embeddings for extending of WordNet with syntagmatic relations, as it was described above.

5 Grammatical Role Embeddings from Parsed Corpora

In our first experiment we learned subatom semantic embeddings on the basis of dependency-parsed corpora. We determined the arguments as wordforms in the text. As an example, for the above mentioned case we used the position of ‘dog’. In order to generalize over the different word forms in the different examples in the corpus we substituted the wordforms for the corresponding argument with a pseudoword form. For example, in the above sentence we generated the following variations with pseudoword forms for the different arguments of the different predicates:

Every *SUBJ_chase* chases some white cat.

Every dog chases some white *OBJ_chase*.

and many more. Having learned embeddings for these pseudowords, we assume that they represent the selectional features for the corresponding grammatical roles of the verbs.

The actual corpus we have used is WaCkypedia_EN corpus — (Baroni et al., 2009). The WaCkypedia_EN corpus was reparsed with a more recent version of the Stanford CoreNLP dependency parser. The dependency of type “collapsed-cc” was selected, which collapses several dependency relations in order to obtain direct dependencies between content words, and in addition propagates dependencies involving conjuncts. For instance,

a parse of the sentence “the dog runs and barks” would result in the relations `nsubj(dog, runs)` and `nsubj(dog, barks)`. This type of dependency allows for a token to have multiple head words.

The head word of each noun phrase subject, as well as direct and indirect object, is then replaced by its predicate role and its governing verb’s lemma (`SUBJ_run`, `SUBJ_bark` — both for the noun ‘dog’). When a token has more than one head word suitable for substitution, copies of the sentence are created for each alternative replacement.

For the relation `has-subj` we use the dependency relations ‘`nsubj`’ and ‘`nsubjpass`’; for the relation `has-dobj` we use the dependency relation ‘`dobj`’; and for the relation `has-iobj` we use the dependency relation ‘`iobj`’. In order to minimize some errors we enforced a condition that the dependency word should be a noun.

6 Experiments and Results

In this section we describe the experimental set up and the results.

Corpora preparation.

The corpora that the algorithms for word embeddings are trained on can contain either natural language text (e.g. Wikipedia or newswire articles) or artificially generated pseudo texts. Such pseudo texts can be the output from the Random Walk algorithm, when it is set to the mode of selecting sequences of nodes from a knowledge graph (KG) — see (Goikoetxea et al., 2015) for generation of pseudo corpora from a WordNet knowledge graph and (Ristoski and Paulheim, 2016) for generation of pseudo corpora from RDF knowledge graphs such as DBpedia, GeoNames, FreeBase. Here we report results only for knowledge graphs based on WordNet and its extensions.

The corpus for training of the embeddings reported here consists of two parts: (1) pseudo corpus generated over WordNet (PCWN); and (2) real text corpora (RTC). PCWN is used to ensure that the embeddings represent features extracted from the knowledge within the WordNet. RTC is used to represent relevant contexts for learning embeddings of pseudo words for subjects, direct objects and indirect objects. As RTC we have used WaCk-

ypedia_EN corpus processed as described in Section 5.

The union of both corpora is used in the experiments. In RTC all the words were substituted with their lemmas. Punctuation marks and numbers were deleted. The PCWN corpus first was generated on the level of synset ids, then for each synset a lemma was selected from the synset randomly. The resulting corpus consists of lemmas and pseudowords for the grammatical roles. We used the Word2Vec tool³ in order to train the embeddings. From the various models we select the one with the best score on the similarity task. This model was trained with the following settings: context window of 5 words; 7 iterations; negative examples set to 5; and frequency cut sampling set to 7. The resulting embedding is lemma and pseudoword embedding. Training on the joint corpus ensure that the noun embeddings and pseudoword embeddings are in the same vector space and thus they are comparable.

Since the synset embeddings are not directly available, we need to calculate those. Thus, for each synset, we obtain its vector by averaging the vectors for all lemmas it can be expressed with (this information is retrieved from WordNet). For grammatical roles, we average the corresponding grammatical role vectors per each lemma in the particular verb synset; in this way, if a particular synset comprises N lemmas, we will average the vectors for $SUBJ_lemma_1$, $SUBJ_lemma_2$, ..., $SUBJ_lemma_N$.

The first experiments with these embeddings showed some, but very small, improvements for the task of Knowledge-based Word Sense Disambiguation. The explanation for these results is that calculating synset embeddings on the bases of lemma embeddings is not good enough because of the high level of ambiguity of lemmas in PWN.

This is why we performed two more experiments with two new versions of the corpora. First, we annotated the RTC with senses using UKB system⁴ for knowledge-based word sense disambiguation. For the PCWN corpus we have used the version generated only using synset ids. In this case the embeddings

³<https://code.google.com/archive/p/word2vec/>

⁴<http://ixa2.si.ehu.es/ukb/>

are directly trained over synsets. Unfortunately, this approach did not improve the results significantly. Our explanation for this is the fact that the annotation with UKB, even with our best knowledge graph from (Simov et al., 2016), is under 68% accuracy. This result is too low for our task. Second, we used the POS annotation for RTC to substitute each word with lemma-POS strings. In this way we differentiated the same lemma used as different parts-of-speech. For PCWN it is straightforward to substitute the synset ids with the combinations lemma-POS. This experiment demonstrated the best results which we report here. From these corpora we trained two embeddings: (1) embedding trained over RTC only⁵. We denote this embedding as RTC; and (2) embedding trained over the joint corpus. We denote this embedding as RTCPCWN.

Selection and Ranking of Candidate Relations.

The candidate relations are selected in the following way. For each verbal synset that has at least one grammatical role embedding we form candidate relations in the following format:

```
u:noun-synset-id v:verb-synset-id
```

where noun-synset-id is any noun synset in PWN. Thus, for each verb we generate more than 74 000 candidate relations. Here is an example:

```
u:00031264-n v:02005948-v
```

for ‘arrive’ (02005948-v) and ‘group’ for (00031264-n).

After the completion of this step, we have all the information necessary to compare synset embeddings with grammatical role embeddings that match verb synsets. The comparison is carried out by calculating the cosine similarity measure. By setting a similarity threshold, the filter can be controlled, so that more or fewer new relations are added to the extended graph. The same procedure is repeated for DOBJ and IOBJ relations. Using this approach for each candidate relation we calculate the cosine similarity measure between the noun synset embedding and the embedding for the corresponding grammatical role. We then

⁵This was suggested to us by one of the reviewers in order to see the impact of adding PCWN.

used the result as a rank over the candidate relations.

Experiments with Knowledge-based Word Sense Disambiguation.

In order to check the usefulness of the added relations, we performed experiments with the UKB system⁶ for knowledge-based word sense disambiguation. The UKB tool requires two resource files to annotate the input text — a dictionary file with all lemmas that can be possibly linked to a sense identifier. In our case WordNet-derived relations were used for our knowledge base; consequently, the sense identifiers are WordNet IDs. For instance, a line from the WordNet extracted dictionary looks like this:

```
predicate 06316813-n:0 06316626-n:0
           01017222-v:0 01017001-v:0
           00931232-v:0
```

First comes the lemma associated with the relevant word senses, after the lemma the sense identifiers are listed. Each ID consists of eight digits followed by a hyphen and a label referring to the POS category of the word. Finally, a number following a colon indicates the frequency of the word sense, calculated on the basis of a tagged corpus. When a lemma from the dictionary has occurred in the analysis of the input text, the tool assigns all the associated word senses to the word form in the context and attempts to disambiguate its meaning among them.

The second resource file required for running the tool is the set of relations used to construct the knowledge graph over which UKB is run. The distribution of UKB comes with a file containing the standard lexical relations defined in WordNet, such as hypernymy, meronymy, etc., as well as with a file containing relations derived on the basis of common words found in the synset glosses, which have been manually disambiguated. The format of the relations in the KG is as follows:

```
u:SynSetId01 v:SynSetId02 s:Source d:w
```

where `SynSetId01` is the identifier of the first synset in the relation, `SynSetId02` is the identifier of the second synset, `Source` is the source of the relation, and `w` is the weight of the relation in the graph. In the experiments reported

⁶<http://ixa2.si.ehu.es/ukb/>

Knowledge Graph	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTC40	50.32	49.51
wn30RTC45	52.60	49.57
wn30RTC47	50.20	48.47
wn30RTC50	50.34	49.63
wn30RTC52	50.58	51.88
wn30RTC55	51.05	51.70
wn30RTC57	51.60	51.52

Table 1: Results about relations ranked by embeddings from POS tagged real text corpus. The improvement for SemCor is **1.04** and for M13 SemeVal is **3.47**.

Knowledge Graph	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTCPCWN35	51.88	49.27
wn30RTCPCWN38	53.68	51.39
wn30RTCPCWN40	53.91	51.45
wn30RTCPCWN42	54.33	50.42
wn30RTCPCWN43	54.08	50.18
wn30RTCPCWN44	52.56	49.93

Table 2: Results about relations ranked by embeddings from POS tagged real text corpus and pseudo corpus. The improvement for SemCor is **2.77** and for M13 SemeVal is **3.04**.

in the paper, the weight of all relations is set to 0.

In our experiments we relied on the following knowledge graphs: **wn30** — a knowledge graph formed from the relations in PWN (baseline); **wn30RTCNN** — a knowledge graph formed on the basis of **wn30** extended by the grammatical role-based relations, ranked by RTC embeddings. The number NN is the rank threshold for selection of the new relations. If NN is 47, then all relations with rank equal or higher than 47 are selected; **wn30RTCPCWNNN** — a knowledge graph formed on the basis of **wn30** extended by the grammatical role-based relations, ranked by RTCPCWN embeddings. The interpretation of NN is the same.

The evaluation of the Word Sense Disambiguation is done over two test data sets: the test part of SemCor as defined in (Simov et al., 2015) and (Simov et al., 2016) and the English part of the test data set for the Multilingual Word Sense Disambiguation⁷ — named here M13 SemeVal. The results are presented in

⁷<https://www.cs.york.ac.uk/semeval-2013/task12/>

Table 1 with improvement of **1.04** for SemCor and **3.47** for M13 SemeVal and in Table 2 with improvement of **2.77** for SemCor and **3.04** for M13 SemeVal. As it can be seen, the results depend on the type of the test corpus: SemCor is a balanced one and hence shows usages of many senses; M13 SemeVal is a smaller one and does not provide so many diverse types of text. All the results show that there is a rank for which there is a highest result, and for lower or higher ranks the result drops. Our explanation of this is that: (1) for higher ranks the number of added relations is smaller and thus their impact on the result is smaller; and (2) for the lower ranks the number of the not-so-good relations is higher. The impact of the PCWN embeddings is with respect to the type of the test corpora. In our view better relations are selected for a wider set of verbs.

Experiments have been performed for evaluating the number of examples in the corpus as well as the quality of learned embeddings. Thus the verbs for which there were less than 10 examples of the corresponding dependency relation in the original corpus, were not taken into account. The results are reported in Ta-

Knowledge Graph	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTCPCWN10-34	52.35	51.39
wn30RTCPCWN10-35	50.64	53.04
wn30RTCPCWN10-36	50.25	50.72
wn30RTCPCWN10-40	50.49	49.45
wn30RTCPCWN10-45	51.15	49.27
wn30RTCPCWN10-50	51.45	48.29

Table 3: Results after cutting less frequent verbs grammatical roles (less ten examples of the corresponding grammatical role in the original corpus). The improvement for SemCor is **0.79** and for M13 SemeVal is **4.62**.

Role	Good	Acceptable	Bad
Subject	68	28	4
Direct object	67	24	9

Table 4: These are the manual evaluation results of the first 100 suggested relations selected via RTCPCWN embeddings for subject and direct object roles.

ble 3. They show that there are improvements for both corpora: **0.79** for SemCor and **4.62** for M13 SemeVal. In our view the very small improvement for SemCor is due to the varying senses in it. This variety makes it more sensitive to the changes in the knowledge graphs with respect to deletion of many new relations. In M13 SemeVal corpus, on the other hand, there were not so many rare senses.

These results, however, succeed to show that the presented approach for selecting syntagmatic relations is quite feasible. Since this evaluation approach seems to be too indirect, we think that more work is necessary to adequately evaluate the grammatical role embedding.

6.1 Manual Inspection

The results were manually evaluated for the first 100 top-ranked subject and direct object relations. A scale was used that classifies the examples into the following groups: **good**, **acceptable**, and **bad**. The labels are correlative. This is possibly due to the fact that most verbs have intransitive and transitive usages. As it can be seen from the table, most relations have been labeled as 'good', then come the 'acceptable' relations, and finally some 'bad' ones.

For both syntactic labels it was observed that the most frequent among the top-

ranked relations are chemistry-oriented domain ones, such as: <dimethylglyoxime, dehydrogenate>. For the 'good' relation one example is as follows <streusel, caramelize>: "The streusel seeps down and caramelizes the apples in the most glorious way".

As acceptable relations we marked mostly ones that are good semantic relations but would not generate reasonable sentences because they are derivationally related. For example: <celebration, celebrate>, <chart, chart>, <oxidation, oxidate>, <measurement, measure>, etc.

As bad example the following relation is considered: <cassareep, splinter>.

Thus manual evaluation also shows that the proposed mechanism of adding syntagmatic relations to PWN is feasible.

7 Conclusion

The paper presents an approach for learning features by subatomic semantic representation. It is useful for addition of syntagmatic relations to WordNet. Our longterm plan is to design a learning approach for each semantic argument of predicates in a logical form. The results here are the first steps in this direction.

In future we plan to do the following: (1) To include more arguments in the learning process like arguments of relational nouns and adjectives. They will impose mutual constraints

on the learned features; (2) To experiment with different algorithms for learning of embeddings such as (Levy and Goldberg, 2014), where it is possible to select arbitrary contexts. Such contexts could be more appropriate for grammatical role embeddings learning; (3) To improve sense annotation in order to respectively improve sense embeddings; (4) To evaluate the subatom embeddings in other tasks such as coreference resolution, neural network word sense disambiguation; (5) To perform tuning to the linguistic knowledge already represented in WordNet, FrameNet and other lexical resources as well as manually annotated corpora, by techniques similar to retrofitting; and (6) To develop compositional semantics over this representation.

Acknowledgements

This research has received partial support by the grant 02/12 — *Deep Models of Semantic Knowledge (DemoSem)*, funded by the Bulgarian National Science Fund in 2017–2019. We are grateful to the anonymous reviewers for their remarks, comments, and suggestions. All errors remain our own responsibility.

References

- Eneko Agirre and David Martinez. 2002. Integrating selectional preferences in WordNet. In *Proceedings of First International WordNet Conference*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Luisa Bentivogli and Emanuele Pianta. 2004. Extending wordnet with syntagmatic information. In *In Second Global WordNet Conference*, pages 47–53.
- Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332.
- Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.
- Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *HLT-NAACL*, pages 1434–1439. The Association for Computational Linguistics.
- Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June. Association for Computational Linguistics.
- Holger Wunsch Lothar Lemnitzer and Piku Gupta. 2008. Enriching GermaNet with Verb-Noun Relations - a Case Study of Lexical Acquisition. In *Proc. of the Sixth International Conference on Language Resources and Evaluation*.
- George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Proceedings of the workshop on Human Language Technology, HLT '93*, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pasquale Minervini, Claudia d’Amato, Nicola Fanizzi, and Floriana Esposito. 2015. Efficient learning of entity and predicate embeddings for link prediction in knowledge graphs. In *Proceedings of the 11th International Workshop on Uncertainty Reasoning for the Semantic Web co-located with (ISWC 2015)*, pages 26–37.
- Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. 2016. Stranse: a novel embedding model of entities and relationships in knowledge bases. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–466.
- Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–99.
- Petar Ristoski and Heiko Paulheim, 2016. *RDF2Vec: RDF Graph Embeddings for Data Mining*, pages 498–514. Springer International Publishing, Cham.
- Kiril Simov, Alexander Popov, and Petya Osenova. 2015. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 596–603.
- Kiril Simov, Petya Osenova, and Alexander Popov. 2016. Using context information for knowledge-based word sense disambiguation. In Christo Dichev and Gennady Agre, editors, *Artificial Intelligence: Methodology, Systems, and Applications*, pages 130–139, Cham. Springer International Publishing.

An Iterative Approach for Unsupervised Most Frequent Sense Detection using WordNet and Word Embeddings

Kevin Patel and Pushpak Bhattacharyya

Department of Computer Science and Engineering

Indian Institute of Technology Bombay

{kevin.patel,pb}@cse.iitb.ac.in

Abstract

Given a word, what is the most frequent sense in which it occurs in a given corpus? Most Frequent Sense (MFS) is a strong baseline for unsupervised word sense disambiguation. If we have large amounts of sense-annotated corpora, MFS can be trivially created. However, sense-annotated corpora are a rarity. In this paper, we propose a method which can compute MFS from raw corpora. Our approach iteratively exploits the semantic congruity among related words in corpus. Our method performs better compared to another similar work.

1 Introduction

Word Sense Disambiguation (WSD) remains to be one of the relatively hard problems in the field of Natural Language Processing. Machine Learning approaches to WSD can be broadly classified into two categories: supervised and unsupervised. Supervised techniques rely on learning patterns from sense-annotated training data. However, such data are hard to come by. SemCor, one of the most common sense-annotated corpus in English language, contains around 700k tokens, 200k of which have been sense-annotated. It is really small as compared to raw corpora such as ukWAC, where the number of tokens is close to 2 billion. On the other hand, unsupervised techniques do not require sense-annotated corpora.

A strong baseline for unsupervised WSD is the Most Frequent Sense (MFS) baseline. While performing sense disambiguation, the baseline completely ignores the context, and simply assigns the most frequent sense to the target word.

In spite of ignoring context, which is one of the main source of information for performing sense disambiguation, the MFS baseline gives re-

ally strong results. This is because of the inherent skew in the sense distribution of the data.

Computing MFS baseline is trivial, if one has access to large amounts of sense-annotated corpora. However, that is not the case as explained earlier. Thus there is a need for uncovering MFS from raw data itself.

Word embeddings collectively refers to the set of language modelling and feature learning techniques, which maps words to real valued vectors (Bengio et al., 2003; Mnih and Hinton, 2007; Collobert and Weston, 2008; Mikolov et al., 2010; Huang et al., 2012; Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington et al., 2014). Do note that most word embedding models only output *one embedding per word*, instead of the ideal case of outputting *one embedding per sense of a word*. Though, some models do exist, which provide one embedding per sense of a word by inferring number of senses either through context clustering approaches (Neelakantan et al., 2015), or by using sense inventory (Chen et al., 2014). For the rest of this paper, we mean *one embedding per word* models, when we use the phrase word embeddings.

The field of Natural Language Processing is increasingly seeing the use of word embeddings for various problems, and MFS is no exception. Bhingardive et al. (2015) showed that pretrained word embeddings can be used to compute most frequent sense.

In this paper, we propose an iterative approach for extracting most frequent sense of words in a raw corpus. The approach uses word embeddings as an input. Thereby, in order to obtain MFS from some raw corpus, one need to apply the following two steps:

1. Train word embeddings on the raw corpus.
2. Apply our approach on the trained word embeddings.

The key points of this paper are:

- Our work further strengthens the claim by (Bhingardive et al., 2015) that word embeddings indeed capture most frequent sense.
- Our approach outperforms others at the task of MFS extraction.

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 explains our approach. Section 4.1 details our experimental setup and results. Section 5 provides some error analysis, followed by conclusion and future work.

2 Related Work

Buitelaar and Sacaleanu (2001) present an approach for domain specific sense assignment. They rank GermaNet synsets based on the co-occurrence in domain corpora. Lapata and Brew (2004) acquire predominant sense of verbs. They use Levin’s classes as their sense inventory. McCarthy et al. (2007) use a thesaurus automatically constructed from raw textual corpora and the WordNet similarity package to find predominant noun senses automatically. Bhingardive et al. (2015) exploit word embeddings trained on untagged corpora to compute the most frequent sense. Our work is most similar to Bhingardive et al. (2015) owing to our reliance on word embeddings. We therefore evaluate our approach against theirs.

3 Approach

Our approach relies on the semantic congruity of raw text. Consider the following example: Consider the word *cricket* having two senses **sport** and **insect**, and the word *bat* having two senses **sport_instrument** and **reptile**. Then, if in our corpus, we already know that *bat* is in **sport_instrument** sense for most cases, then in order for the corpus to be semantically congruent, the most frequent sense of *cricket* has to be **sport**.

So, in order to find most frequent sense of all words in the vocabulary of the corpus, we start with the word whose sense is already known. So, the approach begins with monosemous words, for which MFS is trivial. Next, it moves on to bisemous words, and uses the monosemous words sense information to detect most frequent sense. Then it moves on to trisemous words, and use

the hitherto resolved words for detecting most frequent sense, and so on. Thus the approach **iterates** over the degree of polysemy, and uses the computed MFS of words with degree of polysemy 1 to $n - 1$ to compute the MFS of words with degree of polysemy n .

At any point of time, we call the words whose MFS is already established as *tagged words*. For a given word whose MFS is to be computed, we enumerate all senses, and then compute the *vote* for each senses by the top k nearest neighbors who are already tagged. The vote is a product of two measures: the cosine similarity (w_i) between the embedding of the current tagged word and the target word, and the wordnet similarity (s_i) between the MFS of current tagged word (which would have been established in the previous iteration), and the sense for which the vote is being computed. The votes are summed over, and the sense with the highest sum is considered to be the Most Frequent Sense of the target word. The basic flow is illustrated in figure 1.

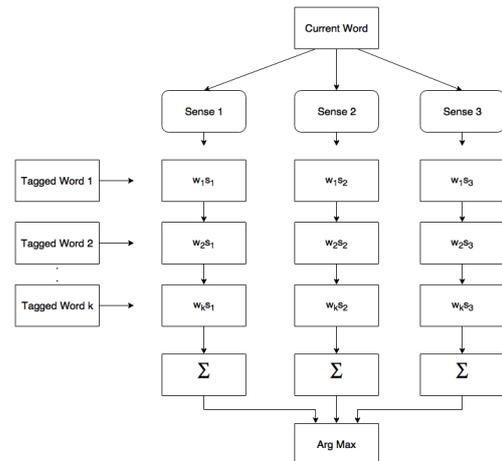


Figure 1: Illustration of our approach

The major parameters in our approach are:

1. **K**: The number of nearest neighbors who will vote. Through experimentation, we found $K=50$ to be a reasonable value.
2. WordNet Similarity measure (s_i): We tried all combinations of the six available similarity measures in Princeton WordNet, namely Path similarity, Leacock Chodorow Similarity, Wu Palmer Similarity, Resnik Similarity, Jiang Conrath Similarity, and Lin Similarity.

Our experiments found the average of normalized Wu Palmer and Lin similarity performs slightly better than other combinations.

3. Vector space similarity measure (w_i): We tried both dot and cosine similarity. Dot performed slightly better. In future, we would try other similarity measures such as Tanimoto coefficient.

4 Experiments and Results

4.1 Datasets

We have used the following datasets for our evaluation:

1. SemCor: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses using WordNet 1.7 to WordNet3.0 mapping by Rada Mihalcea
2. Senseval 2: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses as above
3. Senseval 3: Sense-annotated corpus, annotated with Princeton WordNet 3.0 senses as above

4.2 Evaluating MFS as solution for WSD

Given that MFS is a strong baseline for unsupervised WSD, a good MFS will give good performance on unsupervised WSD. This is what this experiment evaluates. While in theory, our approach can also use embeddings trained on test corpora directly, we use pretrained word2vec embeddings, as they are crucial to Bhingardive et al. (2015) with whom we are comparing. Table 1 shows the results of using MFS for WSD on Senseval 2 and Senseval 3 only for nouns. We report this noun specific result for comparison with (Bhingardive et al., 2015), who have reported results only for nouns. Here, Bhingardive(reported) and SemCor(reported) are the results as reported in the paper. However, their exact experiment settings are not clear from their paper. Thus we used also computed their results in our setting, which are reported as Bhingardive and SemCor respectively.

In addition to this, we also report the performance on all parts of speech, in table 2. Here, Bhingardive(reported) is the result with the parameter configuration for their approach as reported

Method	Senseval2	Senseval3
Bhingardive(reported)	52.34	43.28
SemCor(reported)	59.88	65.72
Bhingardive	48.27	36.67
Iterative	63.2	56.72
SemCor	67.61	71.06

Table 1: Accuracy of WSD using MFS (Nouns)

in their paper. We also tried out different parameter settings for their algorithm, and Bhingardive(optimal) is the best result obtained with optimal parameter setting. It is clear that our approach outperforms both their reported approach and the one with empirically obtained optimal parameters.

Method	Senseval2	Senseval3
Bhingardive(reported)	37.79	26.79
Bhingardive(optimal)	43.51	33.78
Iterative	48.1	40.4
SemCor	60.03	60.98

Table 2: Accuracy of WSD using MFS (All Parts of Speech)

4.3 Evaluating MFS as classification task

Another way to evaluate our approach was to learn MFS from pretrained word embeddings which were trained on large corpora, and compare it with WordNet First Sense (WFS). Table 3 shows how our approach fares against Bhingardive et al. (2015)’s when both the approaches are applied on pretrained word2vec embeddings (trained on Google News Dataset with billions of tokens and released by them).

A similar evaluation can also be done by using true MFS obtained from frequencies in sense-annotated corpora. Tables 4 show the results for the same.

5 Discussion

Even though our approach performs better than Bhingardive et al. (2015), we are not able to cross SemCor and WFS results. The following are the reasons for the same:

- There are words for which WFS doesn’t give *proper* dominant sense. Consider the following examples:
 - *tiger* - an audacious person

Method	Nouns	Adjectives	Adverbs	Verbs	Total
Bhingardive	43.93	81.79	46.55	37.84	58.75
Iterative	48.27	80.77	46.55	44.32	61.07

Table 3: Percentage match between predicted MFS and WFS

	Nouns (49.20)	Verbs (26.44)	Adjectives (19.22)	Adverbs (5.14)	Total
Bhingardive	29.18	25.57	26.00	33.50	27.83
Iterative	35.46	31.90	30.43	47.78	34.19

Table 4: Percentage match between predicted MFS and true SemCor MFS. Note that numbers in column headers indicate what percent of total words belong to that part of speech

- *life* - characteristic state or mode of living (social life, city life, real life)
 - *option* - right to buy or sell property at an agreed price
 - *flavor* - general atmosphere of place or situation
 - *season* - period of year marked by special events
- In some cases, the tagged words actually rank very low in order for them to make a significant impact. For instance, while detecting MFS for a bisemous word, it may happen that the first monosemous neighbour actually ranks 1101, *i.e.* a 1000 polysemous words are closer than this monosemous word. Thus in such cases, the monosemous word may not be the one who can influence the MFS.

6 Conclusion

In this paper, we proposed an iterative approach for unsupervised most frequent sense detection in raw corpus. The approach uses word embeddings. Our results bears similar trends to those of Bhingardive et al. (2015), thereby strengthening the claim that word embeddings do indeed capture most frequent sense. Through 2 different categories of experiments, we established that our method is better than theirs. Since there are no language specific restrictions, we believe that our approach should be easily applicable to other languages. In the future, we would like to experimentally validate this claim.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic lan-

guage model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Sudha Bhingardive, Dharendra Singh, Rudramurthy V, Hanumant Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised most frequent sense detection using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238–1243, Denver, Colorado, May–June. Association for Computational Linguistics.

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of the WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL Workshop*, Pittsburgh. o.A.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL ’12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of pre-

- dominant word senses. *Computational Linguistics*, 33(4):553–590.
- Tomáš Mikolov, Martin Karafiát, Luk Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Inter-speech*, pages 1045–1048.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.
- Andriy Mnih and Geoffrey E. Hinton. 2007. Three new graphical models for statistical language modelling. In Zoubin Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 641–648. ACM.
- Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Conference on Empirical Methods in Natural Language Processing, 2014*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12.

Automatic Identification of Basic-Level Categories

Chad Mills

University of Washington
Seattle, WA, United States
chills@uw.edu

Francis Bond

Nanyang Technological University
Singapore
bond@ieee.org

Gina-Anne Levow

University of Washington
Seattle, WA, United States
levow@uw.edu

Abstract

Basic-level categories have been shown to be both psychologically significant and useful in a wide range of practical applications. We build a rule-based system to identify basic-level categories in WordNet, achieving 77% accuracy on a test set derived from prior psychological experiments. With additional annotations we found our system also has low precision, in part due to the existence of many categories that do not fit into the three classes (superordinate, basic-level, and subordinate) relied on in basic-level category research.

1 Introduction

WordNet organizes concepts into a hierarchy of hypernyms and hyponyms (Miller 1995). While WordNet also identifies other information, such as meronymy, one interesting property that is not currently captured is which concepts represent basic-level categories.

This is an important and valuable property to capture. Brown (1958) first noted that, although there are many terms that could be used to refer to an object at different levels of abstraction, “it often happens that a hierarchy develops in both directions from a middle level of abstraction.” Rosch et al. (1976) called this the basic-level, identifying psychological advantages basic-level categories have as well as psychological tests to find these concepts in a hierarchy. Examples of basic-level categories include *table*, *car*, *tree*, *bird*, *guitar*, *shirt*, *fish*, and *apple* (Rosch et al. 1976).

Unfortunately, though, the process of identification does not scale well and only dozens of

these concepts have been identified in the psychology literature (Rosch et al. 1976, Markman and Wisniewski 1997).

While there has been little work to automate the identification of basic-level categories (discussed in Section 2), knowing the basic-level has been shown to be valuable. Knowing the basic-level helps with word sense disambiguation (Legrand 2006), image searches (Rorissa and Iyer 2008), ad targeting (Wang et al. 2015), accurately measuring the readability of a text (Lin et al. 2009), making search result entity cards more easily consumable (Wang et al. 2015), linking together different domain-specific information classification systems (Green 2006), and user-centered design of image-browsing interfaces (Rorissa and Iyer 2008). We also believe it could help with having a common set of words to work from in building WordNets for other languages, as well as language grounding and many other problem areas.

Given the wide variety of demonstrated applications of this information as well as the opportunity for application in other areas, we attempt to automate the identification of basic-level categories.

We specifically look at heuristics to identify the basic-level noun categories in the Princeton WordNet of English (Fellbaum 1998), hereinafter PWN. One author assigned this task as a project in a class he taught in 2010 and 2011. This work builds on the various techniques students used and combines them with novel rules into a rule-based system to identify basic-level categories.

2 Related Work

2.1 Basic-level categories

Interest in basic-level concepts spans many disciplines, including philosophy (Rand 1966), psy-

chology (Rosch et al. 1976), library and information science (Green 2006), computer science (Wang et al. 2015), and others. While different disciplines have come up with very different theories to explain essentially the same underlying phenomena, they each bear many resemblances given the similarity in phenomena described.

While philosophy provides the foundation on which much of the work is based, and the field even has some work specifically on basic-level categories, the most numerous work on basic-level categories has been in psychology following the work of Rosch et al. (1976).

Rosch et al. (1976) distinguished between three levels of categories: basic-level, superordinate (hypernyms of the basic-level), and subordinate (hyponyms of the basic level). They found many properties of these categories, such as that basic-level categories are the most inclusive level at which a concrete picture of the category as a whole can be formed.

Markman and Wisniewski (1997) offer what may be a more fundamental and clear definition of the basic-level as being the level with the most alignable differences. An alignable difference is a difference in degree rather than kind; for example, cars and motorcycles have a different number of wheels (alignable) but a car carries a jack and a motorcycle does not (non-alignable). *Car* and *motorcycle* here are both taken to be basic-level categories, while *vehicle* is a superordinate and *coupe* is a subordinate. The various subordinates of car (*coupe*, *sedan*, etc.) vary in a handful of ways, but they have more similarities than differences. Cars and motorcycles, on the other hand, have many more differences and many of these are alignable (number of wheels, type of seat, steering controls, acceleration controls, etc.). According to (Markman and Wisniewski 1997), this abundance of alignable differences is a clear indicator that *car* and *motorcycle* are basic-level.

There has been a wide variety of additional research in this area within psychology showing a range of properties, applications, and even several potential issues with basic-level categories. Though before the concept was well-established, Brown (1958) noticed that children learn some middle level of concepts first, which Rosch et al. (1976) later showed was true of basic-level categories. Rosch et al. (1976) also showed basic-level category membership is verified fastest, objects are named faster at the basic-level, and objects are preferentially named with their basic-level category. Studies have shown children learn basic-level categories first, then subordinates,

then superordinates (Jónsdóttir and Martin 1996), with children not even considering a novel noun to potentially be a superordinate until around age 7 (Golinkoff et al. 1995).

At the same time, there are some limitations to these advantages. Adult experts in a domain may be so fluent with the subordinate level in that domain that some of the advantages of the basic-level over the subordinate level become greatly diminished (Tanaka and Taylor 1991). Still, even here the boundary between basic and superordinate concepts is an important one with qualitative differences in how they are represented, such as superordinate concepts (e.g. *furniture*) often referring to groups of entities and basic-level (e.g. *table*) referring to individuals (Murphy and Wisniewski 1989). Some interesting corner cases have also been found with abnormal exemplars, for example with *penguin* having the basic-level advantages but *bird* being the clear basic-level category for most birds (Jolicoeur et al. 1984).

Despite these and other limitations, though, there has been a surprisingly broad variety of research into applications of basic-level categories, as discussed to motivate the problem in Section 1, showing that a system identifying the basic-level would be valuable.

2.2 Identifying basic-level categories at scale

There has been very little work specifically on detecting basic-level categories at scale. The experiments in psychology have around a dozen examples of basic-level categories (Rosch et al. 1976, Markman and Wisniewski 1997).

There have only been a few efforts to use this data to learn patterns and extrapolate to a broader set of basic-level categories, all working with PWN, though some of the psychology literature also points out attributes of basic-level categories that may be helpful.

Farwell (2009) started with all nouns and did some filtering of superordinates and subordinates by depth in the hierarchy. This was followed by a voting scheme to pick the best candidate on each path from the top of the hierarchy to a leaf node, considering how short the word is, how frequently the word is used, and how many words are in the synset all as positive features while having few hyponyms and fewer relationships with other synsets more broadly as negative features (Green 2006). There was no effort to reconcile results from nearby paths down the hierarchy, though, and the list of basic-level categories generated

was fed into a downstream system to map information systems together, with no evaluation of the categories themselves.

Another effort focused on word sense disambiguation, with Izquierdo et al. (2007) using a simpler approach that filtered out the lower levels of the hierarchy and searched up the hypernym tree exclusively looking for a synset with a large number of PWN relations. These features were already included by (Green 2006), and here as well the evaluation was only performed on the applied system and an evaluation was not performed on this basic-level category identification system as such. Izquierdo et al. (2007) did make one important distinction, though, between basic-level categories and the similarly-named base concepts. Base concepts are a set of concepts core to many relations and tend to occur relatively high in the hierarchy (Izquierdo et al. 2007). On the other hand, while there is certainly overlap, basic-level concepts tend to occur closer to the middle of the hierarchy and tend to have less relations (Izquierdo et al. 2007).

Lin et al. (2009) attempted to identify basic-level categories by looking for words that are shorter than their hyponyms and where the word is frequently contained within its hyponyms as a compound. Again this was only evaluated in the application of measuring text readability, and like the other experiments they used all the available data for forming the rules without holding aside any data for an independent evaluation.

3 Data

We are aware of two major lists of basic-level categories as well as corresponding superordinates and some subordinates.

The original experiments that started much of the work in this area (Rosch et al. 1976) include nine superordinate taxonomies for their first two experiments. For the three of these superordinates falling in the biological taxonomy, the experimental results showed the presumed superordinate level (*tree, fish, bird*) is actually the basic-level. So, for these three groups the taxonomy was shifted down one level (e.g. basic to supordinate) and new superordinates (*plant, animal, animal*) were added to ensure the experimental results were accounted for. Additionally, eight additional basic-level categories were used in their later experiments 3-4 (Rosch et al. 1976), so these were also added. Markman and Wisniewski (1997) also provide a large list of superordinates,

basic-level categories, and subordinates, though there is overlap with the aforementioned list.

A summary of the lists is shown in Table 1.

Level	Rosch	Markman	Combined
Superordinate	8	24	24
Basic-level	29	80	92
Subordinate	45	25	68

Table 1: Categories with known classification by level

This is the data used for training and evaluating our system. The details of how the data is split up for that purpose is discussed in Section 5.

4 Our Approach

We start with 29 student projects each independently trying to solve this problem, cataloging the types of approaches and rules considered and then combining a slightly-constrained set of these, as well as novel rules, into a combined system.

While the goal is to produce one system by evaluating the collective set of rules, some boundaries are needed to constrain this. For example, one student only considered words also appearing in the ‘adventure’ category of the Brown corpus (Francis and Kucera 1964), a small, categorized corpus of English, which restricts the project beyond the goals of this work. We therefore start with a general approach common to most solutions (Section 4.1), describe the relevant rules (Section 4.2), experimenting to determine which Filtering Rules are more and less effective (Section 5.1), and then combine the more effective rules into a combined system before experimenting with a set of Voting Rules (Section 5.2).

4.1 General Approach

We start with all noun synsets in PWN. The available gold standard labels discussed in Section 3 are all nouns, though it is worth noting some research has indicated it is likely possible to extend the basic-level to other parts of speech (Lemaitre and Heller 2013).

We then take the labeled data from the psychology literature discussed in Section 3, manually map each of the categories to the closest PWN synset when one exists, and the goal becomes to extrapolate from these to other PWN synsets that are also at the basic-level and not at the superordinate or subordinate levels. In the psychology experiments (Rosch et al. 1976, Markman and Wisniewski 1997) this was done with words

whose senses were disambiguated by context, so we operate at the sense level. For our purposes, category and synset will be used interchangeably.

The students were identifying words, not synsets, though each student had to try to map words to synsets to use PWN features before producing a final list of words from there, losing the synset distinctions. For this work, we treat the basic-level as operating at the sense level and ensure our labels for training and evaluation are on PWN synsets to remove this unnecessary complexity.

Essentially everything the students did to identify basic-level categories can be generalized as one of two approaches:

1. filtering out nouns that are not basic-level or
2. on a particular path from the root to a leaf node in the hypernym/hyponym hierarchy, score each node and choose the optimal one as the basic-level on that path

We adopt both of these approaches, first applying a set of Filtering Rules to remove synsets unlikely to be basic-level and then choosing at most one per path based on a set of Voting Rules.

There were a few other extensions students considered, such as taking the top 2000 results with a provided sorting function, but since we do not want to assume a particular number of basic-level categories we do not incorporate these approaches into our system. Many students also deduped their final list, dealt with lemmatization, chose which word in a synset to use to represent the synset, and other issues that are not necessary when operating at the synset level and thus were omitted here.

4.2 Rules

We have cataloged the rules students used, along with our own novel rules, generalizing them and parameterizing rules where possible to enable experimenting with different thresholds. Note some rules focus on a word since students were not working on synsets, so for these rules we follow the convention most students followed in mapping words to synsets by taking the first lemma in the synset as the word for applying these rules.

The list of Filtering Rules is shown in Table 2, and the Boolean Voting Rules used for voting schemes to pick the best synset left in a chain after filtering are shown in Table 3. Parameter ranges used by students, or examples in cases where there are long lists of parameters, are shown after the rule. Ranges are given in interval notation to avoid boundary condition ambiguity.

Filtering Rules
1. Filter words with a set of suffixes (-ing, -ment, ... [59 total])
2. Filter words with a set of prefixes (un-, th-)
3. Filter words of length n or greater [7, 16]
4. Filter words of length n or fewer [1, 4]
5. Filter space-separated compound words
6. Filter hyphenated words ('-')
7. Filter joined compounds (e.g. racetrack)
8. Filter words with numbers
9. Filter words with symbols
10. Filter words with more adjective than noun senses
11. Filter words with more adverb than noun senses
12. Filter words with over 1 more verb than noun sense
13. Filter words that are not substrings in immediate subordinate nodes
14. Filter words containing any word at a higher level
15. Filter stopwords
16. Filter plural words
17. Filter words with no vowels
18. Filter words with over n vowels [1]
19. Filter capitalized words
20. Filter synsets with average depth $((\min+\max)/2, \text{recursive})$ outside the range a to b [4.2, 9)
21. Filter synsets with hyponym depth $(\min+\max)/2$ outside the range a to b [1.1, 2.2)
22. Filter synsets with $\text{avg_depth}/(\text{avg_depth}+\text{avg_height})$ outside the range a to b [.74, .91]
23. Filter the top n levels of the hierarchy [2-7]
24. Filter nodes with n levels below them (5)
25. Filter synsets with an average depth $((\max+\min)/2)$ of $\leq n$ (5.4)
26. Filter the bottom n levels of the hierarchy [1, 3]
27. Filter synsets n or more levels deep [9, 15]
28. Filter siblings of synsets with 0 hyponyms
29. Filter nouns with a to b hyponyms [0,2], [5,inf)
30. Filter synsets in the Brown corpus with frequency $< n$ (1-10)
31. Filter synsets in the Brown corpus with frequency $> n$ (40)
32. Filter all synsets under abstraction.n.06
33. Filter all synsets except those under set S (combinations of physical_entity.n.01, thing.n.08, substance.n.01, process.n.01)
34. Filter all words in the CHILDES corpus

- | |
|---------------------------------------------------------------------------------|
| 35. Filter words in the CMU Pronouncing Dictionary with > 9 phonemes |
| 36. Filter all synsets with n or more siblings having no hyponyms |
| 37. Filter all synsets with at least p percent of siblings having no hyponyms |
| 38. Filter synsets with less than n siblings |
| 39. Filter words not in the Childes corpus |

Table 2: Filtering rules

Voting Rules
40. Top frequency in the chain (sum of lemma frequencies in synset)
41. Top frequency in the chain in SEMCOR and frequency $\leq n$ (60)
42. Word length between a and b [3, 7]
43. Synset is of depth a to b in the hierarchy [6, 10]
44. The word appears in Dolch's Word List
45. The word appears in compound nouns
46. Maximum % of children including the term as a compound in the chain
47. The synset has hyponyms
48. The highest value in the chain for (frequency in brown + 1)/15 + (compounds in hyponym subtree containing word + 1)/5
49. Highest frequency in Brown + Gutenberg corpora combined in the chain
50. Maximum word length in chain
51. Maximum number of meronyms in the chain
52. Minimum word length in chain

Table 3: Voting rules

Several resources are used in the Rules listed in Table 2 and Table 3. The Brown corpus (Francis and Kucera 1964) is a one million word corpus of American English. The CHILDES corpus (MacWhinney 2000) is a collection of transcripts of early language acquisition. The CMU Pronouncing Dictionary (Weide 1998) is a machine-readable English pronunciation dictionary which maps words to phonetic translations. SEMCOR (Landes et al. 1998) is a PWN sense-tagged corpus. Dolch’s Word List (Dolch 1948) is a list of 510 words commonly spoken by kindergarteners. The Gutenberg Corpus is a subset of the public domain books available on Project Gutenberg (Gutenberg n.d.) and made available by the Natural Language Toolkit (Loper and Bird 2002).

5 Experiments

For the purpose of evaluation, we mapped the gold standard labels mentioned in Section 3 to

synsets in PWN. Some categories, such as *green seedless grapes* and *double knit pants*, did not have corresponding PWN synsets and were discarded. The labels also included four superordinates under which the psychology experiments and PWN had substantial incompatibilities, and these were also discarded. For example, whereas one superordinate in psychology experiments was taken to be *exercise equipment* (Markman and Wisniewski 1997), the three basic-level categories underneath this mapped to very different hypernym trees in PWN: *sports equipment*, *exercise device*, and even *athletic facility*.

We then divided the mapped categories into a train, development, and test set. This division was done manually at the superordinate level rather than completely randomly because there are several hypernyms with many basic-level categories labeled underneath them and having those split across sets may result in reporting better-than-real-world results as a result of learning location-specific patterns. Instead, splits have been made manually at higher levels in the hypernym hierarchy, though the available labels leave out significant portions of the PWN hypernym hierarchy so this is still imperfect. The number of categories in each set is shown in Table 4:.

	Train	Dev	Test	Total
Superordinate	7	8	9	24
Basic-level	29	24	25	78
Subordinate	10	22	18	50
Total	46	54	52	

Table 4: Summary of the labels for the experiments

5.1 Filtering Rule Experiments

Our first step was to set parameters on each individual Filtering Rule (Table 2) on the train set and select the promising rules based on their performance on the development set. The filtering rules are designed to provide accurate filtering to remove many non-basic categories before applying voting rules where the system can be more robust to errors by combining multiple rules. Filtering rules were tuned on the train set to not filter out any basic-level categories but to filter out as many superordinates and subordinates as possible.

Of the 39 proposed filtering rules, only 15 could be tuned to avoid filtering out basic-level categories while also filtering out subordinate or superordinate categories. These rules then generalized poorly to the development set, with only 5

rules performing at that same standard, though another 3 rules were kept which had worked on the train set and which did not filter anything out in the development set.

We also considered rules that filtered out a small number of basic-level categories in the train set while also filtering out a large number of non-basic-level categories, but these made even more mistakes on the development set and the mistakes did not overlap well. As a result, we left the decisions with imperfect filters to the Voting Rule portion of the system.

The final Filtering Rules chosen are shown with their parameters in Table 5.

1. Filter words with suffixes -ment or -age
10. Filter words with more adjective than noun senses
19. Filter capitalized words
21. Filter synsets with hyponym depth $(\min + \max) / 2$ outside the range [1,3.5]
23. Filter the top 6 levels of the hierarchy
24. Filter nodes with 7 levels below them
36. Filter all synsets with 65 or more siblings having no hyponyms
37. Filter all synsets with at least 92% percent of siblings having no hyponyms

Table 5: Chosen Filtering Rules with Parameters

5.2 Voting-Rule Experiments

The Voting Rules (Table 3) are applied to categories not already filtered by Filtering Rules. These are applied along each chain from the bottom to the top of the hypernym hierarchy. Like Filtering Rules, these rules are also applied to a category although evaluated in the context of a chain.

Using a greedy search starting with the most accurate Voting Rules, we identified a set of the rules which together enabled high accuracy on the development set. This combination is listed in Table 6.

40. Top frequency in the chain (sum of lemma frequencies in synset)
47. The synset has hyponyms
49. Highest frequency in Brown + Gutenberg corpora combined in the chain
51. Maximum number of meronyms in the chain

Table 6: Selected Voting Rules

We determined that by using these rules together, and only selecting categories with three of these Voting Rules being fulfilled, high accuracy

could be obtained on the development set. This does limit the number of basic-level categories that can be selected to one in each chain from the bottom to the top of the hypernym hierarchy. However, with three of the four rules only being fulfilled for one node in the chain, it is possible not to select a basic-level category in some chains.

6 Evaluation

Our system’s overall performance on the test data is listed in Table 7.

	Accuracy
Superordinate	100%
Basic-level	84%
Subordinate	44%
Overall	77%

Table 7: System Effectiveness

Accuracy is measured as the percentage of categories filtered (or not filtered) correctly based on the test data. Our system did well at filtering out superordinates, made a moderate number of mistakes filtering out basic-level categories, and was least successful at filtering out subordinates.

Just as when tuning the Filtering Rules on the development set, there was a substantial degradation in performance when extrapolating to the test set. Results on the development set, including subsystem breakdowns, are shown in Table 8.

The Filtering Rules provide the most substantial portion of the impact as measured on the development set with 77% accuracy, while the Voting Rules improved accuracy by 17 points to 94%. Comparing this to the results on the test data from Table 7, though, the system only performed as well as the Filtering Rules component did on its own on the development set. The generalization to the test data was better than expected for superordinates, worse than expected for the basic-level, and substantially worse than expected for subordinates.

	Filtering Rules	Filtering + Voting Rules
Superordinate	62%	88%
Basic-level	100%	92%
Subordinate	68%	100%
Total	77%	94%

Table 8: Accuracy on Development Set by Subsystem

The labels used here are sparse and non-random, collected from psychological research papers which had experimental reasons to control the ratios of subordinates to basic-level and superordinate categories. As an additional measure of system performance, we manually annotated a random set of 250 categories predicted to be at the basic-level by our system to estimate the precision of the system. Unfortunately, the estimated precision was only 10.4% (26 of 250). This annotation was done by two annotators using as a standard the property from Rosch et al. (1976) that basic-level categories are the most inclusive level at which a concrete picture of the category as a whole can be formed (previously mentioned in Section 2.1). This was chosen because it was a simple mental test to perform unlike many of the other properties and it was a pattern observed in all of the basic-level categories from that experiment (Rosch et al. 1976). That same experiment was one of the underlying sources of our labeled data. The inter-annotator agreement is 92% and the kappa score is 59%. Disputes were resolved through discussion.

Our system predicts 13,082 synsets are basic-level. Using our accuracy on the basic-level as a measure of recall, combined with our estimated precision, we estimate that there are around a total of 1,620 basic-level categories in PWN. This is a quantity we have not previously seen estimated.

Judging by the examples predicted as basic-level in our estimate of precision, there are some systematic errors in cases where the system predicts a category is basic-level but it turns out not to be. The most interesting type of mistake accounted for just over half of the mistakes. In this case, the categories were not basic-level but also were not clearly superordinates or subordinates either, at least as described in the psychology literature. In the psychology experiments, the focus is primarily on physical objects and organisms. Rather, there were many examples where the word was a noun describing an action (e.g. violence), denoting a relation (e.g. proportion), or denoting a role in a more complicated semantic frame (e.g. defalcation). It is possible we are being too restrictive in our labeling, but it appears to us that there are many nominal categories which describe things that belong to classifications other than superordinate, basic-level, and subordinate. This suggests the low precision is not just due to the prevalence of subordinates relative to basic-level categories (which is an issue). In addition, though,

much of the imprecision may be due to phenomena outside our limited theoretical label space.

We are making our list of predicted basic-level categories available for download at http://e22pii.com/research/files/GWC2018/predicted_basic_level_categories_synsets.txt. The labels we used, mapping labeled words in the psychology literature to PWN synsets, is also available at <http://e22pii.com/research/files/GWC2018/labels.txt>

7 Conclusion

We built a rule-based system to automatically identify basic-level categories using PWN. We were effective at including most basic-level categories and excluding superordinates, but not as effective at excluding subordinates.

We were 77% accurate overall at classifying our limited test data derived from psychological experiments. However, we have evidence that suggests these labels are based on a simplistic view that divides categories into 3 groups which do not appear to cover the full range of phenomena described by nouns. Outside of this limited test data, we manually annotated a sample of our system's predicted basic-level categories and found low precision with the majority of the mistakes outside of these three groups. This suggests that for greater broad-coverage accuracy it may be necessary to model cross-part-of-speech relationships and other phenomena that do not fit nicely in the existing label space.

In the future, we hope to refine and scale a labeling process using mechanical turk to build a larger and less-biased training set. We hope to rely on several of the tests in the psychology research, although modeling additional phenomena may require extending these tests. Additionally, we hope to build a machine learning-based system to turn the many weak rules we have into features that can help improve system performance, as well as to evaluate the system on a much larger test set with this rule-based system as a baseline.

References

- Roger Brown (1958). "How shall a thing be called?" *Psychological review* 65(1): 14.
- Edward William Dolch (1948). *Problems in reading*, Garrard Press.
- D. Dorr B. Habash N. Helmreich S. Hovy E. Green R. Levin L. Miller K. Mitamura T. Rambow O. Farwell (2009). "Interlingual annotation of multilingual text

- corpora and FrameNet." *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS* **200**: 287-318.
- Christiane Fellbaum (1998). *WordNet : an electronic lexical database*. Cambridge, Mass, MIT Press.
- N Francis and H Kucera (1964). "Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Department of Linguistics, Brown University, Providence, USA)." *icame. uib. no/brown/bcm. html* (accessed 12 October 2010).
- Roberta Michnick Golinkoff, Margaret Shuff-Bailey, Raquel Olguin and Wenjun Ruan (1995). "Young children extend novel words at the basic level: Evidence for the principle of categorical scope." *Developmental Psychology* **31**(3): 494.
- Rebecca Green (2006). "Vocabulary alignment via basic level concepts. Final Report 2003 OCLC/ALISE Library and Information Science Research Grant Project." *Dublin, OH: OCLC Online Computer Library, Inc. Retrieved May 14*: 2008.
- Project Gutenberg (n.d.). from <http://www.gutenberg.org>.
- Rubén Izquierdo, Armando Suárez and German Rigau (2007). *Exploring the automatic selection of basic level concepts*. Proceedings of RANLP, Citeseer.
- Pierre Jolicoeur, Mark A Gluck and Stephen M Kosslyn (1984). "Pictures and names: Making the connection." *Cognitive psychology* **16**(2): 243-275.
- María K Jónsdóttir and Randi C Martin (1996). "Superordinate vs basic level knowledge in aphasia: A case study." *Journal of Neurolinguistics* **9**(4): 261-287.
- Shari Landes, Claudia Leacock and Randee I Tengi (1998). "Building semantic concordances." *WordNet: an electronic lexical database* **199**(216): 199-216.
- Steve Legrand (2006). "Word Sense Disambiguation with Basic-Level Categories." *Advances in Natural Language Processing. Ed. Alexander Gelbukh, Research in Computing Science* **18**: 71-82.
- Guillaume Lemaitre and Laurie M Heller (2013). "Evidence for a basic level in a taxonomy of everyday action sounds." *Experimental brain research* **226**(2): 253-264.
- Shu-Yen Lin, Cheng-Chao Su, Yu-Da Lai, Li-Chin Yang and Shu-Kai Hsieh (2009). "Assessing text readability using hierarchical lexical relations retrieved from WordNet." *Computational Linguistics and Chinese Language Processing* **14**(1): 45-84.
- Edward Loper and Steven Bird (2002). *NLTK: The natural language toolkit*. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1, Association for Computational Linguistics.
- Brian MacWhinney (2000). *The CHILDES project: The database*, Psychology Press.
- Arthur B Markman and Edward J Wisniewski (1997). "Similar and different: The differentiation of basic-level categories." *Journal of Experimental Psychology: Learning, Memory, and Cognition* **23**(1): 54.
- George A. Miller (1995). "WordNet: A Lexical Database for English." *Communications of the ACM*. **38**(11): 39.
- Gregory L Murphy and Edward J Wisniewski (1989). "Categorizing objects in isolation and in scenes: What a superordinate is good for." *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**(4): 572-586.
- Ayn Rand (1966). "Introduction to Objectivist Epistemology (II)." *The Objectivist*(08/66).
- Abebe Rorissa and Hemalata Iyer (2008). "Theories of cognition and image categorization: What category labels reveal about basic level theory." *Journal of the American Society for Information Science and Technology* **59**(9): 1383-1392.
- Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson and Penny Boyes-Braem (1976). "Basic objects in natural categories." *Cognitive psychology* **8**(3): 382-439.
- James W Tanaka and Marjorie Taylor (1991). "Object categories and expertise: Is the basic level in the eye of the beholder?" *Cognitive psychology* **23**(3): 457-482.
- Zhongyuan Wang, Haixun Wang, Ji-Rong Wen and Yanghua Xiao (2015). *An Inference Approach to Basic Level of Categorization*. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM.
- Robert L Weide (1998). "The CMU pronouncing dictionary." URL: <http://www.speech.cs.cmu.edu/cgibin/cmudict>.

African Wordnet: facilitating language learning in African languages

Sonja Bosch

University of South Africa (UNISA)
Pretoria, South Africa.

boschse@unisa.ac.za

Marissa Griesel

University of South Africa (UNISA)
Pretoria, South Africa.

griesm@unisa.ac.za

Abstract

The development of the African Wordnet (AWN) has reached a stage of maturity where the first steps towards an application can be attempted. The AWN is based on the expand method, and to compensate for the general resource scarceness of the African languages, various development strategies were used. The aim of this paper is to investigate the usefulness of the current isiZulu Wordnet in an application such as language learning. The advantage of incorporating the wordnet of a language into a language learning system is that it provides learners with an integrated application to enhance their learning experience by means of the unique sense identification features of wordnets. In this paper it will be demonstrated by means of a variety of examples within the context of a basic free online course how the isiZulu Wordnet can offer the language learner improved decision support.

1 Introduction

The development of the African Wordnet (AWN) containing wordnets for five African languages, namely Setswana (TSN), isiXhosa (XHO), isiZulu (ZUL), Sesotho sa Leboa (NSO) and Tshivenda (VEN), has reached a stage of maturity where the first steps towards an application can be attempted.

The aim of this paper is to investigate the usefulness of the current isiZulu Wordnet¹ in an application such as language learning. Against the background of a multi-lingual scenario of eleven official languages of South Africa, the

National Development Plan 2030 (NDP) enjoins all South Africans to learn at least one indigenous language as part of nation-building and social cohesion (South African Government, 2017). Such an imperative calls for prioritizing the development of language learning courses. Currently, very limited material exists to support language learners who wish to focus on improving their skills in their own time, without much cost involved and with actual real-world examples.

Most of the material developed to engage learners of an African language, are either taught in time-consuming (university) classes or after purchasing expensive software with generic course content and teaching style. An exception is a set of basic free online courses **LEarn To Speak an African Language** (*LetsAL*) (University of South Africa, 2010) that were developed for first time language learners of some of the indigenous South African languages, namely Setswana, isiXhosa, isiZulu, Sesotho and Sesotho sa Leboa.

Although the African languages can still be regarded as under-resourced (cf. the resource audit performed by Grover, Van Huyssteen and Pretorius, 2011) the African Wordnet (AWN) project reported in Bosch and Griesel (2017), has played a significant role in filling the gap as available source of data for further human language technology and linguistic research. The individual African languages wordnets, with less than 20,000 synsets per language, are still relatively small in comparison to some of the large wordnets such as the Princeton WordNet² (117 659 synsets for English) and the FinnWordNet³ (120 449 synsets for Finnish). Nevertheless, we investigate the possibility of using the AWN effectively as support for language learners in a computer assisted language learning (CALL) environment. Our focus is on isiZulu.

¹ Available for download from <https://rma.nwu.ac.za/index.php/resource-catalogue/african-wordnet-isizulu.html>. Please note that this catalogue will soon be transitioned to the newly constituted South African Centre for Digital Language Resources (SADiLaR) <https://rma.nwu.ac.za/index.php/about-sadilar/>

² <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

³ <http://www.ling.helsinki.fi/en/lt/research/finnwordnet/news.shtml>

In this paper, the interaction between the AWN and *LetsAL* will be explored as a first step towards moving to a more integrated CALL system, with a focus on improving user interaction. A brief background to both the AWN and *LetsAL* will be provided, before the contents of the two resources is assessed. We also discuss ways to integrate the AWN into the current *LetsAL* system, with a view on moving towards an intelligent CALL (iCALL) application as proposed in Bosch and Griesel (2013). Fast-tracking the expansion of the AWN with *LetsAL* data is also shown to be effective, even with minimal additional resources.

2 Background

In this section, we discuss the status quo of the African Wordnet as potential language learning support resource and provide background on the *LetsAL* courses.

2.1 The African Wordnet Project

The African Wordnet Project (AWN) deals with the development of aligned wordnets for African languages spoken in South Africa (i.e. languages belonging to the Bantu language family) as multilingual knowledge resources with the long-term purpose of including a wide variety of related languages also from other parts of Africa. Bosch and Griesel (2017) discuss the development strategies implemented for building a first version of the AWN for five languages in parallel.

The expand model was followed from the onset since, as stated by Ordan and Wintner (2007), this model provides a tested structure for building a new resource and is therefore typically the choice for less resourced languages. During the first development phases, the AWN used the extended Common Base Concepts list from the EuroNet Project⁴ as well as the Core Concepts list designed for the BalkaNet Project⁵ to extract English synsets for linguists to include and translate into the African languages concerned. However, it soon became clear that a more localised approach was needed. The seed lists mentioned above contain many concepts that are not lexicalised in the African context.

As the development team became more experienced, and appropriate lexical resources became available, more localised support could be given in the form of frequency-based seed terms

and semi-automatic linking of lemmas from bilingual wordlists and the PWN (Princeton University, 2017). Opportunities to harvest usage examples from online corpora also contributed to promising results.

Throughout the development, the AWN used the DEBVisDic editor tools (DEBVisDic: WordNet editor and browser, n.d.) which are distributed as freeware and were recently re-launched as a web application (Rambousek and Horak, 2016).

2.2 The current *LetsAL* environment

In an initiative to actively promote African languages in anticipation of the Soccer World Cup that took place in South Africa in 2010, a modest beginning was made with the development of free online courses that focus on basic language skills (Mischke, n.d.).



Figure. 1. Learn To Speak an African Language (*LetsAL*)

The so-called *LetsAL* (**LE**arn **T**o **S**peak an **A**frican **L**anguage) courses are aimed at first time language learners, who are offered basic lessons covering 10 general themes, including greetings and courtesies; asking for help; numbers, days, months, seasons; question words, quantities, weather; banks, taxis and restaurants; transportation and finding your way; touring and socializing; at the filling station; the human body and ailments; as well as shopping and sport. Each theme is explored via a list of appropriate vocabulary and phrases accompanied by translations, a short dialogue as well as a video to contextualize the content. Noteworthy cultural customs are also shared, such as who greets first or the role of a traditional doctor. This Open Educational Resource (OER) created by the University of South

⁴ See <http://www.ilc.uva.nl/EuroWordNet/>

⁵ See <http://www.dblab.upatras.gr/balkanet/>

Africa (2010) is actively used by banks and private schools for language learning purposes.

When compared to similar products involving international languages, several areas begging for further development become apparent, e.g. inclusion of real-time interaction between learners; games and fun content included for assessment purposes; etc. (see Bosch and Griesel, 2013). Currently the courses also do not offer links to other external resources. It is in this area where the AWN, already freely available for the five languages, could play a central role in future improvements to these courses.

3 Incorporating AWN in *LetsAL* as additional reference material

The advantages of incorporating all available natural language processing (NLP) resources for a language into a CALL system, is that learners are offered an integrated application to enhance their understanding of the subject matter. In a related study, Winiwarer (2011) describes COLLIE – a collaborative language learning and instruction environment for Japanese foreign language learners. This system combines advanced NLP tools such as machine translation and complex analysers with the English PWN and the Japanese Wordnet to provide learners with translations of Japanese webpages, as well as detailed information on the word level. “All this information is very useful as decision support for selecting the word sense, reading, and English translation of a Japanese word” (Winiwarer, 2011:3761).

While NLP support is still limited for isiZulu, it is important that we integrate the resources that are freely available and use them to the best advantage of language learners. The learner improved decision support that can be offered by the isiZulu AWN, which typically covers a variety of semantic relations including synonymy and antonymy, along with usage example sentences and definitions, will be demonstrated by means of a selection of examples within the context of *LetsAL*. Each of the example words already occur in the *LetsAL* course material, but only with the English translation and no further information on meaning nuances or associated misconceptions in isiZulu. To illustrate the need for more disambiguating context and the type of information that a language learner might also find useful, we combined information from the isiZulu AWN and the relevant information from the English PWN (Princeton University, 2017) in tables

1 – 10. The significance of these examples is discussed, before a suggestion for improving the current *LetsAL* environment with similar information is made at the end of this section.

IsiZulu and most of the languages belonging to the Bantu language family are known as tone languages in which pitch variation plays a role in conveying lexical as well as grammatical distinctions⁶ (cf. Poulos and Msimang, 1998:543, and Heine and Nurse, 2000:152). The two basic tone levels that can be distinguished in isiZulu, namely high (H) and low (L), may be marked by means of placing acute and grave accents above the syllables of a word, or by placing the symbols H and L after the word, e.g.

indòdà (man) OR *indoda* HLL (man)

In the following pairs of nouns and verb stems it is illustrated how tone distinguishes meaning:

ìthàngá (thigh) HLH
ìthàngà (pumpkin) HLL

ìnyàngá (moon; month) HLH
ìnyàngà (herbalist) HLL

-dùmà (be tasteless) HL
-dùmá (roar, be famous) LH

Since tone is not, however, marked in the standard orthography of isiZulu, language learners are often confronted with a difficult choice between two meanings of (seemingly) the same word. It should also be noted that tones are not absolute. Tones documented in dictionaries usually refer to tones of the word as it occurs in isolation or in sentence final position. In other sentence positions, tones may undergo various changes.

3.1 Example 1: *ithanga*

A case of homonymy which is an example of potential confusion for language learners is the noun *ithanga* which occurs in the human body theme of *LetsAL*, with the English translation of “thigh”. In a different context such as vegetables, the meaning of *ithanga* is completely different and unrelated, namely “pumpkin”. The differentiated meanings of the orthographic form of *ithanga* (thigh/pumpkin) are illustrated in Tables 1 and 2.

⁶ We only concentrate on lexical distinctions in this paper.

POS: n; ID PWN 2.0: ENG20-05243922-n; ID PWN 3.1: 05569882
Synonyms: <i>ithanga</i> :1 Definition: <i>isitho somuntu esiphakathi kwedolo ne-nqulu</i> Usage: <i>umdlalikazi wathleka emcimbini wama-Grammy esho ngelokwe eliveza lonke ithanga</i> Domain: anatomy
Synonyms: thigh :1 Definition: the part of the leg between the hip and the knee Domain: anatomy

Table 1. *ithanga* (anatomy domain)

POS: n; ID PWN 2.0: ENG20-07263505-n; ID PWN 3.1: 07751486
Synonyms: <i>ithanga</i> :2 Definition: <i>isitshalo esimila phansi esinombala ophuzi</i> Usage: <i>kanti-ke abangani abathathu bapheka ithanga elikhulu kakhulu</i> Domain: gastronomy
Synonyms: pumpkin :2 Definition: usually large pulpy deep-yellow round fruit of the squash family maturing in late summer or early autumn Domain: gastronomy

Table 2. *ithanga* (gastronomy domain)

3.2 Example 2: *inyanga*

The *LetsAL* example *inyanga* is a noun with two related meanings “moon” and “month” though in different domains, namely *time_period* and *astronomy* respectively, as shown in Tables 3 and 4. It is significant that the same orthographic noun has a third meaning “herbalist” (in the *medicine* domain) which however is unrelated to the former two meanings, therefore representing a homonymous relationship (cf. Table 5).

POS: n; ID PWN 2.0: ENG20-14348156-n; ID PWN 3.1: 15234209
Synonyms: <i>inyanga</i> :1 Definition: <i>isikhathi sezinsuku ezingamashumi amathathu</i> Usage: <i>inyanga yesibili manje</i> Usage: <i>ngiphumule inyanga eyodwa ngemuva kwama-Olympics</i> Domain: <i>time_period</i>
Synonyms: calendar month :1, month :1 Definition: one of the twelve divisions of the calendar year Usage: he paid the bill last month Domain: <i>time_period</i>

Table 3. *inyanga* (*time_period* domain)

POS: n; ID PWN 2.0: ENG20-08772174-n; ID PWN 3.1: 09381255
Synonyms: <i>inyanga</i> :1 Definition: <i>isathalayithi yoMhlaba ekhanyisa esibhakabhakeni</i> Usage: <i>inyanga iphuma ebusuku</i> Usage: <i>wabona inyanga eyisiliva iqala ukuphakama phezulu esibhakabhakeni</i> Domain: astronomy
Synonyms: moon :1 Definition: the natural satellite of the Earth Usage: the average distance to the moon is 384,400 kilometers Usage: men first stepped on the moon in 1969 Domain: astronomy

Table 4. *inyanga* (astronomy domain)

POS: n; ID PWN 2.0: ENG20-09516232-n; ID PWN 3.1: 10191128
Synonyms: <i>inyanga</i> :1 Definition: <i>umuntu onolwazi lokwelapha izifo ngemithi</i> Usage: <i>uGcabashe umbikele ukuthi akaphilile udinga ukubona inyanga</i> Usage: <i>inyanga ithi ifuna ukuthenga umuthi omhlophe</i> Domain: medicine
Synonyms: herbalist :1, herb doctor :1 Definition: a therapist who heals by the use of herbs Domain: medicine

Table 5. *inyanga* (medicine domain)

3.3 Example 3: *siza/lekelela*

Synonymy, the central relation encoded in word-nets, is represented in Table 6 by means of the the *LetsAL* examples *siza*, *lekelela* (help, assist), two isiZulu verbs that are exchangeable in most contexts.

POS: v; ID PWN 2.0: ENG20-02472355-v; ID PWN 3.1: 02553283
Synonyms: <i>lekelela</i> :1, <i>siza</i> :1 Definition: <i>ukwelekelela noma ukusiza ekwenzeni okuthile</i> Usage: <i>le nhlango ilekelela ekufundiseni izingane eziqhamuka emakhaya</i> Usage: <i>iholo lakhe lisiza ukuthi akwazi ukuphilisa umndeni</i>
Synonyms: help :1, assist :1, aid :1 Definition: give help or assistance; be of service Usage: Everyone helped out during the earthquake Usage: Can you help me carry this table? Usage: She never helps around the house

Table 6. *siza – lekelela* (synonymous verbs)

3.4 Example 4: *luhlaza/vuthiwe*

Adjectives in wordnets are typically organized in terms of antonymy (Princeton University, 2017). The *LetsAL* examples *luhlaza* (raw) in Table 7 and *vuthiwe* (cooked) in Table 8 reflect the semantic contrasts involved between the two adjectives. For the purposes of wordnet construction we adhere to the English part-of-speech term “adjective”, although this category includes so-called adjective as well as relative and verb stems as noun qualifiers in isiZulu (also see Le Roux et al., 2008:276).

POS: a; ID PWN 2.0: ENG20-00589776-a; ID PWN 3.1: 00622052
Synonyms: <i>luhlaza</i> :4 Definition: <i>ukudla okungavuthiwe</i> Usage: <i>Babengawusebenzisi umlilo, babemane bayifukuthe bayidle luhlaza inyama</i> Domain: gastronomy -->> [near_antonym] <i>vuthiwe</i> :1
Synonyms: raw :3 Definition: not treated with heat to prepare it for eating Domain: gastronomy -->> [near_antonym] cooked:1

Table 7. *luhlaza* (raw)

POS: a; ID PWN 2.0: ENG20-00586933-a; ID PWN 3.1: 00618376
Synonyms: <i>vuthiwe</i> :1 Definition: <i>kuphekiwe kulungele ukudliwa</i> Usage: <i>ukudla okuvuthiwe kudayiswa emgwaqweni</i> Domain: gastronomy -->> [near_antonym] raw:3
Synonyms: cooked :1 Definition: having been prepared for eating by the application of heat Domain: gastronomy -->> [near_antonym] raw:3

Table 8. *vuthiwe* (cooked)

For language learners, it would also be useful to take note of the additional antonymous relations of *luhlaza* (green, unripe) and *vuthiwe* (ripe, mature) to avoid confusion, as illustrated in Tables 9 and 10.

3.5 Implementation

We propose that the current *LetsAL* environment be enriched with the data from the AWN in much the same way as suggested above and in the COLLIE project. Synonyms, usage examples, definitions and other semantic relations, as described in Tables 1 to 10, can all assist

POS: a; ID PWN 2.0: ENG20-01442460-a; ID PWN 3.1: 01497045
Synonyms: <i>luhlaza</i> :5 Definition: <i>okungavuthiwe kwezithelo</i> Usage: <i>abanye badla izithelo eziluhlaza</i> -->> [near_antonym] <i>vuthiwe</i> :2
Synonyms: green :3, unripe :1, unripened :1, immature :4 Definition: not fully developed or mature; not ripe Usage: unripe fruit Usage: fried green tomatoes Usage: green wood -->> [near_antonym] ripe:1, mature:4

Table 9. *luhlaza* (green, unripe)

POS: a; ID PWN 2.0: ENG20-01441835-a; ID PWN 3.1: 01496321
Synonyms: <i>vuthiwe</i> :2 Definition: <i>okulungele ukudliwa noma ukuphuzwa</i> Usage: <i>izitshalo zikabhekilanga zivuthiwe lapho umbala ngemuva kwesihloko ushintsha</i> -->> [near_antonym] <i>luhlaza</i> :5
Synonyms: ripe :1, mature :4 Definition: fully developed or matured and ready to be eaten or used Usage: ripe peaches Usage: full-bodies mature wines -->> [near_antonym] green:3, unripe:1, unripened:1, immature:4

Table 10. *vuthiwe* (ripe, mature)

learners to understand not only the broad meaning of words in context, but can also point out subtle differences and potential language specific pitfalls such as those involving orthography.

Figure 2 shows a mockup of the value-added website, including a pop-up window with information for the isiZulu word *ithanga* (thigh) in the anatomy domain. A link to the PWN (Princeton University, 2017) or the isiZulu AWN entry (at the top of the pop-up window) gives learners access to the full synset without cluttering the *LetsAL* environment. The domain, usage example(s) and definition(s) are also shown. If any other disambiguating or relevant relations such as polysemy or antonymy are identified for a particular synset, the information will also be presented in the pop-up circle. In the example (Figure 2), a hypernym for *ithanga* is *isitho* (limb). We deliberately avoid the use of the terms “synset” and “hypernym” since users of *LetsAL* are assumed to be general language learners who might not be familiar with these terms.

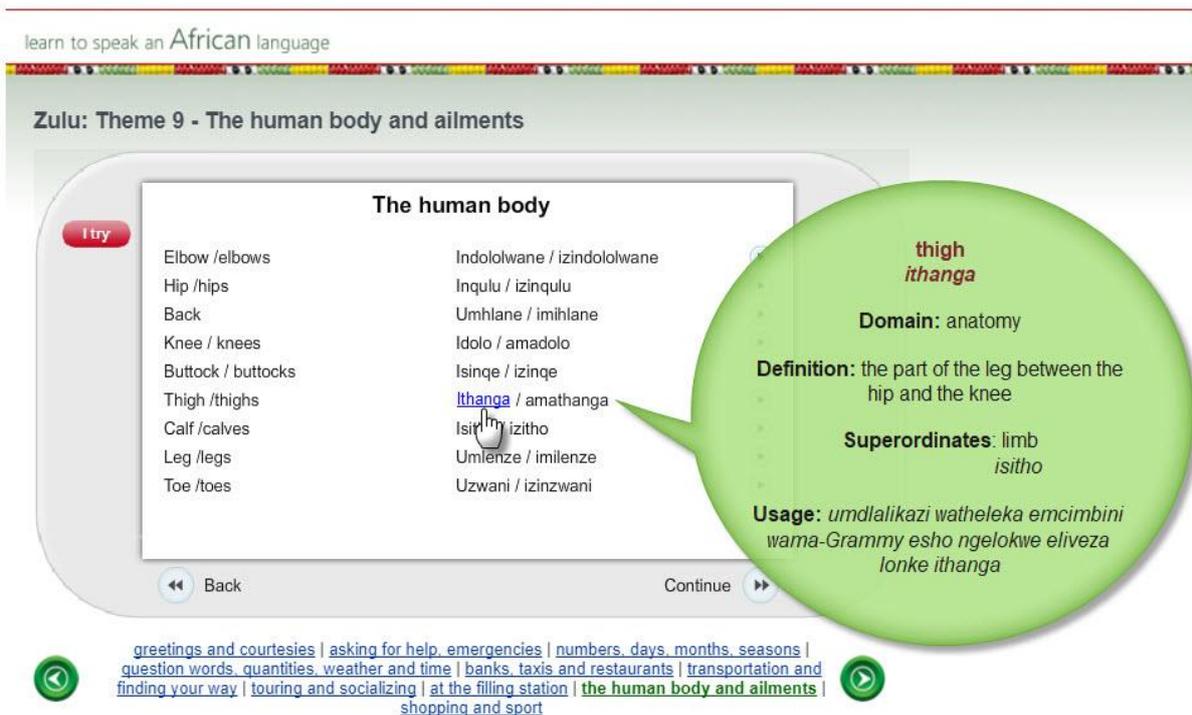


Figure 2. AWN data enriching the *LetsAL* learner experience

4 Expanding the AWN with *LetsAL*

Marrying the AWN and *LetsAL* to some extent, offers valuable advantages for users of both resources. Language learners will benefit greatly from having the additional information included in the AWN readily available in the *LetsAL* interface (see Section 3). In turn, the AWN will also become a more balanced resource by including fundamental meanings, such as those presented in a beginner course for L2 or foreign language learners.

To measure the amount of overlap between the two resources, a list of unique words was first extracted, and function words removed. This resulted in a list of 486 shared words. Of this list, 94 were already included in the isiZulu AWN and therefore also removed from our experiment, resulting in a list of 392 words from the *LetsAL* content that are not covered in the AWN yet. This clearly shows that there is still some work to be done on the isiZulu AWN to achieve a balanced coverage of the most basic vocabulary, as taught to language learners. To this end, methods previously employed in the AWN to speed up development of new synsets, namely semi-automatic linking of information from bilingual wordlists to information from the PWN. In short, this method uses the minimum amount of additional resources to extract potential synsets from the English for each word in the African language. A human expert then assesses the validity

of each link before it is included in the AWN. The results when applying this method to the isiZulu *LetsAL* data can be seen in Table 11.

POS category	Lemmas in <i>LetsAL</i>	Possible links identified	Correct matches added to AWN
Nouns	259	130	119
Verbs	126	51	46
Adjectives	9	7	7
TOTAL	392	188	172

Table 11. Results of the linking experiment

A human expert provided a bilingual list (English – isiZulu) of meaningful stems, derived from all the vocabulary and dialogues included in the *LetsAL* course for isiZulu. The English stems with their associated synsets were extracted from PWN (Princeton University, 2017) and limited to those with attributes “sense:1”. We also split the data on POS categories assigned by a linguist to limit the choices necessary in the validation step. The lemmas that were already included in the AWN were not presented for validation again as the goal of this experiment was rather to improve the coverage of the AWN, than the depth of already included synsets. That important aspect will be covered in a next pass on the data. Following this method grew the isiZulu AWN with 172 additional synsets in a matter of a few hours – now totaling 10 954 synsets.

5 Conclusion and future work

The *LetsAL* infrastructure lends itself perfectly to inclusion of data from the AWN as additional on-line reference source. Incorporating this feature for isiZulu into the live system will be our primary focus, where after the same improvement will be made to the other three languages for which a wordnet already exists (namely isiXhosa, Setswana and Sesotho sa Leboa). The AWN project is about to enter a new development stage and it is therefore planned to add Sesotho as a sixth language. As soon as a wordnet is available, it can be incorporated into the *LetsAL* course.

Growing the AWN is also a priority and the project team is looking at the best method by which to select new synsets for inclusion. The comparison with *LetsAL* presented here will serve as valuable input for the next phase of development, especially the low initial coverage of basic terminology, as shown in Section 4. The team is also performing comparisons with what is currently included in the AWN and a base list of terms for the African languages compiled by Snider and Roberts (2004). The so-called SIL-CAWL contains 1700 words in various categories such as *Man's Physical Being* and *Environment*. A further aim in the next phase of development of the AWN will be to fast-track inclusion of usage examples. For this experiment, an open-source corpus management system named NoSketch Engine (Rychlý, 2007) was used to manually look up usage examples in three small, but freely available online corpora for isiZulu, created in the Wortschatz project at the University of Leipzig⁷.

Future work will include optimising a semi-automatic process by which developers of the AWN are presented with the best candidate sentences from the corpora, to be edited and included as usage examples. Roughly 4 000 isiZulu synsets in the AWN do not have any usage examples added yet, so speeding up development in this category is essential.

Another promising resource that needs to be explored for possible inclusion in an iCALL system such as this, is ImageNet (2016)). This extension of PWN 3.0 includes a large image database organised according to the WordNet hierar-

chy to further disambiguate the meaning of each node. The example in Table 2 (*pumpkin*) could for instance be further enhanced by adding a link to its ImageNet counterpart (see <http://image-net.org/synset?wnid=n12158443>). Images could also be used in an assessment component to *LetsAL* or to base interactive games on – two important aspects that are envisaged for future work as well.

Acknowledgements

The authors would like to thank: the South African National Research Foundation, National HLT Network, Department of Arts and Culture, the University of South Africa's Women-in-Research Fund as well as the South African Centre for Digital Language Resources for providing funding in the various phases of the AWN project; as well as reviewers and conference participants for valuable inputs to the paper.

References

- Sonja E. Bosch and Marissa Griesel. 2013. *Language Learning in a Modern African Context: Enhancing the User's Experience within an ODeL and Mobile Framework*. Presented at the 6th International Conference for ICT in Language Learning. 14-15 November 2013. Florence, Italy. ISBN 978-88-6292-423-8. <https://conference.pixel-online.net/ICT4LL2013/common/download/Paper/pdf/125-ELE10-FP-Bosch-ICT2013.pdf> Accessed on 18 September 2017.
- Sonja E. Bosch and Marissa Griesel. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator* 38(1), a1351. <https://doi.org/10.4102/lit.v38i1.1351>
- DEBVisDic: WordNet editor and browser, n.d. https://deb.fi.muni.cz/proj_debvisdic.php. Accessed on 25 October 2017.
- Aditi Sharma Grover, Gerhard B. Van Huyssteen and M.W. Pretorius. 2011. South African human language technology audit. *Language Resources and Evaluation*, Vol. 45, pp. 271-288.
- Bernd Heine and Derek Nurse. 2000. *African Languages: An Introduction*. Cambridge: Cambridge University Press.
- ImageNet. 2016. <http://www.image-net.org> Accessed on 12 February 2018.
- Language Resource Management Agency (RMA). 2013. <http://rma.nwu.ac.za/> Accessed on 18 September 2017.
- Jurie Le Roux, Koliswa Moropa, Sonja E. Bosch and Christiane Fellbaum. 2008. Introducing the African

⁷ See <https://nlp.fi.muni.cz/trac/noske> for more details on NoSketch Engine and <http://wortschatz.uni-leipzig.de/en> for information on the Wortschatz project. Access to the isiZulu corpora in NoSketch Engine is via http://cql.corpora.uni-leipzig.de/?corpusId=zul_mixed_2014.

- Languages Wordnet, in Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen (eds) *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary, 2008, pp. 269-280. Szeged: University of Szeged, Department of Informatics. <http://www.inf.u-szeged.hu/projectdirs/gwc2008/>. Accessed on 19 September 2017.
- Gerda Mischke. n.d. Virtual classroom to teach a free online course in African languages. <http://www.oerafrica.org/ResourceDownload.aspx?id=37668>. Accessed on 26 August 2013.
- George Poulos and Christian T. Msimang. 1998. *A linguistic analysis of Zulu*. Cape Town: Via Afrika.
- Princeton University. 2017. WordNet – A lexical database for English. <https://wordnet.princeton.edu/> Accessed on 18 September 2017.
- Adam Rambousek and Aleš Horák. 2016. DEBVisDic: Instant Wordnet building. In V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016 (GWC2016)*, Bucharest, Romania, January 25–29, pp. 317–321.
- Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 2007. p. 65-70. ISBN 978-80-210-4471-5.
- Keith Snider and James Roberts. 2004. SIL comparative African word list. *The Journal of West African Languages* 31(2):73–122. <http://www-01.sil.org/silewp/2006/silewp2006-005.pdf> Accessed on 14 August 2017.
- South African Government. 2017. <https://www.gov.za/issues/national-development-plan-2030>. Accessed on 13 September 2017
- University of South Africa. 2010. Learn To Speak An African Language: Free Online Language Learning Courses. <http://www.unisa.ac.za/sites/corporate/default/Unisa-Open/OER-@-Unisa/Learn-to-speak-an-African-Language>. Accessed on 26 August 2017.
- Werner Winiwarter. 2011. COLLIE – Towards a Collaborative Language Learning and Instruction Environment. In T. Bastiaens & M. Ebner (Eds.), *Proceedings of ED-MEDIA 2011--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 3756-3765). Lisbon, Portugal: Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/p/38400/> Accessed on 14 August 2017.

Hindi Wordnet for Language Teaching: Experiences and Lessons Learnt

Hanumant Redkar¹, Rajita Shukla², Sandhya Singh¹,
Jaya Saraswati¹, Laxmi Kashyap¹, Diptesh Kanojia¹,
Preethi Jyothi¹, Malhar Kulkarni¹ and Pushpak Bhattacharyya¹

¹Indian Institute of Technology Bombay, Mumbai, India.

²Bennett University, Greater Noida, India.

{hanumantredkar, rajita.shukla38, sandhya.singh}@gmail.com,

{jayasaraswati, laxmi.kashyap, dipteshkanojia}@gmail.com,

{preethijb, malharku and pushpakbh}@gmail.com

Abstract

This paper reports the work related to making Hindi Wordnet¹ available as a digital resource for language learning and teaching, and the experiences and lessons that were learnt during the process. The language data of the Hindi Wordnet has been suitably modified and enhanced to make it into a language learning aid. This aid is based on modern pedagogical axioms and is aligned to the learning objectives of the syllabi of the school education in India. To make it into a comprehensive language tool, grammatical information has also been encoded, as far as these can be marked on the lexical items. The delivery of information is multi-layered, multi-sensory and is available across multiple digital platforms. The front end has been designed to offer an eye-catching user-friendly interface which is suitable for learners starting from age six onward. Preliminary testing of the tool has been done and it has been modified as per the feedbacks that were received. Above all, the entire exercise has offered gainful insights into learning based on associative networks and how knowledge based on such networks can be made available to modern learners.

1 Introduction

A Wordnet is a large digital lexical database of a language in which information is organised around cognitive synonym sets or synsets (Fellbaum, 1998). The underlying basis of such organization are the word association studies in psycholinguistics, which proved that our mental lexicon is structured on associations, *i.e.* an appear-

ance of one entity entails the appearance of the other in the mind. Thus, it was found that subjects respond quicker than normal to the word ‘nurse’ if it follows a highly associated word such as ‘doctor’ (Church and Hanks, 1990). This property of the mental lexicon is structurally built in the Wordnets and manifests itself in the lexical and semantic relations which are encoded in it. Thus, a Wordnet is a ready resource of vocabulary of a language, which captures associative learning in its structure. Conventional sources of vocabulary learning, such as the dictionaries and thesauri, do not have these relations due to the very nature of their composition. This is the motivation to present Hindi Wordnet as a tool for vocabulary learning and teaching.

The second motivation is the fact that education is undergoing rapid digitalization. Innovative instruction techniques, which can cater to the tenets of anywhere, anytime, any size learning, flipped classroom approach and blended learning environments, are the need of the hour. Such technology based learning solutions can help in better learner engagement in classrooms and can also be used in informal teaching–learning environments, as the delivery of knowledge is in a multi-sensory mode and is available across various digital platforms. In the form of a digital resource, the language learning aid that is presented here also seeks to redress the long standing problem of the burden of the school bag of young learners in India. This pedagogical application of Hindi Wordnet will address all the above issues and provide a resource which will cater to the various gap areas of learning.

The rest of the paper is organized as follows: Section 2 provides the related work. Section 3 discusses the need for association capturing in language learning, which provides the basis for using Wordnet as a language learning resource. Section 4 presents the digital aid developed for Hindi language teaching and learning. Section 5 discusses the process, experiences and lessons learnt while

¹<http://www.cfil.t.iitb.ac.in/wordnet/webhwn/index.php>

developing this aid. Section 6 briefs the field test and user feedback followed by conclusion and future work in section 7.

2 Related Work

In digital educational technology, the literature shows that psychological aspects of language learning has been explored under various conditions due to continuous advancements on the technological front. The research indicates that multi-modal learning has always resulted in better retention (Dale, 1969). When the information enters the system through various senses, it helps the brain to circumvent the limited processing capabilities of each individual senses and allows for greater total information to be processed (Clark and Paivio, 1991). With technology in place, the ease of multi-modal learning environments have been studied in different settings (Mayer and Moreno, 2003; Moreno and Mayer, 2007; Shams and Seitz, 2008; Sankey et al., 2010). Mobile Assisted Language Learning (MALL) is also being explored as mobile technologies are becoming an integral part of lifestyle. The findings (Yang, 2013) show that MALL has not reached its potential and it is moving towards being the new stage of Computer Assisted Language Learning (CALL). Pedagogical experts have stressed on the need for improving various approaches to enhance the willingness of the learners for self-directed technology to maximize the technology potential for language learning (Lai et al., 2016). The use of gamification is seen as an effective pedagogical strategy which can engage and motivate the learner to learn in a relaxed environment, which is fundamental to any learning (Werbach and Hunter, 2012; Figueroa Flores, 2015).

In language learning, using semantic network relations for learning new word helps in better understanding of its meaning (Lin, 1997). The wordnet, a semantic based rich lexical resource, has been used for various language learning applications such as - the semantic and lexical relations between synsets enables the learners to know the connotations of a word along with its various possible contexts (Hiray, 2015). A gamification system based on wordnet was used to assess the depth of word knowledge of a learner (Brumbaugh, 2015). A system was experimented for similar looking and near synonyms word learning based on wordnet for English language learners (Sun et al., 2011).

3 Language Learning through Association Capturing

3.1 Need for association capturing

Association capturing lies at the core of language learning. The term association is used here to refer to the connection or relation between ideas, concepts, or words, which exist in the human mind and manifest in such a way that an appearance of one entity entails the appearance of the other in the mind (Sinopalnikova, 2004). Learning or instruction strategies must be able to encapsulate this association for the creation of better pedagogical techniques, as associative networks help not only in understanding new knowledge but also to retain it firmly in the mind.

The understanding of how the meaning of a word is understood and retained in the human mind is of crucial importance as vocabulary plays an important role in all the competencies of language learning, such as speaking, reading, and writing. Methods of vocabulary learning too have moved beyond the traditional ways which were based on the behavioristic theory (Demirezen, 1988) of language learning to the modern methods of vocabulary learning (Nation and Newton, 1997). The latter are based on the communicative theory (Brown, 2000), where understanding a word involves committing to memory its form, capturing its meaning and finally knowing how and where to use it.

The semantic network theory (Collins and Loftus, 1975; Collins and Quillian, 1972; Rips et al., 1973; Smith et al., 1974) states that a word's meaning is defined as "whatever comes to mind when someone says the word" or "you shall know a word by the company it keeps" (Firth, 1957). It implies that a word's meaning is represented in the mental lexicon by a set of associations of that word with other words. Thus, the meaning of a word is understood as collections of associated concepts. Also, it has been proved that the syntactic category of a word and the associated words that come to the mind is the same (Fillenbaum and Jones, 1965). Since it is rare in any discourse for adjacent words to be from the same syntactic category, therefore this cannot be explained as association by being contiguous. This association is because a word's meaning is represented in the mental lexicon by a set of nodes and the links between them. Here the nodes represent concepts whose meaning the network is trying to capture, and the links represent relationships between concepts.

These nodes and links translate into word relationships and meaning relationships or lexical and semantic relations. These lexical and semantic relations have been found to be cognitive universals (Lin, 1997), *i.e.*, these relations are found in all languages. Exploring the various relations in the semantic field such as hypernymy, hyponymy, *etc.*, consolidates a new word's position in the student's mind.

3.2 Associative network based Wordnet for language learning

As stated above, the learning and teaching of vocabulary with associative networks is helpful as new words are presented in semantically related group. Thus, it is best when adjectives are taught in clusters around antonymic pairs, nouns are taught as hyponyms of other nouns or with synonyms and verbs are presented as groups of troponyms or entailments. For learners, meaning is captured and retained when, for example, it is said that a horse is a kind of an animal (hypernymy relation), uniform is opposite of diverse (antonymy relation), snore is a part of sleep (entailment relation) and cultivate has as its object land (argument relation). Such associations are not captured in conventional dictionaries as much of the structural information is omitted from them and can only be provided by semantic networks like Wordnets. A classroom teacher will certainly not be able to provide associative information to students, by merely using a conventional dictionary. A thesaurus does give synonyms, but lack in lexico-semantic relations. Moreover, both the dictionaries and thesauri do not give the grammatical features that have been added to the teaching aid (discussed in section 5.4).

Various tests have proven that amongst the plethora of web-based resources for language teaching and learning, which have come up due to the exponential growth of the Internet and the World Wide Web, Princeton WordNet has emerged as one of the most reliable, authentic and useful sites (Hiray, 2015). In first of two such evaluations, three criteria, which were, (i) **Non-commerciality** – such sites were freely available, (ii) **Adequacy** – those that covered all the 570 word families featuring in the AWL (Academic Word List) 1 and (iii) **Authenticity** – the academic vocabulary exercises belong to the site itself. The result of this evaluation has marked 14 useful sites in which the Princeton WordNet stood 4th.

Another evaluation was done based on four different criteria, which were, (i) **Number of Recommendations** – those recommended by most of the ten ESL resources, (ii) **Authority** – authors of the sites should be related to the field of language, (iii) **Simplicity** – in terms of presentation, be user-friendly and material classified as per different levels of learning and (iv) **Currency** – the site is regularly updated. Here a total of 6 useful sites were marked, out of which Princeton WordNet received a ranking of 4. These results can be projected on Hindi Wordnet as well, since structurally Wordnets are similar. Thus, empirically, a strong case for using wordnet for vocabulary teaching is made.

4 A Digital Aid for Hindi Language Teaching and Learning

In order to employ association capturing in learning and to cater to the needs of rapid digitization of education, the vast lexical database of Hindi Wordnet has been transformed for pedagogical purpose in the form of an e-learning tool - *Hindi Shabdmitra*². In this tool, Hindi wordnet data is modified / simplified and further augmented with audio-visual features and grammatical properties, and is presented in a learner-friendly format for language teaching and learning. Depending upon the understanding level of a user and the purpose of use, this digital aid follows the selective information presentation approach. Henceforth, in this paper, this e-learning tool will be referred as *Digital Aid*.

4.1 Why selective information presentation?

The *Digital Aid* caters to various types of users, *viz.*, school students, teachers, parents, language learners, content managers, proofreaders, natural language processing researchers, mobile/web service providers, tourists, *etc.* As per the need of these users the information content has been moulded, keeping in sight the cognitive load that the user may be able to cope with. This can avoid unnecessary learning efforts, wastage of time and the burden of information overloading. In *Digital Aid*, this is achieved through the multi-layered information rendering approach, suitable for both formal and informal learning environments.

²[urlhttp://www.cfilt.iitb.ac.in/hindishabdmitra/](http://www.cfilt.iitb.ac.in/hindishabdmitra/)

4.2 Multi-modal learning - a psychological aspect

Humans are genetically programmed to communicate and understand things through a multi-sensory learning system. The brain processes the inputs received from different senses to comprehend concepts. As per the principles of multimedia learning (Gilakjani et al., 2011), multi-modal learning leads to comprehensive learning of a concept with higher retention rate. They have shown to stimulate thinking and learning as compared to only text based computer interfaces (Clark and Paivio, 1991; Mayer and Moreno, 2003; Moreno and Mayer, 2007; Shams and Seitz, 2008; Sankey et al., 2010).

For language learners, having a multi-modal e-learning tool provides motivation for learning the second language along with the independence to learn at one's own pace (Lai et al., 2016). It also provides the confidence of learning and handling a digital device with ease from an early age.

In *Digital Aid*, information is provided in the form of text as well as audio-visual inputs. The textual information pertains to the gloss (original Hindi wordnet or simplified), word usage, synonyms, grammatical features, lexico-semantic relations, ontological information. It also has audio pronunciations of words and pictures/illustration of concept, *etc.* This helps in learning and understanding the concept with great ease.

4.3 Multi-layered presentation - an incremental learning approach

The presented *Digital Aid* is designed keeping in mind various aspects of language teaching and learning. It is a five layered model where selective information is rendered in every layer, depending upon the type/need of the user and his/her cognitive competence. This multi-layered model is structured in a level-wise and class-wise manner.

Level-wise information presentation

The level-wise module is designed for informal setup where a user can select any of the five levels depending upon his level of expertise. The information is rendered level-wise as follows:

- **Level 1 (Beginner):** Level 1 is meant for users who are new language learners. Here information such as simplified concept definition, word usage, synonyms, grammatical features is selectively presented. The gloss is simplified so that the beginners can easily

understand and learn the concept. Audio pronunciation of a searched word and a picture or an illustration of a given concept is provided. Apart from this, the corresponding English WordNet synset is provided. For those words which do not have a corresponding synset it is given a bilingual mapping (Singh et al.,).

- **Level 2 (Intermediate):** In level 2, the intermediate users are targeted. Here, users are expected to have basic knowledge of Hindi. In this level, all the information from level 1 is rendered and additionally, more grammatical features such as gender for nouns, antonym for verb and adverb, type of verb, countability for adjective, type of adjective, spelling variation, *etc.* are appended.
- **Level 3 (Proficient):** In level 3, additional information is presented which is necessary for the intended user. Here, instead of simplified gloss and simplified example sentence(s), the Hindi wordnet's original gloss is rendered. This is because, it is expected that these users have a good grip over Hindi and can understand the complexity of a language. From this level onward it is very important that the concept should be clearly explained, so that the learners can understand the fine grained difference between two synsets. Figure 1 shows a screenshot of the tool rendering a word भजन (*bhajana*, hymn) at Proficient level.



Figure 1: Screenshot of the *Digital Aid* rendering a word भजन (*bhajana*, hymn) at Proficient level

- **Level 4 (Advanced):** At level 4, other semantic relations such as hypernymy, hyponymy, *etc.* are introduced along with all the information presented at level 3. Here, all the available synsets in Hindi wordnet and all the grammatical features are rendered.

- **Level 5 (Expert):** This is the highest level in this *Digital Aid* in which all the information available in Hindi Wordnet, along with grammatical features, ontological information, semantic and lexical relations are rendered. The expected target group here is teachers, researchers, language learners, *etc.*

Class-wise information presentation

The main purpose of developing this *Digital Aid* is to target school curriculum as prescribed by the various school boards in India. In this tool, the syllabus for Hindi prescribed in the CBSE Board³ has been selected, as it has a wide reach across India. The tool is devised to assist school teachers in teaching and students in learning Hindi vocabulary available in their curriculum. Once a particular class and chapter is selected, all the corresponding words are listed for learning in the interface.

4.4 Learning outcome - what difference does Hindi wordnet make?

There are many tools which are available for language learning. However, the *Digital Aid* has unique features which can lead to additional learning outcomes. Using the lexical and semantic relations encoded in Hindi wordnet, users can learn different senses or meanings of a word (polysemy), know about different relations like hypernymy (is-a), meronymy (part-of), troponymy, entailment, *etc.* These are the unique features which are present in the wordnet and are rarely found in traditional dictionaries/thesaurus. These will help in the understanding and retention of concepts. Besides this, they can learn the concepts of synonyms (words having similar meaning) or antonyms (words having opposite meaning), learn to associate a concept with a picture, get gender information of a word for formulating a correct sentence, develop a wide vocabulary which can aid in creative writing. Since the tool will also have the corresponding English word/synset it can be very useful in doing simple translation of text. The user can take the help of this tool in identifying parts of speech (POS) of words and learn the usage of words through the example sentences. This can be of great assistance in cases of idioms, *etc.* All this will be made available to the learners and teachers for their use through this tool. Thus, using Hindi wordnet has created a huge difference in language learning.

³<http://cbse.nic.in/newsite/index.html>

5 Process, Experiences and Lessons Learnt

As a part of the effort to align the project with the school education in India, the following key activities have been performed. The process, experiences and lessons learnt have been recorded here:

5.1 Word collection

In the word collection activity, the words from Hindi textbooks by NCERT⁴ have been collected as these books are followed by majority of schools across the country and also in some schools in other countries. Therefore it has maximum number of students studying the same textbooks, thus improving the scope of tool's coverage. In this process, the words which are not available in Hindi wordnet, but are present in textbooks, are collected and added in tool's database. Simultaneously, these words are added to Hindi wordnet, thus expanding the Hindi Wordnet vocabulary. Some types and examples of collected words are:

- **Proper Nouns:** Words like name of persons, places, *etc.* are proper nouns. For *e.g.*, name of a person: नागार्जुन (*naagaarjuna*) and name of a place: हिमाचल प्रदेश (*himaachala pradesh*).
- **Rhyming words:** In poetry, many rhyming words are used to make them interesting and fun to sing for kids. All such words do not necessarily have a proper meaning. For *e.g.*, गमगम गमगम (*gamagama gamagama*, runs).
- **English words:** Some English words included into Hindi vocabulary. For *e.g.*, लेमन (*leman*, lemon).
- **Idioms and Proverbs:** Some idioms and proverbs are also collected. For *e.g.*, मुँह में पानी भर आना (*muñha meM paanii bhara aana*, mouth watering).
- **Name of the Games:** Some common game names not present in the Hindi wordnet, are also added. For *e.g.*, पकड़म-पकड़ाई (*pakaDama-pakaDaaaii*, catch-catch or catching-catch - a popular game among kids).
- **Object:** Some lesser known objects were also found in the text-books are also collected. For

⁴<http://ncert.nic.in/>

e.g., रामानंदी चंदन (*raamaanaMdii chaMdana*, a kind of Sandal).

- **Lesser known Indigenous words:** Words from the text-books which are native to the land, and do not belong to any particular language. For e.g., पछाई - (*paChaaai*, a breed of domestic animals).
- **Productive words:** Words which can be produced by adding a suffix or prefix to generate a list of words which carry a similar sense are called productive words. So far such words have not been added to the Hindi wordnet. However, in *Digital Aid*, productive words have been added separately (as an appendix) for a better coverage w.r.t. the textbooks. For e.g., 'नुमा' (*numaa*) as a suffix means "like a". It can be used to form a productive word as बेलननुमा (*belananumaa*, like a roller).

5.2 Gloss simplification

To make the Hindi wordnet a suitable digital aid catering to various levels of learners, it was apparent that the gloss of many synsets in the Hindi Wordnet was somewhat complex for the understanding of a language learner at a beginner stage. For the ease of the target user base, the "gloss simplification" subtask have been formulated. Gloss simplification activity was carried out by the lexicographers. An example of gloss simplification is as follows. For a word हिम्मत (*himmat*, courage), the original Hindi wordnet gloss is:

मन की वह दृढ़ता जो कोई बड़ा काम करने में प्रवृत्त करती है या जिसके कारण हम निडर होकर किसी खतरे आदि का सामना करते हैं (*mana kii vaha dRiDhataa jo koi baDaaa kaama karane meM pravRitta karatii hai yaa jisake kaaraNa hama niDara hokara kisii khatare aadi kaa saamaanaa karate haiM*, that perseverance of mind which motivates us to do some great work, or because of which we face fear and danger)

Such a gloss, being too elaborate and difficult to follow at the beginner's level, has been simplified to: मन की ताकत (*mana kii taakata*, strength of mind).

A case which posed a challenge in the gloss simplification was the word रंग (*raMga*, colour), for which the Hindi Wordnet gloss is: किसी वस्तु आदि का वह गुण जिसका ज्ञान केवल आँखों द्वारा होता है (*kisii vastu aadi kaa vaha guNa jisakaa GYaana kevala aaKhoM dvaaraa hotaa hai*, That attribute of an object, etc., which is perceived only through eyes).

Now, रंग (*raMga*, colour) is such an everyday word that it was quite tough to find an easy-to-understand definition for it, hence its English translation (कलर / colour) has been provided. The English word is highly in use in the daily language and also occurs frequently in written form too, so it could be readily added to the Hindi wordnet data, thus solving the issue of all such words.

5.3 Picture depiction

As rightly mentioned in a famous idiom 'a picture is worth a thousand words', a complex concept can be easily explained by a picture or an illustration. Kanojia et al. (2016) tried to automatically collect images for IndoWordNet⁵, but due to the lack of tagged images openly available for use, enough images could not be collected. In Hindi Wordnet, there are several concepts which are hard to explain using the gloss. For example, the concept of a word 'milk' in Hindi is explained as वह सफेद तरल पदार्थ जो स्तनपायी जीवों की मादा के स्तनों से निकलता है (*vaha sapheda tarala padaartha jo stanapaayii jiivoM kii maadaa ke stanoM se nikalataa hai*, a white nutritious liquid secreted by mammals and used as food by human beings).

This gloss seems to be difficult for level 1 and 2 learners to understand due to the presence of some difficult words, which would require definitions in turn. However, as shown in figure 2 below, this can be easily understood with the help of a picture. Hence, pictures and illustrations have been used to depict a concept. Also, pictures help in differentiating the fine grained senses found in Wordnet.



Figure 2: Picture depicting the concept of a word दूध (*duudha*, milk)

In the process of picture depiction, most of

⁵<http://www.cfilt.iitb.ac.in/indowordnet/>

the concepts are grouped together and illustrated so that they can be reused for similar concepts with minor changes. Also, antonyms, hypernyms-hyponyms, meronyms-holonyms were illustrated together. At the initial phase, Level 1 and level 2 concepts are illustrated so that their users, being beginners, can easily understand a given concept using an illustration. This will be followed by illustrations of higher level concepts.

5.4 Grammatical feature marking

At each levels of *Digital Aid*, the grammatical features of a given word is rendered. These features are marked by the lexicographers during the process of gloss simplification for level 1 & 2. From level 3 onwards, the feature marking is carried out during word collection process as gloss is not simplified for higher levels. During the process, each word is marked with the grammatical properties corresponding to its POS category. Some of the grammatical features are as follows:

Nouns are either countable or uncountable. They can belong to any of these categories: Proper Noun, Abstract Noun, Common Noun, Collective Noun. When a noun is a compound, it may belong to one of these categories: तत्पुरुष (*tatpuruSha*), कर्मधारय (*karmadhaaraya*), द्विगु (*dvigu*), अव्ययीभाव (*avyayibhaava*), द्वंद्व (*dvaMdva*), or बहुव्रीही (*bahuvriihii*) (Redkar et al., 2016).

The Verbs are either Transitive or Intransitive. The different types of verbs are: Simple verb, Con-junct verb, and Compound Verb. These verbs may also be Causative verb. Kinds of Adverbs that feature in this tool are of Manner, Place, Time and Quantity. Similarly, the Adjectives are categorized as Qualitative, Numeral, Quantitative, Pronominal.

5.5 Audio pronunciation

Cognitive theories of multimedia learning (Mayer, 2002) indicate that audio cues are effective aids in a learning scenario, and also help in retaining the material learned (Bajaj et al., 2015). To help in more effective learning, we intend to include audio pronunciations for all the words across the five levels in *Digital Aid* described in Section 4.3. Manually recording pronunciations for all the words is a tedious task. These recording efforts could be minimized by using text-to-speech (TTS) systems to automatically synthesise speech for most of the words. However, one cannot be sure about the quality of these synthesised clips. We built multiple TTS systems and systematically analyzed the

quality of the resulting synthesised clips, with the help of lexicographers.

We use the data provided by the IndicTTS⁶ (Prakash et al., 2014) forum to create a TTS synthesis system which generates speech audio for a given word; we will refer to this system as **Model 1**. We use the voices *Hindi - Female* and *Marathi - Female* provided by FestVox⁷ (Black and Lenzo, 2000) and Festival Framework⁸ (Black and Taylor, 1997) for Hindi Speech Synthesis and name these systems **Model 2** and **Model 3**, respectively. We also use the tool⁹ available on the IndicTTS forum website to generate a final set of audio samples and refer to it as **Model 4**. (**Model 1** was trained using the IndicTTS data while **Model 4** is a pre-trained model hosted at the forum website mentioned above.)

We synthesised audio for the words corresponding to “Levels 1 and 2” using all four above-mentioned TTS systems. We chose a random sampling of 535 words and generated synthesised outputs from all four models; these outputs were presented to two lexicographers for further analysis. The lexicographers were asked to independently rate the audio clips on the following scale:

- **unusable (#0):** This rating corresponds to audio clips which are either completely distorted, or too noisy for the user to comprehend.
- **usable (#1):** This rating corresponds to audio clips which are moderately usable and suggests that the user can comprehend the underlying words, but can be synthesized better.
- **good (#2):** This rating corresponds to audio clips that are really good and clearly convey the words.

For each of the 535 words, the lexicographers were also asked to mark which of the four synthesised clips they liked the most.

The evaluation results are shown in Table 1. This clearly shows that **Model 1** was marked as the most liked audio clip most often, while **Model 4** performed the best in terms of producing the most number of usable audio clips (obtained by summing clips with ratings #1 and #2).

⁶<https://www.iitm.ac.in/donlab/tts/>

⁷<http://www.festvox.org/index.html>

⁸<http://www.cstr.ed.ac.uk/projects/festival/>

⁹<https://www.iitm.ac.in/donlab/tts/demo.php>

	#0	#1	#2	#1+#2	Most Liked
Model 1	79	55	99	154	101
Model 2	37	78	112	190	90
Model 3	72	86	58	144	51
Model 4	55	117	107	224	70

Table 1: Results of manual evaluation of synthesized speech clips. The values indicate the number of times (*i.e.*, count) an audio from a particular model was rated as per the scale described.

A qualitative analysis of the synthesized clips highlighted the following issues, particularly with respect to the clips that were marked “unusable”: i) Flap or tap sounds (‘ड’ *Da*, ‘ढ’ *Dha*) were pronounced incorrectly, ii) Intonation of the audio for heavy syllables was at times incorrectly rendered and for words such as ‘एकदम’ (*ekadama*), the pronunciation had a specific stress pattern which should have ideally been neutral, thus making it sound unnatural, iii) There were also a few examples of unnecessary lengthening of a vowel. For example, in बीमारी (*biimaarii*), there was unnecessary stress on ‘बी’ (*bii*) and hence it was lengthened, iv) Incorrect syllable breaks were observed in some words. For example, नापसंद (*naa-pasand*, non-favourite), was pronounced as नाप-संद (*naapa-saMda*), which is incorrect, v) It was also noted that sometimes consonant clusters were mispronounced. E.g. कुत्ता - (*kuttaa*) - dog, was incorrectly pronounced as कु-ता (*ku-taa*) or कुत-ता (*kuta-taa*).

6 Field Test and User Feedback

The prototype of the *Digital Aid* was initially demonstrated in three local schools. Two schools were following the CBSE board curriculum where students were learning Hindi as primary language from class 1 onwards. The other school was following the curriculum of the state board¹⁰ where students from class 5 were learning Hindi for the first time as a second language. The feedback was sought from primary language learners as well as second language learners for content, ease of handling the application, classroom impact and overall experience by teachers and students. It was observed that the digital aid helped teachers in explaining concepts clearly with the help of images and simplified concepts even for primary language learner. It was also noticed that class 5 students

¹⁰<https://mahahsscboard.maharashtra.gov.in>

from the state board easily understood the concepts though they were learning Hindi for the first time. The aid assisted teachers in better classroom management, especially with the help of illustrations and reduced effort of reiterating the concepts for better retention and having the standardized pronunciation by native Hindi speakers. The application has been improved based on the feedback received. Some of the suggested changes were *viz.*, include spelling variations, give additional grammatical features, provide English word, *etc.* Accordingly, the suggested changes were implemented. The presented *Digital Aid* is now ready for the next round of field trials.

7 Conclusion and Future Work

The paper presents how a lexically rich resource like Hindi WordNet is suitably modified and enhanced for developing a digital aid for language teaching and learning. The *Digital Aid* presented here is a multi-modal multi-layered Hindi language learning aid which can be used for formal and informal learning environments such as schools and non-government organizations involved in education. While developing this digital aid, the process followed, the experiences earned, the challenges faced and the lessons learnt are recorded in this paper. The *Digital Aid* has been tested successfully during field trials and the work has been appreciated by teachers and students. With the help of this aid a better understanding and retention of concepts has been achieved, which is helped in a large way by illustrations and clear pronunciations. This has led to better classroom management and increased interest in leaning.

In future, *Digital Aid* can be expanded to the other Indian languages. Gamification and evaluation techniques will be incorporated.

Acknowledgements

The authors would like to thank and acknowledge the support and help by the members of Center for Indian Language Technology (CFILT)¹¹ and *Hindi Shabdmitra* team. The funding agency, Tata Center for Technology and Design (TCTD)¹² has been instrumental and supportive throughout the development of this *Digital Aid*.

¹¹<http://www.cfilt.iitb.ac.in>

¹²http://www.tatacentre.iitb.ac.in/digital_aid.php

References

- Jatin Bajaj, Akash Harlalka, Ankit Kumar, Ravi Mokashi Puneekar, Keyur Sorathia, Om Deshmukh, and Kuldeep Yadav. 2015. Audio cues: Can sound be worth a hundred words? In *International Conference on Learning and Collaboration Technologies*, pages 14–23. Springer.
- Alan Black and Kevin Lenzo. 2000. Building voices in the festival speech synthesis system.
- Alan W. Black and Paul A. Taylor. 1997. The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK. Available at <http://www.cstr.ed.ac.uk/projects/festival.html>.
- H Douglas Brown. 2000. Principles of language learning and teaching.
- Heidi Brumbaugh. 2015. *Self-assigned ranking of L2 vocabulary: using the Bricklayer computer game to assess depth of word knowledge*. Ph.D. thesis, Arts & Social Sciences.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- James M Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational psychology review*, 3(3):149–210.
- Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.
- Allan M Collins and M Ross Quillian. 1972. Experiments on semantic memory and language comprehension.
- Edgar Dale. 1969. Audiovisual methods in teaching.
- Mehmet Demirezen. 1988. Behaviorist theory and language learning. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 3(3).
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Jorge Francisco Figueroa Flores. 2015. Using gamification to enhance second language learning. *Digital Education Review*, 27:32–54.
- Samuel Fillenbaum and Lyle V Jones. 1965. Grammatical contingencies in word association. *Journal of Verbal Learning and Verbal Behavior*, 4(3):248–255.
- Raymond Firth. 1957. 2. a note on descent groups in polynesia. *Man*, 57:4–8.
- Abbas Pourhossein Gilakjani, Hairul Nizam Ismail, and Seyedeh Masoumeh Ahmadi. 2011. The effect of multimodal learning models on language teaching and learning. *Theory & Practice in Language Studies*, 1(10).
- Amit C. Hiray. 2015. *Teaching and Learning of EAP Vocabulary: A Web-based Integrative Approach at the Tertiary Level in India*. Ph.D. thesis, Dept. of HSS, IIT Bombay.
- Diptesh Kanojia, Shehzaad Dhuliawala, and Pushpak Bhattacharyya. 2016. A picture is worth a thousand words: Using openclipart library for enriching indowordnet. In *Eighth Global WordNet Conference*. GWC 2016.
- Chun Lai, Mark Shum, and Yan Tian. 2016. Enhancing learners' self-directed use of technology for language learning: the effectiveness of an online training platform. *Computer Assisted Language Learning*, 29(1):40–60.
- Chih-Cheng Lin. 1997. Semantic network for vocabulary teaching. *Journal of Computer Assisted Learning*, (42):43–54.
- Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52.
- Richard E Mayer. 2002. Multimedia learning. *Psychology of learning and motivation*, 41:85–139.
- Roxana Moreno and Richard Mayer. 2007. Interactive multimodal learning environments. *Educational psychology review*, 19(3):309–326.
- Paul Nation and Jonathan Newton. 1997. 19 teaching vocabulary. *Second language vocabulary acquisition: A rationale for pedagogy*, page 238.
- Anusha Prakash, M Ramasubba Reddy, T Nagarajan, and Hema A Murthy. 2014. An approach to building language-independent text-to-speech synthesis for indian languages. In *Communications (NCC), 2014 Twentieth National Conference on*, pages 1–5. IEEE.
- Hanumant Redkar, Nilesh Joshi, Sandhya Singh, Irawati Kulkarni, Malhar Kulkarni, and Pushpak Bhattacharyya. 2016. Samāsa-kartā: An online tool for producing compound words using indowordnet. In *8th Global WordNet Conference*.
- Lance J Rips, Edward J Shoben, and Edward E Smith. 1973. Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, 12(1):1–20.
- Michael Sankey, Dawn Birch, and Michael Gardiner. 2010. Engaging students through multimodal learning environments: The journey continues. In *Proceedings ASCILITE 2010: 27th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Curriculum, Technology and Transformation for an Unknown Future*, pages 852–863. University of Queensland.
- Ladan Shams and Aaron R Seitz. 2008. Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11):411–417.

- Meghna Singh, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, and Pushpak Bhat-tacharyya. Mapping it differently: A solution to the linking challenges.
- Anna Sinopalnikova. 2004. Word association thesaurus as a resource for building wordnet. In *Proceedings of the 2nd International WordNet Conference*, pages 199–205.
- Edward E Smith, Edward J Shoben, and Lance J Rips. 1974. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, 81(3):214.
- Koun-Tem Sun, Huang Yueh-Min, and Liu Ming-Chi. 2011. A wordnet-based near-synonyms and similar-looking word learning system. *Journal of Educational Technology & Society*, 14(1):121.
- Kevin Werbach and Dan Hunter. 2012. *For the win: How game thinking can revolutionize your business*. Wharton Digital Press.
- Jaeseok Yang. 2013. Mobile assisted language learning: review of the recent applications of emerging mobile technologies. *English Language Teaching*, 6(7):19–25.

An Experiment: Using Google Translate and Semantic Mirrors to Create Synsets with Many Lexical Units

Ahti Lohk¹, Mati Tombak¹ and Kadri Vare²

¹Department of Software Science, Tallinn University of Technology, Tallinn, Estonia

²Department of Computer Science, University of Tartu, Tartu, Estonia

<{ahti.lohk, mati.tombak}@ttu.ee, kadri.vare@ut.ee>

Abstract

One of the fundamental building blocks of a wordnet is synonym sets or synsets, which group together similar word meanings or synonyms. These synsets can consist either one or more synonyms. This paper describes an automatic method for composing synsets with multiple synonyms by using Google Translate and Semantic Mirrors' method. Also, we will give an overview of the results and discuss the advantages of the proposed method from wordnet's point of view.

1 Introduction

Three important aspects need to be considered while composing a wordnet (Lohk, 2015): what type of a lexical resource to use, which building model (Vossen, 1998) to implement and what is the level of automation. Wordnets can be built manually, semi-automatically, automatically and can be based on different bilingual or monolingual resources or corpora. This means that synsets can also be created either manually or (semi)automatically and wordnet builders have to decide if a synset contains one or many synonyms; the latter mentioned is a quite difficult task.

Finding and determining synonyms can often be complicated, for example in Estonian wordnet there are two different synsets: 'hypogastrium' and 'abdomen' which belong to one synset (Orav et al., 2011). Synonyms can be identified from monolingual explanatory dictionaries (Blondel and Senellart, 2002) and bilingual dictionaries, text corpora, lexico-syntactic patterns and neural networks (Nguyen et al., 2017); from Wikipedia, spectral clustering and from multi-layered neural networks (Zhang et al., 2017). Also from parallel

corpora (Dyvik, 2004) and from using translations of other wordnet's synsets (Lindén and Niemi, 2014). This paper fills the gap of identifying multi-member synsets by using Google Translate.

Despite that Google Translate has around 70 different languages in its system, in this experiment we only deal with Estonian and English languages. However, throughout all experiment we exploit three linguistic data resources:

- all unique lexical units from the synsets in Princeton Wordnet¹ (version 3.1) (PWN) (Fellbaum, 1998)
- all unique lexical units from the synsets of Estonian Wordnet² (EstWN) (version 72) (Orav et al., 2011)
- Google Translate³ translations and source languages synsets connected with translations (See Figure 1).

1.1 Research questions

The first and most important question this paper address is that how to use Google Translate for identification of multi-membered synsets (synsets with many lexical units). Answer shortly, to form these synsets all unique lexical units from PWN synsets are extracted and then automated queries to will be sent to Google Translate. Afterwards, Semantic Mirroring method will be used on source language (firstly English) and equivalents of the target language (firstly Estonian). As a result, multi-membered synsets' pairs will be identified.

Another important question is the linguistic outcome of this method – how results can be used in building, quality and consistency checking of wordnets. Answer shortly, these automatically composed multi-membered synsets can be used to

¹ <https://wordnet.princeton.edu/>

² <http://www.cl.ut.ee/ressur-sid/teksaurus/teksaurus.cgi.en>

³ <https://translate.google.com/>

validate synsets already present and to create new synsets or add missing members to a synset already present.

2 Previous work

Semantic Mirroring method was initially introduced by Norwegian researcher Helge Dyvik (Dyvik, 2004). Among other things, he used semantic mirrors' method for automatic creation of Norwegian Wordnet. This method helped him to discover both synonym sets and semantic relations (mostly *hyperonymy*) successfully from parallel corpora.

To the best of our knowledge, there haven't been any attempts to discover synsets by using Google Translate. However, Google Translate is being used as a "dictionary" to translate PWN glosses to in Macedonian Wordnet (Saveski and Trajkovski, 2010) or to translate multiword expressions from PWN to Arabic (Attia et al., 2010).

3 Method description

In this section, we formalize the method of synonym sets' pairs for source and target languages mathematically as well as we explain this formalization through an example. The method described here follows the idea of the Semantic Mirrors' method.

3.1 Mathematical formalization

Let w be a word in a source language (input) and $translate(w)$ be a set of Google translations of w .

For each $t \in Translate(w)$ let $Row(t)$ be a row of synonyms of t and

$$W = \bigcup_{t \in Translate(w)} Row(t).$$

Let FS be the set of frequent source words from W , i.e., words which occur in at least two different rows of synonyms.

$$FS = \{s : \exists t_1 t_2 \in Translate(w) [(s \in Row(t_1)) \& (s \in Row(t_2))]\}$$

Let FT be corresponding subset of $Translate(s)$:

$$FT = \{t : \exists s \in FS (s \in Row(t))\}$$

The result is the collection of pairs of sets $\langle S, T \rangle$, where $S \subseteq FS$, $T \subseteq FT$ and

$$S = \{s : \exists t \in T (s \in Row(t))\}$$

$$T = \{t : \exists s \in S (s \in Row(t))\}$$

Binary relation $s \in Row(t)$ defines Galos' connection between power sets of FS and FT . (Pasquier et al., 1999). Every element $\langle S, T \rangle$ is a fixpoint (closed set with frequency ≥ 2).

3.2 Complementary explanation

To get a clearer picture of the method, we complement mathematical formalization (Sec. 3.1) with a screenshot of the results of the Google Translate (Figure 1) and frequency table (Table 1) with synsets' pairs that are composed based on this screenshot in Figure 1.

According to Figure 1, input word w is underlined. Translations of the word w are shown in the first column: $\{idee, m\ddot{o}te, ettekujutus, m\ddot{o}iste, plaan, armavus, kava, aade\}$. For each translation word the set of the row of the (source language) synonyms are given. For example $Row(idee) = \{idea, concept, notion, thought, point\}$.

The screenshot shows the word 'idea' underlined. Below it, a list of translations is provided, each with a corresponding set of synonyms in the source language. The translations are: idee, mōte, ettekujutus, mōiste, plaan, arvamus, kava, and aade. Each translation is followed by a list of synonyms in the source language, such as 'idea, concept, notion, thought, point' for 'idee'.

Translation	Synonyms (Source Language)
idee	idea, concept, notion, thought, point
mōte	idea, thought, point, sense, mind, purport
ettekujutus	idea, imagination, notion, fancy
mōiste	concept, notion, idea
plaan	plan, map, blueprint, schedule, program, idea
arvamus	opinion, view, judgment, guess, idea, voice
kava	plan, scheme, program, schedule, design, idea
aade	ideal, idea, thought

Figure 1. Screenshot of the results from the Google Translate

Frequency	Set of FS	ENG-EST synsets' pairs
3	thought	{idea, thought} - {idee, mōte, aade}
3	notion	{idea, notion} - {idee, ettekujutus, mōiste}
2	concept	{idea, concept} - {idee, mōiste}
2	point	{idea, point} - {idee, mōte}
2	plan	{idea, plan} - {plaan, kava}
2	schedule	{idea, schedule} - {plaan, kava}
2	program	{idea, program} - {plaan, kava}

Table 1: Frequency table with source and target language synsets' pairs

The set of frequent source words for the example

$$FS = \{idea, thought, notion, concept, point, plan, scedule, programm\}$$

The set of frequent target words:

$$FT = \{idee, m\ddot{o}te, aade, ettekirjutus, m\ddot{o}iste, plaan, kava\}$$

The *Result(idee)* is the collection of pairs of sets:

```
{idea, schedule, program, plan }, {plaan, kava}}
{idea, thought}, {idee, mõte, aade}}
{idea, notion}, {idee, ettekujutus, mõiste}}
{idea, concept}, {idee, mõiste}}
{idea, point}, {idee, mõte}}
```

4 Overview of the experiment

Google Translate categorizes translations and synonym sets for source language’s words: translations are distinguishable by the length of the bar underneath word *noun* (see Figure 1).

The longest bar indicates to a *common translation* (two times in this case), middle length indicates to *uncommon translation* (one time in this case), and the shortest bar presents the *rare translations* (five times in this case).

Based on the outputs of the queries, our experiment is divided into two approaches. The first approach counts only common categories, the second approach deals with all categories of the output.

4.1 First approach – common translation

Assuming that uncommon and rare translations do not form a set of exact synonyms, we start with our experiment using only common translations and synonym sets.

Firstly unique lexical units from PWN (version 3.1) and secondly all unique lexical units from EstWN were chosen as input (version 72). If we use translations from both languages, it is possible to discover synsets, which can stay hidden (even with a language as English which has a large vocabulary) if using only translations from one language.

4.2 Second approach – common, uncommon and rare translations

According to Table 1, we see that it is possible to compose synsets even when all translation categories are involved. Current approach provides, of course, new words that can be added into a wordnet. However, it is not clear what will be the number of new words. Also, it is yet to determine how much of the new synsets are equal or similar to wordnet’s synsets. Hereby, a new synset is similar to wordnet’s synset when its all members are part of a wordnet’s synset or at least two its members are part of wordnet’s synset.

4.3 Data from EstWN and PWN and queries

For the experiment, we extracted all the lexical units from EstWN and PWN synsets and compiled them into **two** unique lists of words. The first list contains 101.732 words from EstWN and second one 147.035 from PWN. While implementing both approaches (Section 4.1 and Section 4.2), our program performed 2 x 101.732 queries in list one to Google Translate and 2 x 147.035 queries in list two respectively. We have to admit that if we had saved results for every query, then it would have been possible to reduce the number of queries twofold.

5 Results of the experiment

One of the general results is the synset to synset translations, which can be exploited to check and compare the translation equivalents in wordnets. EstWN is composed manually and often the translation from Estonian to English is complicated to find, here are the synonyms produced to English useful.

5.1 Results of the first approach

input	output		
	eng-est synsets’ pairs	unique words in synsets	not represented words in wordnet
101.732 est words	1.799	Estonian	252
		English	144
147.035 eng words	1.137	Estonian	340
		English	77
summary	2.520	Estonian	532
		English	208

Table 2: Results considering only *common translation* category

If we use Estonian words as input and in output take into account only common translation, the result is 1.799 synset pairs between Estonian-English (see Table 2). For English input, the result is 1.137 synset pairs between English-Estonian. Moreover, while uniting both outputs of the languages, the result is 2.520 synset pairs between English-Estonian. Both results yield to overlap of 416 synset pairs.

The method provides us new words (lexical units) missing from EstWN and PWN that can be added to both wordnets. For the quick analysis, we applied tools of Python package EstNLTK⁴ to

⁴ <https://estnltk.github.io/estnltk/1.4/>

find lemmas and word forms for new words (lexical units). As a result, we identified 527 different lemmas out of 532 words (see Table 2), approximately 50% were nouns, 18% verbs and ca 19% adjectives. Remaining 13% of words were mainly adpositions and adverbs.

eng-est synsets' pairs	language	exact match	all LUs in a wn synset	at least two LUs in a wn synset	no match
1.799	est	109	454	223	1.013
	eng	145	507	143	1.004
1.137	est	69	309	36	723
	eng	97	293	144	603
2.520	est	147	637	260	1.476
	eng	192	658	262	1.408

Table 3: Comparing resulting synsets with EstWN and PWN synsets (only *common* category)

The proposed method can identify new synsets there, where initially lexical units have not been in the same synsets. For example, the automatically produced synset was ‘tavaliselt, üldiselt’ (usually, generally) and this synset can be added to EstWN, since it does count as a new concept. According to Table 3 “*exact match*” refers to a case, where synsets composed during the experiment are equal to some synset in wordnet – both synsets contain the same lexical units. The column “*all LUs in a wn synset*” describes a situation where all lexical units of produced synsets are as a subset of some synset in a wordnet. The column “*at least two LUs in a wordnet*” refers that two produced synset members act as a subset of some synset in a wordnet. The last column of the table shows statistics about these produced synsets with no synset members being as a subset for multi-membered synsets in a wordnet.

5.2 Results of the second approach

Compared to the first approach (Table 2) the second approach (Table 4) produces three times more synset pairs. Also, the amount of unique lexical units is larger as well as the words not present in both wordnet(s).

Similarly to the first approach, we determined the lemmas and word forms for words not present in EstWN and identified 1915 lemmas out of 1940 words (see Table 4): approximately 45% of words were nouns, 20% verbs, and 20% adjectives. The majority of remaining 15% words were, again, adpositions and adverbs.

The similarity of these two approaches is that the English input increases unique Estonian words not yet present in EstWN.

input	output			
	eng-est synsets' pairs	unique words in synsets		not represented words in wordnet
101.732 est words	6.549	Estonian	7.690	1.003
		English	7.384	611
147.035 eng words	7.640	Estonian	9.050	1.805
		English	7.619	434
summary	9.122	Estonian	9.556	1.940
		English	8.440	724

Table 4: Results considering all Google Translate categories: *common*, *uncommon* and *rare*.

Also, it can be observed that around 2.5 times more new Estonian synsets are produced in Table 4 (two last rows). Moreover, the difference between new words in Table 2 and 4 is even four times.

eng-est synsets' pairs	language	exact match	all LUs in a wn synset	at least two LUs in a wn synset	no match
6.549	est	312	1.437	658	4.094
	eng	357	1.253	1.077	3.814
7.640	est	281	1.238	1.020	4.955
	eng	414	1.471	860	4.749
9.122	est	330	1.493	1.238	6.064
	eng	480	1.715	1.314	5.616

Table 5: Comparing resulting synsets with EstWN and PWN synsets (all three translation categories: *common*, *uncommon* and *rare*)

While using the second approach, the method also produces synsets with translations from the rare category. For example, we obtain three different synsets for the Estonian word ‘kallis’ - darling in one sense, expensive in another sense, and noun honey in the third sense. The honey-sense is missing from EstWN.

6 Discussion and Conclusion

For Google Translate unique lexical units from synsets of EstWN and PWN were given as an input, because wordnets (at least EstWN and PWN) represent among other words the core vocabulary of languages, which can be sensibly used exactly in this experiment. The second reason to use

namely wordnets as an input to Google Translate is that the data is adequately comparable since they represent the same vocabulary. As a result, we received a lot of synsets with many lexical units (or synonyms). We considered these synsets to be correct and suitable, where at least two members are also synset members in a wordnet. Our experiment showed that the majority of the synsets do not fill this requirement – they consist new words, or they are completely different from the synsets in wordnet. For example, there are synsets containing words from different part-of-speeches or synsets combining different senses. Many synsets include possible hyperonym (and hyponyms), for example, Estonian ‘komm, komvek, maistus’ (candy, sweets), where ‘maistus’ (sweets) acts more as a hyperonym for ‘komm’ (candy). On the other hand, it is possible to complement synsets already present in EstWN with the synset members identified by the current method.

Our method identifies a significant amount of new words, which can be included into EstWN and to the PWN. Here it should be noted that from the new words 50% are nouns, 20% verbs, and 20% adjectives. If we compare these percentages with the Estonian input words, which are accordingly 80%, 8% and 6% (the rest are mainly adverbs), then we can assume that Google Translate was able to produce significantly more new words for verbs and adjectives than for nouns

6.1 Future works

The first and foremost work that has to be done is to analyze received synsets and new words. At the moment, it is clear that many of synsets contain synonyms that are not correct or their grammatical categories (such as adposition and comparative form of an adjective) are not used in wordnet. For the same reason, not all of the new words do not fit into wordnet. On the other hand, received synsets are useful to improve the quality of EstWN and PWN. Regardless, this analyzing work is still ahead.

Secondly, our experiment exploited only words from synsets present in wordnets since they represent the majority of most commonly used nouns, verbs, adjectives, and adverbs. The next step would be to use all three categories (*common*, *uncommon*, *rare*) of translation synonym sets from Google Translate as an input for semantic mirroring method. This approach enables to make use of the data and vocabulary used in Google Translate even more.

Thirdly, while one of the most common critics on wordnet has been the granularity of senses; this method can help to reduce the amount too fine-grained senses. As seen from the outcome, it clusters together senses with similar meaning, which could, in turn, can be implied in some language technology application.

Reference

- Attia, M., Toral, A., Tounsi, L., Pecina, P., Genabith, J., 2010. Automatic extraction of Arabic multiword expressions, in: Proceedings of the 2010 Workshop on Multiword Expressions: From Theory to Applications. pp. 19–27.
- Blondel, V.D., Senellart, P.P., 2002. Automatic Extraction of Synonyms in a Dictionary, in: Proceedings of the SIAM Workshop on Text Mining. Arlington, Texas, USA, pp. 1–7.
- Dyvik, H., 2004. Translations as Semantic Mirrors: From Parallel Corpus to Wordnet. *Language and Computers* 49, 311–326.
- Fellbaum, C., 1998. A Semantic Network of English Verbs, in: *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA, pp. 69–104.
- Lindén, K., Niemi, J., 2014. Is It Possible to Create a Very Large Wordnet in 100 Days? An Evaluation. *Language Resources and Evaluation* 48, 191–201.
- Lohk, A., 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Tallinn University of Technology, Tallinn, Estonia.
- Nguyen, K.A., Walde, S.S. im, Vu, N.T., 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. *ArXiv Prepr. ArXiv170102962*.
- Orav, H., Kerner, K., Parm, S., 2011. Snapshot of Estonian Wordnet (in estonian). *Keel Ja Kirjand.* 2, 96–106.
- Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., 1999. Discovering frequent closed itemsets for association rules, in: *International Conference on Database Theory*. Springer, pp. 398–416.
- Saveski, M., Trajkovski, I., 2010. Automatic Construction of Wordnets by Using Machine Translation and Language Modelling, in: *13th Multiconference Information Society*. Ljubljana, Slovenia.
- Vossen, P., 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32, 73–89.
- Zhang, L., Li, J., Wang, C., 2017. Automatic Synonym Extraction Using Word2Vec and Spectral Clustering, in: *2017 36th Chinese Control Conference (CCC)*. Presented at the 2017 36th Chinese Control Conference (CCC), pp. 5629–5632. doi:10.23919/ChiCC.2017.8028251

Context-sensitive Sentiment Propagation in WordNet

Jan Kocon
G4.19 Research Group
Wrocław University
of Science and Technology
Wrocław, Poland
jan.kocon@pwr.edu.pl

Arkadiusz Janz
G4.19 Research Group
Wrocław University
of Science and Technology
Wrocław, Poland
arkadiusz.janz@pwr.edu.pl

Maciej Piasecki
G4.19 Research Group
Wrocław University
of Science and Technology
Wrocław, Poland
maciej.piasecki@pwr.edu.pl

Abstract

In this paper we present a comprehensive overview of recent methods of the sentiment propagation in a wordnet. Next, we propose a fully automated method called Classifier-based Polarity Propagation, which utilises a very rich set of features, where most of them are based on wordnet relation types, multi-level bag-of-synsets and bag-of-polarities. We have evaluated our solution using manually annotated part of plWordNet 3.1 emo, which contains more than 83k manual sentiment annotations, covering more than 41k synsets. We have demonstrated that in comparison to existing rule-based methods using a specific narrow set of semantic relations our method has achieved statistically significant and better results starting with the same seed synsets.

1 Introduction

Princeton WordNet (Miller, 1995) has been expanded with sentiment annotation in several projects. However in all these approaches only a very limited part of Princeton WordNet was manually annotated, and the annotation for the remaining part was automatically extended by propagation algorithms, e.g. WordNet-Affect (Strapparava and Valitutti, 2004) or SentiWordNet (Esuli and Sebastiani, 2006), see also Sec. 3. Manual emotive annotation was done for plWordNet (Maziarz et al., 2016) (a wordnet of Polish) on several times larger scale. In the most contemporary version more than 54 000 lexical units (i.e. word senses) are described by sentiment polarity, basic emotions and fundamental human values, cf. (Zaśko-Zielińska et al., 2015). Only nouns and adjectives are annotated, but the manual annotation coverage

of these two part-of-speech categories is almost 24%. Having this large amount of metadata we started to look at methods of automated expansion of such information in a wordnet. Most of the existing solutions are based on a set of handcrafted rules for transferring the polarity along different types of wordnet relations. The proposed method does not require manually designed rules as they are discovered automatically.

2 Background

Lexicons are an important, inherent part of sentiment analysis and opinion mining systems. There are three general approaches to compile sentiment lexicon i.e. *corpus-based* approach: *dictionary-based* and *manual* (Liu, 2015). Manual approaches are laborious and time-consuming, so there is a great need for fast, automated methods of the construction of sentiment lexicons especially for low-resourced languages. The first built lexicons were limited only to simple word lists with positive and negative examples of words. However, the polarity of words often varies across their senses due to the semantic ambiguity. We assume that a sense-based sentiment lexicon may enable more accurate estimation of the sentiment polarity of complex phrases or sentences. One of the possible ways to construct a sense-aware sentiment lexicon is to use a wordnet (i.e. a dictionary-based approach). Approaches of this kind of generally aim at extending a small set of seed words with known polarity using lexical relations of a wordnet, e.g. hypernymy, synonymy, antonymy, etc.

Most of the existing solutions rely on a simple polarity propagation from annotated synsets (*seeds*) to their not annotated neighbours, and mostly utilise only specific subset of relations like hypernymy, hyponymy, similarity and antonymy (Maks and Vossen, 2011). These approaches do not take into account the full structure of WordNet

or even wider contexts of synsets (e.g. n -th level relations). A common approach to construct a non-English sentiment lexicon is a simple translation of SentiWordNet (Esuli and Sebastiani, 2006) polarity annotations to another language.

Simple rule-based propagation prepared for one language does not necessarily perform well for other languages, because wordnets for different languages may differ strongly, e.g. in the number of relation instances and a different semantic structure. On the other hand, corpus-based solutions require a high quality systems for word sense disambiguation. A good method for sentiment propagation should be adaptable to the structure of any wordnet with the least human effort.

3 Related Works

There is a vast amount of methods to construct sentiment lexicons, but most of them were evaluated only for English, on Princeton WordNet (Miller, 1995). One of the major sentiment lexicons for English – SentiWordNet – was introduced in (Esuli and Sebastiani, 2006), and in (Baccianella et al., 2010) its extended version was described. The main objective was to construct a large lexical resource with sentiment polarity of lexical meanings rather than words.

One of the approaches based on a non-English wordnet was evaluated in (Maks and Vossen, 2011). The authors compared three methods:

1. Simple polarity transfer from SentiWordNet (Esuli and Sebastiani, 2006) using translation equivalents between Princeton WordNet and Dutch WordNet (Piek Vossen and VanderVliet, 2008);
2. Automatic polarity propagation using only Dutch WordNet;
3. Combined approach using transfer method from SentiWordNet and polarity propagation over the Dutch WordNet.

The first method resulted in a general performance decrease in comparison to SentiWordNet from 62% to 58% of overall precision, recall and F-score. The second method was based on iterative label propagation with rules using lexical relations from WordNet. Factors such as seed set size, its composition and number of iterations had a great impact on propagation performance. When high-quality pre-selected seed synsets are used,

the obtained performance is significantly higher. One of the drawbacks of their approach is the simplicity of seed selection criteria. The best results were achieved using a mixed dataset derived from a large sentiment lexicon – the General Inquirer Lexicon (Stone, 1966). The performance reached 75% of F-score, precision and recall. The authors concluded that the size of a seed set is the most important factor, but the quality of the seeds also matters. Almost the same performance was achieved by combining transfer method with propagation (74%). The results may also suggest that simple transfer methods are not perfect, but combining multiple approaches with transfer methods may bring us a promising result.

Extended research on the polarity propagation for non-English wordnets was presented in (Maks et al., 2014). Authors applied the same propagation algorithm to five wordnets for different languages. The propagation method was similar to the methods used in their previous works. Words and their polarity extracted from the well-known General Inquirer Lexicon were translated with a machine translation service and manually mapped to the corresponding synsets in particular wordnets. The seed set consisting of synsets with known polarity was expanded using wordnet relations to cover the entire networks. The resulting lexicons varied significantly in their size and precision score. The conclusion was that the way the wordnets are built seems to affect propagation performance.

(Mahyoub et al., 2014) is a first attempt to build an Arabic sentiment lexicon on a basis of Arabic WordNet. Propagation procedure involves an expansion step which is expanding the sentiment lexicon by iteratively reaching concepts of the wordnet and scoring step evaluating the sentiment score of reached concepts according to their distance from the seeds. A task-based evaluation was applied. The acquired polarity scores were incorporated into features for sentiment classification task evaluated on Arabic corpora.

There were several attempts to construct a large sentiment lexicon for Polish in an automated way e.g. (Haniewicz et al., 2013; Haniewicz et al., 2014). (Haniewicz et al., 2013) attempted to build a polarity lexicon from web documents. They utilized plWordNet (yet without sentiment annotation) as a general resource to develop domain-aware polarity lexicons. A large semantic lexicon

with over 70,000 concepts from Web reviews was built where each term in this lexicon was described by a vector of sentiment values, representing the polarity of this term in various domains. plWordNet was utilised to identify semantic relations between acquired terms. To determine their polarity a supervised learning with Naive Bayes and SVM was applied. This approach was extended in (Haniewicz et al., 2014) where the semantic lexicon was expanded to 140,000 terms. To enlarge the lexicon the authors used a simple rule-based propagation with an adaptation of Random Walk algorithm.

SentiWordNet construction in its recent stages was generally based on the glosses from Princeton WordNet. (Misiaszek et al., 2013) proposed a lexicon construction method for wordnets, for which a simple transfer method could not be easily applied or external sources of knowledge such as tagged and disambiguated glosses are not available. This approach was based on relational propagation scheme with local, collective classification method, namely Iterative Classification Algorithm (ICA) for determining polarity of synsets. The training features for the classifier were obtained using only a neighbourhood of annotated synsets, consisting of nodes with known polarity. They manually annotated specific synsets in the wordnet and used them as seeds for the propagation process. However, the details of the feature extraction were not specified and there was no evaluation for their approach.

In (Kulisiewicz et al., 2015) the propagation was performed by using an adaptation of Loopy Belief Propagation (LPA) on Princeton WordNet 3.0. Three different variants of the LPA have been tested and evaluated. The evaluation was carried out in two ways. Firstly, the authors compared their results with polarity scores from SentiWordNet (*Mean Square Error*), but skipping the *Objective* class. Secondly, evaluation was done by comparison with polarity of words existing in the General Inquirer Lexicon. The resultant performance was ambiguous, and the main conclusion was that semantic relations within wordnet may not be a well correlated with the sentiment relations.

4 Classifier-based Polarity Propagation

We propose a fully automated method called *Classifier-based Polarity Propagation* (henceforth CPP) with a very rich set of features. In Sec-

tion 5.1 we compare the results obtained by CPP with rule-based and relation-based method called *Seed Propagation* and its best configuration presented by Maks and Vossen (2011).

4.1 Polarity Transfer from Units to Synsets

We analysed the contemporary annotation of plWordNet to see how diverse synsets are in terms of units polarity. In contrast to SentiWordNet the manual annotation in plWordNet is done on the level of lexical units (Zaśko-Zielińska et al., 2015). Available values for polarity are: *strong negative*, *weak negative*, *neutral*, *weak positive*, *strong positive*, *ambiguous*. One annotator can assign only one of these values for a single lexical unit.

Currently there are more than 83k annotations covering more than 54k lexical units and 41k synsets. About 22k of the polarity annotations are different than neutral and these annotations cover 13k lexical units and 9k synsets (22% of all synsets including annotated units). We found that 1.5k of these synsets were annotated with different polarity across their units. If we exclude neutral units, only 345 of them have varying polarity strength (e.g. synset that contains two lexical units annotated as strong positive and one annotated as weak positive). If we exclude both neutral and ambiguous annotations, there is only 41 synsets having conflicting, opposite polarity of their units (synsets that have both positive and negative units), and it is only 3.8% of all polarized synsets (synsets that do not contain any neutral units - 9164).

The acquired statistics show that synsets are strongly homogeneous in terms of the lexical units polarity, so we decided to move annotations from unit-level to the synset-level. In order to simplify the problem we decided to project these values to only three: *positive*, *negative*, *neutral*. For each annotation value we assigned the following weights: 2 for *strong* variants, 1 for weak variants, neutral and ambiguous. Then we recounted the number of annotations in each synset including assigned weights. For example, if we have a synset with a set of its lexical units like {*strong negative*, *negative*, *strong positive*, *neutral*}, we have total weight for *positive* category equal to 2, *negative* category equal to 3 (2 + 1) and *neutral* category equal to 1. We decided to assign only one polarity class to each synset – the one having the largest

Relation	Occurrences [%]
hyponymy	34.72
hypernymy	34.72
fuzzynymy	9.40
similar_to	3.20
feature_value	3.03
meronymy	1.86
holonymy	1.49
collection_meronym	1.29
collection_holonym	1.23
type	1.06
member	1.06
taxonomic_meronym	1.00
taxonomic_holonym	0.99
SUM	95.06

Table 1: Frequency (as part of the whole number of relations) of the selected relations in plWordNet.

weight. In the given example the assigned polarity will be *negative*. If we have two classes of the same weight, we apply the following rules to solve this discrepancy:

- $\{positive, neutral\} \rightarrow positive$
- $\{negative, neutral\} \rightarrow negative$
- $\{positive, negative\} \rightarrow neutral$

4.2 Features

We analysed the existing structure of plWordNet to select the most common relations. The results are presented in Table 1. We took a subset of relations which covers more than 95% of all relation instances in plWordNet.

Each synset is described by a set of features, where the feature value is represented as *bag-of-words* containing synsets or polarities. Each feature type is a set of 4 variables:

- *Relation* – one of the 13 relations given in Table 1.
- *Direction* – the direction of the relation, the described synset can be a *source* or *target* of the given relation.
- *Word* – there are two types of *words* in *bag-of-words* model: *synset_ID* (any number) and *synset_polarity* (one of the following numbers: $-1, 0, 1$; it represents 3 polarity classes: *negative, neutral, positive*).

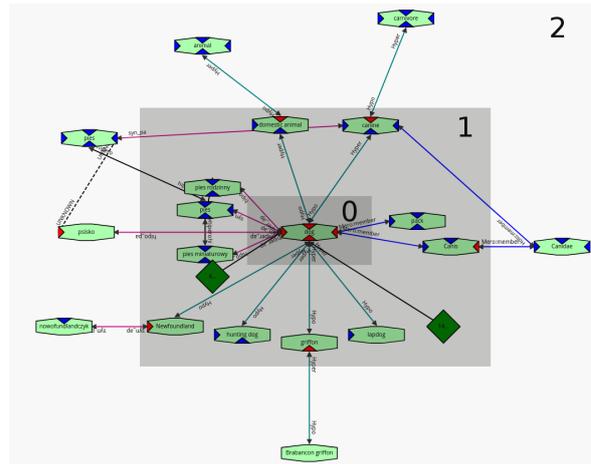


Figure 1: Example of synsets at the specific level (1 and 2) with respect to the synset at level 0.

- *Level* – the first level means synsets in *direct* relation to the described synset, the second level are synsets in direct relation with synsets from the first level, but excluding synsets from the first level. The example is presented in Figure 1.

There are 13 relations, 2 directions, 2 word types and 2 levels, which in total gives $13 \cdot 2^3 = 104$ types of features. For example a feature of the type *hyponym_source_id_level_2* contains all IDs of synsets which are sources of all hyponym relation instances, for which the target is any synset at the 1st level (see Figure 1).

4.3 Classifier

Having a set of annotated synsets and 104 bags of words as features for each synset, we utilised `TfidfVectorizer` module from `scikit-learn`¹ Python machine learning package. This feature extraction method allows to convert a collection of elements to a matrix of TF-IDF features. Each synset belongs to one of three following classes: *positive, negative, neutral*. Transformed data is used to train a predictive model. We used Logistic Regression from `scikit-learn` package as a classifier.

4.4 Propagation

With a trained classifier we perform propagation for the remaining, unlabelled part of plWordNet. At the beginning we treat our seeds as a set of synsets at level-0 (see figure 1). Each next iteration is a classification of synsets at the 1st level,

¹<http://scikit-learn.org>

using annotated synsets from the other levels. We prepared the solution using one of the following approaches applied to each iteration:

- naive – we get the graph order of the remaining synsets to be classified,
- sorted – before each iteration we sort synsets at the 1st level by the number of relations with synsets which already have the polarity value assigned (descending order).

5 Experiments and Results

5.1 Experimental Set-up

The developed method assumes that the propagation is performed only for synsets. However, existing polarity annotations in the plWordNet refer only to lexical units, thus some pre-processing was required. First, we used a simple generalization function to assign the polarity to the synsets, depending on the polarity of their units (see Section 4.1) and projecting a 5-degree scale of polarity to a 3-degree scale. Then, to evaluate the lexicon we prepared a large graph of plWordNet, consisting of generalized synsets.

5.2 Evaluation Procedure

The evaluation procedure utilises full plWordNet with 43k synsets annotated with sentiment polarity (positive, negative, neutral). Annotated synsets were divided once into 10 parts and 9 parts (about 40,400 synsets) were treated as a seed set for baseline (or learning set for CPP) and the last part (about 3,600 synsets) as a test set. For each method and configuration we performed 10-fold cross-validation.

We implemented a simple rule-based seed-driven propagation method described in (Maks and Vossen, 2011) to obtain a *baseline* (henceforth BASE). Then we compared the results with CPP in two variants described in Section 4.4: naive (CPP-N) and sorted (CPP-S).

5.3 Results and Discussion

Table 2 presents the results obtained during experiments. We calculated precision (P), recall (R) and F-measure (F) for separate classes of polarity: negative (NEG), positive (POS) and neutral (NEU). We compared differences between two pairs: {BASE, CPP-N} and {CPP-N, CPP-S}. In Tab. 2 we highlighted results for which differences were statistically significant. We anal-

Measure	BASE	CPP-N	CPP-S
P-NEG	84.01	84.58	84.73
P-NEU	92.18	93.75	93.66
P-POS	69.20	83.11	82.95
R-NEG	68.63	75.82	75.90
R-NEU	95.80	97.02	96.97
R-POS	64.64	68.41	67.80
F-NEG	75.52	79.91	79.81
F-NEU	93.95	95.34	95.35
F-POS	66.77	74.99	74.61

Table 2: Precision (P), recall (R) and F-score (F) for separate classes of polarity. BASE results are compared to CPP-N and CPP-S. Statistically significant differences are emphasised.

ysed the statistical significance of differences using paired-differences Student’s t-test with a significance level $\alpha = 0.05$ (Dietterich, 1998).

Naive solution (CPP-N) is significantly better than BASE in all test cases except precision for class *negative*. The order of neighbours classified in each iteration is not important in this case, because there was no significant difference between CPP-N and CPP-S variants.

6 Conclusions

The results prove that the proposed method performs better in almost all cases comparing to simple rule-based methods which transfer known polarity from seeds to other parts of wordnet. Surprisingly for us, the solution with sorting synsets in each iteration in descending order by the number of neighbours with known polarity did not provide any increase of propagation quality. We think that the further work should be concentrated on training the classifier after each iteration and in this scenario sorting before classifying should be beneficial.

Acknowledgments

Work co-financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and in part by the National Centre for Research and Development, Poland, under grant no POIR.01.01.01-00-0472/16.

References

- [Baccianella et al.2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, volume 10, 01.
- [Dietterich1998] Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.
- [Esuli and Sebastiani2006] Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.
- [Haniewicz et al.2013] Konstanty Haniewicz, Wojciech Rutkowski, Magdalena Adamczyk, and Monika Kaczmarek. 2013. Towards the lexicon-based sentiment analysis of Polish texts: Polarity lexicon. In Costin Bădică, Ngoc Thanh Nguyen, and Marius Brezovan, editors, *Computational Collective Intelligence. Technologies and Applications: 5th International Conference, ICCCI 2013, Craiova, Romania, September 11-13, 2013, Proceedings*, pages 286–295. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Haniewicz et al.2014] Konstanty Haniewicz, Monika Kaczmarek, Magdalena Adamczyk, and Wojciech Rutkowski. 2014. Polarity lexicon for the Polish language: Design and extension with random walk algorithm. In J. Swiątek, A. Grzech, P. Swiątek, and J. M. Tomczak, editors, *Advances in Systems Science: Proc. of the Int. Conf. on Systems Science 2013 (ICSS 2013)*, pages 173–182. Springer.
- [Kulisiewicz et al.2015] Marcin Kulisiewicz, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki, 2015. *On Sentiment Polarity Assignment in the Wordnet Using Loopy Belief Propagation*, pages 451–462. Springer International Publishing, Cham.
- [Liu2015] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 1.
- [Mahyoub et al.2014] Fawaz H.H. Mahyoub, Muazam A. Siddiqui, and Mohamed Y. Dahab. 2014. Building an arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University - Computer and Information Sciences*, 26(4):417–424. Special Issue on Arabic NLP.
- [Maks and Vossen2011] Isa Maks and Piek Vossen. 2011. Different approaches to automatic polarity annotation at synset level. In *Proceedings of the First International Workshop on Lexical Resources*, pages 62–69.
- [Maks et al.2014] Isa Maks, Ruben Izquierdo, Francesca Frontini, Rodrigo Agerri, Piek Vossen, and Andoni Azpeitia. 2014. Generating polarity lexicons with wordnet propagation in 5 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- [Maziarz et al.2016] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Pawel Kedzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268.
- [Miller1995] George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.
- [Misiaszek et al.2013] Andrzej Misiaszek, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki, 2013. *Relational Propagation of Word Sentiment in WordNet*, pages 137–140. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Piek Vossen and VanderVliet2008] Roxane Segers Piek Vossen, Isa Maks and Hennie VanderVliet. 2008. Integrating lexical units, synsets and ontology in the cornetto database. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.
- [Stone1966] Philip J. Stone. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.
- [Strapparava and Valitutti2004] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.
- [Zaško-Zielińska et al.2015] Monika Zaško-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP'2015*, pages 721–730, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities

Bolette S. Pedersen¹, John McCrae², Carole Tiberius³, Simon Krek⁴

University of Copenhagen, Denmark¹, National University of Ireland Galway, Ireland², Dutch Language Institute, The Netherlands³, Jožef Stefan Institute, Slovenia⁴
bspedersen@hum.ku.dk¹, john@mccr.ae², Carole.Tiberius@ivdnt.org³, simon.krek@guest.arnes.si⁴

Abstract

The paper describes objectives, concept and methodology for ELEXIS, a European infrastructure fostering cooperation and information exchange among lexicographical research communities. The infrastructure is a newly granted project under the Horizon 2020 INFRAIA call, with the topic Integrating Activities for Starting Communities. The project is planned to start in January 2018.

1. Background

Reliable and accurate information on word meaning and usage is of crucial importance in today's information society. The most consolidated and refined knowledge on word meanings can traditionally be found in dictionaries – monolingual, bilingual or multilingual.

Dictionaries are not only vast, systematic inventories of information on words, they are also important as cultural and historical artefacts. In each and every European country, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited.

Consequently, the lexicographic landscape in Europe is currently rather heterogeneous. On the one hand, it is characterised by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, complicating reuse of this valuable data in other fields, such as natural language processing, linked open data and the Semantic Web, as well as in the context of digital humanities. On the other hand, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. This forms a major obstacle to more ambitious, inno-

vative, transnational, data-driven approaches to dictionaries, both as tools and objects of research.

In 2013, the European lexicographic community was brought together for the first time in the European Network of e-Lexicography (ENeL) COST action (www.elexicography.eu). This initiative was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. This networking initiative, which ended in October 2017, started with 34 members from 20 countries but grew to 285 members from 31 countries. In the context of this network, a clear need emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, artificial intelligence, NLP and digital humanities.

As a synthesis of the ENeL efforts, a consortium was established in 2016 and in August 2017 the proposal for an infrastructure under the name ELEXIS was selected for funding in the Horizon 2020 INFRAIA call with the topic Integrating Activities for Starting Communities. The project is planned to start in January 2018.

In the following sections we will outline the objectives, concept and methodology of the infrastructure, and finally we will sketch out some foreseen wordnet related research tasks in the project concerning sense clustering and multilingual linking.

2. Objectives

The main objectives of ELEXIS can be summarized as follows:

- to foster *cooperation* and *knowledge exchange* between different research communities in lexi-

cography in order to reduce the gap between lesser-resourced languages and those with advanced e-lexicographic experience; and

- to work with strategies, tools and standards for *extracting, structuring* and *linking* of lexicographic resources;
- to facilitate *the access* to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders;
- to encourage to an *open access culture* in lexicography, in line with the European Commission Recommendation on access to and preservation of scientific information.

ELEXIS is based on the conviction that lowering the barrier for retrieving and analysing multilingual lexicographic data across Europe cannot be accomplished in the long term without lowering the barrier for providing lexicographic data to research infrastructures. As a result, the following impacts are pursued:

- efficient (open) access to high quality lexicographic data for researchers, institutions and stakeholders from different fields;
- a common platform for building, sharing and exploiting knowledge and expertise between lexicography and computational linguistics which will facilitate cross-disciplinary fertilisation and a wider sharing of information, knowledge and technologies across and within these fields. The platform will thus aim at bridging the gap between lesser-resourced languages and those with advanced e-lexicographic and/or computational linguistic experience;
- the creation of a scalable, multilingual and multifunctional, language resource. By integrating and linking lexical content and interlinking it with other structured or unstructured data - corpora, multimodal resources, etc. - on any level of lexicographic description, the project will strive towards creating a multilingual and multifunctional language resource incrementally enriching the available information;
- the inter- and multidisciplinary nature of lexical data will help researchers ask new questions and pursue new avenues of research.

3. ELEXIS Participants

The ELEXIS consortium includes the following 17 participants:

1. "Jožef Stefan" Institute, Slovenia
2. Lexical Computing, Czech Republic
3. Dutch Language Institute, The Netherlands
4. Sapienza University of Rome, Italy
5. National University of Ireland, Galway, Ireland
6. Austrian Academy of Sciences, Austria
7. Belgrade Center for Digital Humanities, Serbia
8. Hungarian Academy of Sciences, Research Institute for Linguistics, Hungary
9. Institute for Bulgarian Language »Prof Lyubomir Andreychin«, Bulgaria
10. Faculty of Social Sciences and Humanities, Universidade Nova de Lisboa, Portugal
11. K Dictionaries, Israel
12. Consiglio Nazionale delle Ricerche - Istituto di Linguistica Computazionale "A. Zampolli", Italy
13. The Society for Danish Language and Literature, Denmark
14. University of Copenhagen, Denmark
15. Trier University, Center for Digital Humanities, Germany
16. Institute of Estonian Language, Estonia, and
17. Real Academia Española, Spain

4. Concept and Methodology

ELEXIS will build on the existing expertise and knowledge of partners in the fields of lexicography, computational linguistics and artificial intelligence in an interdisciplinary effort to make existing lexicographic resources available on a significantly higher level compared to their availability as stand-alone resources, which is to a certain degree the current state of affairs.

These resources are in fact results of long-term projects in which literally thousands of person years were and continue to be dedicated to their compilation in national and regional projects, and in most cases they represent the most consolidated and refined knowledge on word meanings in individual languages. A tremendous effort is needed for their compilation, and this implies the necessity to control the contents in order to ensure both the continuation of consistent language description and maximum quality of the results. Furthermore, and resulting from current isolation of efforts, these resources are typically encoded in incompatible data structures. Both issues con-

tribute to the fact that the data from these resources is currently not fully accessible for extensive, interoperable computer use.

On the other hand, the language technology (LT) community, for their part, created an overwhelming number of different types of lexical resources over the last thirty years, which are used for natural language processing tasks. These include corpora, lexicons, glossaries (used in machine translation), machine-readable dictionaries, lexical databases, and many others. One of the important issues that will be addressed by ELEXIS is the fact that the impressive results of the LT community have only to a limited degree found their way into the practical work of creating lexicographic resources in the past. This can be largely attributed to the lack of a common platform for building, sharing and exploiting knowledge and expertise between computational linguistics and lexicography, which is one of the goals of the ELEXIS infrastructure.

4.1 Supporting lexicographic process and language description

To support the lexicographic process and to contribute to lexicographically-oriented language description ELEXIS will work towards:

- developing methods and tools for the automatic processing and extraction of data from corpora and other (multimodal) resources for lexicographic purposes;
- developing methods and tools for the inclusion of extracted data into interlinked (open) lexicographic data;
- developing methods, guidelines and tools enabling the use of crowdsourcing and citizen science in the lexicographic process;
- elaborating on the guidelines and solutions for handling copyright and authorship protection to enable inclusion of extracted data into the lexicographic workflow.

4.2 Supporting natural language processing

To support the natural language processing community, several steps are needed to make existing lexicographic resources globally available. ELEXIS will:

- develop methods, guidelines and tools for harmonisation of dictionary formats, building on the existing standards within the lexicographic and NLP community;

- develop methods and tools for automatic segmentation and identification of dictionary structure, enabling interlinking of dictionary content;
- develop methods and tools for interlinking, maintenance, reuse, sharing and distribution of existing lexicographic resources;
- define evaluation and validation protocols and procedures (lexicographic data seal of compliance);
- elaborate on the guidelines and solutions for handling copyright and authorship protection to enable open access to lexicographic data in LOD framework.

Therefore, in contrast with previous more NLP-oriented efforts spanning from computational lexicographical projects like EAGLES (Calzolari et al. 2002), PAROLE/SIMPLE (Lenci et al. 2000) to a current infrastructure on language resources and technology like CLARIN, ELEXIS will develop methods and tools to produce collections of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms. In this context, focus will be on defining interoperability standards and data services in close cooperation with the existing CLARIN and DARIAH infrastructures.

4.3 Methodology

Lexicography as a field has a long tradition of refining semantic description of individual languages in comprehensive monolingual dictionaries, or performing detailed contrastive analysis between two or more languages in bilingual and multilingual dictionaries. However, these resources are currently not used to a sufficient degree within existing and emerging language technologies. They are almost completely absent in linked (open) data clouds and Semantic Web technologies, and are to some degree “digitally invisible”.

In the last decade the new field of e-lexicography emerged, which can be seen in initiatives such as the ENeL COST action (<http://www.elexicography.eu/>), the eLex conference series (<https://elex.link/>), or the Globalex workshop at LREC 2016 (<http://ailab.ijs.si/globalex/>). Globalex is the first

initiative which includes all continental lexicographic associations: EURALEX, ASIALEX, AFRILEX, AUSTRALEX and the Dictionary Society of North America. The field of e-lexicography is dedicated to creating digitally-born dictionaries defined as lexical resources intended for human users but intentionally moving away from the paper medium and exploring the almost infinite possibilities of the new digital environment, with a view to take human-oriented lexical description to entirely different levels. In this context, machine learning, data mining and other computational techniques are starting to find their way into lexicography. Combining both traditional lexicographic knowledge and expertise with computational linguistics, while engaging also wider language communities in the process, creates huge potential for the development of the field.

4.4 Lexicography and »semantic bottleneck«

Lexicographic resources contain quality information about general vocabulary and more difficult types of language phenomena such as highly polysemous words or semantically opaque multi-word expressions (idioms, phraseology), which are rather inconsistently covered in LT-oriented resources. These phenomena represent a bottleneck in achieving precision and computational efficiency of NLP applications. This can be seen also from efforts such as PARSEME COST action (<http://typo.uni-konstanz.de/parseme/>) which was devoted to the role of multi-word expressions in parsing. Word sense disambiguation as part of content analytics, text understanding and computer reasoning remains another complex task for computational processing of text, and is still largely unsolved, especially for languages other than English. Typically, resources such as Wikipedia, Wiktionary, wordnets or framenets are used for word sense disambiguation tasks, collected in the (L)L(O)D cloud (<http://linked-data.org/>, <http://www.linguisticlod.org/>). Knowledge bases and complementary applications such as BabelNet (<http://babelnet.org/>), Babelfy (<http://babelfy.org/>), Cyc (<http://sw.opencyc.org/>) or wikifiers (<http://www.wikifier.org/>) have been developed to enrich text processing with semantic information. ELEXIS proposes enriching the existing linked data clouds and knowledge bases with data available in existing and new lexicographic resources, which are currently not used for solving these tasks.

4.5 Standards in lexicography and NLP

There are several reasons for the negligible incorporation of lexicographic data in LT so far. The first is almost non-existent interoperability and use of common standards in lexicography. In past decades there were several important efforts to harmonise and standardise linguistic resources, including lexicographic resources. These include first initiatives such as EAGLES/ISLE, Multext(-East), PAROLE, SIMPLE, CONCEDE etc. in the 1990s. From these efforts, standards emerged such as Text Encoding Initiative (TEI - <http://www.tei-c.org/>), Lexical Markup Framework (LMF - <http://www.lexicalmarkupframework.org/>), and others, most of them under the umbrella of the Terminology and other language and content resources ISO/TC 37 standard.

The standardisation process was much more successful with resources directly dedicated to computer use, such as corpora, lexicons, lexical databases, wordnets, ontologies etc., but standards were less successful in case of lexicographic resources initially intended for human users.

4.6 Availability of lexicographic data

Although early digitisation projects involving lexicographic resources date back to the 1980s (Boguraev and Briscoe, 1989), or in case of English even the 1960s (Urdang, 1966), and even if the 1990s saw massive digitisation of existing dictionaries, including works like the Oxford English Dictionary, general (open) access to lexicographic data is extremely limited. The main reason for this is the massive effort necessary to compile such resources, either by national language institutions, or by commercial companies in the case of “commercial languages” with a sufficient number of speakers.

The effort needed consequently implies the necessity to control the contents, resulting in the need to resolve intellectual property right issues before this data can be included in open access infrastructures.

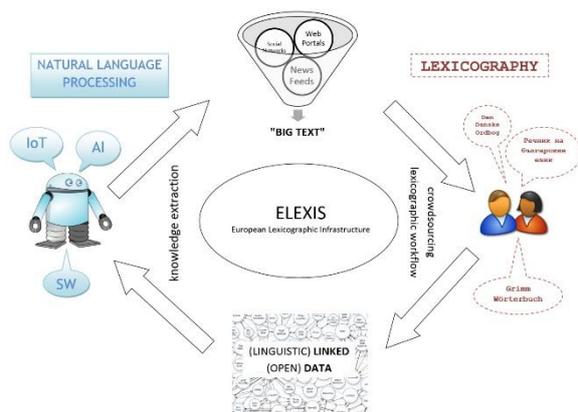
The ELEXIS infrastructure will dedicate serious efforts to handle IPR issues related to lexicographic data and enable their integration as linked data. In the last decade, initiatives promoting open access to the results of publicly funded projects (Open Research Data Pilot etc.) and the increasing wealth of (open) data available on the Web (Wikipedia, Wiktionary etc.), also instigated new trends within lexicography,

particularly the move towards e-lexicography. This new trend is not yet supported by an infrastructure where quality semantic data from dictionaries could be linked, shared, distributed and stored on a massive scale. Therefore, the objective of contributing quality semantic data in the digital age means that the proposed project will work towards enabling existing lexicographic resources to be included seamlessly into the Linked (Open) Data family (see Picture 1).

4.7 Virtuous cycle of e-lexicography

As was established in ENeL surveys, the results of the LT work are rarely used in lexicography, which is one of the important issues addressed by the ELEXIS infrastructure. This can be largely attributed also to the lack of an infrastructure enabling sharing knowledge and expertise between LT and lexicography. Ideally, the part of the virtuous cycle starting from NLP towards lexicography will produce proto-dictionary content in a completely automatic process with the use of machine learning, data mining and information extraction techniques focusing on massive amounts of data in various modalities available on the Web.

ELEXIS aims to develop methods and tools to produce such collections of structured data in an automated process where the data can be used as a starting point for further processing of the collected material either by traditional lexicographic process or through crowdsourcing platforms.



Picture 1: Virtuous cycle of e-lexicography

5 ELEXIS and Wordnets

5.1 Sense clustering and predominance information

Based on achievements from BabelNet (Navigli et al. 2012) and other works as found in Izquierdo et al. 2009, McCarthy et al. 2016, and Pedersen et al. forthcoming, ELEXIS will work on developing principled methods for sense clustering which are preferably semi-automatic. This also involves wordnet sense inventories which in many cases incorporate ontological typing but are on the other hand not organised in main- and sub-senses (in opposition to most dictionaries).

The project will include frequency and predominance information of senses in this work with the overall aim of improving word sense disambiguation and other NLP-related tasks. The intention is to work towards making existing lexical resources including wordnets more operational and practically useful in NLP by focusing on the organisation of the sense inventory.

Frequency and predominance information of senses is however not information which is directly accessible for all the involved languages at the current stage. Therefore, an initial task will be to develop methods to process these data for the less-resourced European languages.

5.2 ELEXIS and the WordNet Interlingual Index

The Global WordNet Association has proposed an Interlingual index of concepts (Bond et al., 2016), in which synsets from any wordnet can be identified with a single unique identifier, enabling interlingual linking of wordnets. It is clear that these goals correspond well with those of the ELEXIS project and it is expected that the benefits of these tools will be offered also to the wordnet community.

As a minimal step to enable this, the XML LMF format of the Global Wordnet Association¹ will be supported as a valid input and output format to the tools developed in the context of ELEXIS. Thus the linking tools that will establish cross-lingual similarity between concepts will be applicable to wordnets and thus this will be used to detect duplicate concepts between different wordnets and ameliorate the task of introducing new interlingual identifiers. Secondly it is hoped that the knowledge extraction components of the ELEXIS infrastructure will be integrated

¹ <http://globalwordnet.github.io/schema>

into the lexicographic procedures for new wordnet creation, and we intend to demonstrate this by integrating the crowd-sourced procedure used to create the Colloquial Wordnet (McCrae et al., 2017) within the ELEXIS infrastructure. In particular, this resource selects potential neologisms by NLP analysis of Twitter in order to detect terms that appear to be emerging in the English language.

Finally, it is expected that ELEXIS will encourage a closer interaction between the BabelNet project (Navigli & Ponzetti 2012) and other wordnets. In particular, BabelNet and the Interlingual Index will be linked so that users may access the data through either interface and new concepts in either resource can be integrated automatically. Furthermore, we will define metadata such that the licensing and sources for information can be clearly and unambiguously identified.

Acknowledgement

This work was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

- Boguraev B. and Briscoe T. (Eds.). (1989). *Computational Lexicography for Natural Language Processing*. Longman Publishing Group, White Plains, NY, USA.
- Bond F., P. Vossen, J. McCrae, Ch. Fellbaum (2016). CILI: the Collaborative Interlingual Index, in: *Proceedings of the 8th Global WordNet Conference 2016 (GWC2016)* in Bucharest, Romania, January 27-30.
- Calzolari N., Zampolli A., Lenci. (2002). Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. In: *CICLing 2002: Computational Linguistics and Intelligent Text Processing* pp 264-279
- Izquierdo, R., A. Suárez, and G. Rigau. (2009). An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* pp 389-397. The Association for Computational Linguistics.
- Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli. (2000). SIMPLE: A general framework for the development of multilingual Lexicons. *International Journal of Lexicography*, 13(4), 249-263
- McCarthy, D., M. Apidianaki & K. Erk (2016). Word Sense Clustering and Clusterability. In: *Computational Linguistics*, Vol. 42, no. 2.
- McCrae J.P., Wood I., Hicks A. (2017) The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In: Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., Hellmann S. (eds.) *Language, Data, and Knowledge*. LDK 2017. Lecture Notes in Computer Science, vol. 10318. Springer, Cham.
- Navigli, R. and S. P. Ponzetto. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193: 217-250.
- Pedersen, B.S., M. Agirrezabal, S. Nimb, S. Olsen, I. Rørmann (forthcoming). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of Global WordNet Conference 2018*, Singapore.
- Urdang L. (1966). The Systems Designs and Devices Used to Process: The Random House Dictionary of the English Language. *Computers and the Humanities* 1 (2).

Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic Analysis in the Domain of Theology

Laura Slaughter
University of Oslo
Oslo, Norway
laurasla@ifi.uio.no

Wenjie Wang, Luis Morgado da Costa[♦], Francis Bond
[♦]Global Asia, Interdisciplinary Graduate School,
School of Humanities,
Nanyang Technological University, Singapore

Abstract

We aim to support digital humanities work related to the study of sacred texts. To do this, we propose to build a cross-lingual wordnet within the domain of theology. We target the Collaborative Interlingual Index (CILI) directly instead of each individual wordnet. The paper presents background for this proposal: (1) an overview of concepts relevant to theology and (2) a summary of the domain-associated issues observed in the Princeton WordNet (PWN). We have found that definitions for concepts in this domain can be too restrictive, inconsistent, and unclear. Necessary synsets are missing, with the PWN being skewed towards Christianity. We argue that tackling problems in a single domain is a better method for improving CILI. By focusing on a single topic rather than a single language, this will result in the proper construction of definitions, romanization/translation of lemmas, and also improvements in use of/creation of a cross-lingual domain hierarchy.

1 Introduction

Sacred texts, including scriptures and exegesis, are the primary source of insight for scholars seeking to understand religious beliefs and practices of past cultures. Scholarly work on ancient manuscripts and papyri has seen an increase in the use of language technologies and automated processing methods. Making texts available in machine-readable formats is a priority within Digital Humanities (DH) projects and new tools are being developed for automated translation and alignment, and also to

determine text similarity or patterns of variation. Wordnets play an integral role in improving performance and have been constructed for processing texts in key ancient languages, including Ancient Greek (Bizzoni et al., 2014), Sanskrit (Kulkarni et al., 2010), and Pre-Qin Ancient Chinese (Zhang et al., 2017).

The ideologies of religious traditions are manifested in their sacred texts, with texts being copied, paraphrased, revised, and dispersed over time as a tradition spreads. Various DH projects and tools are developed to process digital editions of aligned multilingual texts. For example, eTrap,¹ the Electronic Text Reuse Acquisition Project has developed a tool called TRACER for assessing text similarities and detecting reuse of parts of texts in later works (i.e. when a portion of text has been appropriated by a later author). It has been used to identify biblical quotes in Swedish literature (Kokkinakis and Malm, 2015) and also to assess similarity of early Christian Coptic texts (Miyagawa et al., 2016, Manuscript submitted for publication.). TRACER makes use of BabelNet² which is based on modern language wordnets but does not include ancient languages, leading to less than satisfactory results.

There are now several multi-lingual corpora available in digital formats. For example, the well-known SAT Daizōkyō Text Database (Nagasaki, 2008), a repository making available Buddhist canonical texts in Sanskrit, Tibetan, Chinese, and Japanese, are available for scholars studying the flow of Buddhism from India to China and then into Japan. Projects such as these are expanding their ability to process multi-lingual texts.

Processing of multi-lingual texts in ancient

¹<http://www.etrapp.eu>

²<http://babelnet.org>

languages is hindered by the inaccuracies of wordnets. An example of this, related to Ancient Greek, is described in Berti et al. (2016). One of the main problems is that wordnets built by bootstrapping from modern languages can be highly erroneous and also miss concepts that are not lexicalized in modern languages (Peters et al., 1998). Bond et al. (2016) proposed the Collaborative Interlingual Index (CILI) as a means to solve just this problem by creating a repository of cross-lingual concepts connecting wordnets.

Future versions of the Open Multilingual Wordnet (OMW) (Bond and Foster, 2013) will use CILI, instead of the current Princeton WordNet (Fellbaum, 1998), to link individual wordnets cross-lingually. This will make it easier to create new concepts that deviate from language specific concept hierarchies, and will invite new ways to link and investigate meaning across languages. The study of sacred texts using DH methods will greatly benefit from CILI, and along with the wordnets for ancient languages will improve results when working with multi-lingual corpora.

This paper discusses the initial steps to organize concepts across wordnets for the case of supporting scholarly work on sacred texts. In the next section, we present results from a survey of the literature in theology that gave us insight into the domain. We then outline specific observations related to coverage, definitions, and structure in Princeton Word Net (PWN). The final section will discuss our proposal that specific domains, such as religion/spiritualism, can be enriched through domain-focused multilingual wordnets.

2 Concepts Related to Theology

The simple question concerning which concepts are most relevant to scholars studying sacred texts is a pertinent one. In an attempt to understand what scholars writing journal articles refer to as a concept within the field of theology, we conducted a search within the theology literature database, ATLA Religion Database® (ATLA RDB®). This database covers the following areas: Bible, archaeology, and antiquities; human culture and society; church history, missions, and ecumenism; pastoral ministry; world religions and

religious studies; theology, philosophy, and ethics. There are over 1.8 million records, 667,000+ journal article records, 267,000+ essay records, 607,500+ review records, and 300,500+ book records. The search strategy combined the keyword “concept” in the title and “comparative study” as a subject, making use of ATLA’s subject index. The total number of relevant articles retrieved was 116.

The results of this effort are shown in Figure 1. Relevant journal articles were included if they argued that a specific concept was expressed in two or more religious traditions. A wide-range of concepts (e.g. forgiveness) are seen in this literature and also named entities (e.g. God). Terms are usually presented in their original language, and a suitable translation is presented for the English-speaking reader. There were terms provided in diverse languages: Arabic, Hebrew, Sanskrit. Some of the Sanskrit terms are currently standard in the modern English language such as *karma* and *nirvana* although not necessarily with the original meanings.

This review provided us with a starting point. We can begin to think of connections to upper-level ontology categories. Only a few of the concepts from the literature review are related to acts or practices. These are things that a person must actively do and they have endpoints. Forgiving and repentance are acts, or rather processes, with an outcome (for forgiveness, this is alleviation of anger and resentment). Actions can be further divided, some of these are internal processes, for example, deciding and ridding oneself of vengeful thoughts, while others are external actions (including rituals) or deeds having a specific meaning for the doer. Many of the concepts are not actions but experiences that are felt, such as awakening, emptiness, oneness or the experience of humility. Other abstract concepts discussed in the literature are: freedom, justice, and evil. These are non-physical but are observable when acted out.

In this project, we can begin to describe the types of concepts of interest to this domain. We also plan to catalog specific named entities that are found within texts and compare them with those available in existing wordnets, for example, a list of specific places (e.g.

Hell), individuals, supernatural beings, mythical beasts, or objects with magical properties. Some of these entities are no doubt available in wordnets (e.g. Jesus), and we consider how these might be related to other general categories (e.g. prophet).

One question that does come to mind, given the volume of discussion in the literature that resulted in Figure 1, is whether it is at all feasible to connect “theological concepts” in CILI from such diverse languages. This work will be challenging due to the wide variety of concepts expressed in different traditions and how these are interpreted within various cultures. We do expect that it is a feasible task to find cross-lingual equivalents based on prior work on human spirituality. Boyer’s (2016) research explains religious concepts in terms of evolved cognitive dispositions and states that these are universal in human minds. He writes “human beings seem disposed to entertain thoughts about non-physically present agents, this includes their thoughts about absent or deceased persons, but also about mythical heroes, fictional characters and a variety of superhuman agents with, usually, counter-intuitive physical capacities but standard mental processes, such as gods, spirits, ancestors, shadows and the like.” The anthropologist Donald Brown (2004) produced a compilation of traits that he found were common to all human cultures. From his list, belief in supernatural/religion was one of these traits and there are several others that might be considered elements of common human spirituality, including beliefs about death, divination, empathy, imagery, magic, and moral sentiments.

Nevertheless, given the wide-range of human beliefs, even within a single language there may be problems in defining religious/spiritual concepts, so we can’t anticipate that cross-linguistic consensus will naturally emerge. For this reason, it may be the case that some domains are better tackled in domain-focused multilingual wordnets. This would facilitate the proper construction of definitions, romanization/translation of lemmas, and also a better use of/creation of cross-lingual domain hierarchy that could be targeting CILI directly and not each individual wordnet.

3 Examination of Wordnet Synsets

We did two key tasks in order to examine the available synsets in PWN that are relevant to support scholarly work on sacred texts. First, we examined glosses of synsets and possible additions to CILI. Second, we looked at existing categories within PWN: synsets connected to WordNet Domains³ and hierarchy within PWN.

3.1 PWN Synset Glosses

From Figure 1, we looked for an equivalent English synset for the concept *sunyata*, शून्यता in the Sanskrit WordNet which was translated and discussed as *emptiness* in the literature. *Sunyata*, शून्यता has two available senses in the Sanskrit WordNet, one is connected to the PWN synset for *emptiness* (PWN3.1:14478672-n having the gloss “*the state of containing nothing*” and the other is linked to a different translation altogether for *lack, deficiency, or want* (PWN3.1:14472871-n “*the state of needing something that is absent or unavailable*”, but neither of these would be accurate according to definitions given by Buddhists who explicitly state that *emptiness* is not defined as “*containing nothing*” (e.g., Suzuki, 2002) and it is never translated in theological terms as *lack, deficiency, want*.

We also looked at the concept of *offering*, which has two noun synsets. One is the contribution (PWN3.1:13270373-n “*money contributed to a religious organization*”), and the other the act of giving (PWN3.1:01041498-n “*the act of contributing to the funds of a church or charity*”). We found both definitions to be narrower than what we had expected. The former synset’s definition restricts offerings to just money, even though other items (such as food) can be offered, and the latter’s definition restricts the act to that of only contributing to the church/charity, and only to their funds.

In Hinduism and Buddhism, *daana* (Sanskrit/Pali: *dāna दान*; *gift-giving* or *generosity*) is the cultivation as well as the practice of generosity and giving. The “gift” in question can be alms; contributions to monasteries and temples, to charity, to the needy; hospitality. Sanskrit has various words

³<http://wndomains.fbk.eu>

- | | | | |
|---------------------|----------------------|--------------------------------|------------------------|
| • anamnesis | • free will | • known by God | • reality |
| • awakening | • justice | • law—natural law | • redemption |
| • best place | • god—creator spirit | • light | • repentance |
| • charisma | • God—supreme being | • love— God’s love | • revelation |
| • contemplation | • Godhead | • loving-kindness ^a | • rina ^c |
| • covenant concept | • good works | • meditation | • salvation |
| • creation | • grace—divine grace | • mercy | • self |
| • devil | • heaven | • messiah concept | • sin |
| • divine action | • hebdomadal | • miracle | • soul |
| • divine personhood | • Holy Spirit | • new creation | • spirit |
| • divine providence | • hospitality | • nirvana | • spiritual perfection |
| • duality | • just war | • no-self, also no-I | • submission |
| • dyadic nature | • justice | • nothingness ^b | • time |
| • evil | • heaven | • paradise (after-life) | • ubuntu ^d |
| • exile | • Hell | • power—supernatural powers | • universal savior |
| • forgiveness | • higher self | • progressive solemnity | • wilderness |
| • free choice | • karma | • punitive justice | • wisdom |
| • freedom | • kingdom, of God | | |

^a“chesed” in Hebrew

^btranslation of “ayin” in Hebrew

^cdebts owed to persons, gods, and ancestors

^dmeans personhood, humanness

Figure 1: A Sample of Concepts from ATLA Religion Database®

to describe different types of such offerings. The word *Daana/dāna* (दान) is not found in Sanskrit Wordnet. This would not have been covered by the existing *offering* synsets in PWN. Another related term to *daana*, *Paropakāra* (परोपकार), meaning benevolence or charity, is not found in either the Sanskrit Wordnet nor in PWN. Sanskrit: *Bhiksha* (भिक्षा) is linked to two PWN synsets, *handout* (PWN3.1:01092266-n) and *beggary* (PWN3.1:07202656-n), though it’s meaning is closer to an existing PWN synset for *alms* (PWN3.1:01092041-n).

3.2 PWN: Hierarchy and Classifications

The concepts themselves are not always consistent in how they are placed in the wordnet structure and linked to one another. For example, the synset for *Kuan Yin* is defined as “a female *Bodhisattva*”, and *Avalokitesvara* as “a male *Bodhisattva*”. However, only the latter is linked to the synset for *Bodhisattva*.

We also found that related synsets are not always linked to one another. Following up on the previous examples, *Kuan Yin* and *Avalokitesvara* are different forms of the same *Bodhisattva* (with the former being the East Asian Buddhism variant), but this is not reflected in the relation between the two synsets, nor in their definitions. Also, the two synsets for

Buddha, one specifically for the historical Buddha (Gautama Buddha), and the other for the concept of a perfectly enlightened being, are also not linked to each other or to *Buddhism*. Synsets for other major figures of various religions — including Jesus Christ and Mohamad — are similarly not linked to their respective religions.

Another concept is that of *spiritual beings*, which has *god* and *satan* as instances, with *angel*, *deity*, *fairy*, etc, as hyponyms. The instances are specific to Christianity (or the Abrahamic religions), and other spiritual beings from other religions are not linked to this synset. We see that the *spiritual being* synset has the hyponym *deity*. Reasonably, the synset for *god* should instead be made a hyponym of *deity*, instead of being an instance of *spiritual being*, albeit that being possible as well. The position of the synset *satan* could likewise be reconsidered. The biblical *Satan*, or the Devil, had been an angel and is now a “fallen angel”. Should the *satan* synset be considered under *angel*, or a new hyponymous *fallen angel* concept? This brings to surface the issue of specificity when it comes to the position of the existing synsets in the hierarchy, and how fine-grained such classifications should be made.

In addition to the hierarchy and structure within PWN discussed above, work has been

done to link synsets to domain categories. In Gella et al. (2014), a mapping between WordNet Domains, WordNet topics, and Wikipedia gives us a coarse alignment between WordNet and Wikipedia. The WordNet Domains contain about 200 domain labels that were selected from dictionaries and then structured into a taxonomy based on the Dewey Decimal Classification (Scott, 1998). All of the PWN 1.6 synsets were assigned domains in this project. In all, 2055 synsets are assigned to the domain *Religion*. These do provide a starting point, but the domain labels themselves were assigned roughly based on the conceptual relationships already in PWN and, as we have shown above, there are many issues that must be addressed. The majority of the labeled synsets for Religion are linked to Christian theology, and more specifically to Roman Catholic Christianity. We need to evaluate the accuracy of the labels. *Paradise* (PWN3.1:05636722-n) is assigned the label *Christianity* though this is the term generally used in Islam, and the definition given within PWN is not Christian-specific, “*the abode of righteous souls after death*”.

Another question is how to deal with potential relationships between cross-lingual synsets and whether the domain labels assigned reflect useful categories for scholars. Lefebure (1997) equates the Buddhist concept *sunyata* (शून्यता) with *grace*. Looking at *grace* in PWN, there are two synsets with domain labels for Christian theology, (PWN3.1:14481629-n “*a state of sanctification by God; the state of one who is under such divine influence*”) and the other is (PWN3.1:04847946-n “*the free and unmerited favor or beneficence of God*”). Both *grace* and *sunyata* are states, but the overall discussion of the relationship between these states is complex.

4 Discussion

In the above section, we provided examples for a wide-range of domain-related issues we uncovered in this domain. Synsets are missing from PWN and other wordnets for key concepts in theology (e.g. *emptiness*). We’ve seen badly formed definitions; those that are too narrow and overly restrictive in addition to having varying details. The hierarchical struc-

ture of PWN in this domain is inconsistent. Links to specific instances are missing (e.g. *Kuan Yin*). We have questions concerning the level of granularity and how to deal with fuzzy relationships related to dogma.

We propose a cross-lingual wordnet within the domain of theology; the process of connecting synsets adds to CILI. We believe that creation of such a wordnet will provide a methodology for similar needs in other domains as well as insight that helps correct problems in the single-language wordnets. We start by engaging with experts, scholars who study religious texts, to help with defining religious/spiritual concepts. The proposed method is to make use of the OMWEdit tool (Da Costa and Bond, 2015), a web-based system that is capable of multilingual browsing and editing concepts within OMW. The tool is freely available under an open license and can be used to annotate a corpus. Scholars examining texts in the target languages will be asked to contribute. The goal is to get scholars to provide and annotate parallel texts. These scholars will help us to add new entries to CILI, understand the extent of missing synsets, and make clear the relationships between synsets when it is not possible to identify equivalence.

Wordnets we ultimately wish to examine as part of the proposed work include: Ancient Greek (Bizzoni et al., 2014), Latin (Minozzi, 2009), Sanskrit (Kulkarni et al., 2010), Pre-Qin Ancient Chinese (Zhang et al., 2017), and Quranic Arabic (AlMaayah et al., 2014). However, we start on a smaller scale, with proposed work based on specific corpora and limiting the number of languages with a focus on depth-first before breadth. Future work will also incorporate on-going results providing CILI definition guidelines (Seppälä, 2015) and the model for diachrous lexical variants from DICOLOD, the Clariah project⁴.

Acknowledgments

This research was partially supported by the joint research project on *Multilingual Semantic Analysis* between Fuji Xerox Corporation, Japan and Nanyang Technological University, Singapore.

⁴<https://github.com/cltl/clariah-vocab-conversion/blob/master/dicolod-documentation.pdf>

References

- Manal AlMaayah, Majdi Sawalha, and M Abushariah. 2014. A proposed model for Quranic Arabic WordNet. In *Proc. 2nd Workshop on Lang Resources and Eval for Religious Texts, 31 May*. LRA, pages 9–13.
- Monica Berti, Gregory R Crane, Tariq Yousef, Yuri Bizzoni, Federico Boschetti, and Riccardo Del Gratta. 2016. Ancient Greek WordNet meets the dynamic lexicon: The example of the fragments of the Greek historians. In *Proc. 8th Global WordNet Conf.* pages 34–8.
- Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory R Crane. 2014. The making of Ancient Greek WordNet. In *Proc. 9th International Conf on Lang Resources and Eval (LREC'14)*. pages 1140–7.
- Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proc. 51st Annual Meeting of the Assoc for Comp Ling.* Sofia, pages 1352–62.
- Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proc. 8th Global WordNet Conf.* pages 50–7.
- Pascal Boyer. 2016. Explaining religious concepts. *Mental Culture: Classical Social Theory and the Cognitive Science of Religion* page 164.
- Donald E Brown. 2004. Human universals, human nature and human culture. *Daedalus* 133(4):47–54.
- Luis Morgado Da Costa and Francis Bond. 2015. OMWEdit- the integrated open multilingual wordnet editing system. *Proc. ACL-2015, System Demonstrations* pages 73–8.
- Christiane Fellbaum. 1998. *WordNet*. MIT Press.
- Spandana Gella, Carlo Strapparava, and Vivi Nastase. 2014. Mapping WordNet domains, WordNet topics and Wikipedia categories to generate multilingual domain specific resources. In *Proc. 9th International Conf on Lang Resources and Eval (LREC 2014)*. pages 1117–21.
- Dimitrios Kokkinakis and Mats Malm. 2015. Detecting reuse of Biblical quotes in Swedish 19th century fiction using sequence alignment. *Corpus-Based Research in the Humanities (CRH)* pages 79–86.
- Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing Sanskrit Wordnet. In *Proc. 5th Global Wordnet Conf (GWC 2010)*. pages 287–294.
- Leo D Lefebure. 1997. Awakening and grace: Religious identity in the thought of Masao Abe and Karl Rahner. *CrossCurrents* pages 451–472.
- Stefano Minozzi. 2009. The Latin WordNet project. In *Latin Ling Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik.* volume 137, pages 707–716.
- So Miyagawa, Marco Büchler, and Heike Behlmer. 2016, Manuscript submitted for publication. Computational analysis of text reuse/intertextuality: The example of Shenoute Canon 6. In *Proc. 11th International Congress of Coptic Studies. Orientalia Lovaniensia Analecta..* Leuven: Peeters.
- Kiyonori Nagasaki. 2008. A collaboration system for the philology of the Buddhist study. *Digital Humanities 2008 Book of Abstracts* pages 262–3.
- Wim Peters, Piek Vossen, Pedro Díez-Orzas, and Geert Adrians. 1998. Cross-linguistic alignment of wordnets with an inter-lingual-index. *Computers and the Humanities* 32(2/3):221–251.
- Mona L Scott. 1998. *Dewey Decimal Classification*. Libraries Unlimited.
- Selja Seppälä. 2015. An ontological framework for modeling the contents of definitions. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 21(1):23–50.
- Daisetz Teitaro Suzuki. 2002. *Mysticism: Christian and Buddhist*. Courier Corporation.
- Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. Pqac-wn: constructing a wordnet for Pre-Qin Ancient Chinese. *Lang Resources and Eval* 51(2):525–45.

Estonian Wordnet: Current State and Future Prospects

Heili Orav

University of Tartu

heili.orav@ut.ee

Kadri Vare

University of Tartu

kadri.vare@ut.ee

Sirli Zupping

University of Tartu

sirli.zupping@ut.ee

Abstract

This paper presents Estonian Wordnet (EstWN) with its latest developments. We are focusing on the time period of 2011–2017 because during this time EstWN project was supported by the National Programme for Estonian Language Technology (NPELT¹). We describe which were the goals at the beginning of 2011 and what are the accomplishments today. This paper serves as a summarizing report about the progress of EstWN during this programme. While building EstWN we have been concentrating on the fact, that EstWN as a valuable Estonian resource would also be compatible in a common multilingual framework.

1 Estonian Wordnet: Project Progress

Estonian Wordnet is a lexical-semantic resource describing Estonian words and their lexical relationships. The history of EstWN starts already in 1998 when Estonian team joined the EuroWordNet (EWN) project (see also Vossen 1998). Back at 1998 the only available example was Princeton WordNet (PWN) (Fellbaum 1998), so the EWN project followed the same principles. The EWN added a completely new component – multilinguality – the possibility to link different languages via a central InterLingualIndex (ILI) that was based on PWN version 1.5 at that time.

At the beginning of 2011 the EstWN had reached around 40 000 concepts (including 10 000 synsets taken over automatically), by September 2017 there are around 85 000 concepts with 230 664 semantic relations and 135 497 senses in EstWN.

Over the years EstWN project has been mainly supported by the National Programme for Estonian Language Technology, the first programme lasted from 2006–2010 and the second one from 2011–2017. We greatly appreciate that the Estonian government has realized that it is crucial to support the creation of Estonian language re-

sources so that the Estonian language is able to survive in the digital world among the larger languages.

There are two main directions in EstWN project – to add new and missing concepts and to improve the quality of existing data – for example performing the systematic revision of English equivalents and semantic relations or complementing EstWN with extra-information like sentiment, domain (see Bentivogli 2004) etc. Recently some wordnets have employed sentiment (opinion) information and also in EstWN 57 000 synsets have been automatically annotated with SentiWordNet's (see Baccianella et al. 2010) data. In addition to SentiWordNet, we have incorporated sense annotated vocabulary from the dictionary made for emotion detection (this vocabulary is manually tagged by linguists, see Pajupuu et al. 2016). Besides to the negative-positive-neutral scale, there is also contradictory-tag in this vocabulary, for example, *emotional*, *receptive* could be both positive or negative, depending on context. In the future, we plan to get sentiment tags for all synsets in the latest version of EstWN. In the long run, we expect that EstWN will be implemented more frequently as a language technology resource and for linguistic studies as well. Another important foresight is to belong into a unified global linguistic data infrastructure. While building EstWN we still follow general PWN principles and structure to enable linking, but at the same time, the EstWN should remain as language-specific as possible.

1.1 Where do new synsets come from?

Our team started to compile EstWN from translating base concepts and then we extended EstWN with the knowledge from different lexicons, corpora etc. Since EstWN has been mostly manual work of different people, then the semantic relations reflect largely human subjectivity. We have included vocabulary from dictionaries like Estonian Explanatory Dictionary, Orthological Dictionary, different terminology dictionaries, word frequency lists of corpora of written Estonian. Since general vocabulary of Estonian

¹ National Programme for Estonian Language, <https://www.keeletehnologia.ee/en>.

is covered, then we have moved on to special terminology. Although Martin Benjamin (2017) has written that “too many specialist terms would make PWN so unwieldy that the resource would become dysfunctional for users trying to sift through numerous esoteric senses” we continue to add vocabularies from different domains for the purpose of more broader usage of EstWN. Also, several students have contributed their work of the bachelor’s thesis to improve EstWN – for example, the vocabulary of veganism, climate, transportation etc has deeply studied and semantic relations inside chosen vocabulary have been thoroughly examined. The computer game Alias which draws information from EstWN is also useful for feedback of the new and missing words and senses (we talked about it on last conference (Aller et al. 2016)).

1.2 Automatically generated synsets

At some point during the project, it seemed sensible to construct some part of the resource automatically. Only a few attempts have been made to increase the database (semi)-automatically before 2011. We have to admit, that these attempts haven’t been overly successful and there are still problems to deal with.

Firstly, we included words that were missing from word sense disambiguation corpus but ended up with lots of proper names and words belonging already to some existing synset. Then synsets from the Dictionary of Synonyms were transferred automatically, but these synsets needed many corrections because the distinction between synonym and near-synonym was not clearly visible. Also, a lot of dialectal and archaic words were included, but not systematically or consistently.

Ideally, we would want to have a broad coverage of vocabulary. That was the reason for our attempt to add automatically nominalizations, especially words with the suffixes *-ja* (equal to *-er* suffix in English) and *-mine* (equal to *-ing* suffix in English). In this way, almost 10 000 synsets were added. Unfortunately, very many of these derivations are not valid because both one internal and one external relation were generated automatically – internal with *xpos_hyponym* relation linked to a verb and external *equal_hyperonym* relation to a verb. This lead into a confusing situation, because both relations are not accurate and more importantly link only to another part of speech, which does not follow the principles of wordnet. For example, the verb

synset ‘say, state, tell’ got automatically several *xpos_hyponyms* (all following synset are nouns):

lisamine, täiendamine ‘adding’
andmine ‘giving’
deklareerimine, kuulutamine ‘declaring’
hõikamine, hõiskamine ‘whooping’
protestimine ‘protesting’
esitamine ‘presenting’
kordamine ‘repeating’
vastamine ‘answering’.

Another problem occurred while transferring these derivations into EstWN – although the verb as a derivation base can have multiple senses, then the derived nouns with *-mine* and *-ja* suffix don’t share the same senses – not syntactically and not semantically. For example, the word *andma* ‘to give’ has 14 senses in EstWN, but derivations *andmine* ‘giving’ and *andja* ‘giver’ are used only in some of these 14 senses. The revision of automatic derivations is quite challenging since they also miss definitions. We still deal with these derivations manually – either fix the set of relations and add definitions or delete the invalid concepts completely.

Because of rich Estonian morphology many derivations are possible, like adverbs which are easily derived from other word classes, for example, *ahne* ‘greedy’ – *ahnelt* ‘greedily’ (Kerner et al. 2010). However, the described experiments have made us cautious about fully automatic enlargements, since the manual correction is unreasonably time-consuming. Of course, we are open to implementing proven automatic extension methods, which measure up to the quality of manual work.

1.3 How to define synsets – general challenges

It is widely known that definitions are difficult to write and take a lot of time even in one’s mother tongue, yet they provide clarity both for native speakers and foreigners (Benjamin 2017). Because a lot of synsets in EstWN are missing definitions, we have to provide them a proper one, if possible. The problem of definitions originates from our existing dictionaries of Estonian – we can find a lot of tautology – an unnecessary repetition of meaning. None of the dictionaries we have used contain information about hierarchical concepts. The explanatory dictionary features information about hypernym (also synonyms, near-synonyms or antonyms) for some headwords in definitions, but this information is, unfortunately, unsystematic and can be rather confusing.

In Estonian, it is possible (and common) to rewrite concepts with compound words, since patterns of compound word formation are productive in Estonian (Kerge 2016). Again, the problem of tautology arises if a synset contains a compound word, for example, *hüpertoon-ia+haige* ‘hypertonia+sick person’, *hüpertoonik, kõrgvererõhu+haige* – ‘person, who suffers from hypertonia’. A good definition is meant to paraphrase the concepts, but tools (i.e words) seem to be missing. Lew (2015) has pointed out, that surprisingly people look up the explanation of meaning firstly through synonyms, so it might be more helpful in some cases to pay attention to synset members rather than to a (bad) definition. Similarly, from the Estonian Text Simplification application (Peedosk 2017) appeared that for the better understanding of a concept it is essential to be able to choose between foreign word and native word (encephalitis vs. *ajupõletik* ‘inflammation of the brain’ or *kõht* ‘belly’ vs. *abdoomen* ‘abdomen’). Native words are often more informative to native speakers, whereas foreign word is understandable to foreigners (and through the foreign word they are able to learn and understand the native word).

2 EstWN odyssey from ILI1.5 to PWN3.0 and to CILI

Since we wanted EstWN to be linked to the Global WordNet Association repository with Collaborative Interlingual Index (CILI), the first step was to update the old ILI1.5 to the latest PWN3.0 version. As said before, different wordnets are generally similar but still need some effort to combine in a common interoperable multilingual framework (Bond, Piasecki 2017). As follows we describe our efforts and challenges of the CILI-linking process from the wordnet builders point of view.

EstWN was connected to ILI1.5 almost 20 years, and on 2017 we could finally update ILI1.5 to PWN3.0 thanks to our new wordnet editing tool – WordNet WorkBench². The first ILI version (1.5) contained more than 90 000 concepts, yet it was often difficult to determine equal synonyms from Estonian to English. ILI1.5 missed suitable senses, especially regarding adjectives and adverbs. Another problem was that a lot of definitions were missing from ILI and it

was complicated to decide the exact meaning of the ILI synsets. PWN 3.0 is of course much richer with different concepts to choose from, so we started to correct English equivalents systematically – changing other ILI-relations into more precise equal synonym relation.

In order to share the data with Open Multilingual Wordnet project, we still have to link EstWN’s synsets to CILI, since the reference to CILI is the obligatory attribute of synset.

At the moment in EstWN 22 345 synsets have the external reference with relation type ‘eq_synonym’ to PWN 3.0 and thereby are mapped to CILI. Number of CILI-links which are not linked with EstWN is 95 314. This number includes also 7556 proper names, connected with PWN via instance-relation. Thus other (approx 65 thousand) synsets require work in order to either find a relation with appropriate concept from CILI or in the future to define a new concept with a new definition and propose them to CILI.

It is also widely known, that some mistakes are inevitable and the solution is the manual correction of errors. Next, we describe the process of improving English part of EstWN through the English equivalents. Since it is complicated and unreasonable to check English equivalents from the first entry in EstWN, we composed different types of lists³, which we considered to be problematic.

From these lists different types of mistakes occurred, for example, 940 English synsets were connected to 1881 Estonian synsets via the eq_synonym relation, which indicates that these synsets need to be either corrected or united. Some examples:

- Small variations in spelling – like between singular and plural (for example *helilaine(d)* – ‘acoustic wave(s)’) or spelling error between *diakoniss* and *diakoness* – ‘deaconess’).
- Indistinguishable senses which are dealt as mistakes and were united to one synset (for example *final* ‘finale’ ja *kooda* ‘coda’ as music terms; *brie* and *brii* (as Estonian adaption of the name of Brie cheese)).

² The tool is freely available, please contact EstWN team for further information. For detail see Jentson et al. forthcoming.

³ For example, list of eq_has_hyperonym relation with frequency more than 4 times of usage, list of eq_near_synonym with frequency more than 2 times of usage etc.

After the linking process to CILI was completed, then other general types of errors were found from the composed lists, for example:

- Some cases where *eq_near_synonym* and *eq_has_hyperonym* have been in confusion, for example, English concept ‘folk singer’ has 12 *near_synonym* and 13 *has_hyponym* in Estonian and therefore with *kerjuslaulik* ‘beggar singer’ being *eq_near_synonym* to ‘folk singer, jongleur, minstrel, poet-singer, troubadour’ and *rüütliulik* ‘troubadour’ being linked with *eq_hyperonym* relation to ‘folk singer, jongleur, minstrel, poet-singer, troubadour’.
- 8411 cases, where the Estonian synset has an external link to English concept in the different part of speech, for example, adjective *nunnalik* ‘nun-like’ is connected via ILI with noun *nun*. The Estonian word *nunnalik* ‘like a nun’ is rich with nuances (different across cultures, looks, behavior, attitudes, mentalities) and it is complicated to link this particular Estonian adjective to English adjective. So the only way is to link it to a noun.
- One English synset may have too many hyponyms in EstWN, for example, ‘denizen, dweller, habitant, indweller, inhabitant’ has 42 hyponyms.
- We counted synsets which use the same *eq_near_synonym* more than 2 times and we got 347 such. For example, ‘district, dominion, territorial dominion, territory’ has *eq_near_synonym* relation 7 times in EstWN.
- Mistranslations: the meaning of the word often depends on context (see e.g Wittgenstein 2005) - English concepts don’t fit into Estonian context and vice versa. Lexical caps can be roughly:
 - referential (missing concept, as snow for African people) and
 - lexical (missing word or expression, for example, onomatopoeic words in English and culture-specific words like *kama* (Estonian food made from grain)).

As no lexicon can cover all words and senses there are lot’s of concepts which are lexicalized in language but haven’t found their way to a lexicon or wordnet yet. For example, the Estonian concept *piimasupp*, ‘milk soup’ in English, which is lexicalized also in English but is missing currently from PWN3.0. Same on the contra-

ry, Estonian synset may have several *near_synonym* links to English synset, for example *härria, isand, saks* has link of *near_synonym* to ‘landlord’ and ‘gentleman’ and has *hyperonym* link to ‘man of means, rich man, wealthy man’ – in Estonian concept, different nuances are mixed from all three English concept. One possible solution is offered by Frankenberg-Garcia (2015) who emphasized that correct translation should be shown with 4-5 examples of usages (i.e to show broader context) or with clear definitions to understand nuances of differences.

The remarks above summarized and discussed only some challenges of our wordnet building, and not the whole project, which is still in progress.

3 Future plans

The EstWN project has most definitely achieved the initial goals of the project and at the end of this NPELT program, there is an appropriate time to set new goals and plan future activities. EstWN project has several quite challenging stages ahead: we continue to increase the size of EstWN with a special focus on the quality. Another direction is to find applications for EstWN – it has been proven for EstWN, that via these applications it is possible to perform different types of quality checks. We have to look more into the topic of the compound words because EstWN is missing some of the mostly used compounds. For compound extraction a corpus will be used, and compounds which occur more than 10 times in this corpus are considered as possible candidates as new concepts or senses.

The new editing tool WordNet WorkBench enables us to create, change or delete semantic relations, so we can create (and rename) new semantic relations valid for Estonian and adopt relations from other resources, for example, domain relation from PWN. Also, we plan to integrate domain labels from WordNetDomains automatically; of course we have to validate if the domains initially created for English apply also in the context of Estonian.

Summing up, we can say that EstWN has reached a level where it can be used in several language technology applications and in research as a valuable language resource.

References

- Aller, Sven; Orav, Heili; Vare, Kadri; Zupping, Sirli. (2016). Playing Alias – efficiency for wordnets(s). – *Proceedings of the 8th Global WordNet Confer-*

- ence [GWC 2016]: Bucharest, Romania, January 27–30, 2016. Ed. by V. Barbu Mititelu, C. Forascu, C. Fellbaum, P. Vossen. Bucharest: Alexandru Ioan Cuza University of Iași, pp. 16–21; <http://jiangbian.me/papers/2016/gwc2016.pdf> (15.09.2017).
- Baccianella, Stefano; Esuli, Andrea; Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. – *LREC 2010 Proceedings: LREC 2010, Seventh International Conference on Language Resources and Evaluation*. <http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf> (15.09.2017).
- Benjamin, Martin. 2017. Inside Baseball: Coverage, Quality, and Culture in the Global WordNet. – *Proceedings of the Workshop on Challenges for Wordnets*, http://ceur-ws.org/Vol-1899/CfWNs_2017_proc9-paper_5.pdf (15.09.2017).
- Bentivogli, Luisa; Forner, Pamela; Magnini, Bernardo; Pianta, Emanuele. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Geneva, Switzerland, August 28, pp. 101-108.
- Bond, Francis; Piasecki, Maciej. 2017. Introduction: Contemporary Challenges for Development and Application of Wordnets. – *Proceedings of the Workshop on Challenges for Wordnets*, http://ceur-ws.org/Vol-1899/wordnet_preface.pdf (15.09.2017).
- Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Frankenberg-Garcia, Ana. 2015. Dictionaries and Encoding Examples to Support Language Production. *Oxford University Press International Journal of Lexicography*. Repository URL: <http://epubs.surrey.ac.uk/808172/> (15.09.2017).
- Jentson, Indrek; Orav, Heili; Vare, Kadri; Kahusk, Neeme. Forthcoming. LiLT Paper on Estonian Wordnet. *Special Issue on Linking, Integrating and Extending Wordnets. Linguistic Issues in Language Technology – LiLT*. Volume 10, Issue 4 Sep 2017.
- Kerge, Krista. 2016. Word-formation in the individual European languages: Estonian. – *Word-Formation. An International Handbook of the Languages of Europe*. Ed. by P. O. Müller, I. Ohnheiser, S. Olsen, F. Rainer. Berlin, New York: De Gruyter. (Handbooks of Linguistics and Communication Science ; 40.5), pp. 3228–3259.
- Kerner, Kadri; Orav, Heili; Parm, Sirli. 2010. Semantic Relations of Adjectives and Adverbs in Estonian WordNet. – *LREC 2010 Proceedings: LREC 2010, Malta, Valletta, May 17-23, 2010*. ELRA, pp. 33–37.
- Lew, Robert. 2015. *Dictionaries and Their Users. International Handbook of Modern Lexis and Lexicography*. Springer-Verlag Berlin Heidelberg.
- Lohk, Ahti, Orav, Heili; Vare, Kadri; Võhandu, Leo. 2016. Experiences of lexicographers and computer scientists in validating Estonian Wordnet with test patterns. – *Proceedings of the 8th Global WordNet Conference [GWC 2016]: Bucharest, Romania, January 27-30, 2016*. Ed. by V. Barbu Mititelu, C. Forascu, C. Fellbaum, P. Vossen. Bucharest: Alexandru Ioan Cuza University of Iași, 184–191.
- Pajupuu, Hille; Altrov, Rene; Pajupuu, Jaan. 2016. Identifying polarity in different text types. – *Folklore. Electronic Journal of Folklore*, 64, 25–42.
- Peedosk, Martin. 2017. *Applying Estonian Digital Resources and Technologies in a Text. Simplification Program*. University of Tartu, BA thesis, https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=58269&year=2017 (15.09.2017).
- Vossen, Piek (ed.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands.
- Wittgenstein, Ludwig (2005). *Filosoofilised uurimused*. Tartu: Ilmamaa.

Further expansion of the Croatian WordNet

Krešimir Šojat, Matea Filko

University of Zagreb
Zagreb, Croatia
ksojat@ffzg.hr
msrebaci@ffzg.hr

Antoni Oliver

Universitat Oberta de Catalunya
Barcelona, Catalonia (Spain)
aoliverg@uoc.edu

Abstract

In this paper a semi-automatic procedure for the expansion of the Croatian Wordnet (CroWN) is presented. An English-Croatian dictionary was used in order to translate monosemous PWN 3.0 English variants. The precision values of the automatic process is low (about 30%), but the results proved valuable for the enlargement of CroWN. After manual validation, 10,884 new synset-variant pairs were added to CroWN, achieving a total of 62,075 synset-variant pairs.

1 Introduction

The building of the Croatian Wordnet has begun in 2004 at the Institute of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb. The Croatian WordNet is a lexical database built through the expand model (Vossen, 1998). The development of the Croatian Wordnet (CroWN) can be divided into two major phases (CroWN 1.0 vs. CroWN 2.0 / 3.0). Both versions are available for download and on-line queries. CroWN 1.0 (Raffaelli et al., 2008) was built completely manually. The main objective in this phase of the project was to translate and adapt the so-called basic concept sets extracted from the WN version 1.5 and used in the multilingual projects EuroWordNet (EWN) and BalkaNet (BN). For each synset a meaning definition was translated and adapted. Each synset in CroWN 1.0 is also accompanied by one or more examples of contextual usage. Synsets contain *literals* or *synset variant pairs* of the same part of speech. CroWN 1.0 comprises 10,000 synsets. 8500 of these are from the basic concept sets of EWN and BN. Approximately 1500 noun synsets were added using the same procedure. Although rich in information and data, CroWN 1.0 is a relatively small resource.

In order to make it more useful in various NLP tasks, the second phase of the project was primarily oriented toward its enlargement. CroWN 2.0 and CroWN 3.0 (Oliver et al., 2015; Oliver et al., 2016) were built by using different automatic approaches. These versions of the lexicon are the result of joint work between two research teams from Zagreb and Barcelona. CroWN 2.0 and 3.0 contain only synset-variant pairs in Croatian, i.e. meaning definitions and examples of contextual usage have not been translated (yet). CroWN 2.0 and CroWN 3.0 are available at the Open Multilingual Wordnet website¹.

In this paper we present a semi-automatic method that was used for further expanding of CroWN, i.e. for the creation of its version 3.1.

The paper is structured as follows: in section 2 we describe the algorithms and procedures applied in the creation of versions 2.0 and 3.0 and provide some statistics regarding the number of synsets, POS distribution etc. Section 3 deals with the procedure and resources applied in the experiment presented in this paper. In section 4 results are discussed as well as advantages or potential disadvantages of the method applied here. Section 4 brings concluding remarks and the outline of future work.

2 Versions of the Croatian Wordnet

At this time, CroWN is the only resource for Croatian that deals with lexical semantics and also provides multilingual links to similar resources via The Open Multilingual Wordnet project. As mentioned, CroWN 2.0 and CroWN 3.0 are the result of joint work between two research teams from Zagreb and Barcelona. The 2.0 version of the CroWN was developed using the WN-Toolkit² (Oliver, 2014), a set of Python programs for the

¹<http://compling.hss.ntu.edu.sg/omw/>

²<http://sourceforge.net/projects/wn-toolkit>

automatic creation of wordnets following the expand model. The WN-Toolkit implements 3 different strategies for wordnet creation:

1. Dictionary-based strategy - bilingual dictionaries are used to translate English variants associated with each synset. The strategy can deal only with monosemous English variants, i.e. variants associated with a single synset.

2. BabelNet-based strategy - the data from the BabelNet (Navigli and Ponzetto, 2010) file was extracted in order to obtain the data for CroWN.

3. Parallel-corpus-based strategy - in order to extract a target language wordnet, at least the English part of a parallel corpus should be sense tagged with PWN synsets. As such resources are rare and not easily available, two additional procedures were used for the creation of such a corpus: machine translation of sense-tagged corpora and automatic sense tagging of the English part of the parallel corpus.

Another line of work in CroWN 2.0 was oriented towards the enlargement of verbal synsets in CroWN. In CroWN 1.0 nouns make up almost 75 % of the whole lexicon (7391 noun synsets vs. 2318 verb synsets). The goal was to make CroWN a more balanced and representative resource for Croatian by enlarging the number of verbs. For this purpose we used CroDeriv (Šojat et al., 2013)³, a large derivational database of Croatian verbs. The data was extracted and matched with PWN automatically. A more detailed account of the procedure and results is given in (Oliver et al., 2015). As in all other procedures described here, all candidates for synsets were manually checked and corrected if necessary. Taking into account that every automatic processing of data is followed by a manual revision, all procedures discussed here can be considered as semi-automatic.

With all these strategies we reached the 70.63 % of the Core synsets ((Boyd-Graber et al., 2006)). Finally, we manually populated CroWN 2.0 with the remaining 1,456 synsets, thus reaching 100 % of the Core WordNet.

For the creation of the version 3.0 we used a new version of the WN-Toolkit. It implements several strategies for mapping lexical resources (Wikipedia, Wiktionary and Omegawiki). An extensive account of this procedure is given in (Oliver et al., to appear).

³croderiv.ffzg.hr

In table 1 the number of synsets and synset-variant pairs in each of the three versions is presented. More details will be given in the subsections below.

Version	Synsets	Synset-variants
V 1.0	10,026	31,367
V 2.0	23,137	47,931
V 3.0	25,658	51,168
V 3.1	31,614	62,075

Table 1: Number of synsets and synset-variant pairs in different versions of the CroWN

In the following section we explain the process of further extension of the CroWN V. 3.0 and the creation of the new V. 3.1.

3 Experimental part

3.1 Automatic creation of synset-variant candidate pairs

For the new extraction we have used the EH dictionary⁴. This is an on-line dictionary, and the source file is provided by the authors under request. The EH dictionary comprises 186,098 entries. The dictionary is a plain text file containing two columns: an English word and a Croatian word, with no POS information included, as in the following fragment:

```
mother majka
mother materinski
mother posiniti
```

However, correct information about the POS of each word is vital for the method applied here. We have therefore used the Croatian Morphological Lexicon ((Tadić and Fulgosi, 2003))⁵ to automatically attach POS information to the dictionary entries. The data in this morphological lexicon is structured as follows (*majka* – mother; *materinski* – maternal; *posiniti* – to adopt as son).

```
majka majka Ncfsn
materinski materinski Afpmsny
posiniti posiniti Vmn
```

With such information we were able to attach the POS information to 79,608 dictionary entries:

```
mother majka n
mother materinski a
mother posiniti v
```

Dictionary entries with the POS information were used to translate monosemous English variants in PWN-3.0. A variant is regarded as monose-

⁴<http://web.vip.hr/zcindori.vip/ehrjecnik/>

⁵hml.ffzg.hr

mous, at least according to WordNet, if it is attached to a single synset. Table 2 shows the number of monosemous and polysemous variants in WordNet for each POS:

	Noun	Verb	Adj.	Adv.
All	117,798	11,529	21,479	4,481
Mono	101,863	6,277	16,503	3,748
Poli	15,935	5,252	4,976	733

Table 2: Monosemous and polysemous variants in PWN 3.0

The translation of the variants enabled the extraction of 62,353 Croatian synset-variant pairs. Table 3 displays the distribution by POS of the extracted data as well as the results of automatic evaluation. The evaluation was performed by comparing the extracted synset-variant pair with CroWN 3.0. In section 3.2.2 a more detailed evaluation is presented.

	Extract.	Eval.	Correct	%
All	62,353	30,123	9,357	31.06
Noun	33,451	17,829	5,803	32.55
Verb	14,230	8,754	2,695	30.79
Adj.	14,048	3,277	794	24.23
Adv.	624	263	65	24.71

Table 3: Extracted synset-variant pairs by POS and automatic evaluation figures

The automatically calculated precision values are low, about 31%. As the numbers indicate, there are 30,123 synset-variant pairs that were evaluated since they are present in the CroWN 3.0 versus 32,230 instances that could therefore not be evaluated. Further, 20,766 synset-variant pairs were evaluated as incorrect. A candidate is marked as incorrect if we have some variant for the given synset in the CroWN 3.0, but no the extracted variant. This extracted variant can be correct, but not present in the CroWN. The subset of pairs evaluated as incorrect can be also manually revised.

3.2 Manual revision and completion

In order to further evaluate the automatically extracted Croatian synset-variant pairs, all the results were revised by hand. During this time-consuming task we wanted to maximize our contribution and to expand CroWN as much as possible. Our revision was hence divided into several steps. First, non-evaluated candidates and

candidates automatically evaluated as incorrect were set apart and evaluated in separate actions. Further, both sets of extracted Croatian synset-variant pairs were arranged according to PWN synset-IDs. Meaning definitions provided for PWN synsets were used as a criterion to evaluate candidates as correct or incorrect. In other words, each candidate was marked either as correct or incorrect on the basis of meaning definitions from PWN. During this process we were adding one or more Croatian variant pairs whenever it was possible. Finally, if none of the candidates for a particular synset was correct, we added new synset-variant pairs by hand as well.

3.2.1 Problems for the automatic approach

Manual evaluation of candidates revealed several problematic cases for the automatic method of expansion applied here. Problems that we faced regard to several aspects:

1. Problems that result from linguistic features of Croatian and American English as well as cultural differences that are reflected in conceptualization and lexicalization. One of the problems that we faced is related to the processing of multi-word expressions. For example, one of the senses of the noun *wall* in PWN is defined as "a difficult or awkward situation". This candidate was translated with Croatian *zid*, a wall (as in *brick wall*). The problem for this and similar examples is that the Croatian noun is normally used in this sense only in idioms, e.g. *naići na zid*, *naći se pred zidom*. In other words, English synsets list literals that are used only as parts of idioms or phrasemes in Croatian.
2. Besides, several problems resulted from the fact that Croatian collocations composed of adjectives and nouns, e.g. *genska ekspresija*, generally act as a single semantic unit, whereas in English synsets only a noun is listed as a literal. Unlike in English, in many cases Croatian candidates were obligatory multi word expressions.
3. Further, we came across numerous cases in which PWN literals cannot be lexicalized in Croatian due to its morphological properties. Although derivation of nouns from verbs is common in Croatian, it is not possible for numerous PWN literals (e.g. there are no derivatives for *skidder*, *slider*, *slipper* defined as "a person who slips or slides because of loss of traction" and *chew*, *chaw*, *cud*, *quid*, *plug*, *wad* defined as "a wad of

something chewable as tobacco”).

4. We also found several examples when concepts represented by PWN literals are lexicalized with completely other lexical means. For example, the closest relatives of the PWN literal *near miss* defined as “an accidental collision that is narrowly avoided” are various Croatian verbal idioms, e.g. *promašiti za dlaku, izbjеći za malo, “miss by a hair’s breadth”* etc.

5. Some concepts from PWN do not exist at all in Croatian, e.g. *dictator*, as “a speaker who dictates to a secretary or a recording machine”, or *show-stopper, showstopper, stopper* as “an act so striking or impressive that the show must be delayed until the audience quiets down”. Since we could not come up with a better solution, in CroWN 3.1 we marked such examples with the tag GAP. The same mark was used for numerous expressions denoting concepts from various domains characteristic almost exclusively for the US. Problems that result from cultural differences pertain to specific terms used in stock market, the US legal system, sports as baseball and American football, cuisine etc. For example, PWN literals *bomber, submarine, torpedo* denote the same type of sandwich eaten in the US. The meaning definition for this synset points out that different names are used in different sections of the United States. Such words are almost impossible to translate or adapt without additional explanations. Candidates from this group exclusively belong to the non-evaluated part of the obtained candidates. The second group of problems pertains to differences between Croatian and English:

6. An issue that poses a challenge to the adopted expand model pertains to cases when PWN literals can be translated only with Croatian words of different POS. For example, adjectival synset containing the adjective *several* should be translated with the adverb *nekoliko*. Similarly, but not so often, PWN literals can be translated only with Croatian suffixoids, i.e. units that are neither words nor morphemes, e.g. *-ology, -ism* etc. E.g., the most accurate translation of the PWN’s *stasis* “an abnormal state in which the normal flow of a liquid (such as blood) is slowed or stopped” is the Croatian suffixoid *-staza*, although word *zastoj* can be used. Further, parts of English compounds are also sometimes listed as literals, e.g. *wort* is defined as: “usually used in combination: ‘liverwort’; ‘milkwort’”, which

makes the processing almost impossible.

7. PWN verbal literals referring to both causative and reflexive senses of English verbs are also highly problematic. In Croatian, as it is common in Slavic, these are different verbs and consequently different lemmas. Lemmas for reflexive verbs include the reflexive pronoun *se* (e.g. *otopiti se* ‘to become melted’), whereas causatives do not co-occur with *se* (e.g. *otopiti* ‘to melt’). Such cases pose a challenge for the construction of verbal synsets in CroWN. On top of that, there is group of reflexive verbs that co-occur with the so-called reflexive particle *se* (e.g. *smijati se* ‘to laugh’). As far as the discussed method of expansion is concerned, there were numerous cases when only infinitives were recognized, while reflexive pronouns or particles were missing.

8. Although phrasal verbs do not exist as a separate category according to Croatian grammars, based on the examples from CroWN, (Katunar et al., 2012) argue that they should be recognized and treated as such. In some cases, the meaning of verbs is altered by co-occurring prepositions, e.g. verb *držati* ‘to think’ vs. *držati do* ‘to value’. The applied automatic approach can account only for infinitives, thus yielding incorrect candidates.

9. Finally, the problem with the automatic approach is that it relies on one-to-one translation and therefore offers all translation equivalents from the dictionary in all their senses. This usually results in one or more correct and one or more incorrect candidates per synset if the word in case is highly polysemous.

However, in many cases new candidates for the already existing synsets were offered, i.e. candidates omitted in previous versions of CroWN. The result is a more diversified language resource.

3.2.2 Evaluation of the methodology

The manual revision of the candidates facilitated the calculation of precision values for two subsets: the non-evaluated candidates and the candidates automatically evaluated as incorrect. Table 4 presents these values. They are similar (in the region of 30 %) as the values shown in table 3 for the automatic evaluation of the non-evaluated subset. The precision values for the incorrect subset are lower, as expected, but in this subset there are still about 15 % of correct synset-variant pairs.

In table 5 the number of synset-variant pairs for each POS for versions 3.0 and 3.1 are shown.

	P	P_N	P_V	P_A	P_R
non-eval.	30.06	29.6	18.98	39.53	-
incorrect	14.11	16.54	11.21	22.92	-

Table 4: Precision figures for the manually evaluated subsets.

	3.0	3.1
Nouns	30,240	38,951
Verbs	17,913	18,645
Adjectives	2,623	4,064
Adverbs	415	415
Total	51,191	62,075

Table 5: Number of synset-variant pairs in version CroWN 3.0 and 3.1.

Once all the new synset-variant pairs had been manually validated and corrected, we could calculate final values of precision for the applied methodology. In table 6, we present these figures, which are in fact very similar to the precision figures of the automatic evaluation in table 3.

	Extract.	Eval.	Correct	P.
All	62,353	46,774	14,682	31.39
Noun	33,451	30,802	9,880	32.08
Verb	14,230	10,111	2,969	29.36
Adj.	14,048	5,598	1,768	31.52
Adv.	624	263	65	24.71

Table 6: Extracted synset-variant pairs by POS and automatic evaluation figures

4 Conclusions and future work

The main goal of the experiment procedure described in this paper was to expand the CroWN 3.0 with a) new synsets, and b) new literals in the existing synsets. The development of CroWN is not financially supported on a regular basis, therefore automatic and semi-automatic procedures for its further expansion are particularly valuable. When dealing with large amount of data, it is easier to manually edit the results of the automatic extraction of candidates than to work from scratch.

The use of the EH dictionary has allowed us to further expand the Croatian Wordnet. In previous works we have used other free lexical resources (namely Omegawiki, Wiktionary and Wikipedia) and a similar methodology. The precision values obtained with EH are much lower than those obtained with other resources. The main reason is the

size of the EH dictionary, which is much larger and provides a lot of translation equivalents for each English word. Some of these translation provide similar meaning that are not suitable for the construction of a wordnet.

Acknowledgments

This research has been carried thanks to the project TUNER TIN2015-65308-C5-1-R (MINECO/FEDER, UE) and the short-term research support of the University of Zagreb.

References

- Jordan Boyd-Graber, Christiane Fellbaum, Daniel Os-herson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the 3rd International WordNet Conference*, pages 29–36. GWC.
- Daniela Katunar, Matea Srebačić, Ida Raffaelli, and Krešimir Šojat. 2012. Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet. In *Proceedings of the LREC 2012*, pages 33–39. ELRA, Istanbul.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the ACL*, ACL '10, pages 216–225, Stroudsburg, PA, USA. ACL. ACM ID: 1858704.
- Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2015. Enlarging the Croatian WordNet with WN-Toolkit and Cro-Deriv. In *RANLP*, pages 480–487.
- Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2016. Automatic expansion of Croatian Wordnet. In Sanda Lucija Udier and Kristina Cergol Kovačević, editors, *Metodologija i primjena lingvističkih istraživanja*, pages 171–185. Zagreb.
- Antoni Oliver, Krešimir Šojat, and Matea Filko. to appear. The Croatian WordNet: CroWN 3.0. *Linguistic Issues in Language Technology - LILT. Special Issue on Linking, Integrating and Extending Wordnets*, 10.
- Antoni Oliver. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. In *Proceedings of the 7th GWC*, Tartu, Estonia.
- Ida Raffaelli, Bekavac Božo, Željko Agić, and Marko Tadić. 2008. Building Croatian WordNet. In *Proceedings of the 4th GWC*, Szeged, Hungary.
- Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. 2013. CroDeriv and the morphological analysis of Croatian verb. *Suvremena lingvistika*, 75:75–96.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian Morphological Lexicon. In *Proceedings of the EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–46. ACL, Budapest.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Linking WordNet to 3D Shapes

Angel X Chang, Rishi Mago, Pranav Krishna, Manolis Savva, and Christiane Fellbaum

Department of Computer Science, Princeton University

Princeton, New Jersey, USA

angelx@cs.stanford.edu, rmago19@lawrenceville.org,
pranavskrishna@gmail.com, msavva@cs.stanford.edu, fellbaum@princeton.edu

Abstract

We describe a project to link the Princeton WordNet to 3D representations of real objects and scenes. The goal is to establish a dataset that helps us to understand how people categorize everyday common objects via their parts, attributes, and context. This paper describes the annotation and data collection effort so far as well as ideas for future work.

1 Introduction

The goal of this project is to connect WordNet (Fellbaum, 1998) to 3D representations of real objects and scenes. We believe that this is a natural step towards true grounding of language, which will shed light on how people distinguish, categorize and verbally label real objects based on their parts, attributes, and natural scene contexts.

Our main motivation is to establish a dataset connecting language with realistic representations of physical objects and scenes using 3D computer-aided design (CAD) models to enable research in computational understanding of the human cognitive process of categorization.

Categorization is the process by which we group entities and events together based on salient similarities, such as shared attributes or functions. For example, the category “furniture” includes tables, chairs and beds, all of which are typical parts of a room or house and serve to carry out activities inside or around the house. Subcategories are “seating furniture,” which includes chairs and sofas and “sleeping furniture,” which includes beds, bunkbeds and futons. Note that some categories have a simple verbal label (a name, like “furniture”), but often category names are compounds (like “sleeping furniture”). Compounding, a universal feature of human language that accounts in part for its infinite generativity, allows us to make

up names on the fly whenever we feel the need to distinguish finer-grained categories, such as “college dorm room furniture.” Of course, not all languages share the same inventory of simple labels.

We form and label categories all the time. Categories help us to recognize never-before-seen entities by perceiving and assessing their attributes and functions and, on the basis of similarity to known category members, assign them to a category (Rosch, 1999). Young children in particular learn to form categories by being exposed to an increasing number of different category members and gradually learning whether they belong to one category or another. Importantly, categories allow us to reason: if we know that beds are made for lying down on and sleeping, encountering a new term like “sleigh bed” will tell us that such a bed is likely to have a flat surface on which a person can lie down. Conversely, seeing a sleigh bed for the first time and identifying this salient feature will prompt us to call it a “bed.” Categorization is so fundamental to human cognition that we are not consciously aware of it; however, it remains a significant challenge for computational systems in tasks such as object recognition and labeling.

Parts, attributes, and natural contexts of objects are all involved in category formation. Objects are made of parts, and parts often imbue functionality, especially in the broad category of “artifacts.” Thus, seat surfaces are a necessary part for functioning chairs. Parts and the functionality they enable are fundamentally intertwined with categorization (Tversky and Hemenway, 1984).

Beyond their concrete parts, objects are perceived to have a general set of attributes. For example, the distinction between a cup, a goblet and a mug relies less on the presence or absence of specific parts and more on the geometric differences in aspect ratio of the objects themselves.

Lastly, real objects occur in real scenes, meaning that they possess natural contexts within which

they are observed. In language, context reflects a variety of aspects beyond functionality, including syntactic patterns and distributional properties. In contrast, physical context is concrete and defined by the sets of co-occurring objects and their relative arrangements in a given scene.

To study these three aspects of category formation, we will connect WordNet at the part, attribute, and contextual level with a 3D shape dataset. The longer term goal of this project is to ask how people distinguish and categorize objects based on how they name and describe the parts and attributes of the objects. We will focus primarily on the category “furniture.”

2 Existing datasets

There has been much prior work linking WordNet (Fellbaum, 1998) to 2D images. The most prominent effort in this direction is ImageNet (Deng et al., 2009), which structures a large dataset of object images in accordance with WordNet hierarchies. Our project differs from ImageNet because even though Imagenet is based on the WordNet hierarchy, the focus of our project is on annotating parts and attributes on 3d models rather than the labeling of images. Following this, the SUN (Xiao et al., 2010) dataset focuses on scene images, and the VisualGenome (Krishna et al., 2016) dataset defines object relations and attributes in a connected scene graph representation within each image. Another line of work focuses on detailed annotation of objects and their parts — a prominent recent example is the Ade20K dataset (Zhou et al., 2016). However, a fundamental assumption of all this work is that objects and their properties can be adequately represented in the 2D image plane. This assumption does not generally hold, as many object parts, spatial relations between objects and a full view of object context are hard to infer from the limited field of view of a 2D image.

More recently, there has been some work that links 3D CAD models to WordNet. The ShapeNet (Chang et al., 2015) dataset is a large collection of CAD representations (curating close to 65,000 objects in approximately 200 common WordNet synsets), whereas the SUNCG (Song et al., 2017) dataset contains CAD representations of 45,000 houses, composed of individual objects (in 159 WordNet synsets). These two datasets are both “synthetic” in the sense that the 3D CAD

representations are designed virtually by a human expert. A different form of 3D representation is obtained by scanning and 3D reconstruction of real world spaces. Recent work introduced ScanNet (Dai et al., 2017) and the Matterport3D (Chang et al., 2017) dataset, which both contain 3D reconstructions of various public and private interior spaces (containing 409 and 430 object synsets respectively).

Though both synthetic and reconstructed 3D data are increasingly available, no effort currently exists to connect such 3D representations to WordNet at the part, attribute, and contextual levels. Such a link of WordNet entries to 3D data can provide much richer information than 2D image datasets. Naturally, 3D representations allow us to reason about unoccluded parts and symmetries, arrangement of objects in a physically realistic three dimensional space, and to account for empty space, a critical property of real scenes which is not observable in 2D images. Moreover, 3D representations are appropriate for computationally simulating real spaces and the actions that can be performed within them. The ability to do this is a powerful tool for investigating and understanding actions (Pustejovsky et al., 2016). Therefore, our project aims to annotate 3D models in one of the existing datasets and link them to the appropriate synset in the WordNet database at the part, attribute, and contextual level.

3 Project description

Our project has so far focused on annotating part and attribute information on 3D CAD objects in SUNCG (Song et al., 2017) and linking them to the corresponding WordNet synsets. We are working with a preliminary categorization of the objects performed in prior work, which establishes their connection to WordNet synsets denoting physical objects. However, we plan to refine the granularity of this categorization by introducing finer-grained categories — e.g. partitioning “doors” into “garage doors” and “screen doors” among others.

We chose this dataset because the 3D objects in SUNCG (approximately 2,500) are used across a large number of 3D scenes (more than 45,000). This means that for each object, we can automatically establish many contextual observations. This property of the SUNCG dataset differs from other common 3D CAD model research datasets such

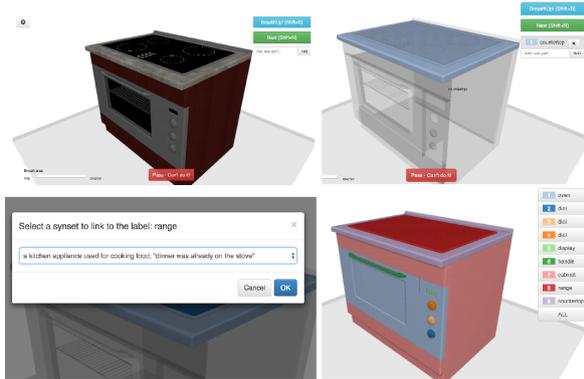


Figure 1: Interface for labeling object parts. Top left: a stove and range object is displayed to a user. Top right: the user paints the “countertop” part. Bottom left: the user links the “range” part to the corresponding WordNet synset. Bottom right: the fully annotated object with parts in different colors.

as ShapeNet (Chang et al., 2015) where each object is de-contextualized. For the latter dataset we would have to additionally compose scenes using available objects in order to acquire observation contexts for the objects.

We first augment the SUNCG objects with part annotations that are linked to WordNet synsets. We defer the assignment of attributes to the same objects as it is an easier annotation task in terms of interface design. To perform the part annotation, we designed an interface with a “paint and name” interaction where the user paints parts of the surface of an object corresponding to a distinct part and assigns a name to that part. The details of the interface and annotation task are described in the following section.

4 Annotation interface

Our interface is designed to allow for efficient annotation of parts by inexperienced workers on crowdsourcing platforms such as Amazon’s Mechanical Turk. The interface is implemented in javascript using three.js (a WebGL-based graphics library). It can be accessed on the web using any modern browser, and does not require specialized software or hardware.

Figure 1 shows a series of screenshots from the interface illustrating the annotation process. The user first sees a rotating “turn table” view that reveals the appearance of the object from all directions. The user then types the name of a part in a text panel and drags over the surface of the object to select the regions corresponding to the part. The process is repeated for each part and the fi-



Figure 2: Partially annotated object with parts highlighted in different colors. The “door panel” label is selected, indicating that this part was just annotated.

nal result is a multi-colored painting of the object with corresponding part names for each color. The partitioning of the object’s geometry into different parts is saved and submitted to a data server upon completion. Figure 2 shows a close-up of the interface while an object is in the process of being annotated.

In order to enable an efficient multi-level painting interaction, the size of the paint brush can be adjusted by the annotator. The object geometry is pre-segmented for several levels of granularity: segmentation into surfaces with the same material assignment, a segmentation with a loose surface normal distance criterion, and finally a segmentation into sets of topologically connected components of the object geometry. This multi-level painting allows the speed of labeling to be adjusted to accommodate both small parts (e.g., door handles) and large parts (e.g., countertops on kitchen islands).

The annotators use freeform text for the part names, requiring that we address the problem of mapping this part name text to a WordNet synset. We implemented a simple algorithm that restricts the candidate synset set to physical objects in the WordNet hierarchy, preferring furniture (since we are dealing with indoor scenes that are predominantly composed of furniture). Given the object category, we can additionally use the meronym relations in WordNet to suggest and rerank possible part synsets.

After applying this algorithm to connect each part name’s text to a WordNet synset, we manually verify and if needed fix the inferred link using an interface that displays the WordNet synset assignment for the given part (in the same view as the part annotation view) and allows the user to select a different synset.

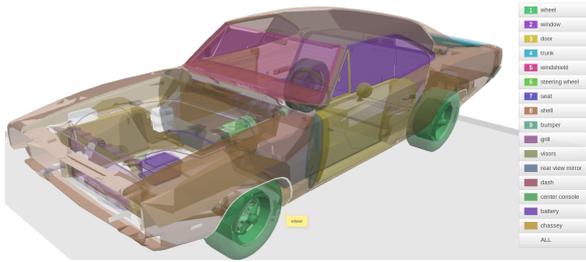


Figure 3: A fully completed part annotation example for a car 3D object. Different colors correspond to distinct parts with corresponding names provided by the annotator on the right.

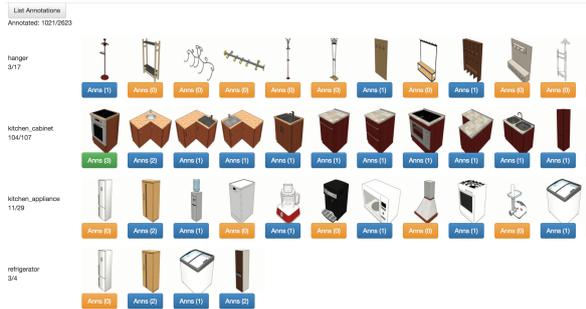


Figure 4: List of 3D models to be annotated. This is the interface through which one selects a model to annotate or can examine a previously annotated model.

5 Initial annotations and statistics

Five persons have worked on the annotation thus far. Annotating one 3D model takes roughly five minutes on average, with time being mostly a function of the complexity of the particular object. To maintain consistency while annotating the objects in the database, the annotators were instructed to name the geometric and functional parts of the object, not decorations or stylistic elements (e.g., a picture of a fish on a lamp shade).

Using the interface described above, we have so far collected more than 100,000 part instance annotations. An example of a car annotated by a crowd worker is shown in Figure 3. SUNCG includes a total of 2547 models, of which we have so far annotated 1021 during the prototyping process for developing our interface (Figure 4). Figure 5 shows several other example objects with their part annotations visualized.

6 Limitations

The initial stages of the annotations were limited by two major factors: the quality of the 3D models, and language in general. A handful of the models in the database had segmentation issues, i.e., any error in how the model was broken up

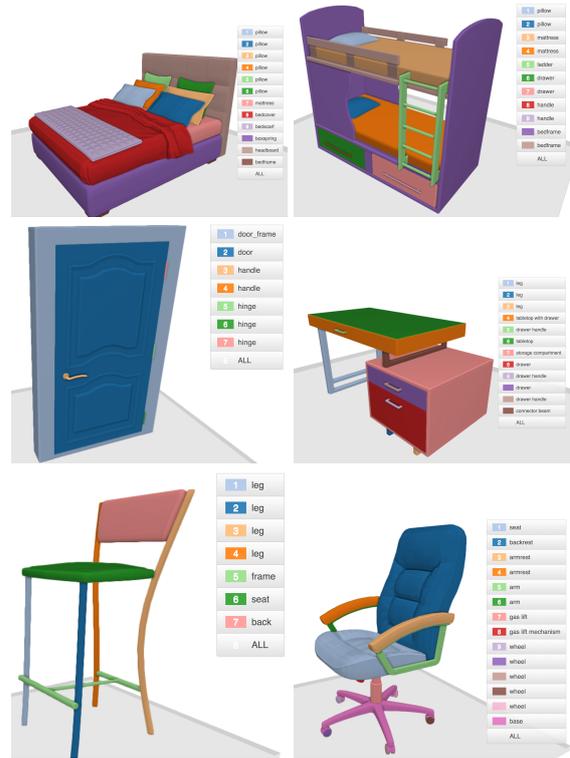


Figure 5: Example objects annotated with parts. Objects are assigned to the following WordNet synset, from top left: `double_bed.n.01`, `bunk_bed.n.01`, `door.n.01`, `desk.n.01`, `straight_chair.n.01`, `swivel_chair.n.01`.

into regions. For example, a segmentation issue for a table is encountered when more than one leg is connected into one region, thus making it impossible even with the smallest brush size to label the individual legs. To solve this problem, the segmentation algorithm must be improved upon. A more widespread problem with the database, however, was the lack of internal structure in many of the models. For example, in book cases with doors, only the doors would be present in the model, while the internal structures — in this case the shelves — were omitted.

Some linguistic factors can cause the annotation to be less than straightforward. For example, should the link be to the British or the U.S. word? If we annotate a coffee cup, should we annotate the piece of cardboard around the cup that is designed to protect our hand from the heat of the coffee using its specific but rare name “zarf,” or should we choose a more common but less specific term such as “sleeve” or even “piece of cardboard?” Moreover, some parts do not have proper labels: what should we call the beams that help stabilize some tables other than “support beams?” See Figure 6 for an example.



Figure 6: An example demonstrating some of the limitations of our annotation system. Parts of the table are identified as “storage support beam” and “shelf support panel” due to lack of a better term.

7 Future Work

Our work so far has focused on developing the infrastructure and annotation interfaces to collect 3D object part annotations at scale. This part data linked to WordNet is of tremendous potential value, which we plan to investigate as our project continues.

A very interesting direction of work is in building contextual multimodal embeddings. Many object parts and attributes are rarely mentioned in language. For example, a stool doesn’t have a back, but people don’t refer to stools as “backless chairs.” Neither do speakers encode the fact that chairs often have four legs or 5 wheels; only non-default exemplars might be labeled in an ad hoc fashion as “five-legged chairs,” for example. Furthermore, the physical contexts of objects (see Figure 7) provides richer information than is found in text. In this regard, the text and 3D modalities are complementary and provide an excellent target for building multimodal distributional representations (Bruni et al., 2014). Multimodal embeddings are a promising semantic representation which has been leveraged for various Natural Language Processing and vision tasks (Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015; Kottur et al., 2016).

Another direction for future work is to leverage the object part and attributes and their correspondences to WordNet to go beyond the set of WordNet synsets and automatically induce new senses, along the lines of recent work on sense induction (Chen et al., 2015; Thomason and J. Mooney, 2017). For example, we have found that WordNet synsets do not have good coverage of some fairly modern categories of objects that we observe in



Figure 7: An example of the same nightstand object (outlined in blue) in two different 3D scene contexts. A contextual embedding afforded by the full 3D representation of the scene within which the nightstand is observed would be a powerful way to analyze and disentangle different usage contexts for common objects.

our 3D object datasets, including iPads, iPhones and various electronic devices such as game consoles.

References

- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.
- Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An information-rich 3D model repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.
- Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. 2015. Sense discovery via co-clustering on images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5298–5306.

- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*.
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.
- Satwik Kottur, Ramakrishna Vedantam, Jose M. F. Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.
- James Pustejovsky, Tuan Do, Gitit Kehat, and Nikhil Krishnaswamy. 2016. The development of multimodal lexical resources. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 41–47.
- Eleanor Rosch. 1999. Principles of categorization. *Concepts: core readings*, 189.
- Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pages 721–732.
- Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*.
- Jesse Thomason and Raymond J. Mooney. 2017. Multi-modal word synset induction. In *IJCAI*.
- Barbara Tversky and Kathleen Hemenway. 1984. Objects, parts, and categories. *Journal of experimental psychology: General*, 113(2):169.
- Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*.

Multisłownik: Linking plWordNet-based Lexical Data for Lexicography and Educational Purposes

Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
maciej.ogrodniczuk@ipipan.waw.pl

Joanna Bilińska

University of Warsaw
j.bilinska@uw.edu.pl

Zbigniew Bronk

Institute of Computer Science
Polish Academy of Sciences
zbigniew.bronk@ai.pl

Witold Kieras

Institute of Computer Science
Polish Academy of Sciences
wkieras@ipipan.waw.pl

Abstract

Multisłownik is an automated integrator of Polish lexical data retrieved from multiple available online sources intended to be used in various scenarios requiring access to such data, most prominently dictionary creation, linguistic studies and education. In contrast to many available internet dictionaries Multisłownik is WordNet-centric, capturing the core definitions from Słowosieć, the Polish WordNet, and linking external resources to particular synsets. The paper provides details of construction of the resource, discussed the difficulties related to linking different logical structures of underlying data and investigates two sample scenarios for using the resulting platform.

1 Introduction

Multisłownik (Pol. *multidictionary*) is a linguistic integration platform for Polish lexical data retrieved from multiple available online sources intended to be used in various research and educational scenarios. The difficulty of such setting is clear: lexical data is created for different purposes resulting in various underlying structures and representation formats, tailored to specific requirements of each subfield of linguistics. For instance, morphological dictionaries may not differentiate word senses when inflectional patterns of each sense is the same; in turn, when they are different, senses can be assigned properly but at the same time usage examples from corpora restricted to a given sense may be difficult to retrieve.

The paper presents an attempt of creating such linked resource for Polish using computational

methods. Section 2 presents similar attempts for other languages, Section 3 describes the data sources used, Section 4 documents the decisions made during the process of data linking, Section 5 provides two sample scenarios based on the integrated data and Section 6 summarizes the paper and presents the work in progress.

2 Related Work

In contemporary lexicography there can be seen a tendency to integrate dictionaries into portals¹ mainly provided as a source of information for the ordinary users rather than linguists and researchers. Usually the idea of such portals is to give maximum data big publicity as possible with a minimal effort.

As compared to FRAN, a Slovenian dictionary of a similar type², gathering in-house lexical resources available to the Fran Ramovš Institute of the Slovenian Language ZRC SAZU, the initial assumption was that external resources will be used as well. The reason for such a decision was a desire to present Polish vocabulary in an extensive way which seemed to be impossible while using only open resources or those published by a single unit. Unlike in Slovenia, the main Polish linguistic sources were prepared by various publishing houses and research centres. However, because of the authors' rights, not all the dictionaries could be used in the same way. Therefore some of the dictionary data is only presented as references and information whether the searched word can be found in a given dictionary. By default FRAN presents results 'dictionary by dictionary'

¹See e.g.: <https://en.oxforddictionaries.com/>, <http://www.termania.net/>, <http://dictionaryportal.eu/>.

²See <http://fran.si>.

ordering them from the general one (with definitions) through etymological and historical to more specialised ones (e.g. spelling dictionary, medical lexicon or the dictionary of climber’s language).

Online dictionary of the PWN publishing house³ offers a similar approach to Polish: entries from dictionaries of several types are presented “as is” on a single Web page together with language use comments, encyclopaedia entries and corpus-based examples. Even less used-friendly Dictionary Portal of such type⁴ mainly facilitates searches in various dictionaries providing references to source entries.

Multisłownik combines the concepts of a dictionary portal and a general dictionary trying emulate a traditional dictionary. Therefore the query results are presented in a form of an automatically generated dictionary-like entry.

3 Sources of Lexical Data

Multisłownik integrates three different kinds of lexical resources:

1. traditional dictionaries created by philologists and meant for human readers only, either web-based or digitalized
2. electronic datasets created by computational linguists for both human users and automatic processing in NLP implementations
3. community-based lexical collections developed online.

The main two sources of lexical entries, forming the core of Multisłownik, are plWordNet (Piasecki et al., 2009)⁵ and Grammatical Dictionary of Polish (Saloni et al., 2012; Saloni et al., 2015; Woliński and Kieraś, 2016)⁶. Several others contributing to its content are: Polish language version of Wikipedia and Wikisource, Walenty valency dictionary (Przepiórkowski et al., 2014) and National Corpus of Polish (Przepiórkowski et al., 2012, NKJP)⁷. Various other lexical datasets are linked to each entry.

We briefly characterize these sources below showing their lexical potential and pointing out

³See <http://sjp.pwn.pl>.

⁴See <http://dictionaryportal.eu/>.

⁵See <http://plwordnet.pwr.wroc.pl/wordnet/>.

⁶Pol. Słownik gramatyczny języka polskiego, SGJP, see <http://sgjp.pl>.

⁷Pol. Narodowy Korpus Języka Polskiego, see <http://nkjp.pl>.

their most important features hindering integration.

3.1 plWordNet

plWordNet (Piasecki et al., 2009) is a lexico-semantic network reflecting the lexical system of Polish inspired by Princeton WordNet (Miller, 1995)⁸. It contains sets of synonymous lexical units (synsets) interconnected with lexico-semantic and derivational relations such as synonymy, hypo-/hypernymy or mero-/holonymy. plWordNet is currently the largest wordnet in the world and contains 178K synsets, 259K word senses and over 600K relations.

Apart from a very rough assignment of part-of-speech category (one of: noun, verb, adjective, adverb) to each lexical unit, plWordNet does not cover any other grammatical information such as grammatical gender for nouns or aspect for verbs. Some of this information may be derived from relations such as verb–noun *mpar_VN* relation linking verbs and derived gerunds. Currently plWordNet does not cover numerals and uninflected parts of speech.

3.2 SJP.pl

SJP.pl is a Web-based dictionary created by Polish enthusiasts of word games (mainly Scrabble). It aggregates vocabulary from various contemporary printed dictionaries, including spelling and foreign words dictionaries, and classifies them as permitted or non-permitted in word games. Currently it contains ca. 200,000 lexemes. SJP.pl is being developed by the community of its users. As the list of forms noted in SJP.pl is distributed under the terms of open source license it is also used as a data source for spell-checkers. Apart from inflectional forms SJP.pl entries usually also contain short definitions. For Multisłownik it serves mainly as a supplementary source of lexical and grammatical data, especially when the word searched by the user is not present in SGJP.

3.3 Grammatical Dictionary of Polish

Inflectional information is based on The Grammatical Dictionary of Polish (Pol. *Słownik gramatyczny języka polskiego*, SGJP) (Saloni et al., 2012; Woliński and Kieraś, 2016). SGJP is the largest existing linguistically elaborated data set of Polish inflectional morphology, from the very

⁸See <http://wordnet.princeton.edu>.

beginning developed as an electronic dictionary, now in its third edition turned into Web-based linguistic resource. SGJP serves as a main source of grammatical information for widely used morphological analyzer Morfeusz (Woliński, 2006; Woliński, 2014), as well as for the new general dictionary know as The Great Dictionary of Polish (Pol. *Wielki słownik języka polskiego*), currently under development (Żmigrodzki, 2007).

The integration of morphological data with plWordNet senses is hindered by high inflectional variation of Polish lexemes.

3.4 National Corpus of Polish

The National Corpus of Polish (Przepiórkowski et al., 2012) is the most prominent corpus of general Polish, providing a balanced representation of contemporary Polish. For Multisłownik it offers real usage examples. To ensure that they represent extensive variety of possible usage of the word it looks for corpus examples for all the possible non-syncretic forms from the inflectional paradigm of the word. For each such form a corpus frequency is also provided.

The corpus data is limited only to NKJP as the largest and most representative corpus of Polish available. Still, closing the dataset in in 2010 makes it less and less up to date each year. As a consequence, NKJP does not reflect the newest Polish vocabulary such as the word *prekariat* ‘prekariat’ which appears in 1-billion-word data set only twice while its actual frequency in daily and weekly newspapers is much higher in the recent years.

3.5 Wiktionary and Wikipedia

Wiktionary⁹ and Wikipedia¹⁰ are open-source, multilingual, community-developed dictionary and encyclopaedia fully available to download in XML format. For Multisłownik they are used as additional sources of lexemes, inflection forms, definitions, examples, collocations, information on pronunciation and etymology.

3.6 Other Linked Sources

Multisłownik also provides information about the presence of a search word in various other lexical resources unable to integrate directly due to licence or format constraints. The list of

⁹See <https://pl.wiktionary.org/wiki>.

¹⁰See <https://pl.wikipedia.org/wiki>. Note: due to its character, Wikipedia covers mostly nominal entries.

such resources is extremely heterogeneous. It contains both specialized linguistic dictionaries, both digitalized versions or paper dictionaries and Web-based developments as well as community-based lexical databases. The list of linked sources varies from well known general dictionaries such as PWN dictionaries (Słownik Języka Polskiego PWN, Słownik Wyrazów Obcych PWN, Doroszewski’s classical dictionary, available as scanned pages¹¹, through the electronic Dictionary of 17th & 18th Century Polish (Instytut Języka Polskiego PAN, 2010) to various resources capturing the newest vocabulary, both academia-based (such as the entries from the Language Observatory of the University of Warsaw¹²) and community-based, e.g. urban slang dictionaries¹³. Other sources include the Great Dictionary of Polish (Żmigrodzki, 2007), dictionaries of Polish personal and place names¹⁴ and dictionaries of synonyms, antonyms and crossword definitions¹⁵.

Their integration was motivated by practical reasons put forward by lexicographers: it saves user’s time and effort used for searching the word in all these sources separately.

4 Integration

Integration of multiple dictionary resources, heterogeneous by nature, poses various problems due to diverse representation and scope of lexical properties, different levels of detail and incompleteness of coverage of lexical entries. For online resources this situation gets additionally hindered by their constant change: new entries are added to lexicons, models are getting restructured and new data sources appear regularly. Based on all these assumptions we believe that the close integration of resources in such setting (such as combining them into a common LMF¹⁶ resource) is a myth — the complexity of such resource would need to exceed the complexity of its parts, already very high for most of the resources. Our approach is differ-

¹¹See <http://doroszewski.pwn.pl/>.

¹²See <http://nowewyrazy.uw.edu.pl/>.

¹³See e.g. Słownik miejski, <http://www.miejski.pl/>.

¹⁴See <http://nlp.actaforte.pl:8080/Nomina/Nazwiska> and <http://nlp.actaforte.pl:8080/Nomina/Miejscowosci>.

¹⁵<http://synonimy.net>, <http://antonimy.net>, <http://krzyzowka.net>.

¹⁶Lexical Markup Framework, an ISO 24613:2008 standard for machine-readable dictionary lexicons (Francopoulo, 2013).

ent and assumes interfacing related sources rather than absorbing them into a single common 'super-resource'.

At the same time a common point of reference is needed to serve as the core of the integration; for Multisłownik we decided it to be Słowosieć, the Polish WordNet (Piasecki et al., 2009), further referred to as plWordNet, the most extensive freely available semantic resource offering lexeme-to-sense mapping. plWordNet contains extensive description of lexical-semantic relations for Polish with interlinked synsets and short definitions, currently featuring over 300K lexical relations, 320K synsets and 1.2M inter-synset relations. In Multisłownik it serves as the main source of lexemes and semantic information.

Since plWordNet and SGJP make the most prominent resources covering respectively semantic and grammatical layers, comparison of these resources was of vital importance. As for the data set, SGJP contains 150K entries which do not have their counterparts in plWordNet (not taking into account negated adjectives, representing in SGJP as separate entries). On the other hand, plWordNet contains 20K entries absent from SGJP. plWordNet contains many multiword lexical units (over 30% of the total number) while SGJP does not cover any multiword entries apart from hyphenated entries such as *vis-a-vis* or *ping-pong* and a small sample of words functioning today only as parts of fixed phraseological expressions. Homonymy is the main problem of linking plWordNet data to SGJP; the set of homonyms contains 3450 nouns, 926 adjectives and 586 verbs.

The integration process starts with plWordNet taking over its semantic domains, lexical relation and synset relation types. SGJP is the main source of grammatical data and other resources are used to populate the entry.

Figure 1 presents a simple Web application interfacing Multisłownik platform. Sections provide information about pronunciation and etymology of the entry, its plWordNet senses with SGJP inflection variants assigned properly, related words retrieved from Wikidictionary, concordance from NKJP and information on presence of the lexeme in available online sources.

Information on pronunciation is presented in two formats: IPA and AS. For each sense its domain, definition, example and selected semantic

relations as well as English translation are presented. Grammatical information covers grammatical class, selective categories and inflection pattern symbol. Inflection section presents selected inflectional forms:

- for nouns – singular genitive and locative and plural nominative and genitive
- for adjectives – singular nominative feminine and neutral and plural nominative masculine
- for verbs – selected personal forms.

Syntax information is presented according to Walenty model and annotation. Frequency data and NKJP-based quotations are currently dynamically retrieved using PELCRA search engine.

5 Possible Usage Scenarios

The aggregation platform is intended to reflect a standard dictionary, therefore the results are presented in a form similar to a dictionary entry and reflect its microstructure. Each entry provides a number of slots for information: headword, pronunciation, etymology, senses/definitions, grammar information (inflectional patterns), translations into English, derived words and collocations, concordances with quantitative data from the NKJP. An important part are links to online dictionaries of surnames, geographical names, antonyms, synonyms, city slang vocabulary and new vocabulary which makes getting information about the contents of other sources, popularity or importance of lemmata very straightforward.

5.1 Lexicographic Scenario

Multisłownik is by its nature a highly heterogeneous resource on many levels: it integrates synchronic and diachronic dictionaries, specialist and general purpose dictionaries, scientific-driven and crowd sourced lexical databases. Thus it does not provide a sound lexicographic description but it can serve as an instant support for a professional lexicographer working in the field of extending a specific dictionary or a linguistic text annotation.

Since Polish is a highly inflectional language, morphological resources are crucial to almost any natural language processing task. For this reason grammatical data sets need constant development especially in reference to new vocabulary. A lexicographer working on this task needs to determine both grammatical features of the lexical entry (such as gender for nouns and aspect for verbs)

Multisłownik

🔍

PLÝWAK

Wymowa
IPA: ˈpwɪvak, AS: pɥɪvak Źródło: Wikisłownik

Pochodzenie
pol. pływać Źródło: Wikisłownik

Znaczenia, charakterystyka gramatyczna

◊ Rzeczownik męskoosobowy, wzór odmiany: **B3k**
Charakterystyczne formy fleksyjne:
lp D.: *pływaka*
lm M.: *pływacy*
depr.: *pływaki*

Źródło (z tabelą odmiany): [SGJP](#)

pływak₁ – (osoba) «mężczyzna uprawiający pływanie» Źródło: [Słowosiec](#)

Mikołaj był najszybszym pływakiem w polskiej reprezentacji olimpijskiej.

Synonimy: -

Hiperonimy: [wodniak₁](#)

Hiponimy: [nurek₂](#), [kraulista₁](#), [żabkarz₁](#), [grzbiecista₁](#), [delfinista₁](#), [zmiennista₁](#) [mniej ▲]

Tlum. ang.: *swimmer₁*

Agens|subiekt: [pływać₂](#)

Żeńskaść: [pływaczka₁](#)

Derywacyjność: [pływactwo₁](#)

Synonimia międzyparadygmatica dla relacyjnych: [pływacki₁](#)

◊ Rzeczownik męskozwierzotny, wzór odmiany: **B3k**
Charakterystyczne formy fleksyjne:
lp D.: *pływaka*
lm M.: *pływaki*

Źródło (z tabelą odmiany): [SGJP](#)

pływak₂ – (wytwór) «boja» Źródło: [Słowosiec](#) [więcej ▼]

pływak₃ – (wytwór) «splawik wędkarski» Źródło: [Słowosiec](#) [więcej ▼]

pływak₄ – (wytwór) «szczelny zbiornik, wypełniony powietrzem lub gazem, służący do utrzymania samolotu na powierzchni wody» Źródło: [Słowosiec](#) [więcej ▼]

pływak₆ – (wytwór) «element sieci rybackiej utrzymujący ją w pozycji pionowej i na zamierzonej głębokości» Źródło: [Słowosiec](#) [więcej ▼]

pływak₇ – (wytwór) «zakończenie przyrządu służącego do oznaczania poziomu cieczy w jakimś zbiorniku» Źródło: [Słowosiec](#) [więcej ▼]

◊ Rzeczownik męskozwierzotny, wzór odmiany: **B3k**
Charakterystyczne formy fleksyjne:
lp D.: *pływaka*
lm M.: *pływaki*

Źródło (z tabelą odmiany): [SGJP](#)

pływak₅ – (zwierzę) «drapieżny chrząszcz z rodziny pływakowatych» Źródło: [Słowosiec](#) [więcej ▼]

Wyrazy pokrewne

- rzecz. pływanie (n), pływalnia (f), pływaczka (f)
- czas. pływać, płynąć
- przym. pływacki, pływakowy

Źródło: Wikisłownik

Związki wyrazowe

[pływak żółtobrzeżek](#), [pływak lapoński](#), [pływak szerokobrzeżek](#), [zatapiacz pływaka](#), [świąd pływaków](#)

Frekwencja form i cytaty z NKJP (zrównoważonego)

pływak: 307

... rok, czas po Październiku, wydawał się **świt**. Zanurzyłam się jak **pływak** w morze nadziei. Nie spostrzegłam, że na dźwięk słowa "odwilż" moje...

... Geralt wyrwał się, rzucił w przód rozgarniając śmieci piersią jak **pływak** wodę, rąbnął z całej siły, z góry, z mocą naparł na ostrze...

... zewnętrznego, „z przestrzeni” powiedzielibyśmy my. Pomarańczowy **pływak** od sieci rybackiej. Ciężki, z polistyrenu, o dwu wielkich uszach....

... serca. Daleko, daleko jarzył się czerwienią jeszcze jeden czepek, **pływak** podniósł rękę w geście pozdrowienia. Poznałem Sergiusza Michalkowa....

... jej kolana. Kobieta popłynęła na dół, walcząc z wichrem niby **pływak** z rozchukanymi balwanami. Brązowy papier pomknął dalej i stara...

... wykonywał równocześnie gwałtowne ruchy rękami i nogami, jak **pływak**, ale w niczym nie poprawiało to jego rozpaczliwej sytuacji....

Występowanie w słownikach i encyklopediach

Źródło	Zawartość	Linki
Słowosiec	semantyka, powiązania, tłumaczenia	pływak
SGJP	gramatyka, fleksja	pływak
SJP.PL	znaczenia, formy odmiany	pływak
Wikisłownik	znaczenia, wymowa, etymologia, fleksja, tłumaczenia	pływak
Wikipedia	znaczenia, opis encyklopedyczny	Pływak
Walenty	składnia	-
WSJP	znaczenia, odmiana, etymologia...	-
Sł. nazwisk	odmiana nazwisk i ich występowanie w Polsce	Pływak
Sł. miejscowości	występowanie i odmiana nazw polskich miejscowości	-
PWN (SJP+SO+SJPDor.)	znaczenia, odmiana, ortografia, cytaty	pływak
SJPDor.	znaczenia, cytaty, odmiana	pływak
Slang miejski	znaczenia slangowe	-
svvii.pl	słownictwo polskie XVII i XVIII w.	-
Oberwatorium UW	najnowsze słownictwo polskie	-
Synonimy.NET	synonimy	-
Antonimy.NET	antonimy	-
Krzyżówka.NET	definicje do krzyżówek	-

Figure 1: Test front-end of Multisłownik

and some specific word endings. Consider for example a noun PARKOUR ‘a training discipline’, which does not appear in the *Grammatical Dictionary of Polish*. Since she is dealing with an obvious loanword the lexicographer needs to determine, whether the noun declines or it has all its forms homonymous. If it declines, some alternative word endings need to be determined, such as *-u* or *-a* in genitive singular (both are possible). Also a grammatical gender needs to be assigned (could be either neuter or masculine inanimate). Since the word refers to a rather niche sport activity, a regular lexicographer cannot rely on her own experience and needs to consult some external lexical resources. By simply typing the word *parkour* in Multisłownik’s search bar the lexicographer gains access to

1. basic definition (provided by plWordnet)
2. characteristic inflectional forms and hypothetical gender value (provided by Multisłownik’s own heuristic algorithms)
3. usage examples for four different inflectional forms including their frequencies (found in the National Corpus of Polish).

Based on these informations a proper grammatical description of the word can be formulated and included in the dictionary.

On the other hand a human annotator conducting a morphological, syntactic or semantic text annotation needs a constant access to large lexical data sets supporting her work. Text samples often do not provide a sufficiently large context to determine the proper meaning of a text token or the annotator simply does not have enough specialist knowledge to determine i.e. a lemma of a word. Consider a locative phrase *w Sycowie* (“in Syców/Sycowo”) in which a proper name can be lemmatized either as SYCÓW or SYCOWO. Both endings (*-ów* and *-owo*) are correct and both are very common in Polish names of settlements, both form a locative case form ending with *-owie* but only one of the resulting base forms actually exists and refers to a small town Syców in southwestern Poland. The proper lemma can be easily determined in Multisłownik in which a proper names’ declension dictionary is integrated.

5.2 Educational Scenario

Although the platform is aimed at the linguistically- and lexicographically-aware user, it can also be an attractive source of informa-

tion for wider audience, for instance high school pupils. Searching for random words can be a good start point to teach the students what is the dictionary microstructure and how it can differ between dictionaries. After this stage we plan to present the dictionary by looking up the words. We would suggest following queries for teaching purposes, aiming to present the platform to the young people:

1. Check the word KAFAR and PROMULGOWAĆ in Google and in Multisłownik — what are the differences, information given, which source gives you more information on the lemma in the first hit (without further clicking)?
2. What is GEN.PL of MECZ or DAT.SG of MUCHA? (results from the grammatical dictionary)
3. What are the possible lemmata for the word form “danie” (the grammatical dictionary)
4. Which animals groups are called STADO? (the National Corpus of Polish)
5. Who is KALETNIK (plWordNet)
6. What are the other words derived from SEKRET (plWordNet)
7. What are the antonyms of the word SEKRET? (the dictionary of antonyms)
8. Is the form “Dania” in “Dania jest piękna” and “Dania hiszpańskie są smaczne” pronounced in the same way? (Wikisłownik)
9. What is the difference in meaning of NYGUS in general Polish and in the city slang? (plWordNet, slang dictionary)
10. Is the word form ŁABADŹ always incorrect? (dictionary of surnames and 16–17 century dictionary)
11. What is the origin of the words KSIEŻYC and ŁABEDŹ? (Wikisłownik)
12. Is there a place (city, town, village) called “Łabędź” in Poland? (dictionary of surnames)
13. What does the word TRZECIOTEŚCIK mean? (language observatory)
14. What are the synonyms for the DOM? (plWordNet)
15. Which case is “tysiącpięćsetletniemu”? (grammatical dictionary)

The classes on using the dictionary portal would be even more attractive to students when cross-words or other word games (e.g. Scrabble) are used as search targets. One of such activities could be deciphering a coded information with the usage of Multisłownik conducted in a following way:

- Formulating a question that needs to be answered.
- Providing the coded answer with some or all characters replaced with numbers connected to the questions that lead to decoding the secret characters.
- Possible types of questions:
 - “The last letter of the synonym of the word SEKRET that ends with letter T”.
 - “What is the origin of the word KUŚNIERZ? The first letter of the original language name is the secret character number X”.
 - “Is there a surname Łabądz in Polish? If yes, the secret letter is N, if no, the secret letter is C”.

6 Conclusions and Further Steps

Multisłownik already proved useful in many scenarios related to combining lexical information by offering a simple yet practical method of referring to multiple sources at the same time.

The most obvious further direction for extension of Multisłownik is adding more data; it occurs that even resources less relevant to the current task, e.g. numerous historical corpora can help lexicographers retrieve usage examples from historical texts to trace back the change of word meanings.

Another type of interesting functionality of Multisłownik would be searching for so called “cultural traces” of a given word. Apart from offering the user extensive dictionary-based grammatical and semantic information also references of a given word or phrase to important artwork (e.g. its presence novel and movie titles, lyrics of popular song or famous quotes) could be tracked. This would require building much larger datasets based on library catalogues, movie databases and Wikiquote, integrated and sorted according to its impact on both high and popular culture.

Acknowledgments

The work reported here was carried out within the research project financed by the Polish National Science Centre (contract number 2014/15/B/HS2/00182) and was partially financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

- Gil Francopoulo. 2013. *LMF. Lexical Markup Framework*. ISTE - Wiley.
- Instytut Języka Polskiego PAN. 2010. *Słownik języka polskiego XVII i 1. połowy XVIII w. [En. Dictionary of 17 century and 1st half of 18 century Polish]*. Warszawa.
- Piotr Żmigrodzki. 2007. O projekcie Wielkiego słownika języka polskiego. *Język Polski*, 5(LXXXVII):265—267.
- George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.
- Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej.
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw.
- Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2785–2792, Reykjavík, Iceland. ELRA.
- Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. 2012. *Słownik gramatyczny języka polskiego*. Warszawa, 2. edition.
- Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. 2015. *Słownik gramatyczny języka polskiego*. 3. edition, online publication.
- Marcin Woliński and Witold Kieraś. 2016. The online version of Grammatical Dictionary of Polish.

In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2589–2594, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining 2006 Conference*, pages 511–520, Wisła, Poland, June.

Marcin Woliński. 2014. Morfeusz Reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1106–1111, Reykjavík. European Language Resources Association.

Putting Figures on Influences on Moroccan Darija from Arabic, French and Spanish using the WordNet

Khalil Mrini

Ecole Polytechnique Fédérale de Lausanne
Switzerland
khalil.mrini@epfl.ch

Francis Bond

Nanyang Technological University
Singapore
bond@ieee.org

Abstract

Moroccan Darija is a variant of Arabic with many influences. Using the Open Multilingual WordNet (OMW), we compare the lemmas in the Moroccan Darija Wordnet (MDW) with the standard Arabic, French and Spanish ones. We then compared the lemmas in each synset with their translation equivalents. Transliteration is used to bridge alphabet differences and match lemmas in the closest phonological way. The results put figures on the similarity Moroccan Darija has with Arabic, French and Spanish: respectively 42.0%, 2.8% and 2.2%.

1 Introduction

Locally known as Darija and referred to as a dialect, the Moroccan variant of the Arabic language is spoken by the overwhelming majority of Moroccans (HCP, 2014) with small regional differences. The Moroccan Darija Wordnet (MDW) (Mrini and Bond, 2017) was released as part of the Open Multilingual WordNet (OMW) (Bond and Foster, 2013), thereby linking all the languages in the OMW to Moroccan Darija.

Morocco has a complex language situation. Its two official languages are Arabic, the basis of Moroccan Darija, and, since 2011, Tamazight. The North African Kingdom has gained its independence in 1956 from colonial France and Spain, and both countries have had linguistic influence on Moroccan Darija through loanwords.

Moroccan Darija is used in day-to-day informal communication (Ennaji, 2005) and doesn't have the prestige associated with Arabic or French, which are the languages used in education. In the 2014 census (HCP, 2014), it was reported that Morocco's literacy rate is at 67.8%, making it one of the lowest in the Arab World. A 2010 study (Magain, 2010) found that although the reasons of high

illiteracy rates in the Arab World are varied and subject to controversy, one of them was the “*disconnect between high Arabic used as the medium of instruction in schools and the various dialects of Arabic spoken in Arab region*”.

This paper aims at putting figures on the influences of the Arabic, French and Spanish languages on Moroccan Darija. Accordingly, on top of the MDW, the Arabic (Black et al., 2006; Abouenour et al., 2013), French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets were used.

To gauge the influence of these languages on Moroccan Darija, the word distance between each Moroccan lemma and its corresponding lemma in the other language was computed. This is done using the sense-based property of the WordNet. Transliteration helped to bridge the difference between the Arabic and Latin alphabets and the difference of use of the Latin alphabet between the European languages studied. The language-dialect similarities were computed for both the automatically linked Moroccan synsets, as well as the manually validated ones.

2 Related Work

In this section, we describe relevant aspects of the MDW. Then, we provide a review of studies on language-dialect and dialect-dialect similarity, as well as a review of methods to compute word-to-word similarity.

2.1 The Moroccan Darija Wordnet

The Moroccan Darija Wordnet (Mrini and Bond, 2017) was developed using an *expand* approach with the vocabulary being extracted from a bilingual Moroccan-English dictionary (Harrell, 1963). There were 12,224 Moroccan synsets automatically connected to the Princeton WordNet (Fellbaum, 1998), with 2,319 of them being manually verified.

During the process of developing the MDW, a Latin-based Moroccan alphabet was set, using as basis the one used in the bilingual dictionary, as well as the colloquial alphabet used in daily written communication between Moroccans. The MDW alphabet assigns one sound to one letter, and this facilitates transliteration to other languages.

2.2 Linguistic Similarity and Dialects

A similar case to Moroccan Darija is the Maltese language. Brincat (2005) recounts that it was first considered an Arabic dialect, but an etymological analysis of the 41,000 words of a bilingual Maltese-English dictionary shows that 32.41% of them are of Arabic origin, 52.46% of Sicilian or Italian origin and 6.12% of English origin. This heterogeneous mix is probably one of the reasons that Maltese is the only colloquial Arabic dialect that emancipated to become a full-fledged language (Malette, 2011). Aquilina (1972) established a wordlist of Maltese Christian words of Arabic origin, as well as an earlier detailed etymological study comparing Maltese and Arabic (Aquilina, 1958).

Scherrer (2012) proposes a simple metric to measure the similarity between different dialects of Swiss German in a corpus. It is based on the Levenshtein distance (Levenshtein, 1996), also known as the edit distance. The latter is a string metric measuring the difference between two sequences. It represents the minimum number of characters to modify to make both sequences identical. Those modifications can be single-character insertions, substitutions or deletions. Heeringa et al. (2006) propose an evaluation of distance measurement algorithms for dialectology, in which they use a normalised version of the edit distance so that it is comprised between 0 and 1. Inkpen et al. (2005) propose normalising the Levenshtein distance by dividing by the length of the longest string. That is the method that we will be using to compute word-to-word similarity.

3 Estimating Influences

To compare Moroccan Darija with Arabic, French and Spanish, we will perform word-to-word comparisons. These comparisons must be phonologically accurate despite alphabet differences.

3.1 Computing Word Distance

The similarity is assessed through looking at lemmas across different languages that share the same sense. That means that in a given OMW synset with one or more Moroccan lemmas, the latter will be compared to the synset's Arabic, French and Spanish lemmas.

Spaces and underscores in multiword expressions would count as letters, and there may be word order differences, thus resulting in inaccuracies. Therefore, multiword expressions were excluded from the comparisons.

We compare the three languages first to all the Moroccan synsets that were automatically linked in Mrini and Bond (2017), and then only to the ones manually validated and included in the first release of the MDW.

3.2 Transliteration of the MDW Alphabet

We can compute the normalised distance between two words, but we also need to bridge the difference of alphabets between the languages studied. We do this through a process of transliteration. The transliteration used from Moroccan Darija was specific to each language to which it was compared and proposed phonological correspondences. The purpose of these phonological correspondences is to be able to recognise what words are cognates or borrowings. It is at this point difficult to distinguish if a pair of similar lemmas are cognates or the result of a borrowing.

The transliteration process is complicated, as, if it is strict, its accuracy may be low. This is why numerous options were considered for the transliteration of each Moroccan letter. This way, all the possible transliterations of a word are considered for comparison. The number of possible transliterations for a lemma is the product of the lengths of each set of possible transliterations of each letter contained in the lemma. The flexibility of the transliteration process is optimistic, as the smallest distance resulting from any transliteration option is the one considered for computing the overall cross-lingual average distance.

3.2.1 Transliteration to Arabic

The Arabic WordNet (Black et al., 2006) has words in the Arabic alphabet with irregular diacritics, meaning that a short vowel in a word may or may not be illustrated by diacritics. Each diacritic is considered to be a separate character in the string that represents the Arabic lemma. Therefore, two

Darija	Transliterations		
	Arabic	French	Spanish
a	ا, ة, ي, ة, ϕ	a	a, á
b, ḅ	ب	b, p, v	b, p, v
d	د, ظ, ض, د	d	d
ḍ	ظ, ض	d	d
e	ϕ	e, é, è, ê	e, é
ã	ا, ϕ	a, e, é, è, ê	a, e, é
f	ف	f, ph	f
g	ق, گ	g	g
8	غ	r	r
h	ه	h	h
7	ح	h, ϕ	h, j, ϕ
i	ي, ة, ي, ϕ	i	i, í
ĩ	ي, ϕ	i	i, í
j	ج	j	y
k	ك	k, c	k, c
l, ḷ	ل	l	l
m, ṃ	م	m	m
n	ن	n	n
o	و, ة, و, ϕ	o	o, ó
q	ق	q, k, c	q, k, c
r, ṛ	ر	r	r
s	س, ص	s	s, c, z
ş	ص	s	s, c, z
š	ش	ch	ch
t	ط, ث, ت, ت	t	t
ṭ	ظ, ط	t	t
u	و, ة, و, ϕ	ou, u	u, ú
w	و, ة, و, ϕ	w, ou	u, ú
x	خ	kh	j
y	ي, ϕ	y	y, i, í, ll, ϕ
z, ẓ	ز	z	z
2	ϕ	ϕ	ϕ
3	ع	a, ϕ	a, ϕ

Table 1: Transliteration from Moroccan Darija to Arabic, French and Spanish

transliterations were necessary. On the one hand, the diacritics on Arabic WordNet lemmas were erased. On the other hand, each Moroccan character was transliterated as per Table 1. The emphatic Arabic characters were included in both emphatic and dotless *b*'s, *d*'s, *s*'s and *t*'s, but non-emphatic Arabic characters were not included in the possible transliterations of *b*'s, *d*'s, *s*'s and *t*'s. Diacritics having been removed, transliteration of Moroccan vowels to short vowels was represented by the possibility of removing them (ϕ).

3.2.2 Transliteration to French and Spanish

The transliterations of Moroccan Darija to lemmas of the French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets were also made to be as flexible as possible. In French, all accents on *e*'s were considered. Likewise, the accents used for stressed syllables in Spanish were considered for all vowels. For both languages, the Moroccan *b* could be transliterated as either *b*, *p* or *v*, as there is a near-absolute absence of *p* and *v* in Moroccan Darija. The Spanish pronunciation of *ll*, *z* and *j* differs from the French one and therefore they were mapped to different letters in the Moroccan alphabet. Furthermore, some Moroccan letters were matched to two French letters, as the French pronunciations of *ou* and *ph* respectively match the Moroccan *u* and *f*, and Morocco's official transliteration of the /x/ sound is *kh*. Like the Arabic Transliteration code above, each character was transliterated individually. The transliterations to French and Spanish are also in Table 1.

4 Results and Discussion

The aggregated results of the word-to-word comparisons gave an estimation of the linguistic influences on Moroccan Darija. We obtained results for the Moroccan synsets that were automatically linked and the ones that were manually validated.

Each of the Arabic, French and Spanish wordnets had a certain number of links to Moroccan synsets for which they had at least one available single-word lemma. To these synsets, a certain number of lemmas were associated. In both comparisons, the Moroccan lemmas were matched for pairwise comparisons. Based on that number of synsets matched, an average normalised Levenshtein distance was given for both languages. Examining the results, we decided that synsets should only be counted as a match if they had at least 60% similarity.

Comparison with:	Arabic	French	Spanish
Number of links to Moroccan synsets	7,958	11,605	10,167
– excluding synsets with only multi-word expressions	6,702	9,954	8,612
Average normalised Levenshtein distance	0.4619	0.7337	0.7521
Number of synsets with one or more word pairs at least 60% similar	2,816	278	188
Percentage of synsets with one or more word pairs at least 60% similar	42.02%	2.79%	2.18%

Table 2: Results of the comparisons of automatically linked synsets of Moroccan Darija with Arabic, French, Spanish

4.1 Comparison based on Automatically Linked Moroccan Synsets

The results of the comparisons of the automatically linked Moroccan synsets with each language are given in Table 2.

4.1.1 Cross-lingual Similarity Scores

The results show that, on average, a Moroccan Darija word is 53.81% similar to its Arabic translation, 26.63% similar to its French translation and 24.79% similar to its Spanish translation, the similarity being 1 minus the distance. The similarity method used is akin to related work on semantic similarity (Ciobanu and Dinu, 2014). The average normalised distance was computed by averaging the lowest normalised Levenshtein distance found in any lemma pair in each comparison of a Moroccan synset to the WordNet synset matches, with all Moroccan synsets having equal weights in the average.

If the confidence scores were used as weights in the average normalised Levenshtein distance, then Moroccan Darija would be on average 52.99% similar to Arabic, 24.02% similar to French and 22.25% similar to Spanish. Some of the similarities may be random, this is why a threshold must be empirically established, such that word pairs which similarity has crossed the threshold are visibly similar. On establishing a threshold of 60% similarity, the similarity numbers dwindle faster for French and Spanish than for Arabic.

4.1.2 Similarity with Arabic

Moroccan Darija and Arabic share an average normalised Levenshtein distance of around 0.4619. This number puts a figure on the similarity between Moroccan Darija and Arabic.

For comparison, the same method of comparison can be applied to other pairs of languages.

This way, it can be determined that Portuguese (de Paiva et al., 2012) and Galician (Gonzalez-Agirre et al., 2012) are the closest case to Moroccan Darija and Arabic with an average Levenshtein distance of 0.4760. The former two languages are considered independent languages. These comparisons show how blurry the line is between a dialect or variant and an independent language, especially within the continuum of Arabic dialects (Greene, 2013). From these results, Moroccan Darija can be seen as distinct enough from Arabic to possibly be considered a language of its own.

4.1.3 Similarity with French and Spanish

Out of the 278 synsets that were more than 60% similar to Moroccan Darija for French and the 188 ones for Spanish, there were 95 common synsets. Therefore, some non-negligible part of the similarity of French and Spanish with Moroccan Darija is due to the similarity between French and Spanish. Future work would allow to distinguish the linguistic influence represented by each of these common synsets.

4.1.4 Moroccan Lemmas of Unknown Origin

Taking the Moroccan synsets connected to the Arabic, French and Spanish wordnets, the lemmas that were among any of the lists of word pairs that were more than 60% similar were eliminated. Therefore, this resulted in a set of 2,736 Moroccan synsets of unknown origin. Among these, there are words of Arabic origin such as “*deqq*” (from the Arabic verb for “to block”) and “*nzel*” (from the Arabic verb for “to go down”). Some words are of French or Spanish origin such as “*serbisa*” (from the Spanish noun for “beer”). These were probably due to errors in linking the Moroccan synsets to the WordNet.

A sizeable proportion is of Tamazight origin,

Comparison with	Average distance			At least 60% similarity		
	Arabic	French	Spanish	Arabic	French	Spanish
The 12,224 synsets that form the total	0.4619	0.7337	0.7521	42.02%	2.79%	2.18%
The 617 manually validated synsets	0.4393	0.7544	0.7721	47.00%	3.08%	2.92%

Table 3: Comparison of average Levenshtein distances between different sets of synsets and comparison of percentage of number of synsets that are at least 60% similar between different sets of synsets

such as “*degdeg*” (from the Tamazight verb for “*to smash*”) and “*seqsi*” (from the Tamazight verb for “*to ask*”). The influence of Tamazight is very visible on the words that start with “*ta-*” and end in “*t*” such as “*tazellajt*” and “*tabennayet*”. The study of the Tamazight influence will most likely require the creation of a Tamazight WordNet.

4.2 Comparison based on Manually Validated Moroccan Synsets

In order to investigate the effect of linking errors, we perform the same comparison on the 2,319 manually verified synsets contained in the current release of the MDW. Then we filtered them to obtain the synsets with links to each of the Arabic, French and Spanish wordnets. Therefore this set used for validation contains 617 Moroccan synsets.

The average Levenshtein distances and the percentages of synsets that are at least 60% similar are in Table 3. The difference in figures between the manually validated Moroccan synsets and the automatically linked ones proved small enough to say that the linking noise was not an issue.

5 Summary

In this paper, we attempted to put figures on the similarity between Moroccan Darija and each of Arabic, French and Spanish.

Transliteration was used to bridge the alphabet gap and perform phonological comparisons. The methods used were flexible and the comparisons exploited all possible transliterations for each letter. Transliteration was one-way from the Moroccan Darija Wordnet (Mrini and Bond, 2017) for the French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets, but was both ways for the comparison with the Arabic WordNet (Black et al., 2006). The word-to-word distance was computed using Levenshtein distance (Levenshtein, 1996), which was normalised

(Heeringa et al., 2006) using the biggest word length in the word pair (Inkpen et al., 2005). Multiword expressions were ignored for the comparisons.

The comparisons using the automatically linked Moroccan synsets gave that Moroccan Darija has an average normalised Levenshtein distance of 0.4619 with Arabic, 0.7337 with French and 0.7521 with Spanish. The percentage of synsets with word pairs that were at least 60% similar is 42.02% for Arabic, 2.79% for French and 2.18% for Spanish. There remained 2,763 Moroccan synsets of unknown origin out of those linked to the OMW. Some have origins in Arabic, French or Spanish due to errors in linking, whereas others were found to have links to Tamazight.

The comparisons using the manually validated Moroccan synsets yielded an average normalised Levenshtein distance of 0.4393 with Arabic, 0.7544 with French and 0.7721 with Spanish, with the percentage of synsets with word pairs that were at least 60% similar is 47.00% for Arabic, 3.08% for French and 2.92% for Spanish. The results for the normalised Levenshtein distance can be considered as a validation, but the number of word pairs that were at least 60% similar is too small to give a clear validation.

The similarity between Moroccan Darija and Arabic is closest to the one between Portuguese (de Paiva et al., 2012) and Galician (Gonzalez-Agirre et al., 2012), that are two independent languages. This shows that Moroccan Darija may be considered a language of its own. Nonetheless, there is no case of WordNet dialect-language or variant-language comparison to confirm this hypothesis.

References

- Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2013. On the evaluation and improvement of Arabic wordnet coverage and usability. *Language Resources and Evaluation*, 47(3):891–917.
- Joseph Aquilina. 1958. Maltese as a mixed language. *Journal of Semitic Studies*, 3(1):58.
- Joseph Aquilina. 1972. Maltese christian words of arabic origin.
- W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, M. Bertran, and C. Fellbaum. 2006. The arabic wordnet project. In *Proceedings of LREC 2006*.
- Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013, Sofia*, page 1352–1362.
- Joseph M Brincat. 2005. Maltese—an unusual formula. *MED Magazine*, 27.
- Alina Maria Ciobanu and Liviu P. Dinu. 2014. An etymological approach to cross-language orthographic similarity. application on romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1047–1058.
- Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMap technical report, Escola de Matemática Aplicada, FGV, Brazil.
- Moha Ennaji. 2005. *Multilingualism, cultural identity, and education in Morocco*. Springer Science & Business Media.
- Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. MIT Press.
- Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.
- Robert Lane Greene. 2013. Arabic: A language with too many armies and navies? *The Economist*.
- Richard S. Harrell. 1963. A dictionary of moroccan arabic: Moroccan-english. *Georgetown University Press*.
- Haut Commissariat au Plan du Maroc HCP. 2014. Recensement de la population.
- Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances*, pages 51–62. Association for Computational Linguistics.
- Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.
- Vladimir Levenshtein, editor. 1996. *Binary codes capable of correcting deletions, insertions, and reversals*.
- Shawn Magin. 2010. Illiteracy in the arab region: A meta study. *GIA Lens*, 2.
- Karla Mallette. 2011. *European Modernity and the Arab Mediterranean: Toward a New Philology and a Counter-Orientalism*. University of Pennsylvania Press.
- Khalil Mrini and Francis Bond. 2017. Building the moroccan darija wordnet (mdw) using bilingual resources. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP), Casablanca, Morocco*.
- Benoît Sagot and Daria Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco*.
- Yves Scherrer. 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EAACL 2012 Joint Workshop of LINGVIS and UNCLH, Avignon, France*, pages 63–71.

pyiwn: A Python-based API to access Indian Language WordNets

Ritesh Panjwani[†], Diptesh Kanojia^{†,♣,*}, Pushpak Bhattacharyya[†]

[†]Indian Institute of Technology Bombay, India

[♣]IITB-Monash Research Academy, India

^{*}Monash University, Australia

[†]{riteshp, diptesh, pb}@cse.iitb.ac.in

Abstract

Indian language WordNets have their individual web-based browsing interfaces along with a common interface for IndoWordNet. These interfaces prove to be useful for language learners and in an educational domain, however, they do not provide the functionality of connecting to them and browsing their data through a lucid application programming interface or an API. In this paper, we present our work on creating such an easy-to-use framework which is bundled with the data for Indian language WordNets and provides NLTK WordNet interface like core functionalities in Python. Additionally, we use a pre-built speech synthesis system for Hindi language and augment Hindi data with audios for words, glosses, and example sentences. We provide a detailed usage of our API and explain the functions for ease of the user. Also, we package the IndoWordNet data along with the source code and provide it openly for the purpose of research. We aim to provide all our work as an open source framework for further development.

1 Introduction

WordNets are extensively used in many sub-tasks for Natural language Processing (NLP) (Knight and Luk, 1994; Tufiş et al., 2004). They are a rich semantic lexicon which are accessible, free-to-use and fairly accurate. They have been used in cross-lingual information retrieval (Gonzalo et al., 1998), word sense disambiguation (Sinha et al., 2006), question answering (Pasca and Harabagiu, 2001) etc. Princeton WordNet (Fellbaum, 1998) or the English WordNet was the first to come into existence. EuroWordNet (Vossen, 1998) followed with a common structure for 12 European

languages. Indian language WordNets originated with the advent of Hindi WordNet (Narayan et al., 2002) and based on an expansion approach, the rest of them were created. They form a common lexico-semantic resource called IndoWordNet (IWN) (Bhattacharyya, 2010). India has more than 22 languages and 18 of these have constituent WordNets under a common roof - IndoWordNet. A considerable effort has gone into the creation of a perfect Application Programming Interface (API) for English WordNet and many of these are available for use, publically. Although, we do not see that kind of push for a common API for Indian language WordNets. NLP research for Indian languages has seen tremendous growth in the recent past and wordnets are a crucial resource especially in the context of NLP methodologies based on knowledge bases.

“With our work, we aim to provide an accessible, robust, easy-to-use API for Indian language WordNets.”

Additionally, this will help emerging wordnets acquire our open-source framework and adapt to it for creating a simple python based API, thus helping NLP for their own language.

2 Motivation

Efforts to create a lexical semantic network for Indian languages began with Hindi WordNet¹ (Narayan et al., 2002), and based on the concept of pivotal expansion, IndoWordNet² (Bhattacharyya, 2010) was created. NLP for Indian languages is gaining traction among the computer scientists in India, and a robust framework which is readily available for use is much needed. We believe that such an API bundled with the IndoWordNet data could be really helpful to the NLP community.

¹<http://www.cfilt.iitb.ac.in/wordnet/webhwn/index.php>

²<http://www.cfilt.iitb.ac.in/indowordnet/>

Princeton WordNet or the English WordNet API is available for use via NLTK³ (Bird et al., 2009) in Python. Bond et al. (2016) collaborate many WordNets and provide open access for using the wordnet data aligned with them. These wordnets from all over the world are linked via their linkages to English WordNet. Indian language wordnets are also linked to English, but only 25000 out of 40000+ synsets. Until the time all the synsets are unlinked; a separate API for browsing through Indian languages is required, and a common API might not sufficiently cover the data available with IndoWordNet.

Hence, we build this API with an aim that IndoWordNet data should also readily available in an easy-to-use framework. Python facilitates pre-built libraries and datasets for NLP via NLTK. TensorFlow by Google (Abadi et al., 2016) is also built on Python, and other classic Machine Learning algorithms are available for use via the *sci-kit learn* (*sklearn*) library (Pedregosa et al., 2011). Hence, we choose Python for implementing the API and build a framework using it.

3 Related Work

The Java WordNet Library⁴ has been extensively used for research across various domains in NLP (Chauhan et al., 2013; Zesch et al., 2008; Gurevych et al., 2012). *extJWNL*⁵ extend JWNL and provides command-line support, and Maven⁶ support among many other features in their API. Emerging WordNets like Sinhala WordNet (Welgama et al., 2011) employ JWNL to create an API for their WordNet. Java API for WordNet Searching (JAWS) (Spell, 2009) is another such implementation. The MIT Java WordNet Interface (JWI)⁷ is also available for the same purposes and is available under the Creative Commons 4.0 License⁸. Finlayson (2014) presents an extensive evaluation of the APIs available in Java for accessing Princeton WordNet. All of the work above has been done for Java, and is available for Princeton WordNet. A Python based toolkit, ESTNLTK (Orasmaa et al., 2016) includes Esto-

³<http://www.nltk.org/>

⁴<http://jwordnet.sourceforge.net/handbook.html>

⁵<http://extjwnl.sourceforge.net/>

⁶<https://maven.apache.org/>

⁷<https://projects.csail.mit.edu/jwi/>

⁸<https://creativecommons.org/licenses/by/4.0/>

0/

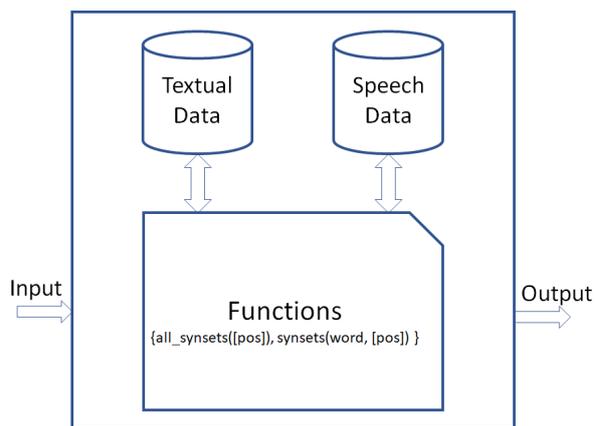


Figure 1: Basic flow of the *pyiwn* API

nian WordNet developed under the EuroWordNet project (Vossen, 1998).

Previously, efforts had been made to create an API for IndoWordNet but they are not Python based. Prabhugaonkar et al. (2012) describe a two-layered architecture of a web-based API created using PHP. It requires one to download data separately and is inconvenient to deploy; we also come across hard-coded paths while trying to deploy their API. A Java-based API⁹ is available for download on the Hindi WordNet web interface, and also requires one to separately download the database for Hindi WordNet. Redkar et al. (2016) claim to have built an API for WordNets universally but their work is not publicly accessible, and no references to their implementation could be found. Hence, we work on an API which would contain NLTK like functionality, should be robust, readily available, and more importantly easy-to-use.

4 API Design

We choose Python for implementation due to its widespread use in the NLP community and aim to align our work with NLTK. With this in mind, we keep the design of our API similar to that of NLTK WordNet Interface¹⁰. Our API provides access to synsets and their relational connections such as hypernymy, hyponymy, meronymy, *etc.* with other synsets for all the languages mentioned in the Table 1. The basic flow of the API is visualized in Figure 1. The API is available as an open source project on GitHub¹¹.

⁹<https://goo.gl/N8GXAU>

¹⁰<http://www.nltk.org/howto/wordnet.html>

¹¹<https://github.com/riteshpanjwani/pyiwn>

Language	Noun	Verb	Adjective	Adverb	Total
Hindi	29807	3687	6336	541	40371
Assamese	9065	1676	3805	412	14958
Bengali	27281	2804	5815	445	36346
Bodo	8788	2296	4287	414	15785
Gujarati	26503	2805	5828	445	35599
Kannada	12765	3119	5988	170	22042
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Malayalam	20071	3311	6257	501	30140
Manipuri	10156	2021	3806	332	16351
Marathi	23271	3146	5269	539	32226
Nepali	6748	1477	3227	261	11713
Odiya	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Sanskrit	32385	1246	4006	265	37907
Tamil	16312	2803	5827	477	25419
Telugu	12078	2795	5776	442	21091
Urdu	22990	2801	5786	443	34280

Table 1: Statistics of the synsets in IndoWordNet

4.1 Data

The data for all the 18 languages is stored in the file system. The organization of the data is described as follows:

- **Synsets.** All the synset data for each language in IndoWordNet is stored in a different file. In each file, on each line, there is a synset with its unique identifier, the synset, gloss, examples, and the part of speech of the synset. There are 4 more files for each language for each of the part of speech: noun, verb, adjective, and adverb. These files contain the synsets for the respective part of speech of the synset.
- **Words.** This type of file contains all the unique words available in the WordNet along with its synset unique identifier and part of speech tag. Similar to Synsets file, there is separate file for each language. Each such file contains all the words in the WordNet of the respective language. There are 4 more files for each language that has words for the respective part of speech of the synset.
- **Synset relations.** This type of file stores various lexico-semantic relations among the synsets. Since, the Indian language WordNets are based on Hindi WordNet, all the relations in Hindi WordNet are also valid for other language WordNets in the IndoWordNet.
- **Ontology nodes.** The next is ontology nodes file which contains a list of nodes that the

synset belongs to in an ontology tree (Fensel, 2001).

Apart from the textual data, we also specifically augment Hindi WordNet with speech data for all the words, glosses and example sentences. The speech data is in Waveform Audio File Format (WAV). This data is obtained using Indic TTS, a text-to-speech synthesis system for Indian languages (Patil et al., 2013).

Our system provides access to IWN data for the languages mentioned in Table 1. The number of synsets present in our database are also present in the table above. Figure 1 shows the architecture for our system.

4.2 Features

Our API provides access to the synset data and its lexico-semantic relations with other synsets for all the languages in IndoWordNet. The API module can imported in the following manner:

```
>>> from pyiwn import pyiwn
>>> iwn = pyiwn.IndoWordNet(lang)
```

The class *IndoWordNet* in the *pyiwn* module takes language (*lang*) as an argument. In this way, an object *iwn* is created to access the synset and speech data from WordNet of that language. The core features provided are described in the following sections.

4.2.1 Synsets

The API returns a *Synset* object for all the functions that are described ahead. The *Synset* object holds the information described in Section 4.1. The synsets can be accessed using in the following ways:

Access to all synsets

```
>>> iwn.all_synsets()
```

This function gives access to all the synsets for the given language.

```
>>> iwn.all_synsets(pos=pos_tag)
```

The above line signifies that the API will give all the synsets for a given language where the *pos_tag* can hold a string value from {noun, verb, adjective, adverb}.

Access to synsets of a given word

```
>>> iwn.synsets(word)
```

This function searches for all the synsets that contain the given *word* and returns a Python list of *Synset* objects that have all the properties of a synset described in the next section.

```
>>> iwn.synsets(word, pos=pos_tag)
```

Similarly, this function is used to filter the results for a given *word* by an optional second argument, *pos_tag*.

4.2.2 Synset properties

The synset has the properties like, head word (first word of the synset), POS tag, gloss (definition of the synset), examples, lemma names, ontology nodes and relations which is described in detail in Section 4.1. The API has a *Synset* class that has all of these mentioned properties as functions. The below code examples demonstrate the functions of the *Synset* class.

```
# creates a list of Synset objects for the given word
and returns the first Synset object
>>> syn = iwn.synsets(word)[0]
```

```
# returns part of speech tag
>>> syn.pos()
```

```
# returns head word
>>> syn.head_word()
```

```
# returns definition
>>> syn.gloss()
```

```
# returns a list of examples
>>> syn.examples()
```

```
# returns a list of ontology nodes
>>> syn.ontology_nodes()
```

```
# returns a list of lemmas
>>> syn.lemma_names()
```

```
# returns a dictionary of relations
>>> syn.relations()
```

4.2.3 Words

The API also provides functions to access only words of a particular language. The below code examples show the usage of this functionality.

```
# returns a list of all the words in the given
language
>>> iwn.all_words()
```

```
# returns a list of all the words in the given
language filtered by given an optional argument,
pos_tag
>>> iwn.all_words(pos=pos_tag)
```

4.2.4 Speech

The speech data for Hindi words can also be accessed via the API using the following function that takes a *word* as an argument and returns the WAV file object.¹²

```
# returns a WAV file object for a given word
>>> iwn.word_speech(word)
```

```
# For the speech of the glosses and exam-
ples, first create a list of Synset objects for the
given word and returns the first Synset object
>>> syn = iwn.synsets(word)[0]
```

```
# returns a WAV file object for a given gloss
>>> syn.gloss_speech(gloss)
```

```
# returns a list of WAV file objects for a given list
of examples
>>> syn.examples_speech(examples)
```

4.2.5 Morphological Analyzer

The role of morphological analyzers is to find the dictionary form of the word by restoring the changes caused by inflectional or derivational morphology of the language (*nationalism* → *nation*) (Buckwalter, 2002).

The API provides a function that takes in a word of a given language and returns the dictionary form of the word (lemma) which can then be passed on to other functions in the API.

```
# returns a lemma for the given word
>>> iwn.morph(word)
```

This functionality enables us to find the related synsets for many morphological variants of the

¹²<https://docs.python.org/2/library/wave.html>

words. This is really helpful in case of morphologically rich languages like Marathi. Currently, this feature is available only for the languages Hindi and Marathi. We plan to include the morphological analyzer for other languages in the future.

5 Conclusion & Future work

We provide an API for accessing IndoWordNet using Python. WordNet package in NLTK is already widely used amongst NLP researchers; we provide them with a similar functionality for Indian language WordNets and bundle the data along with. We also provide audio data in a separate package for the Hindi language. Currently, we only provide with functionalities such as displaying synset data, browsing through relational connections of a synset with other synsets, and access to speech data for the Hindi language. We believe our work will help the NLP community by providing them with a robust and easy-to-use framework for Indian languages.

In future, we plan to add functionalities like getting the top-level relational synset, the path-length of longest and shortest relational synset, finding all parent/child synsets with respect to a lexico-semantic relation. We also plan to create voice models using a speech synthesis system for other Indian languages or use pre-built voice models to generate audios for Indian languages and augment our API with the audio data. In addition to this, we aim to provide morphological analysis for more languages other than Hindi and Marathi as a feature for the ability to search concepts using the inflectional form of a word. We hope our work helps the Indian NLP diaspora further their research and gain more insights.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

P Bhattacharyya. 2010. Indowordnet. lexical resources engineering conference 2010 (Irec 2010). *Malta, May*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.

Tim Buckwalter. 2002. Buckwalter {Arabic} morphological analyzer version 1.0.

Rashmi Chauhan, Rayan Goudar, Robin Sharma, and Atul Chauhan. 2013. Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*, pages 397–402. IEEE.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Dieter Fensel. 2001. Ontologies. In *Ontologies*, pages 11–18. Springer.

Mark Alan Finlayson. 2014. Java libraries for accessing the princeton wordnet: Comparison and evaluation. In *Proceedings of the 7th Global Wordnet Conference, Tartu, Estonia*, volume 137.

Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. 1998. Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.

Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. Estnltk - nlp toolkit for estonian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Marius Pasca and Sanda Harabagiu. 2001. The informative role of wordnet in open-domain question answering. In *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*, pages 138–143.

- Hemant A Patil, Tanvina B Patel, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, GR Kasthuri, T Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, et al. 2013. A syllable-based framework for unit selection synthesis in 13 indian languages. In *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, pages 1–8. IEEE.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.
- Neha R Prabhugaonkar, A Nagvenkar, and R Karmali. 2012. Indowordnet application programming interfaces.
- Hanumant Redkar, Sudha Bhingardive, Kevin Patel, Pushpak Bhattacharyya, Neha Prabhugaonkar, Apurva Nagvenkar, and Ramdas Karmali. 2016. Wwds apis: application programming interfaces for efficient manipulation of world wordnet database structure. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Manish Sinha, Mahesh Reddy, and Pushpak Bhattacharyya. 2006. An approach towards construction and application of multilingual indo-wordnet. In *3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea*.
- Brett Spell. 2009. Java api for wordnet searching (jaws). URL <http://lyle.smu.edu/~tspell/jaws>.
- Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1312. Association for Computational Linguistics.
- Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.
- Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe, and Tissa Jayawardana. 2011. Towards a sinhala wordnet. In *Proceedings of the Conference on Human Language Technology for Development*.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.

Sinitic Wordnet: Laying the Groundwork with Chinese Varieties Written in Traditional Characters

Chih-Yao Lee

National Taiwan University
Taipei, Taiwan.
chihyaolee@gmail.com

Shu-Kai Hsieh

National Taiwan University
Taipei, Taiwan.
shukai@gmail.com

Abstract

The present work seeks to make the logographic nature of Chinese script a relevant research ground in wordnet studies. While wordnets are not so much about words as about the concepts represented in words, synset formation inevitably involves the use of orthographic and/or phonetic representations to serve as headword for a given concept. For wordnets of Chinese languages, if their synsets are mapped with each other, the connection from logographic forms to lexicalized concepts can be explored backwards to, for instance, help trace the development of cognates in different varieties of Chinese. The Sinitic Wordnet project is an attempt to construct such an integrated wordnet that aggregates three Chinese varieties that are widely spoken in Taiwan and all written in traditional Chinese characters.

1 Introduction

As with Romance languages descending from Classical Latin that stand on their own in present days, Sinitic languages¹, or major descendants of Archaic Chinese, have developed into fully-fledged languages without or with very limited mutual intelligibility (Tang and van Heuven, 2007). However, thanks to a shared logographic writing system that has not seen drastic changes in modern times², speakers of distinct Chinese languages can use a common set of logographic characters to communicate.

As language is not merely a vehicle for the expression of thought, but the way to thought itself,

¹The term “Sinitic” was chosen to suggest that the varieties of Chinese are distinct languages rather than different dialects of a same language.

²At least not until after the 1950s, when the Chinese Character Simplification Scheme was introduced in China.

writing systems not only represent a language, but reflect and record its ever-changing nature. As both linguists and wordnet builders, we see a great potential for wordnets to assist in lexical-semantic studies across Chinese languages, synchronic and diachronic alike, and serve as a handy repository where logograph-based searches are enabled.

In this paper, we present the initial version of a new resource named “Sinitic Wordnet”, which not only includes the lexicons of Mandarin, Southern-Min and Hakka, but makes use of Collaborative Interlingual Index to link them to other wordnet projects.

2 Methodology

In this section, we explain how synsets were organized based on the dictionaries and how they were interlinked afterwards.

2.1 Conversion of Individual Lexicons into Wordnets

We retrieved from the website of gov-zero³ machine-readable versions of Mandarin-to-Mandarin, Southern-Min-to-Mandarin, and Hakka-to-Mandarin dictionaries compiled by the Ministry of Education, Taiwan. The statistics of the three dictionaries are given in Table 1.

Dictionary Type	Entry Count
Mandarin-to-Mandarin	166,119
Southern-Min-to-Mandarin	20,377
Hakka-to-Mandarin	15,487

Table 1: Entry counts of the three dictionaries.

Assuming that (nearly) synonymous word senses were glossed largely the same (Sinha et al., 2006), we started by using sense glosses as the

³More commonly referred to as gov, gov-zero is a civic tech community that promotes the ideas of open government, open data, civic participation, and new media in Taiwan.

unique identifier for a synset entry. By means of comparing the similarities of sense definition between every two pairs of synsets, we were able to merge entries of synsets that are similar, if not identical, in meaning. After automated matching and merging, the resulting synsets were manually checked. Table 2 gives the numbers of synsets derived from each of the three dictionaries.

Dictionary Type	Synset Count
Mandarin-to-Mandarin	25,761
Southern-Min-to-Mandarin	3,158
Hakka-to-Mandarin	2,400

Table 2: Synset counts of the three lexicons.

2.2 Alignment of Individual Wordnets

To align synsets from the individual wordnets, we resorted to pattern-matching in three pieces of information that may (or may not) be included in an entry:

1. **Sense definition:** a gloss, (near-)synonym or translation equivalent (in the case of bilingual dictionaries). As we were able to compare the degree to which two sense definitions are alike to organize synsets within an individual wordnet, by the same token, we could map between synsets of different wordnets by computing their similarities based on synset glosses. Also, in Southern-Min-to-Mandarin and Hakka-to-Mandarin dictionaries, some of the definitions are not really glosses, but simply translation equivalents in Mandarin. We used those Mandarin equivalents as links to Southern-Min and Hakka. While this link is from sense to lemma rather than between senses, we selected the first and usually the most salient sense of the lemma to be the represented concept.
2. **Example words:** when the lemma of an entry has usages as bound-morpheme, there can be compound words to illustrate how the lemma combines with others. Along with such example words in the two bilingual dictionaries, there are usually Mandarin equivalents to Southern-Min and Hakka, respectively. Again, by choosing the first sense of the translation in Mandarin, we were able to establish links between the three languages.

3. **Multilingual translation:** in a separate section of the dictionaries, there are translations to European languages (including French, German and Spanish) as well as between the three Chinese varieties. Once again, the first sense of the translation words was used to connect the three lexicons.

2.3 Mapping with Princeton WordNet

In order to facilitate integration with other resources as well as enable queries in English, Sinitic Wordnet has its synsets mapped with those of Princeton WordNet (Fellbaum, 2010). The mapping was done by bilinguals of English and one of the Chinese varieties.

3 Results

In this section, we show how it is possible to track in Sinitic Wordnet concepts that are encoded in different logographs, and vice versa. Also, an entry is given the Turtle format for the sake of illustration.

3.1 Sinitic Wordnet as Bridge between Concepts, Synsets and Logographs

When converting dictionaries into wordnets, we focused on concepts as expressed by sense glosses written in Mandarin, grouped them into synsets according to how similar their describing texts were, and mapped them with counterparts in Princeton WordNet. Now that the foregoing steps have been completed, the entire procedure can be examined in the opposite direction to help discover lexicalization patterns as well as to observe the way word senses distribute from variety to variety and determine whether new ones have developed in one language, or whether old ones have ceased to exist in another.

Take for example the concept POT. If one is curious about how this idea is encoded in traditional Chinese characters across different varieties, they can run a query using English words (e.g. *pot*) or phrases (e.g. *cooking vessel*) that may express the concept. As shown in Figure 1, a query of “pot” would lead to the English synset {pot}, which is in turn linked with equivalents from each of the Chinese lexicons. To encode the concept POT, Mandarin chooses ‘鍋’ (guō) over ‘鼎’ (tiǎn) and ‘鑊’ (vók), which are respectively adopted in Southern-Min and Hakka.

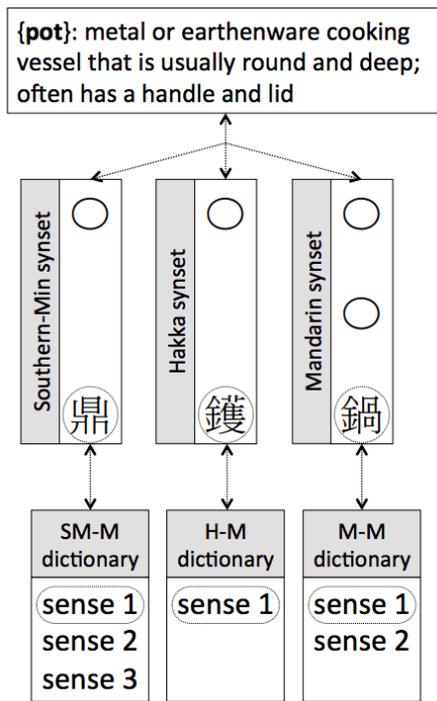


Figure 1: From a concept to synsets, from a synset to lemmas as represented by different holographs.

Reversely, it is equally possible to look at what distinct meanings a single logograph carries in different varieties, as illustrated in Figure 2, where the logograph ‘鼎’ (tiǎn) is taken as query to look for synsets whose consisting members are represented by the same character.

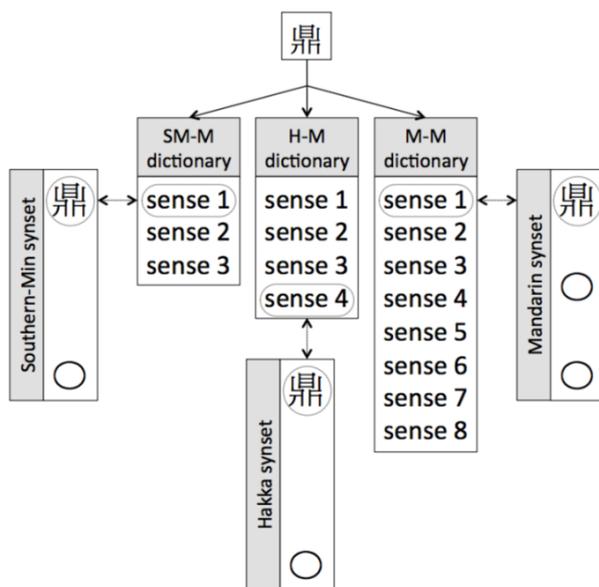


Figure 2: From a holograph to synsets in different lexicons.

3.2 Sinitic Wordnet as Linked Data

To improve its interoperability with other lexical resources, Sinitic Wordnet has been converted in RDF format using the *lemon* model (McCrae et al., 2011; McCrae et al., 2012). Figure 3 shows what a *lemonized* sense looks like in Turtle format⁴.

```
@prefix owl: <http://www.w3.org/2002/07/
  owl#> .
@prefix rdf: <http://www.w3.org
  /1999/02/22-rdf-syntax-ns#> .
@prefix lemon: <http://www.lemon-model.
  net/lemon#> .
@prefix wordnet-ontology: <http://
  wordnet-rdf.princeton.edu/
  ontology#> .
<http://lope.linguistics.ntu.edu.tw/swn/
  mandarin/dong4wu4/052268> a lemon
  :LexicalEntry ;
  lemon:canonicalForm <#CanonicalForm>
  ;
  lemon:sense <#1> ;
  wordnet-ontology:part_of_speech
  wordnet-ontology:noun .
<#CanonicalForm> a lemon:Form ;
  lemon:writtenRep @cmn .
<#1> a lemon:LexicalSense ;
  lemon:reference <http://lope.
  linguistics.ntu.edu.tw/swn/
  mandarin/2068> ;
  wordnet-ontology:gloss
  ;
  @cmn ;
  owl:sameAs <http://wordnet-rdf.
  princeton.edu/wn31/100015568-
  n> .
```

Figure 3: The first sense of *dong4wu4* in Turtle.

4 Publishing the Resource

Once the wordnets and their mappings derived from this project are made more tidy, we will release the data under an open license in order to ensure that it can be put into use as widely as possible. Before that, we have made the resource available by integrating it with two best practices in the WordNet community, namely with the Linguistic Linked Open Data Cloud and the Collaborative Interlingual Index.⁵

4.1 Publishing the Resource as Linked Data

By way of synset mapping, Sinitic Wordnet not only has its consisting lexicons interlinked, but also links directly to Princeton Wordnet. As shown in Figure 3, there is an outward link to Princeton WordNet because the synset referenced

⁴<http://www.w3.org/TR/turtle/>

⁵<http://lope.linguistics.ntu.edu.tw/swn>

to by the lexical sense has an equivalent in English. Meanwhile, the links to WordNet serve as key to the Linguistic Linked Open Data cloud (Chiarcos et al., 2013) and interface with other linguistic resources. Moreover, Sinitic Wordnet can be integrated into the Global WordNet Grid when organized by the ontology consisting of 71 Base Types proposed by the Global WordNet Association.⁶ An initial mapping has identified 169 synsets comparable to the Base Types.⁷

4.2 Integrating the Resource with Collaborative Interlingual Index

The Collaborative Interlingual Index (Bond et al., 2016) has been proposed as a method to enable cross-lingual development of wordnets. Chief among the primary objectives of the project is to establish a standard operating procedure by which new synsets can be defined and added to a common repository, resolving compatibility issues that may occur when wordnets for languages other than English introduce concept not lexicalized in English. In order to facilitate the integration of Sinitic Wordnet with the Collaborative Interlingual Index, we are making the full version of the resource available in the Global WordNet Association's recommended formats and release under an open license.

5 Conclusion

Based on monolingual (Mandarin to Mandarin) and bilingual (Southern-Min/Hakka to Mandarin) dictionaries by the Ministry of Education, Taiwan, we have presented a method for developing an integrated wordnet that includes and interlinks the lexicons of Mandarin, Southern-Min and Hakka. The resource was generated semi-automatically, relying on bilinguals of the Chinese varieties to map synsets with Princeton WordNet. To align the synsets, a mixture of methods was employed, including looking for synonyms in sense definitions, and translation equivalents in example words as well as in a section of the dictionaries that gives translation words in European and Chinese languages.

In addition to more thorough check-ups upon the integrity and quality of existing lexicons, our plans for future development include the addition

of Cantonese as spoken in Hong Kong, another Chinese variety that is also written in traditional Chinese characters, and the construction of a web-based graphical user interface for public access of Sinitic Wordnet.

References

- Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index.
- Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications - Volume Part I, ESWC'11*, pages 245–259, Berlin, Heidelberg. Springer-Verlag.
- John McCrae, Guadalupe Aguado-de Cea, Paul Buiteelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.
- Manish Sinha, Mahesh Reddy, and Pushpak Bhat-tacharyya. 2006. An approach towards construction and application of multilingual indo-wordnet. In *3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea*.
- Chaoju Tang and Vincent J van Heuven. 2007. Mutual intelligibility and similarity of chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*, 24(1):223–234.

⁶http://w.globalwordnet.org/gwa/ewn_to_bc/BaseTypes.htm

⁷<http://lope.linguistics.ntu.edu.tw/swn/gwn/>

Synthesizing Audio for Hindi WordNet

Diptesh Kanojia^{†,♣,*}, Preethi Jyothi[†], Pushpak Bhattacharyya[†]

[†]Indian Institute of Technology Bombay, India

[♣]IITB-Monash Research Academy, India

^{*}Monash University, Australia

[†]{diptesh, pjyothi, pb}@cse.iitb.ac.in

Abstract

In this paper, we describe our work on the creation of a voice model using a speech synthesis system for the Hindi Language. We use pre-existing “voices”, use publicly available speech corpora to create a “voice” using the Festival Speech Synthesis System (Black, 1997).

Our contribution is two-fold: **(1)** We scrutinize multiple speech synthesis systems and provide an extensive report on the currently available state-of-the-art systems. We also develop voices using the existing implementations of the aforementioned systems, and **(2)** We use these voices to generate sample audios for randomly chosen words; manually evaluate the audio generated, and produce audio for all WordNet words using the winner voice model. We also produce audios for the Hindi WordNet Glosses and Example sentences.

We describe our efforts to use pre-existing implementations for WaveNet - a model to generate raw audio using neural nets (Oord et al., 2016) and generate speech for Hindi. Our lexicographers perform a manual evaluation of the audio generated using multiple voices. A qualitative and quantitative analysis reveals that the voice model generated by us performs the best with an accuracy of 0.44.

1 Introduction

WordNets have proven to be a rich lexical resource for many NLP sub-tasks such as Machine Translation (MT) and Cross-Lingual In-

formation retrieval (Knight and Luk, 1994; Richardson and Smeaton, 1995). They are lexical structures composed of synsets and semantic relations (Fellbaum, 1998). Such a lexical knowledge base is at the heart of an intelligent information processing system for Natural Language Processing and Understanding. The first WordNet was built in English at Princeton University¹. Then, followed the WordNets for European Languages² (Vossen, 1998), and then IndoWordNet³ (Bhattacharyya, 2010).

IndoWordNet consists of 18 Indian Languages with an average of 27000+ synsets for all the languages and 40000+ for the Hindi Language. It uses Hindi WordNet⁴ (Narayan et al., 2002) as a pivot to link all these languages and contains more than 25000 linkages to the Princeton WordNet. Cognitive theories of multimedia learning (Mayer, 2002) indicate that audio cues are effective aids in a learning scenario, and also help in retaining the material learned (Bajaj et al., 2015).

“Our goal is to enrich the semantic lexicon of Hindi WordNet by augmenting it with word audios generated automatically using a speech synthesis voice model.”

Manually recording pronunciations for all the words is a tedious task. These recording efforts could be minimized by using text-to-speech (TTS) systems to automatically synthesize speech for all the words. However, one cannot be sure about the quality of these synthesized clips. We build multiple TTS systems and systematically analyze the quality of the resulting synthesized clips, with the help of

¹<http://wordnet.princeton.edu>

²<http://www.illc.uva.nl/EuroWordNet/>

³<http://www.cfilt.iitb.ac.in/indowordnet/>

⁴<http://www.cfilt.iitb.ac.in/wordnet/webhwn/>

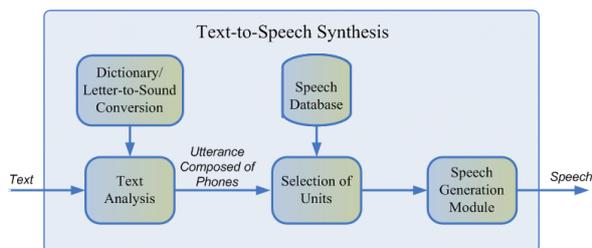


Figure 1: A Unit selection based Concatenative Speech Synthesis System

lexicographers. We envision that this addition to Hindi WordNet will further its use in the education domain, for students and language enthusiasts alike.

1.1 Speech Synthesis: An Introduction

There are four basic approaches to synthesizing speech: 1) waveform concatenation, 2) articulatory synthesis, 3) formant synthesis, and 4) concatenative synthesis. Concatenative synthesis produces a very natural-sounding synthesized version of the utterances. There can be glitches in the output owing to the nature of automatic segmentation of the waveforms, but the speech produced sounds natural indeed. Apart from the first method *i.e.* waveform concatenation, all approaches to speech synthesis are based on the source-filter model. The synthesis method can be broken down into two components, consisting of a model of the source (models of periodic vibration and models of noise supra-glottal sources) and a model of the vocal tract transfer function. In articulatory synthesis, computational models of the articulators are constructed that allow the system to simulate various configurations that human speech organs can attain during speech production. Acoustic-phonetic theory is used to compute the transfer function for vocal tract shape. In formant synthesis, formant transitions across consonants and vowels must be modeled closely. These transitions are most important in identifying the consonant. Designing these set of rules is still a difficult task. The simplest approach to synthesis bypasses most of the problems since it involves taking real recorded/coded speech, cutting it into segments, and concatenating these segments back together during synthesis. It is called concatenative synthesis.

2 Related Work

A significant amount of work has been done in the area of Speech Synthesis or Text-to-Speech conversion for English, Japanese, Chinese, Russian (Takano et al., 2001; Zen et al., 2007; Zen et al., 2009; Wang et al., 2000; Sproat, 1996). Text-to-Speech conversion systems for Indian Languages have also emerged in the recent past (Patil et al., 2013). Although these systems, which are already available, do not produce the most “natural” sounding output, but they are usable to an extent. Manual evaluations of the speech synthesis systems built for the Hindi Language show that there is still a need for better text processing and additional phonetic coverage (Kishore et al., 2003; Raj et al., 2007). Bengu et al. (2002) create an online context sensitive dictionary using Princeton WordNet and implement a Java based speech interface for the Text-to-Speech (TTS) engine. Kanojia et al. (2016) automatically collect images for IndoWordNet and augment them to the web interface, but due to the lack of tagged images openly available for use, they do not collect enough images. To the best of our knowledge, there has been no other work specifically in the direction of synthesizing audio for WordNet words or Synthesizing audio for Indian Language WordNets.

3 Our Approach

Among the three main sub-types of concatenative synthesis, we choose to perform unit selection synthesis and build cluster units of the speech data recorded by a human voice. We use the Festival system to create a synthetic voice for Hindi. We followed the documentation of the Festival Framework along with FestVox⁵ implementation to train a voice on Hindi Speech Corpora provided by the IndicTTS Consortium⁶ for research purposes. Figure 1 displays a generic speech synthesis system which uses the concatenative synthesis or unit selection corpus-based speech synthesis to generate speech given an input text.

⁵<http://festvox.org/>

⁶<https://www.iitm.ac.in/donlab/tts/index.php>

3.1 Dataset

We use the Female Voice - Hindi and Female Voice - English dataset provided by the IndicTTS forum to train our system. The dataset is publicly available for the purpose of research. We download the complete dataset i.e. 7.22 hours of Audio with English and 5.18 hours of monolingual audio. We also download the dictionary provided on the website for providing it to the synthesis system as input. We use a total of 2318 Female Hindi sentence utterances downloaded from IndicTTS consortium, and 1378 word audios manually recorded by us to train the voice model.

3.2 Architecture and Methodology

While training input to the system is a corresponding speech-text parallel corpus, where a WAV file containing audio is aligned to its corresponding text using an ID, a textual unit such as a word or a phrase is given as an input in the testing phase. The output is an audio waveform stored in the WAV format.

The system needs a syllable dictionary for a letter to sound conversion and We generate one which contains unique words and in parallel has corresponding syllabification of the word with the beginning and ending clearly marked. For *e.g.*, The Hindi word “*kamaane*” which means “To Earn” would be represented as:

(“कमाने” nil (((“क_beg”) 0) ((“मा_mid”) 1) ((“ने_end”) 0))),

0 for lower stress, and 1 for the high stress.

Such a system also requires the utterances composed of phones including a considerable amount⁷ of recorded speech.

Both the requirements above can also be generated programmatically from a given text corpora which corresponds to a speech database/corpora. Although, the recorded speech needs to be in parallel correspondence to the recorded audio files (usually in a WAV audio file format - 16KHz, Mono Channel).

⁷conventionally, hours of speech is required

3.3 Implementation Details

We implement the Unit Selection based method for generating audio and try to use a pre-implemented neural network based method to generate audio. Although Hindi audio could not be generated using the latter, but we successfully generate audio using the former method.

3.3.1 Unit Selection based Concatenative Synthesis

We perform Unit selection synthesis using a large corpus of recorded speech labeled with the text being spoken (Details of the corpus in implementation details). During such a corpus creation, each recorded utterance is segmented into some or all of the following: a) individual phones, b) diphones, c) half-phones, d) syllables, e) morphemes, f) words, g) phrases, and h) sentences.

A specially modified speech recognizer set to a “forced alignment” mode with some manual correction is typically used to divide the speech corpus into segments (utterances). It uses visual representations such as the waveform and spectrogram, to divide the speech. An index of such units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At run time, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree (HTS System uses HMM and looks at posterior probability and prior probability to decide the best chain.)

Since the output of our work would be used to generate pronunciations of a word/phrase/short sentences, we need a natural sounding voice and hence choose to build cluster units of the recorded speech data available.

3.3.2 Neural Network based RAW Audio generation

We also use pre-implemented models from around the web to reproduce TTS systems, but as quoted at many places, such systems require huge amounts of data and exorbitant amounts of time to generate even smallest of

the samples. We use a WaveNet implementation (basveeling/wavenet)⁸ to generate RAW audio for a piano music dataset and generate audio using it.

Due to various errors in the implementation when trying to use it to generate audio based on text, we could not use this implementation for any form of Text-to-Speech generation.

3.3.3 Other Experiments

We also use other pre-trained voices available on the FestVox website to generate audio for comparison with the audio generated via our voice model. We downloaded the following voices:

1. Hindi - Male Voice,
2. Hindi - Female Voice,
3. Marathi - Female, and
4. Marathi - Male.

We generate audio using these voices. A brief record of our survey of various speech synthesis systems available is provided in Table 1. We also use the default Festival diphone-based voice for Hindi provided with the system for comparison. We also survey the other potential speech synthesis frameworks and list them in the table for reference.

Technique	Explored	Voice Models Generated	Usable for Hindi TTS
<i>Festival+FestVox (IndicTTS Data)</i>	<i>Yes</i>	<i>Many</i>	<i>Yes</i>
Flite Voice (Hindi - Female)	Yes	1	Yes
Flite Voice (Hindi - Male)	Yes	0	Yes
Flite Voice (Marathi - Female)	Yes	0	Yes
Flite Voice (Marathi - Male)	Yes	0	Yes
Festival (diphone)	Yes	1	Yes
Wavenet (basveeling)	Yes	1	No
DeepVoice	Yes	0	No
Merlin	No	0	No
MaryTTS	No	0	No
Tacotron	No	0	No
SampleRNN	No	0	No
Char2Voice	No	0	No

Table 1: Our tryst with Speech Synthesis- An overall picture of the area explored

⁸<https://github.com/basveeling/wavenet>

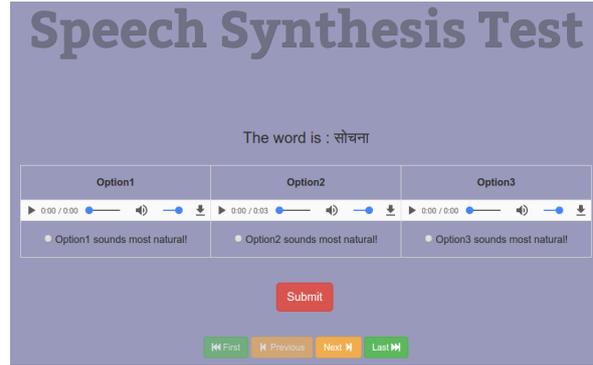


Figure 2: A Unit selection based Concatenative Speech Synthesis System

4 Results & Evaluation

We accumulate 6 usable voice models and produce word audios and randomly sample word audios from them. Among these models, the one which we successfully generated using Unit Selection based Concatenative Synthesis, sounded most natural in a brief overview.

Speech Synthesis evaluation is a subjective issue. Different speech voices are used to train various speech systems, and no agreed upon metric for the quality of such an output has been produced, yet. Quality of production technique is another factor on which speech synthesis depends, and hence the evaluation of speech synthesis systems has been compromised by differences between such factors (production techniques, recording facilities etc.)

Speech Synthesis systems require human annotators for evaluation of their output. The annotation is done based on naturalness and intelligibility of the output. A recent work proposes a novel approach that formulates objective intelligibility assessment as an utterance verification problem using hidden Markov models, thereby alleviating the need for human reference speech (Ullmann et al., 2015). Although nothing exists to assess the naturalness of a speech synthesis output.

We generate word audios for approximately 4000 words using four best voice models. For evaluation of our synthesized data, we create an experiment vaguely based on Turing Test. We randomly choose 30 Hindi Words and also get audio recorded for them with the help of our lexicographers.

We create a PHP-MySQL based web-

	#0	#1	#2	#1+#2	Most Liked
Model 1	79	55	99	154	101
Model 2	37	78	112	190	90
Model 3	72	86	58	144	51
Model 4	55	117	107	224	70

Table 2: Results of manual evaluation of synthesized speech clips

interface show as a screenshot in Figure 2 and crowd-source results. The interface shows a user, three different audio samples, and they were asked to choose the “Most Natural” audio from among them.

We receive a total of 442 responses for 30 word samples. Thus, we assume that 14 people had completed the test. The results of our initial evaluation based on naturalness are as follows: (i) **The mean of our voice model win percentage is over 44%**. We beat both the other voices by an acceptable margin, (ii) **Pre-recorded speech by humans was rated best somewhat less than 30%** of the times, and (iii) Grapheme based synthesized speech scored around **26% on this scale**.

We randomly chose 535 words and generate synthesized outputs from four best models; these outputs were presented to two lexicographers for further analysis. They used the following scale to report the output (i) **unusable (#0)**: This rating corresponds to audio clips which are either distorted, or too noisy for the user to comprehend, (ii) **usable (#1)**: This rating corresponds to audio clips which are moderately usable and suggests that the user can comprehend the underlying words. However, audio clips with this rating can be synthesized better, (iii) **good (#2)**: This rating corresponds to audio clips that are really good and convey the words. For each of the 535 words, the lexicographers were also asked to mark which of the four synthesized clips they liked the most.

The evaluation results are shown in Table 2 which clearly show that **Model 1** was marked as the most liked audio clip most often, while **Model 4** performed the best in terms of producing the most number of usable audio clips (obtained by summing clips with ratings #1 and #2).

A qualitative analysis of the synthesized clips highlighted the following issues, partic-

ularly with respect to the clips that were marked “unusable”: i) Flap or tap sounds (ड, ढ) were pronounced incorrectly, ii) Intonation of the audio for heavy syllables was at times incorrectly rendered and for words such as ‘एकदम’, the pronunciation had a specific stress pattern which should have ideally been neutral, thus making it sound unnatural, iii) There were also a few examples of unnecessary lengthening of a vowel. For example, in बीमारी (*beemari*, sickness), there was unnecessary stress on ‘बी’ and hence it was lengthened, iv) Incorrect syllable breaks were observed in some words. For example, नापसंद (*naapasand*, non-favourite), was pronounced as नाप-संद, which is incorrect, v) It was also noted that sometimes consonant clusters were mispronounced. E.g. कुत्ता - (*kutta*) - dog, was incorrectly pronounced as कु-ता or कुत-ता.

Eventually, we employ the best voice model for generating word, gloss, and example audios. **We generated, using Unit Selection based Concatenative Synthesis, audios for 151831 words, and 40337 synset glosses/example sentence.**

5 Conclusion and Future Work

We present our work on generating voice models using the Festival speech synthesis system. We also describe our efforts to use deep learning based implementations for generating such a model. A survey of the current state-of-the-art techniques available for speech synthesis was also done. We download pre-generated voice models available for Hindi and provide a detailed qualitative and quantitative analysis by comparing them with the voice model generated by us. We evaluate our model via crowd-sourcing and select the best voice model to generate audios for all words, glosses and example sentences for Hindi WordNet. We believe our work will help students and language learners understand the Hindi language, and help them pronounce it as well.

In future, we plan to improve the voice model by analyzing the speech output and incorporate more data for training. We also plan to implement WaveNet and other such neural network based techniques for raw audio generation and training models to produce speech for a given text.

References

- Jatin Bajaj, Akash Harlalka, Ankit Kumar, Ravi Mokashi Punekar, Keyur Sorathia, Om Deshmukh, and Kuldeep Yadav. 2015. Audio cues: Can sound be worth a hundred words? In *International Conference on Learning and Collaboration Technologies*, pages 14–23. Springer.
- G. Bengu, Guyangu Liu, Ritesh Adval, and Frank Shih. 2002. Educational application of an online context sensitive dictionary. *The 17th International Symposium on Computer and Information Sciences*.
- P Bhattacharyya. 2010. Indowordnet. lexical resources engineering conference 2010 (lrec 2010). *Malta, May*.
- Alan Black. 1997. The festival speech synthesis system: System documentation (1.1. 1). *Technical Report HCRC/TR-83*.
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Diptesh Kanojia, Shehzaad Dhuliawala, and Pushpak Bhattacharyya. 2016. A picture is worth a thousand words: Using openclipart library for enriching indowordnet. In *Eighth Global WordNet Conference*. GWC 2016.
- SP Kishore, Alan W Black, Rohit Kumar, and Rameesh Sangal. 2003. Experiments with unit selection speech databases for indian languages. *Carnegie Mellon University*.
- Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.
- Richard E Mayer. 2002. Multimedia learning. *Psychology of learning and motivation*, 41:85–139.
- Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet—a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
- Hemant A Patil, Tanvina B Patel, Nirmesh J Shah, Hardik B Sailor, Raghava Krishnan, GR Kasthuri, T Nagarajan, Lilly Christina, Naresh Kumar, Veera Raghavendra, et al. 2013. A syllable-based framework for unit selection synthesis in 13 indian languages. In *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, pages 1–8. IEEE.
- Anand Arokia Raj, Tanuja Sarkar, Sathish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad, and Alan W Black. 2007. Text processing for text-to-speech systems in indian languages. In *SSW*, pages 188–193.
- Ray Richardson and Alan F Smeaton. 1995. Using wordnet in a knowledge-based approach to information retrieval.
- Richard Sproat. 1996. Multilingual text analysis for text-to-speech synthesis. *Natural Language Engineering*, 2(4):369–380.
- Satoshi Takano, Kimihito Tanaka, Hideyuki Mizuno, Masanobu Abe, and S Nakajima. 2001. A japanese tts system based on multiform units and a speech modification algorithm with harmonics reconstruction. *IEEE Transactions on Speech and Audio Processing*, 9(1):3–10.
- Raphael Ullmann, Ramya Rasipuram, Hervé Boulard, et al. 2015. Objective intelligibility assessment of text-to-speech systems through utterance verification. In *Proceedings of Interspeech*, number EPFL-CONF-209096.
- Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.
- Ren-Hua Wang, Zhongke Ma, Wei Li, and Donglai Zhu. 2000. A corpus-based chinese speech synthesis with contextual dependent unit selection. In *Sixth International Conference on Spoken Language Processing*.
- Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. 2007. The hmm-based speech synthesis system (hts) version 2.0. In *SSW*, pages 294–299.
- Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Toward Constructing the National Cancer Institute Thesaurus Derived WordNet (ncitWN)

Amanda Hicks
University of Florida
Florida, USA
aehicks@ufl.edu

Selja Seppälä
University College Cork
Cork, Ireland
selja.seppala@ucc.ie

Francis Bond
Nanyang Technological University
Singapore
bond@ieee.org

Abstract

We describe preliminary work in the creation of the first specialized vocabulary to be integrated into the Open Multilingual Wordnet (OMW). The NCIt Derived WordNet (ncitWN) is based on the National Cancer Institute Thesaurus (NCIt), a controlled biomedical terminology that includes formal class restrictions and English definitions developed by groups of clinicians and terminologists. The ncitWN is created by converting the NCIt to the WordNet Lexical Markup Framework and adding semantic types. We report the development of a prototype ncitWN and first steps towards integrating it into the OMW.

1 Introduction

The Global Wordnet Grid (GWG) is a platform created to join together wordnets by linking them to a central registry of concepts, using the Collaborative Interlingual Index (CILI) as a pivot. Data in the GWG is linked following an ‘onion model’, with ‘a core of concepts shared by many wordnets’, validated by the community, and axiomatized through ontologies. The core extends to a middle layer with fewer shared wordnets and out to a layer of concepts mapped to only a single wordnet. An external layer contains synsets defined in project wordnets that do not fulfill the CILI inclusion criteria. One of the advantages of the GWG is that the resource is no longer limited to networks of single-word units, but is now open to phrasenets (frequent adjective-noun, noun-prep-noun, and verb-object combinations, as well as proverbs, idioms, and compounds). This feature creates the possibility to link wordnets to domain-specific terminologies, which often include multi-word expressions. The Open Multilingual Wordnet (OMW) is the reference instantiation of the

GWG (Bond et al., 2016) adding the constraint that all member wordnets must be open according to the open definition.¹

To date no specialized terminologies have been included in the OMW. Consequently, there is no established procedure for mapping technical concepts to the CILI nor for determining whether a technical concept ought to be indexed in the CILI. We report a preliminary biomedical wordnet based on the National Cancer Institute Thesaurus (NCIt) called the NCIt Derived Wordnet (ncitWN) and preliminary mappings to the CILI. By mapping the NCIt to the CILI and thereby integrating it into the OMW, we are developing the first specialized vocabulary mapped to the CILI. The two outcomes will be: (i) the NCIt mapped to the CILI and integrated into OMW, but just as significantly (ii) groundwork for a method to reliably integrate open and freely available specialized terminologies with these lexical resources. This work is a first step toward realizing the goals outlined in Smith and Fellbaum (2004).

2 Resources

2.1 The Collaborative Interlingual Index

The CILI is implemented as a collaborative open-source software based on the best-practices of the Semantic Web – persistent IDs, Creative Commons Attribution 4.0 (CC BY) license allowing redistribution, and versioning (Bond et al., 2016). It integrates and extends the list of concepts in the OMW, including all concepts in Princeton WordNet of English (PWN) (Fellbaum, 1998). Each concept in the CILI is described with a unique definition in English. Currently, most of these definitions are derived from PWN 3.0. The CILI is compatible with the two schemas (Wordnet-LMF/lemon) (Vossen et al., 2016; Mc-

¹The Open Definition, <http://opendefinition.org> (October 28, 2017).

Crae et al., 2014) used for encoding individual wordnets. The Semantic Web identifiers conform to the standards being adopted for encoding and integrating biomedical terminologies and ontologies (Ruttenberg et al., 2007; Schuurman and Leszczynski, 2008) and allow the CILI to be linked to ontologies and domain-specific terminologies. The CILI's open collaborative framework includes rules, tools, and safeguards to support high quality, agreed-upon mappings of wordnets to the CILI (Bond et al., 2016).

2.2 Princeton Wordnet

In order to get lemmas and domain information for English, we use the Princeton WordNet of English 3.0 (Fellbaum, 1998). Synsets are grouped into 45 **lexicographer files** which we use as coarse domains (for example, `noun.artifact` contains nouns denoting man-made objects). PWN also has explicit domains linked by the `domain-category` relation, which we intend to use in future work.

2.3 The National Cancer Institute Thesaurus and the UMLS Metathesaurus

The NCIt is a reference terminology developed by the National Cancer Institute that covers over 118,000 concepts and is available in the Web Ontology Language (OWL) (Sioutos et al., 2007). Although initially developed to support research and data management in the domain of cancer, it also includes concepts of general biomedical interest that are not specific to cancer, such as a robust typology of diseases, procedures, and adverse events. Each concept in the NCIt is associated with a unique identifier, a preferred term, and synonyms. Many terms also include an English definition, a description logic definition, and cross-references to other terminologies. The English definitions are developed by groups of clinicians and terminologists. The clinicians are often from the international English speaking community (USA, UK, Australia). The NCIt is released under a special license. We communicated with the creators and maintainers and have ensured that the `ncitWN` and its inclusion in the GWG is in compliance with their license.² Terms are also classified using the Unified Medical Language System (UMLS) Semantic Types: there are 127 semantic types linked in an is-a hierarchy.

²NCI THESAURUS Terms of Use, https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/NCI_THESAURUS_license.txt (October 28, 2017).

The NCIt is included in the Unified Medical Language System Metathesaurus, a biomedical thesaurus that links approximately 200 biomedical terminologies to an index of concepts (Schuyler et al., 1993). In this respect, the UMLS Metathesaurus can be viewed as a domain specific analogue of the Open Multilingual Wordnet (OMW). The UMLS Metathesaurus also includes translations of some of its source vocabularies into languages other than English. It is available in two data formats, the Rich Release Format and the Original Release Format. Semantic types such as “Drug” have been added to the UMLS Metathesaurus to impose more structure and to organize concepts (National Library of Medicine, 2009).

2.4 Wordnet-Lexical Markup Framework

Wordnet-Lexical Markup Framework is a wordnet-implementation of the Lexical Markup Framework (Francopoulo et al., 2006) (LMF), an ISO standard for NLP lexicons and Machine Readable Dictionaries based on the eXtensible Markup Language (XML) format. It encodes linguistic knowledge of the lexicalized concepts represented in the wordnets and supports integration of wordnets with OMW (Morgado da Costa and Bond, 2015; Vossen et al., 2013; Bond and Foster, 2013). Although no domain specific resources have been integrated into the CILI to date, this schema is well suited for the integration of an external resource such as the NCIt. Wordnet-LMF allows for a greater inventory of semantic relations than the NCIt currently contains, including entailment, part-whole relations, and derivations.

3 Methods

3.1 Convert NCIt to Wordnet-LMF

We have written a simple program (in Python 3) to reformat the NCIt as a wordnet (`ncitWN`). It filters out obsolete and retired concepts, creates the necessary metadata, and builds a wordnet. The conversion process is based on a few assumptions, to be tested further: (1) all concepts are lexicalized as nouns, and (2) the child-parent relationship in the thesaurus can be modeled as simple hypernymy.

The UMLS Semantic Types could be modeled as external links or as links within the wordnet (as PWN does). Currently we add them as metadata on each synset (using `dc:type`).

We validate the `ncitWN` data format with (1) the LMF Document Type Definition, which validates

the XML representation of the Wordnet-LMF documents (Vossen et al., 2016) and (2) the OMW’s online tool (Morgado da Costa and Bond, 2015; Tan and Bond, 2011) that detects content violations such as duplicate or missing definitions.

3.2 Map ncitWN to the CILI

We have tested the feasibility of mapping the ncitWN to the CILI using two approaches.

The first, automatic, approach uses the prototype Wordnet-LMF formatted version of NCIt to automatically generate candidate mappings to the CILI using lemma overlap and compatibility of UMLS Semantic Types with WordNet coarse domains. The score is the sum of the Jaccard similarity calculated over lemma overlap with a boost of 0.1 each time there is a match between the wordnet coarse domains and the UMLS Semantic Type, based on a simple table of equivalences.

For example consider the following match:

- *mask* (NCIt) “A protective covering worn over the face, or an apparatus for administering anesthesia or oxygen through the nose or mouth” «Manufactured Object» (C86570)
- *mask_{n:4}* (PWN) “a protective covering worn over the face” «noun.artifact» (i56041)

Here the overlap in lemmas is 100% (*{mask}* vs. *{mask}*) and «Manufactured Object» matches «noun.artifact» so the score is 1.1. The equivalence table was made by first matching only lemmas and, assuming that all 100% matches were good, linking the UMLS Semantic Type and PWN coarse domain. All matches of semantic types with more than 500 examples were taken to be good. An inspection of the less frequent matches showed many to be good, this mapping should be revised in subsequent work.

We manually evaluated a sample of the automatically produced matches with a match score > .75. The annotation scheme is summarized in Table 1. ‘0’ is not used for mapping, but was nevertheless used to annotate candidate matches. These annotations will be used to generate heuristics for refining match scores, thereby expediting the mapping process.

Note that an annotation of ‘0’ does not indicate that there is no relation between the NCIt and PWN term, but only that there is no hierarchical relation. There might be non-hierarchical relations, e.g., lin-

Annotation	Meaning
eq	equivalence
spec	hyponym of
gen	hypernym of
0	no hierarchy relation

Table 1: Annotations for candidate matches from ncitWN synset to PWN synset

guistically derived from, that may be incorporated in future work.

The second approach was a manual analysis of PWN 3.0’s coverage of the NCIt. We randomly selected 94 concepts from the NCIt, stratified according to whether the concept was a root, middle, or leaf node (respectively, 19, 37, and 38 concepts). We then searched for candidate mappings through lemmas in PWN and evaluated the match based on the corresponding definitions in the CILI. The manual coverage analysis was based on NCIt preferred terms and excludes synonyms. Preferred terms that take the form of boolean expressions such as ‘Diagnostic or Prognostic Factor’ were decomposed into their component expressions, which were used for searching candidate mappings. Thus, for ‘Diagnostic or Prognostic Factor’, we restored the elliptical noun and obtained two multiword expressions (MWEs) for which we searched candidate mappings, i.e., ‘Diagnostic Factor’ and ‘Prognostic Factor’.

We distinguish six matching scenarios summarized in Table 2 and illustrate them with examples below.

Annotation	Meaning
0	no match
1	exact match
2	full match
3	partial match of MWE
4	preferred term with partial match
5	not suitable to map to CILI

Table 2: Annotation scheme for candidate matches from NCIt terms to PWN synsets

The coverage analysis was carried out in several steps (see Figure 1). In step 1, we determined whether the NCIt preferred term had a match in the PWN lemmas. If it did not, we annotated it with ‘0’. If there was a match, in step 2, we compared the NCIt and CILI definitions. If both the lem-

mas and the definitions matched, we considered them an exact match ('1'); if the lemmas matched but the NCI definition was either more specific or broader than the CILI definition, the NCI preferred term has a partial map ('4'); if the NCI term and definition were NCI-specific, the concept was not suitable to be mapped to the CILI ('5'). If none of these options applied and the NCI term was an MWE, in step 4, we decomposed the MWE into its parts and searched each word individually. In case of a match, we determined whether the CILI definition for the matched PWN lemma corresponds to the compositional meaning of the word in the NCI MWE. If the meaning and the definition matched, we assigned '1', otherwise '0'. In step 5, we assigned an annotation to the NCI preferred MWE by considering all the individual annotations assigned to each word composing the MWE.

Examples of matching and non-matching cases:

0. NCI *Archaea* (C61092) is not in PWN.
1. NCI *Area* (C25244) and PWN *area_{n:6}* (i63937) have identical definitions.
2. NCI *Breast Cancer Prognostic Factor* (C19601) has no exact match in PWN but its parts do. The individual annotations assigned to each matched part of the MWE ('breast cancer': 1; 'prognostic': 1; 'factor': 1) allow us to assign the global annotation '2' to the preferred term.
3. NCI *Ito Cell Tumor* (C80350) has no exact match in PWN and only two out of the three words composing the MWE are in PWN with the same meaning ('cell': 1; 'tumor': 1; 'ito': 0). These individual annotations allow us to assign the annotation '3' to the preferred term.
4. NCI *Acclimatization* (C68767), defined as "The physiological process through which an organism grows accustomed to a new environment", has a narrower definition than the CILI definition corresponding to PWN *acclimatization_{n:1}*, "adaptation to a new climate (a new temperature or altitude or environment)" (i107289).
5. NCI *NCI Administrative Concept* (C28389) and its definition are specific to the NCI, therefore not suitable for mapping to the CILI.

4 Results

Automatically generating candidate mappings based on lemma overlap and compatibility of UMLS Semantic Types with WordNet domain-category types resulted in 47,464 candidates (out of 118,000), of which 6,028 had a match score $> .75$: this means that either all lemmas overlap or else most lemmas overlap and the domains are compatible. An additional 10,454 matches had a score in the range $.75 > .5$.

To date we have checked 570 of the 6,028 candidates with a match score $> .75$. The results are summarized in Table 3.

Annotation	Number	%
eq - equivalence	369	64.7
spec - hyponym of	21	3.7
gen - hypernym of	33	5.8
0 - no hierarchy relation	147	25.8
<i>evaluated candidates</i>	570	100.0

Table 3: Candidate matches evaluation results

These mappings suggest further heuristics for automatically mapping concepts and refining the match score in future work, thereby expediting mapping and evaluation. Some sample heuristics are listed below.

- Add a score for the similarity of the definitions, e.g., if the Jaccard distance of the definitions is $> .90$, map with 'eq'.
- If the UMLS Semantic Type is 'Manufactured Object' and the PWN synset is a verb, annotate the pair with '0'.

In the manual analysis of PWN 3.0's coverage of the NCI, we found that 20.2% of the NCI concepts had an exact match in PWN (and therefore also the CILI), 11.7% had no match in PWN, and 47.9% had a matching head noun, suggesting a suitable child concept of a synset in PWN. Of the 19 top nodes in the NCI hierarchy,³ three were exact matches and 11 had head nouns that were an exact match in PWN, suggesting a parent/child link.

5 Future Work and Discussion

The coverage analysis and the initial evaluation of the match candidates have brought to light several concrete examples in which guidance is needed to

³We exclude the node 'Retired Concept' from the count.

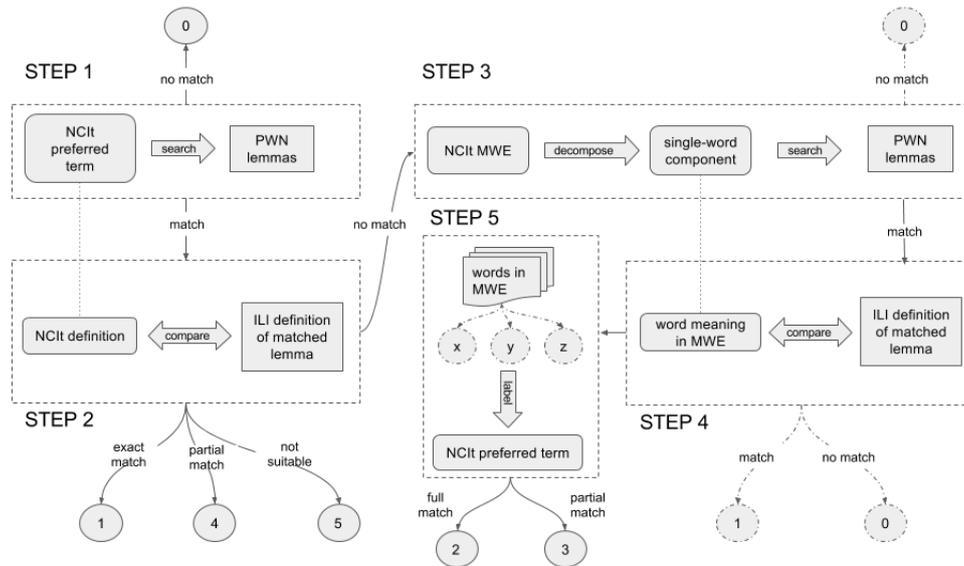


Figure 1: Steps of the manual coverage analysis

integrate specialized terminologies with the CILI. First, the NCI contains dot objects and other cases of systematic polysemy that are sometimes distinguished in WordNet and would therefore have different relevant concepts in the CILI. For example, NCI *Cherry* (C65311) does not have a proper definition but has two UMLS Semantic Types, fruit and plant, suggesting it can refer to a cherry tree or the fruit of a cherry tree. The candidate match in PWN is *cherry*_{n:2} (i103308) which is clearly defined as the tree, not the fruit. A matching strategy for such cases ought to be developed.

Second, we have encountered some cases where the core definition is the same, but exemplars or typical cases are different. In both examples below, an overlay is characterized as something to be applied over an object or surface.

- **Overlay** (NCI) “A device designed to be applied over an object, typically for protection or identification” (C50093)
- **overlay**_{n:2} (PWN) “a layer of decorative material (such as gold leaf or wood veneer) applied over a surface” (i56837)

However, the NCI characterizes an overlay as typically for protection or identification and PWN considers an overlay to be decorative. It is unclear whether these are similar enough to be considered equivalent, whether the NCI concept should be considered a hypernym of the PWN synset (and therefore the corresponding CILI concept), or

whether the typical functions, though not a necessary component of the definition, nuance the meaning sufficiently for no hierarchy relation to be added between the two.

Third, we find that some concepts are probably equivalent, but different definition writing criteria result in a narrower definition in PWN. Consequently, it is unclear whether the NCI concept is a hypernym of the PWN synset.

- **Anovulation** (NCI) “The absence of ovulation” (C34388)
- **anovulation**_{n:1} (PWN) “the absence of ovulation due to immaturity or post-maturity or pregnancy or oral contraceptive pills or dysfunction of the ovary” (i107333)

Fourth, we found that the assumption that all concepts are nouns is not true. Entries such as **unfavorable** are clearly adjectives. Fortunately, the UMLS Semantic Type ‘Qualitative Concept’ and the wordnet coarse domain *adj.a11* both give an indication that it should be an adjective, so we should be able to tell this largely automatically. There are about 1,000 candidate adjectives, and even a few tens of verbs (such as **mutate**), whose definitions tend to start with infinitive **to** in NCI.

- **Unfavorable** (NCI) “Expressing something as negative, undesired or adverse” (C102561)
- **unfavorable**_{a:1} (PWN) “not encouraging or approving or pleasing” (i5455)

- **Mutate** (NCIt) “To undergo or cause genetic mutation” (C28031)
- **mutate**_{v:1} (PWN) “undergo mutation” (i22358)

Finally, we need to decide how to handle multiword expressions that have been annotated with ‘2’ such as **Breast Cancer Prognostic Factor** (C19601). One approach is to create a new concept in the CILI. However, further consideration needs to be given to which concepts are too domain specific to be included in the CILI. Another approach is to map these to the CILI by way of the head word using the hyponym relation. For example, **Breast Cancer Prognostic Factor** (C19601) would be mapped to i75200 by way of PWN **factor**. However, as the number of concepts in the CILI grows, we anticipate that concepts that are not lexicalized in Princeton WordNet will appear in the CILI. For example, the concept prognostic factor may be added to the CILI in the future. In the long term, strategies for detecting and properly aligning such concepts will need to be developed.

We used UMLS Semantic Types, which were created to help disambiguate and cluster senses (McCray et al., 2001), to improve our automatic alignment to PWN coarse domains. PWN contains more detailed domain-category links such as **tobacco**_{n:1} is in the domain of **pharmacy**_{n:1}. We could exploit both them and the hypernyms to improve the automatic mapping. Finally, if all the UMLS Semantic Types can be mapped to synsets, we can link them using **domain-category**. This will enrich the overall graph in ncitWN and facilitate mapping other UMLS terminologies to the CILI.

Acknowledgements

This research was supported in part by the MOE Tier 1 grant *Semi-automatic implementation of clinical practice guidelines in Singapore hospitals* and by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

- Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362. Association for Computational Linguistics, Sofia, Bulgaria. URL <http://aclweb.org/anthology/P13-1133>.
- Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Global WordNet Conference (GWC2016)*, pages 50–57. Bucharest, Romania.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF) for NLP multilingual resources. In *Proceedings of the workshop on multilingual language resources and interoperability*, pages 1–8. Association for Computational Linguistics. URL <http://aclanthology.coli.uni-saarland.de/pdf/W/W06/W06-1001.pdf>.
- John McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing 2014 (LDL-2014)*. Association for Computational Linguistics, Reykjavik, Iceland.
- Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(01):216.
- Luis Morgado da Costa and Francis Bond. 2015. OMWEdit - The Integrated Open Multilingual Wordnet Editing System. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 73–78. Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Beijing, China. URL <http://www.aclweb.org/anthology/P15-4013>.
- National Library of Medicine. 2009. *UMLS Reference Manual*, chapter Chapter 5 - Semantic

- Network. U.S. National Library of Medicine, National Institutes of Health.
- Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, et al. 2007. Advancing translational research with the Semantic Web. *BMC bioinformatics*, 8(Suppl 3):S2.
- Nadine Schuurman and Agnieszka Leszczynski. 2008. Ontologies for bioinformatics. *Bioinformatics and biology insights*, 2:187.
- Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.
- Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. 2007. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.
- Barry Smith and Christiane Fellbaum. 2004. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, page 371. Association for Computational Linguistics.
- Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. Singapore.
- Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference (GWC2016)*, pages 419–26.
- Piek Vossen, Claudia Soria, and Monica Monachini. 2013. *Wordnet-LMF: standard representation for multilingual wordnets*, chapter 4, pages 51–66. John Wiley & Sons, Inc, Hoboken, NJ USA.

Towards a Crowd-Sourced WordNet for Colloquial English

John P. McCrae, Ian D. Wood

Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
john@mccr.ae,
ian.wood@insight-centre.org

Amanda Hicks

Health Outcomes & Policy
University of Florida
Gainesville, FL USA
aehicks@ufl.edu

Abstract

Princeton WordNet is one of the most widely-used resources for natural language processing, but is updated only infrequently and cannot keep up with the fast-changing usage of the English language on social media platforms such as Twitter. The Colloquial WordNet aims to provide an open platform whereby anyone can contribute, while still following the structure of WordNet. Many crowd-sourced lexical resources often have significant quality issues, and as such care must be taken in the design of the interface to ensure quality. In this paper, we present the development of a platform that can be opened on the Web to any lexicographer who wishes to contribute to this resource and the lexicographic methodology applied by this interface.

1 Introduction

The Colloquial WordNet¹, first introduced in (McCrae et al., 2017), is an extension to Princeton WordNet (Fellbaum, 2010; Miller, 1995) that focuses on the use of neologisms and vulgar terminology². The first version of this resource was created primarily by one lexicographer and as such scaling this resource to be able to cover more of the neologisms in English is a significant issue. In this paper, we detail the improvements we have made to the tools that lie behind this resource to enable a more open process for the creation of the resource. We started by detailing the guidelines and methodology for creating the resource and writing new documentation to support lexicographers in their work in annotating the data.

¹<http://colloqwn.linguistic-lod.org>

²We are aware of a similar resource called SlangNet (Dhuliawala et al., 2016) but this does not seem to be publicly available

We also added the possibility to add a confidence so that non-expert lexicographers would be able to provide annotations with some uncertainty. We then improved the interface in order to make it more intuitive for users with little knowledge of the project to use. In particular, we removed a lot of the ‘implicit assumptions’ of the interface that said that if certain options were chosen then other options could not be chosen. Furthermore, we integrated the guidelines in the editor so that lexicographers could easily look up the guidelines at any point where there is uncertainty. Finally, we introduced the idea of queues, where an annotator could add a number of terms, which have been automatically identified as potentially interesting, and these items can be held in the queue for a period of time, before being freed up. This methodology allows multiple lexicographers to collaborate without duplication of effort as each lexicographer’s queue can be kept separate. The candidates that are in the queue are derived from Twitter and we detail the approach that we have taken to preprocessing the corpus and extracting the candidate terms from the result. Finally, we consider the issue of attracting new lexicographers for the resource and detail our plans to use student annotators and the creation of subtasks that may be of particular interest to individual lexicographers. These suggest a wider application of the methodology to more than just creating dictionaries for English neologisms. This project report represents the summary of recent work to create a resource that is more open and will be created by more than one lexicographer.

2 Colloquial WordNet Annotation Methodology

The methodology for creating Colloquial WordNet entries is based on annotating interesting words or short phrases from a corpus of tweets. The lexicographer will be presented with a lemma

and a number of example tweets and is expected to use these in order to write the entry. This is done in three steps: firstly, the lexicographer should check the lemma and examples and make sure he or she is familiar with the term or perform appropriate research in order to find the definition of the term. Then the lexicographer should sort the entry into its status (see Section 2.2), which will influence the method by- which it is further annotated. Then, the main body of the entry is created, in most cases in terms of the senses that define the meaning and any links to other senses.

2.1 Confidence

The first step in the creation of an entry in Colloquial WordNet is the selection of the lexicographer's confidence in the term. We decided to base these categories around the lexicographer's familiarity with the term, and the text guidelines are given as follows:

Very Strong : This is a term I use regularly and know exactly what it means (or the term is clearly an error, incomplete fragment of language or the name of a person, organization, etc.)

Strong : I am clear about the meaning of this term and have heard it used frequently

Medium : I have done a little research and am pretty sure I have found a good definition

Weak : I have guessed from the term and the contexts and think I know what it means

Skip : I don't have a clue about this term and don't want to annotate it

Terms annotated with "skip" are returned to the queue for another lexicographer to handle. All other terms are included and the confidence can be used for other, more experienced lexicographers to check entries which may be weak.

2.2 Entry Status

The status indicates what kind of term this term is, note that "General", "Novel" and "Vulgar" are used for true terms, and "Abbreviation", "Misspelling", "Name", "Not Lexical" and "Error" for terms that will only be included in the ancillary data for Colloquial WordNet.

General : This is a term that should be included in a general-purpose dictionary such as Princeton WordNet. It should be widely and frequently used by native English speakers. Example: "lockpick"

Novel : This term is novel and may not persist in the language. This term should be used for slang, dialectal forms (used only in a particular dialect or social group) and other non-standard usage of English. This should also be used for interjections such as "wow!" or "gosh!" (in this case, the part of speech should be other). Examples: "twerk", "dab", "belieber"

Vulgar : This term is vulgar or obscene and would not be suitable for a general purpose dictionary. Examples: "mindfuck", "paypig".

Abbreviation : This term is an abbreviation; Examples: "IDK", "IMHO"

Misspelling : This term is misspelled; Examples: "agnt", "newjob"

Inflected Form : This term is an inflected form, a simple grammatical variation of a word (e.g.: "running" from the word "run"). Examples: "cats", "the cat"

Name : This term is a name (proper noun) and is not suitable for inclusion in the WordNet. Examples: "Google", "Justin Bieber"

Not Lexical : This is not a proper term. It may be a fragment of text that doesn't make sense as an independent phrase, e.g., "I know a", or it may be a multiword phrase, where the meaning is clearly composed from the constituent words, e.g, "tasty ham", "cheese sandwich".

Error : This is used if the "term" does not seem to be English, e.g., " "

2.3 Entry Details

The entry details are the main work for the lexicographer. For entries, whose status is "General", "Novel" or "Vulgar", the lexicographer will enter the senses as either novel senses with definitions and relations or as synonyms of existing WordNet entries, for which an auto-suggest feature is used to help the lexicographer. This allows the lexicographer to type the lemma of the synonym and then

they are shown the part-of-speech, definition and an Interlingual Index (ILI) ID (Bond et al., 2016; Vossen et al., 2016). In the case the lexicographer chose either “Abbreviation”, “Misspelling” or “Inflected form” the lexicographer simply fills in the lemma that should be used here, i.e., the unabbreviated, noninflected, correctly spelled word. For misspellings and inflected forms this lemma is then queried against existing PWN and Colloquial WordNet entries and if it is not found then it is re-added with the correct lemma to the user’s queue. We require that each new word has at least one link, this is generally to an existing synset in Princeton WordNet, through the Interlingual Index (Vossen et al., 2016; Bond et al., 2016), however it may just be to another existing Colloquial WordNet entry, e.g., “retweet” and “subtweet” to “tweet”.

3 Building an interface for crowd-sourcing

In order to support lexicographers in creating their interface, we have designed an attractive user interface (see Figure 1), that can be used to create new entries in the Colloquial WordNet. The interface is created using Scalatra³, is backed by an SQLite Database⁴ and uses Bootstrap⁵ and Angular⁶ for the user interface. These technology choices were made in order to create an interface with reduced effort.

3.1 Queues

Queues are the main interface that a lexicographer uses to select the terms that they wish to annotate. The lexicographer can choose to add elements to their queue, and these are taken from the most important terms that have not yet been annotated. Once they are entered into the queue they are locked and can only be annotated by this lexicographer for the next 7 days. Lexicographers may remove or extend terms from their queue, and in editing mode, once a lexicographer submits an entry the website automatically redirects them to editing the next entry in their queue or back to the queue page if there are no elements left in their queue.

³<http://scalatra.org>

⁴<https://www.sqlite.org/>

⁵<http://getbootstrap.com/>

⁶<https://angularjs.org/>

3.2 Tweet Collection and Preprocessing

In order to get a sample of current social media language usage, we have been collecting tweets from the “sample” endpoint of the public Twitter streaming API. This provides a continuous stream consisting of a 1% sample of all published tweets. Collection has been ongoing since August 2016, resulting in 435 million English language tweets as of August 2017.

In an attempt to reduce the impact of unintelligible tweets, robots and spam, we apply the following simple rules:

Small Words : We remove tweets if they contain lots of short words. A short word is defined as a word with 1 or 2 characters and we remove the tweet if more than 30% of words are short.

New Lines : We remove tweets with more than two newlines, as these are likely to be advertising or spam.

Non-dictionary words : We check all words against a dictionary of known English words and reject tweets where more than 30% are not in the dictionary. This removes tweets not in English.

Tags : We count the number of words starting with a ‘#’ or ‘@’ and remove it if more than 30% of words start with such a tag.

Tweets matching these rules were mostly either not expressions of natural language or identified as automatically generated tweets. Applying these heuristics substantially reduced the number of tweets, resulting in a collection of 34,776,298 tweet texts as a sample of contemporary social media language usage.

An important feature of this collection is that it spans a whole year. This reduces the effects of high word frequencies associated with specific content associated with large social media coverage (coverage of events such as sports matches, elections, annual television events etc... or tweets that “go viral”). The ongoing and longitudinal nature of the data also permits analysis of *changes* in language usage over time, a topic we intend to investigate in future work.

3.3 Selecting Candidates

Once we have identified the tweets, we attempt to find the words that are most relevant to be anno-

Summary

 <p>awh <cwn-entry-101></p> <p>Edit</p> <p>(other) expression of appreciation emotion → appreciation <i>"an expression of gratitude; he expressed his appreciation in a short note"</i></p>	 <p>awhh <cwn-entry-102></p> <p>Edit</p> <p>(other) Expression of sympathy emotion → sympathy <i>"sharing the feelings of others (especially feelings of sorrow or anguish)"</i></p>
 <p>aww <cwn-entry-104></p> <p>Edit</p>	 <p>aww <cwn-entry-105></p> <p>Edit</p>

Figure 1: The Colloquial WordNet Editor Interface

tated. For this, our primary approach is to use the frequency relative to a background corpus, in particular from a Web Corpus of term frequencies⁷. Our approach chooses the terms that have a high frequency relative to the baseline corpus and in addition we choose terms that mostly occur in all lowercase to remove many of the proper nouns and other terms that are present in tweets. For each of the selected terms we also choose 10 example tweets to help the annotator, these are chosen based on a variation of the GDEX algorithm (Kilgarriff et al., 2008), where in particular we rank tweets based on:

Length If the tweet is between 10 and 25 words.

Blacklisted Words Whether the tweet contains any blacklisted words, such as ‘this’, ‘that’ or ‘http’

Punctuation Whether the tweet starts with a capital letter and ends with a full stop, question mark or exclamation mark.

Frequent Words How many of the words in the tweet are in the top 17,000 words.

These tests give each tweet a score out of 4, with ‘frequent words’ used as a tiebreaker. We greedily choose the top 10 example tweets, in addition requiring that no tweet overlaps by more than 5 words with a previously selected tweet.

⁷<http://norvig.com/ngrams/>

4 Supporting Lexicographers

To facilitate the development of Colloquial WordNet future work involves using linguistics students as annotators and creating subtasks focused on a particular domain of interest so that lexicographers who are proficient with the use of terms that are specific to particular subdomains and communities.

4.1 Gender minority and Pro-Ana Subtasks

We have developed two specialized Twitter corpora in previous projects (Hicks et al., 2015; Wood, 2015), that can also be used to find domain specific terms for addition to Colloquial WordNet and to attract annotators who are interested in and drawn to a specific topic. One corpus, first reported in (Hicks et al., 2015), represents tweets over a period of 49-day period from January 17, 2015 to March 6, 2015 inclusive that contain terms related to gender identity, particularly terms that indicate a transgender or other gender minority identity, (e.g., “transboi”, “FTM”, and “non-binary”). A pilot interface has been created around this corpus using the method described in the previous section to suggest candidate terms for inclusion in the Colloquial WordNet.

The second Twitter corpus, originally report in (Wood, 2015), represents tweets over a period of nearly three years (December 2012–October 2015) that contain hashtags that may indicate

membership of the “pro-anorexia” and eating disorder community (e.g., “#proana”, “#edproblems”, and “#thinspiration”).

While these domain specific subtasks contain community specific neologisms, they also contain general terms that may not already be included in WordNet (e.g. “trans woman” and “queerness”). Many of the candidate terms derived from the gender minority corpus are not specific to gender identity (e.g. “tummy tuck”, “woc” as an abbreviation for woman of color and “tranny” as a synonym of “transmission”). Furthermore, a coverage analysis of WordNet’s gender identity terms showed that adding a small number of wordsenses to WordNet can result in significantly greater coverage of gender identity terms in WordNet due to the prevalence of compositional multi-word expressions used to describe gender identity. (Hicks et al., 2016). We anticipate that these subtasks will also increase coverage of non-domain specific terms while retaining the interest and participation of annotators who are drawn to the topic.

5 Conclusion

In this paper we present the progress in the development of Colloquial WordNet editor and its tools. While there exist many other tools for editing WordNets, e.g., DebVisDic (Horák et al., 2006), SlowTool (Fišer and Novak, 2011), plW-NApp (Derwojedowa et al., 2008) or Wordnetloom (Piasecki et al., 2013), none of these tools meet our goal of being an open Web-based development platform that can be used by any user. The goal of Colloquial WordNet is to be more open, and as such we do not necessarily expect the same level of expertise from our lexicographers or quality in the resulting resource. Instead, we understand Colloquial WordNet to provide a good WordNet-level coverage of English as it used in social media, which will be helpful to handling noisy user-generated text, a problem that has caused significant issues for natural language processing recently (Baldwin et al., 2015). Currently the resource consists of the same 428 entries previously detailed (McCrae et al., 2017), however we now expect to work on expanding the resource. Furthermore, we believe that the exercise of developing the Colloquial WordNet can identify key words that we hope will contribute to the next version of Princeton WordNet and should assist the lexicographers by providing entries that

can be further extended into PWN entries.

Acknowledgments

This work was supported in part by the Science Foundation Ireland under Grant Numbers SFI/12/RC/2289 (Insight) and 16/IFB/4336 and also in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NCTE.

References

- [Baldwin et al.2015] Timothy Baldwin, Young-Bum Kim, Marie Catherine De Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.
- [Bond et al.2016] Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.
- [Derwojedowa et al.2008] Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of Polish WordNet. In *Proceedings of the Global WordNet Conference, Seged, Hungary*, pages 162–177.
- [Dhuliawala et al.2016] Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Slangnet: A wordnet like resource for english slang. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pages 4329–4332.
- [Fellbaum2010] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- [Fišer and Novak2011] Darja Fišer and Jernej Novak. 2011. Visualizing sloWNet. *Proceedings of the Electronic Lexicography in the 21st Century (eLex 2011)*, pages 76–82.
- [Hicks et al.2015] Amanda Hicks, R. Hogan, William, Michael Rutherford, Bradley Malin, Mengjun Xie, Christiane Fellbaum, Zhijun Yin, Daniel Fabbri, Josh Hanna, and Jiang Bian. 2015. Mining Twitter as a first step toward assessing the adequacy of gender identification terms on intake forms. In *Proceedings of the AMIA 2015 Annual Symposium*. American Medical Informatics Association.
- [Hicks et al.2016] Amanda Hicks, Michael Rutherford, Christiane Fellbaum, and Jiang Bian. 2016. An

analysis of WordNets coverage of gender identity using Twitter and the national transgender discrimination survey. In *Proc of the Eighth Global WordNet Conference (GWC)*, volume 2016, pages 122–129.

[Horák et al.2006] Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolný. 2006. DebVisDic—first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International WordNet Conference—GWC 2006*, pages 325–328.

[Kilgarriff et al.2008] Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of Euralex*.

[McCrae et al.2017] John P. McCrae, Ian Wood, and Amanda Hicks. 2017. The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*.

[Miller1995] George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Piasecki et al.2013] Maciej Piasecki, Micha Marciuk, Radosaw Ramocki, and Marek Maziarz. 2013. Wordnetloom: a Wordnet development system integrating form-based and graph-based perspectives. *Int. J. Data Mining, Modelling and Management*, (5).

[Vossen et al.2016] Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.

[Wood2015] Ian Wood. 2015. A case study of collecting dynamic social data: The pro-ana Twitter community. *Australian Journal of Intelligent Information Processing Systems*, 14(3).

WordNet Troponymy and Extraction of “Manner-Result” Relations

Aliaksandr Huminski

Institute of High Performance
Computing, A*STAR, Singapore

huminskia@iphc.a-
star.edu.sg

Zhang Hao

Nanyang Technological University
Singapore

hao.zhang@ntu.edu.sg

Abstract

Commonsense knowledge bases need to have relations that allow to predict the consequences of specific actions (say, if John stabbed Peter, Peter might be killed) and to unfold the possible actions for the specific results (Peter was killed. It could happen because of poisoning, stabbing, shooting, etc.) This kind of causal relations are established between manner verbs and result verbs: manner-result relations.

We offer a procedure on how to extract manner-result relations from WordNet through the analysis of the troponym glosses. The procedure of extraction includes three steps and the results are based on the analysis of the whole set of verbs in WordNet.

1 Introduction

WordNet (WN) as a database is widely used in variety of tasks related with extraction of semantic relations. Verbs in WN are organized hierarchically as troponym-hypernym relations. Meanwhile, the definition of troponym has something in common with the definition of a manner verb suggested by B. Levin and M. Rappaport Hovav.

We consider in more details both types of relations: troponym-hypernym and manner verb-result verb relations.

1.1 Troponym-Hypernym Relation

Verbs in WN are linked through different types of relations – antonym, cause, entailment – but troponym-hypernym relation is a basic and the most frequently found relation among verb synsets (Fellbaum and Miller, 1990). If a hypernym is a verb of a more generalized meaning, a

troponym replaces the hypernym by indicating more precisely the manner of doing something. The troponym-hypernym relations are hierarchical (vertical). Therefore, it makes it possible to create a huge verb net with top synsets that represent the highest hypernyms and branches going down to the bottom with corresponding troponyms. The closer to the bottom, the more specific is the verb synset. There are no further clarifications between different types of troponymy in WN.

As a result, the manner relation is polysemic and many different semantic elements are hidden behind the label ‘manner’ (Fellbaum, 2010). It can be volume as in *talk-whisper*, speed as in *jog-run*, intensity of emotion as in *love-adore-idolize*, etc. The specific manner depends on the semantic field and corresponding dimension.

1.2 Manner Verbs and Result Verbs

The definition of troponym has something in common with the definition of a manner verb suggested by Beth Levin and Malka Rappaport Hovav (2010). They pointed out that a study of the English verb lexicon reveals that within particular semantic domains there can be verbs that describe carrying out activities – manners of doing; and there can be verbs that describe bringing about results. Manner verbs are *walk, jog, stab, scrub, sweep, swim, wipe, yell*, etc. Result verbs are *break, clean, crush, destroy, shatter*, etc.

There are 3 features of manner-result relations that make extraction of them so important for commonsense knowledge bases.

- 1) Manner verbs and result verbs are in causal relations: *stabbing* causes *killing*; *sweeping* causes *cleaning* and etc.
- 2) It is an empirical, not a logical causality with probability less than 100%. Actions represented by manner verbs can fail in achievement of desirable results:

*I wiped the table, but it's not clean.
John shot Peter, but he survived.*

- 3) It is a common situation when several manner verbs cause the same result verb: *sweeping, wiping, blowing* cause *cleaning*.

2 Troponym-Hypernym and Manner-Result Relations

In the WN glossary of terms¹, a troponym is defined as a verb expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner. Having this definition, the obvious question arises: if troponym is defined through the manner, can one state that troponym-hypernym relation equals in manner verb-result verb relation? In other words, is there any correlation between troponym-hypernym relation and manner verb-result verb? The general answer on this question is “no” since there are several types of correspondence that can be unfolded in WN:

- 1) troponym-hypernym relation can be equal “manner verb-manner verb” relation. For example, the verb *stroll* (walk + slow + relaxed) is a troponym for the verb *walk*. But both of them are manner verbs.
- 2) troponym-hypernym relation can be equal “manner verb-underspecified verb” relation. For example, the verb *walk* (move + by steps) is a troponym for the verb *move*. The verb *walk* is a manner verb, the verb *move* is underspecified: it is neither a path verb since it doesn't encode direction, nor a manner verb since it doesn't specify any particular manner. So, it is an underspecified verb taking into consideration that manner-result dichotomy does not fully and exhaustively classify verbs.
- 3) troponym-hypernym relation can be equal “result verb-result verb” relation. For example, the verb *fracture* (break into pieces) is a troponym for the verb *break* (destroy the integrity).
- 4) troponym-hypernym relation can be equal “manner verb-result verb” relation. For example, the verbs *stone*, *lapidate* (kill by throwing stones at) and *poison* (kill with poison) are troponyms for the verb *kill* (cause to die; put to death).

Now, we need to find out the way how to extract the 4th type of correspondence which represents exactly what we are looking for.

3 General Procedure to Extract Manner-Result Relations from WordNet

Manner-result relations are hidden in the WN verb hierarchy. We know for sure that this kind of relations is a subset of troponym-hypernym relations. However, there are not any explicit ways to extract them yet.

Our idea is that manner-result relations can be extracted from the set of troponym-hypernym relations if two conditions, applied to troponym-hypernym relation are valid:

- 1) The hypernym is a result synset;
- 2) In the glosses of its troponyms one of the two templates can be found: “V + by” or “V + with”; where V = hypernym.

For example, if we consider the result synset {*clean, make clean*} as a hypernym, some its troponyms have glosses that satisfy the patterns:

- *sweep* (clean by sweeping)
- *brush* (clean with a brush)
- *steam, steam clean* (clean by means of steaming)

In this case, it can be stated that *sweep, brush, steam, steam clean* are manner verbs for *clean* and the following causality can be constructed:

sweep, brush, steam, steam clean → *clean*

This idea is the basis of the general procedure for manner-result extraction. It includes 3 steps.

3.1 Extraction of Top Verb Synsets

There are 13789 verb synsets in WN 3.1 ordered by troponym-hypernym hierarchical relation.

At this stage, we need to extract synsets located on the top level of the hierarchy. This kind of synsets will be called further “top verb synsets”.

The procedure of extraction is based on the following characteristic of the top verb synsets: they don't have any hypernyms, only troponyms. Using this, all the extracted 13789 verb synsets have been tested whether they have a hypernym. As a result, 564 top verb synsets have been extracted automatically.

3.2 Extraction of Top Result Verb Synsets

Within 564 top synsets we made a manual classification to extract only the result verb synsets. The classification revealed the following 5 classes.

¹ <https://wordnet.princeton.edu/man/wngloss.7WN.html>

- 1) one-level top synsets. This type of top synsets has only one level: the top verb synset itself. It is a substantial portion of top synsets: 203. Example: *admit* (give access or entrance to).
- 2) manner and underspecified verb synsets. Total number: 105. Example of the top manner verb synset: *splash* (strike and dash about in a liquid). Example of the top underspecified verb synset: {*travel, go, move, locomote*}.
- 3) state verb synsets. Total number: 69. Example of the top state verb: *lie* (be lying, be prostrate; be in a horizontal position).
- 4) second order predicates. Total number: 60. Second order predicates govern the other predicate. Example: {*begin, start*} (have a beginning, in a temporal, spatial, or evaluative sense).
- 5) result and change-of-state verb synsets. Total number: 127. We combine these 2 classes of verbs since, as it turned out, change-of-state verbs have manner verbs as troponyms. For instance, the verb *die* has a troponym synset {*suffocate, stifle, asphyxiate*} which obviously contains manner verbs. Example of the result verb synset: {*destroy, ruin*}.

We further analyze the 5th class only. Our assumption was that result verbs as hypernyms can have either result verbs or manner verbs as troponyms. But manner verbs as hypernyms cannot have result verbs as troponyms. They can only have manner verbs as troponyms. Following the assumption, the sequence of troponyms derived from the top result verb hypernym cannot have the subsequence of manner verb as a hypernym and result verb as a troponym. For example, the sequence of 4-level verbs with the top result verb and the bottom manner verb can have the following 3 possible distributions:

- result-result-result-manner
- result-result-manner-manner
- result-manner-manner-manner

The distribution of “result-manner-result-manner” is impossible.

The next step is extraction of manner verbs from the tree with result verb synset on the top.

3.3 Extraction of Manner Verbs through the Patterns in Glosses

At this stage, we look for the manner verbs for each result verb synset through the patterns “V + with” or “V + by” in the glosses of troponyms. If

the synset doesn't contain any patterns we mark it as “NONE”. If the synset contains at least one of the patterns we mark it with its gloss.

As a result, we get a sequence of marked synsets from the top verb synset to the bottom verb synset. If the sequence of all synsets or only the tail of it contains “NONE” we exclude the whole sequence or the tail accordingly from the consideration since there is no manner verbs there. The purpose is to extract all lower synsets that contain the patterns. The procedure of the extraction is automatic.

Following the assumption from 3.2 one can get different types of result-manner sequences. For example, for the top synset {*change, alter, modify*} we will get the following 3 sequences among many others:

{ <i>change, alter, modify</i> }-NONE	{ <i>damage</i> }-NONE	{ <i>frost</i> }-damage by frost
---------------------------------------	------------------------	----------------------------------

The causality *frost* → *damage* can be made from this sequence, where *frost* is manner verb and *damage* is result verb.

{ <i>change, alter, modify</i> }-NONE	{ <i>damage</i> }-NONE	{ <i>burn</i> }-damage by burning with heat, fire, or radiation	{ <i>scald</i> }-burn with a hot liquid or steam
---------------------------------------	------------------------	-----------------------------------------------------------------	--------------------------------------------------

The causality is *scald, burn* → *damage*

{ <i>change, alter, modify</i> }-NONE	{ <i>indispose</i> }-NONE	{ <i>hurt</i> }-NONE	{ <i>injure, wound</i> }-NONE	{ <i>trample</i> }-injure by trampling or as if by trampling
---------------------------------------	---------------------------	----------------------	-------------------------------	--------------------------------------------------------------

The causality is *trample* → *injure, wound, hurt*.

For each *n*-level synset one can get a restricted number of the valid sequences. For example, for each 3-level synset we can get only two valid sequences: “result-result-manner” as in

{ <i>change, alter, modify</i> }-NONE	{ <i>sharpen</i> }-NONE	{ <i>whet</i> }-sharpen by rubbing, as on a whetstone
---------------------------------------	-------------------------	-------------------------------------------------------

and “result-manner-manner” as in

{ <i>damage</i> }-NONE	{ <i>burn</i> }-damage by burning with heat, fire, or radiation	{ <i>scald</i> }-burn with a hot liquid or steam
------------------------	-----------------------------------------------------------------	--------------------------------------------------

As a whole, for the different sublevels of the same top result synset, one can get full variety of valid *n*-level sequences:

{change, alter, modify}-NONE	{shape, form}-NONE	{tabulate}-Manner verb		
{change, alter, modify}-NONE	{shape, form}-NONE	{roll}-Manner verb		
{change, alter, modify}-NONE	{shape, form}-NONE	{draw}-Manner verb		
{change, alter, modify}-NONE	{shape, form}-NONE	{fit}-NONE	{dovetail}-Manner verb	
{change, alter, modify}-NONE	{shape, form}-NONE	{flatten}-NONE	{steamroll, steamroller}-Manner verb	
{change, alter, modify}-NONE	{shape, form}-NONE	{flatten}-NONE	{roll_out, roll}-Manner verb	
{change, alter, modify}-NONE	{shape, form}-NONE	{flatten}-NONE	{roll_out, roll}-Manner verb	{mill}-Manner verb

Table 1. Part of valid n-level sequences from {change, alter, modify} result synset.

To make the table more compact we replaced the glosses that match the patterns to the phrase “Manner verb”. Figure 1. shows the Table 1. in the structural graphical form with glosses.

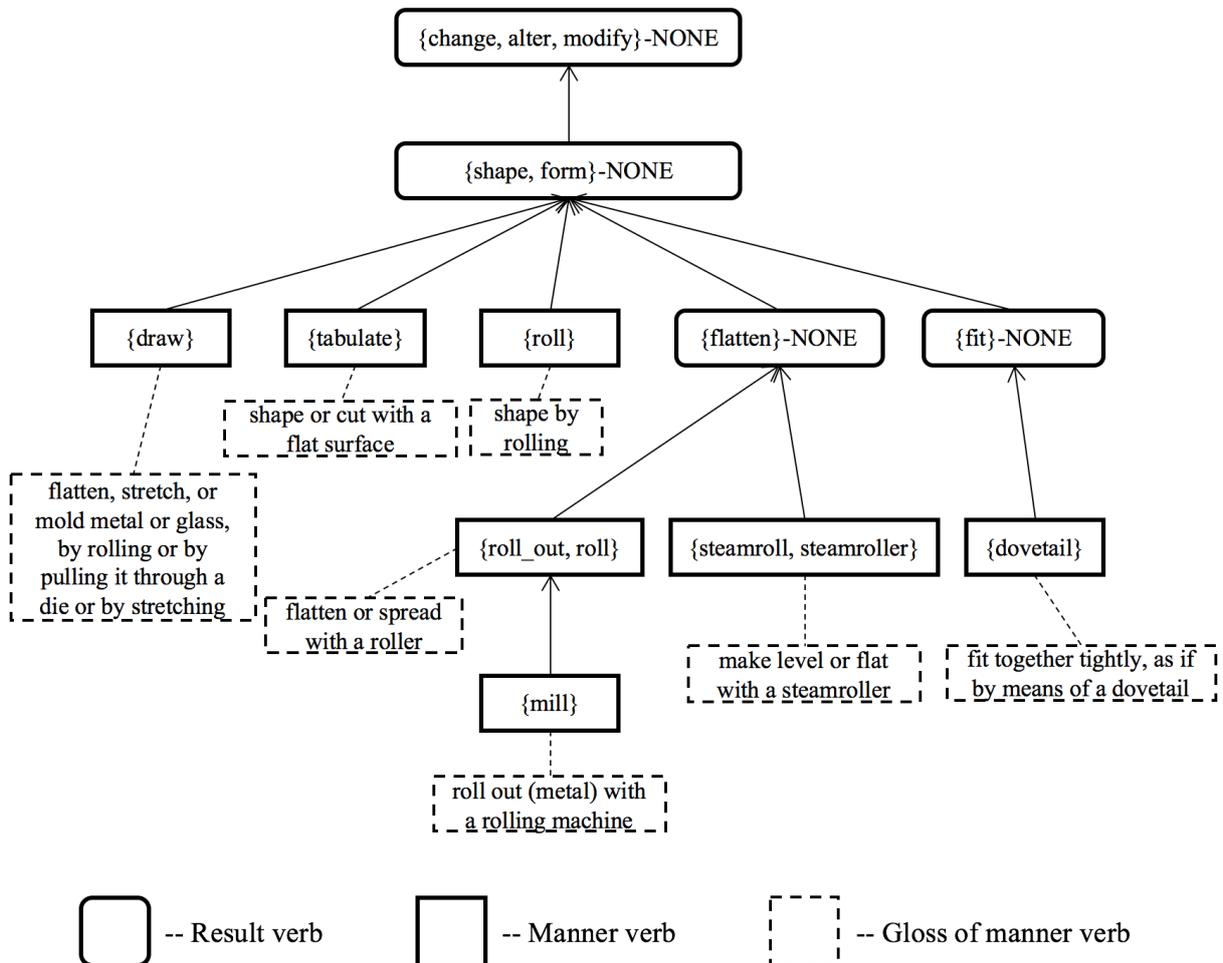


Figure 1. Visualization of the valid n-level sequences.

It is necessary to stress that each line in result-manner causal relation can contain both direct (*frost* → *damage*) and indirect (*scald* → *damage*) causality. Regardless of that, each line is considered as one specific type of causal relations.

After running all the top 127 result verb synsets and counting the lines we got the total number of 1541 lines. It means, 1541 manner-result causal relations have been extracted from WN.

4 Scope of the Results

To evaluate what is the scope of the results we compare them with another type of causal relations that is explicitly presented in WN 3.1: *cause*-relation.

Cause-relation refers to the relation between two verbs V_1 and V_2 where V_1 logically causes V_2 (Fellbaum, 1998). For example, the verb *kill* causes the verb *die*.

Running through 13789 verb synsets in WN 3.1 we automatically extracted 219 verb synsets that contain *cause*-relation. Among them there are 63 verb synsets that cause the same synset. In other words, there are 63 causal relations with absolutely identical left and right sides:

{dry, dry_out} causes *{dry, dry_out}*
{lengthen} causes *{lengthen}*, etc.

It happened because of polysemy in verb meaning. Synsets here are formally identical but represent different meanings of verbs. Since it is hard to use such kind of causality in applications, the real number of the verb synsets that contain *cause*-relation can be reduced to 156.

Comparison of 156 verb synsets containing logical *cause*-relation with 1541 non-logical (empirical) causal relations shows that the scope of the latter relations is significant.

5 Conclusions and Future Work

In this paper, we have described how to extract manner verb-result verb causal relations from WN. The procedure of extraction includes 3 steps: a) extraction of the top verb synsets (total 564), b) extraction of the result synsets and the change-of-state synsets among them manually (total 127), c) running automatically the algorithm “*V + by*” and “*V + with*” on 127 top synsets and getting 1541 types of manner-result causal relations. The results are considered as preliminary ones.

As future work, the algorithm can be elaborated by adding new patterns and tuning the original

ones. For example, the change-of-state verb *die* has a troponym synset *suffocate, stifle, asphyxiate* (be asphyxiated; die from lack of oxygen) which clearly indicates the manner of dying but the gloss doesn't contain the patterns we are working with.

These types of extracted relations can be widely used in commonsense knowledge bases for the prediction of action consequences and unfolding the possible reasons for the results. Commonsense knowledge bases enriched by using this approach can be exploited in dialog systems and the other specific technologies and applications.

References

- Fellbaum C. and G. Miller. 1990. *Folk psychology or semantic entailment? A reply to Rips and Conrad*. The Psychological Review. 97, 565-570.
- Fellbaum, C. 2000. Autotroponymy. In Yael Ravin (ed). *Polysemy: theoretical and computational approaches*. New York, Oxford University Press.
- Levin, B. and M. Rappaport Hovav. 2010. *Lexicalized Meaning and Manner/Result Complementarity*. Ms. Stanford University and The Hebrew University of Jerusalem.
- Fellbaum C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

SardaNet: a Linguistic Resource for Sardinian Language

**Manuela Angioni,
Franco Tuveri**
CRS4

Center for Research and
Scientific Studies in Sardinia
Bld. 1, Piscina Manna,
Pula (CA), Italy.
{angioni,tuveri}@crs4.it

**Maurizio Virdis,
Laura Lucia Lai,
Micol Elisa Maltesi**
University of Cagliari
Cagliari, Italy

virdis@unica.it
llaura@gmail.com
micol-elisa@hotmail.it

Abstract

This paper describes the process of building SardaNet, a linguistic resource for Sardinian language including the different linguistic varieties in Sardinia. SardaNet aims at identifying the semantic relations between Sardinian terms, by manually mapping existing WordNet entries to Sardinian word senses. The work, still in progress, has been developed in collaboration with the University of Cagliari. After discussing some linguistic peculiarities, the paper presents the basic steps of the construction process, the method and the tools involved, the issues encountered during the development and the current version of SardaNet.

1 Introduction

Sardinian territory is characterized by a strong multilingualism, in which it is difficult to trace the precise boundaries between a variant and the other, each characterized by its phonetic, morphological and lexical features.

For a long time, the linguists have been trying to put the distinction between the different linguistic variants spoken in Sardinia, but there is not an unanimously shared theory.

SardaNet examines the Sardinian linguistic variants to which the “Legge regionale 26 del 1997” (Regional law 26 of 1997 for the preservation of linguistic minority) refers: Campidanese, Nuorese and Logudorese, and also the other not-Sardinian variants spoken in the island such as Sassarese, Gallurese, Tabarchino and Algherese.

The ultimate goal is the development of the semantic network related to WordNet¹ (Miller, 1995) and enriched with the peculiar terms and the concepts defined in the Sardinian languages.

SardaNet, in its first preliminary version, has been manually developed, starting with the set of 4689 Common Base Concepts² (CBCs) extended by BalkaNet (Tufis et al., 2004) in the Princeton WordNet 2.0 version, by inserting the corresponding terms in the Sardinian variants.

The work has been developed in collaboration with the University of Cagliari. Two trainees, coordinated by the Prof. Maurizio Virdis, leading expert on Sardinian studies, worked with us in this first phase of the project.

The remainder of this paper is organized as follows: Section 1 introduces the resource and the motivation behind, Section 2 presents an overview of the Sardinian language and its peculiarities and the dictionaries used to build SardaNet. Section 3 describes the method applied by the team involved, highlighting some emerging issues, while Section 4 depicts the building of the resource, the interface used, the LMF format and the mapping CILI. Finally, in Section 5 final remarks and future works directions are presented.

2 The Sardinian Language

Before starting the development of SardaNet, several discussions with Sardinian language

¹ Princeton University "About WordNet." WordNet. Princeton University. 2010. <<http://wordnet.princeton.edu>>

² <http://globalwordnet.org/gwa-base-concepts/>

experts of the University of Cagliari were conducted to better understand the key features of the language.

As it is reported in Virdis (2003a and 2003b), Sardinian is a neo-Latin language, which derives by the evolution of the Latin language, like Italian, French, Spanish, Portuguese and Romanian.

Compared to other neo-Latin languages, Sardinian evidences some peculiarities and even considerable diversities.

The written production in Sardinian language was in fact very plentiful, used in official and documentary written, or in patrimonial and juridical documents, and in particular the *condaghes*. The *condaghe*, from the medieval Sardinian term *kondake* (from the greek *Κοντακιον*), was a kind of administrative document used in the Sardinian *Giudicati* between the 11th and 13th centuries. The *Condaghe* of Santa Maria di Bonarcado, more in details, allows having a valid source for the philological and glottological studies of the Sardinian language, and in particular, for the Arborensis area (Virdis, 1982).

As for the lexicon, Sardinian lexical heritage is an original amalgam of Latin, of ancient and modern Italian, of Catalan and Spanish, often with unique and distinctive creations and interpretations.

The Sardinian has long lived in a state of increasing marginalization by official uses and has been only restricted to familiar and colloquial use. It has been used mainly in low linguistic registers, or most of all for poetic literary compositions, and sometimes it has been also considered a language of marginal use compared to Spanish and Italian.

Currently, the most difficult phase for Sardinian and for dialects in general seems to be overcoming. People are less afraid to speak in dialect, sometimes rediscovering a pride in speaking the native national language. At Italian level, but even more within the European Community, the cultural, historical and anthropological value of European minority languages, such as Sardinian, is becoming more and more important. Political and cultural actions have been launched to save them, and even the regional politic in Sardinia tends to bring Sardinian language back to schools and proposes projects to perform that.

Sardinian is a particular language: there are a lot a variants in relation of the region considered such as Logudorese, Nuorese, Campidanese, and

not native such as Sassarese, Catalano, Gallurese and Tabarchino, as depicted in Figure 1 that displays the geographical distribution of the varieties of Sardinian.

As explained in Virdis (1978), Sardinian language is spoken in Sardinia and only in Sardinia (excluding the large number of emigrants who carry, speak and practice Sardinian language outside of the island). But Sardinian language is not spoken in all Sardinia: in fact, it is necessary to exclude Gallura, where a Southern course dialect is spoken, Alghero where Catalan is spoken, and finally Carloforte and Calasetta where Ligurian is spoken.

The Sassarese has particular historical origins, born in the Middle Age at the time of Pisan-Genovese penetration as a free language due to a contact effect between two linguistic types: the Sardinian and the Italian continental one. Logudorese and Nuorese are mainly spoken in the northern sub-region of the island. Campidanese is the variety of the Sardinian language primarily spoken in South Central Sardinia.

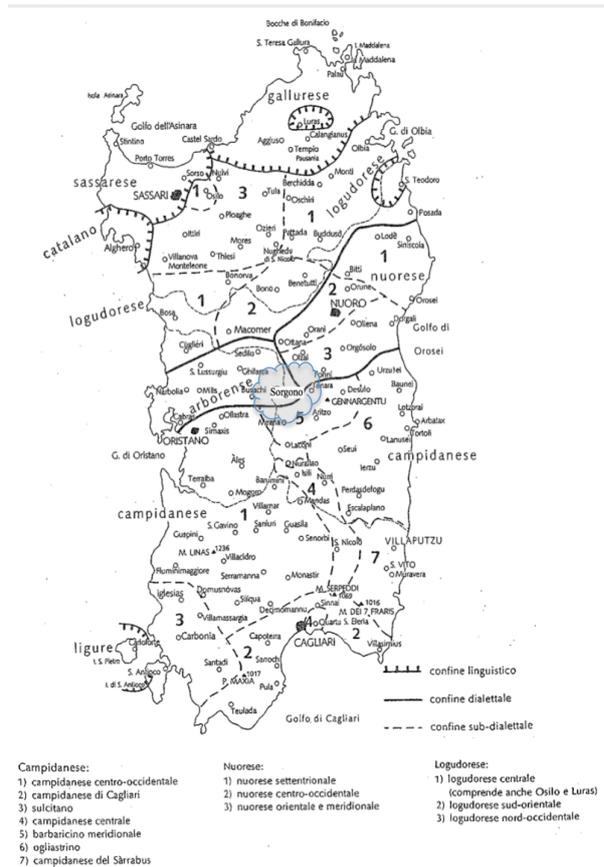


Figure 1. Distribution of the linguistic varieties in Sardinia³.

³ The distribution of Sardinian dialects and sub-dialects (Virdis, 1988).

The language has never had a real unification and never linguistic variety above the others has been imposed. Sometimes who speaks a variant of north Sardinia has some difficulties in understanding a variant of south Sardinia and vice versa. So it is sometimes difficult to communicate.

Due to these peculiarities of Sardinian language, we decided, according to Prof. Maurizio Viridis, to insert in the Sardinian WordNet, for each WordNet entry, the indication of the language variation, considering them as synonyms, and following the expand approach.

2.1 The Sardinian Resources

In the construction of SardaNet we have considered different Sardinian lexical varieties, as shown in Figure 1, which present not only phonetic differences but also a multitude of exclusive lemmas: Logudorese, Nuorese, Campidanese, and the other languages spoken in Sardinia, as Sassarese, Catalano, Gallurese and Tabarchino. At present, Tabarchino and Algerese are not included in SardaNet.

Therefore, several dictionaries have been used, someone available only in paper format, mono linguistic or multi linguistic, mostly related to a single variant of Sardinian, others incorporating in a single dictionary the multiplicity of Sardinian variants.

Among the first resources for the Sardinian language there are the Sardinian Campidanese - Italian dictionary written in 1832 by Vincenzo Raimondo Porru (Porru, 2002) and the “Vocabolariu Sardo-Italianu e Vocabolario Italiano-Sardo” a Sardinian-Italian, Italian-Sardinian dictionary, written by Giovanni Spano (Spano, 1998) in the period between 1851 and 1852.

In the period between July 1934 and April 1947 Pietro Casu (Casu, 2002), dedicated many years to the collection of lexical materials, and wrote a manuscript, the “Vocabolario Sardo Logudorese - Italiano” (Sardinian Logudorese - Italian Dictionary), one of the most important works of Sardinian lexicography for the richness of the phraseology included.

More recently, the “Dizionario Etimologico Sardo” (DES) (The Sardinian Etymological Dictionary) (Wagner, 1964), written in three volumes, is certainly a fundamental work for the study of the Sardinian language. It contains the list of all the most relevant words of the Sardinian, which Wagner compares to

investigate their source and their meaning. However, its consultation is sometimes complicated due to the incompleteness of the general indexes and the phonetic transcription of the lexical material.

The reprint of the DES dictionary edited by Giulio Paulis (Wagner, 2008) has enriched the indexes and has performed a thorough review of the texts, filling out some gaps in the original version. Paulis also wrote “Introduzione a Max Leopold Wagner, Fonetica storica del sardo” (Paulis, 1984), an introduction to the book, related to the historical phonetic of the Sardinian language, written by Max Leopold Wagner.

The “Dizionario Universale della Lingua di Sardegna” (Universal Dictionary of the Language of Sardinia) (Rubattu, 2001), in two volumes, allows instead a simpler and more immediate use. It contains the terms in the various linguistic varieties of Sardinia, Logudorese, Nuorese, Campidanese, Sassarese, Catalano, Gallurese and Tabarchino, whose distribution is shown in Figure 1, with correspondence in English, French, Spanish and German. It is also available on the Sardinian Digital Library⁴.

vrb. èssere a foedhadas									
modu INDICATIVU tempus presente									
campidanese	nucoresu	baronesiu	logudorese						
dèu	seu	dego	soe	zeo	seo	deò	so		
tui	ses	tue	ses	tue	ses	tue	ses		
issu	est	issu	est	issu	est	isse	est		
nosatras	seus	nois	sermus	nois	sermus	nois	sermus		
bosatras	seis	bois	seris	vois	seris	bois	seris		
issus	funt/sunt	issos	sunt	issos	sont	issos	sunt		
modu INDICATIVU tempus imperfectu									
dèu	fiu/femu/furia	dego	fipo	zeo	fipo	deò	fiu/fiu		
tui	fiast/furiast	tue	fis	tue	fis	tue	fis		
issu	fiat/furiat	issu	fit	issu	fit	isse	fit		
nosatras	festus/furiatus	nois	firmus	nois	firmus	nois	firmus/firmis		
bosatras	festus/furiatus	bois	fizis	vois	fizis	bois	fizis		
issus	fiant/furiant	issos	firm	issos	firm	issos	firm		
modu CONGIUNTIVU tempus presente									
chi	sia	dego	sia	zeo	sia	deò	sia		
dèu	sias	tue	sias	tue	sias	tue	sias		
tui	siat	issu	siat	issu	siat	isse	siat		
issu	siat	issu	siat	issu	siat	isse	siat		
nosatras	sias	nois	siamus	nois	siamus	nois	siamus		
bosatras	sias	bois	siazes	vois	siazis	bois	siazes		
issus	siant	issos	siant	issos	siant	issos	siant		
modu CUNDTZIONALE tempus presente									
dèu	ia/emu a èssiri	dego	dia èssere	zeo	dia èssere	deò	dia/ia èssere		
tui	ias	tue	dias	tue	dias	tue	dia/ias		
issu	iat	issu	dias	issu	dias	isse	dia/iat		
nosatras	ias	nois	diamus	nois	diamus	nois	diamus/diamis		
bosatras	ias	bois	diazes	vois	diazis	bois	diazis/diazis		
issus	iant	issos	diant	issos	diant	issos	diant/diant		
issu	iant	issos	diant	issos	diant	issos	diant/diant		
Formas cumpostas: modu ind. r. passau, prus che passau, benidore, cong. e cunditz. passau									
dèu	soe istèu	fipo istèu	deò soe istadu	zeo fipo istèu	fipo istadu	deò soe istadu	fiu istadu		
tui	fiu istèu	fipo istèu	fipo istadu	fipo istèu	fipo istadu	fipo istèu	fipo istadu		
issu	apo a èssiri	apo a èssere	apo a èssere	apo a èssere	apo a èssere	apo a èssere	apo a èssere		
issu	sia istèu	sia istadu	sia istadu	sia istèu	sia istadu	sia istèu	sia istadu		
issu	ia/emu a èssiri istèu	dia/dio èssere istau	dia/dio èssere istadu						
modos INDEFINIOS									
Infinitiu									
èssiri/èssi	èssere	èssere	èssere	èssere	èssere	èssere	èssere		
Participiu passau									
istèu	istèu	istèu	istèu	istèu	istèu	istèu	istèu		
Gerundiu									
sendu/sendi	essendhe	essendhe	essendhe	essendhe	essendhe	essendhe	essendhe		

Figure 2. The conjugation of the verb èssere.

Another reference dictionary is “Su Ditzionariu de Sa Limba e de sa Cultura Sarda” (The Dictionary of the Sardinian Language and Culture) (Puddu, 2015), written entirely in Sardinian language with a partial matching of the

⁴ Dizionario universale della lingua di Sardegna : I e II volume, Sardegna Digital Library: <http://www.sardegna.digitallibrary.it/index.php?xsl=2435&s=17&v=9&c=4459&c1=Rubattu+Antoninu&n=24&ric=1&idtipo=0>

words into five languages: Italian, English, French, Spanish and German. Puddu in his dictionary uses the linguistic variant defined as “Limba de Mesania” (Language of Mesania), a variant located beyond the external arborese border area, around the city of Sorgono, between the two macro-areas Logudorese and Campidanese, as shown in Figure 1.

The Figure 2 shows the conjugation of the verb *èssere* (in English *to be*) in some Sardinian languages as reported in Puddu (2015).

The information needed to build the language resource in the format required by the Global WordNet Association is not always included in the available dictionaries. As for the definitions, for example, Rubattu provides them only for verbs while Puddu puts them but written in the Mesania language.

An indispensable research manual for everyone interested in the Sardinian language and in Romance linguistics in general is the “Manuale di linguistica sarda” (Manual of Sardinian Linguistics) (Ferrer et al., 2017). It presents an overview of the problems of Sardinian linguistics with a detailed introduction of the current linguistic situation in Sardinia completed by a description both of the varieties of Sardinian itself and of the other languages spoken on the island.

3 Methodology

The work, still in progress, was performed manually taking advantage of the involvement and the linguistic expertise of some trainees belonging to the University of Cagliari, coordinated by the Prof. Maurizio Viridis.

The construction of the resource is based on the list of the 4689 Common Base Concepts expanded by BalkaNet from the initial set of 1024 Common Base Concepts developed in the European project EuroWordNet (Vossen, 1998)

SardaNet is created by using the *expand approach*, starting from the multilingual index and translating the English various synsets into the Sardinian language. This approach is more attractive since it maintains the multilingual index as the main structure and central repository of concepts and also allows to automatically using semantic relations already present in the English WordNet.

According to Bond et al. (2016), the majority of wordnets are based on the expand approach,

exactly 28 out of 33 of the wordnets included in the OpenMultilingual Wordnet (OMW)⁵.

3.1 Insertion and Validation

The collaboration with the trainees started by choosing the most suitable dictionaries and resources among those available as described in the Section 2.1. After then, an English term was selected and, for each of its meanings identified by a different synset ID and a gloss, the synonyms in the Sardinian language were assigned in all the variants considered.

The identification of Sardinian terms to be inserted into SardaNet in correspondence of the selected English terms was carried out through the consultation of the various Sardinian dictionaries. They display, besides the Sardinian term and its linguistic varieties, the definition of the same term in Italian or in Sardinian language and sometimes some examples of usage of the term, that help to understand its real meaning, and its translation into several languages, including Italian and English.

Each term inserted in SardaNet has been verified and confirmed by at least two people.

In case of discrepancy the team discussed in order to find an agreement and, when it was not possible, the terms involved were excluded.

3.2 Examples and Issues

During the building of the resource we have sometimes faced the problem of translation equivalence and the lack of correspondence of the Sardinian language with the about 5000 English senses.

As expected, some technical terms do not have correspondence in Sardinian language.

The term *mouse*, as a *hand-operated electronic device (synset ID WN3.1 = 03799022, noun)*, does not have a corresponding sense in the Sardinian language. Other terms, i.e. *website*, *a computer connected to the internet that maintains a series of web pages on the World Wide Web*, could be translate with the Sardinian terms *giassu* (L) and *zassu* (N).

In general, terms belonging to specific domains, such as biology or chemistry, do not have an equivalent term in Sardinian. The English term *state*, as a *chemical state of matter (synset ID WN3.1 =14503199, noun)*, does not have a Sardinian equivalent sense.

The sense of the term *cell*, as *electric_cell, a device that delivers an electric current as the*

⁵ <http://compling.hss.ntu.edu.sg/iliomw/omw>

result of a chemical reaction (synset ID WN3.1 = 02994503, noun), is not present in the available Sardinian dictionaries.

Sometimes the linguists have experienced difficulties to look for the right corresponding sense of some English concepts in WordNet. Despite of this, a concept such as *creating_from_raw_materials*, defined as *the act of creating something that is different from the materials that went into it* (synset ID WN3.1 = 00910607) could be simply translated in the Sardinian verb *bogai* (C).

We observed that the key factor is that many English terms, especially neologisms and technical terms, often not have a correspondence in the Sardinian language. So frequently Italian terms are used instead. On the contrary, there are many common saying, as part of the juvenile language (Ferrer et al., 2017), that hardly find correspondence not only in English but also in Italian.

4 Building SardaNet

In order to build SardaNet, it was necessary to set up an interface able to display for each term in English its synonyms, the corresponding synset IDs, the POS, the definition, and allowing the terms to be included in each of the variants of the Sardinian language.

4.1 The application

Developed in PHP, the application allows the insertion of Sardinian terms starting from the English ones into different ways.

It is an evolution of a previous application developed for FreeWordNet, (Tuveri and Angioni, 2012a; Tuveri and Angioni, 2012b), a linguistic resource, still not released, based on WordNet. FreeWordNet was born as a possible extension of WordNet in Opinion Mining related context.

In FreeWordNet each synset is enriched with a set of properties related to adjectives and adverbs and has a positive, negative or objective value associated. The properties associated to each synset support a better identification of the sentiment expressed in relation to the domain and give more details about the relevant terms or the expressions having an opinion associated. SardaNet inherits the same properties from adjectives proposed in FreeWordNet but they have not been inserted in this first release.

The interface permits to insert any word in English starting from the about 5000 word senses in the set of CBCs.

The starting form, shown in Figure 3, offers 3 different options.

Figure 3. The access interface.

In the first one, followed by an use case, the user “Micol” can modify the entered information by indicating the term.

In the second, the user “Micol” will verify and confirm the items previously entered by another user, by simply submitting the button related to the “Term Validation”. The application allows to show only the terms to be validated, that is, those entered by a person and that needs to be confirmed by at least another person. During this process, the user can also erase incorrectly entered terms or variants.

English Synonyms	Sardinian Terms	Quality	Category	SynsetID	Glosses
cell:	<ul style="list-style-type: none"> cella-CLN cella-GNS cella-CB 	mean_type		0004824	(biology) the basic structural and functional unit of all organisms; cells may exist as independent units of life (as in mammals) or may form colonies or tissues as in higher plants and animals.
cell; electric cell:		mean_artefact		0258439	a device that delivers an electric current as the result of a chemical reaction
cell; jail cell; prison cell:	<ul style="list-style-type: none"> cella-CLN cella-GNS cella-S cella-3 	mean_artefact		02584775	a room where a prisoner is kept

Figure 4. Mapping cell-related terms.

In the third option, the set of the 4689 CBCs is displayed. The user can select each term in English and insert the corresponding terms in the

Sardinian language. The interface will show all the synsets related to the selected word in English in term of polysemy, the definition, the synset ID and the gloss, as the Figure 4 shows. New terms entered or confirmed are saved in the database.

In the Figure 3 the access form to SardaNet is shown. The user “Micol” looks for the noun “cell” in SardaNet.

Figure 4 shows the already inserted Sardinian terms, having possible cellular related meanings, and the information provided by WordNet 3.1. The inserted values can also be changed.

By selecting one of the links referring to the noun “cell”, the user can insert new corresponding nouns in Sardinian language, delete or edit the existing ones, for example by modifying the associated language varieties, as shown below in Figure 5.

cèllula		
Term: cèllula		
A <input type="checkbox"/>	C <input type="checkbox"/>	N <input type="checkbox"/>
S <input checked="" type="checkbox"/>	T <input type="checkbox"/>	G <input checked="" type="checkbox"/>
L <input type="checkbox"/>		
Submit		Delete the Term

Figure 5. How to modify the language varieties of a term.

4.2 Format and CILI Mapping

The used base concepts include the WordNet synsets in version WN2.0. It was therefore necessary to mapping the synsets from this version to WN3.1 version. This work has been carried out in two steps, from version WN2.0 to WN3.0, from WN3.0 to WN3.1.

The mapping from version WN2.0 to WN3.0 of WordNet was done thanks to the work performed by Tufiş et al. (2011), through the resource available at <http://nlptools.racai.ro/>.

The mapping from the WN3.0 to the WN3.1 WordNet version has been made possible thanks to the work of John McCrae, through the *git* project: <https://github.com/globalwordnet/ili>.

Thanks to his work we put together entirely the mapping WN2.0 - WN3.0 - WN3.1, ILI indexes included.

The starting set of about 5000 CBCs is expressed in WN2.0 and a mapping to the WN3.1 version and to the associated ILI indexes has been necessary.

A subset of them has been used in the first edition of SardaNet and we decided to define the resource, right from the beginning, in the LMF

(Lexical Markup Framework) format, as required by the Global WordNet Association.

The mappings are also available on request in the SQL format too.

We try to remain as faithful as possible to the CILI, Collaborative Interlingual Index (Bond et al., 2016), but the building of the LMF evidenced almost an additional requirement related to the SardaNet linguistic variants that has been indicated adding the “Tag” element to the “Lemma” item, built in the following way:

```
<Lemma writtenForm="bóitu" partOfSpeech="n">
  <Tag category="variants">ACGL</Tag>
</Lemma>
```

Figure 6 shows a portion of SardaNet in LMF format with the addition of the linguistic variant.

```
<LexicalEntry id="wn31-cèllula-n">
  <Lemma writtenForm="cèllula" partOfSpeech="n">
    <Tag category="variants">ACGL</Tag>
  </Lemma>
  <Sense id="wn31-00006484-n" synset="wn31-00006484-n">
    <SenseRelation target="wn31-00327929-a" relType="derivation"/>
    <SenseRelation target="wn31-02696036-a" relType="derivation"/>
  </Sense>
</LexicalEntry>
```

Figure 6. The LMF format with the addition of the Sardinian variant.

However, in the first release of the resource in the LMF format we leave out the glosses in the Sardinian language. In fact, there is a scarcity of Sardinian online dictionaries and the most complete available ones not always have a gloss or a sentence defining the usage of the specific term.

An automatic process is not always possible, so we will probably proceed manually with the help of students of Linguistics.

Currently we are also not able to calculate the frequency of use of the Sardinian terms for each synset. So, in the first release of SardaNet it is not provided.

4.3 Current Status of SardaNet

The quantitative data pertaining to the Sardinian WordNet are summarized in the tables below.

Table 1 shows the couple of synsets and Sardinian terms inserted into SardaNet, the total number of distinct synsets and the number of distinct Sardinian terms.

As you can notice, SardaNet includes a lot of terms for each synset. It is due both to several synonyms related to each sense, typical of the

Sardinian language, and to the presence of the several variants of the language.

Synsets -Terms	Distinct Synsets	Terms
21025	1601	9899

Table 1. Synsets and terms in SardaNet.

The following Table 2 reports the distribution of terms and synsets in SardaNet for each part of speech (POS), referred to the number of couple of synsets and Sardinian terms, the number of validated synsets and the number of Sardinian terms.

POS	Synsets Terms	Distinct Synsets	Terms
Nouns	18885	1452	8920
Adjectives	1657	130	685
Verbs	483	19	294

Table 2. Distribution of synsets and terms for POS.

The results above show the prevalence of nouns among the other parts of speech. Translating English nouns into Sardinian nouns seems to be more intuitive and immediate and involves fewer problems than verbs, adjectives and adverbs. The resource does not yet include any adverb.

Variant	Synsets Terms	Distinct Synsets	Terms
C	5622	1584	2787
G	4097	1550	1947
L	8219	1590	3828
N	5738	1581	2738
S	3076	1560	1505

Table 3. Distribution of the Sardinian variants in SardaNet.

As depicted in Table 3, among all the Sardinian variants, Logudorese, Nuorese, Campidanese, Gallurese and Sassarese include in SardaNet the largest number of terms, while Algherese and Tabarchino are not yet considered in the resource.

Despite the application does not calculate the correct percentage of the coverage across the dialects, we found that the total coverage of validated terms in SardaNet is about 16,5%, 775

senses on 4689 of the CBCs. Nevertheless SardaNet contains a total of 1601 senses, 826 not included in the CBCs. These senses come out because the application shows, for each sense included in the CBCs, all the senses related by the polysemy property.

5 Conclusions and Future Works

In its first release, SardaNet includes only a partial set of the 4689 Common Base Concepts expanded by BalkaNet from the initial set of 1024 Common Base Concepts developed in the European project EuroWordNet. So we are first planning to complete all the senses available in the set of CBCs. Although there are many terms, common saying and phrases typical of Sardinian language, they are not currently present in SardaNet. We leave out the glosses in the Sardinian language and the frequency of use of the Sardinian terms for each synset, even if they are required in the LMF format.

Further works will include both the glosses and the frequency of the terms, that will be calculated both manually, using the available dictionaries, and automatically by means of a corpus of Sardinian documents. We are also taking into account to enrich SardaNet with new terms, currently not included in the English WordNet, but characteristic of the Sardinian language.

Acknowledgments

We wish to thank Georgia Sanna for the first joint analysis about the possibility of creating a wordnet for the Sardinian language.

References

- Pietro Casu. 2002. *Vocabolario sardo logudorese-italiano*. Ed. Giulio Paulis, Nuoro, Ilisso, ISBN 88-87825-36-X.
- Francis Bond, Piek Vossen, John McCrae, Christiane Fellbaum. 2016. *CILI: the Collaborative Interlingual Index*, in: Proceedings of the 8th Global WordNet Conference 2016 (GWC2016) in Bucharest, Romania, January 27-30.
- Eduardo Blasco Ferrer, Peter Koch, Daniela Marzo. 2017. *Manuale Di Linguistica Sarda*. Ed. de Gruyter Mouton. ISBN 3110274507, 9783110274509
- George A. Miller. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.

- Giulio Paulis. 1984. *Introduzione a M.L. Wagner, Fonetica storica del sardo (trad. italiana di Historische Lautlehre des Sardischen, Halle, Niemeyer, 1941)*. Casteddu/Cagliari, Gianni Trois Editore, pp. VII-CX.
- Vincenzo Raimondo Porru. 2002. *Nou dizionariu universali Sardu - Italianu*. A cura di Lörinczi Marinella. Ilisso Edizioni, Nuoro.
- Mario Puddu. 2015. *Ditzionàriu de sa limba e de sa cultura sarda*. 2896 p., ed. Condaghes, 2 ed., Collana: Ainas.
- Antoninu Rubattu. 2001. *Dizionario universale della lingua di Sardegna*. EDES. Collana: Lingua e letteratura. 2 voll., 2254 p., EAN: 9788886002394.
- Giovanni Spano. 1998. *Vocabolariu Sardo-Italianu e Vocabolario Italiano-Sardo*. 4 vol. Ed. Giulio Paulis, Nuoro, Ilisso.
- Dan Tufiş, Dan Cristea, Sofia Stamou. 2004. *BalkaNet: Aims, methods, results and perspectives. A general overview*. Science and Technology 7(1/2): 9–43.
- Dan Tufiş, Radu Ion, Verginica Barbu Mititelu, Elena Irimia, Dan Ştefănescu, Cătălin Mihăilă. 2011. *Extending and completing the Ro-WordNet lexical ontology by eliminating the existing semantic conflicts and by validating the differential semantics model based on Ro-WordNet*. Academic report (in Romanian). Bucharest, Romania.
- Franco Tuveri, Manuela Angioni. 2012a. *Definition of a Linguistic Resource for Opinion Mining*. Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science – NLPCS - SciTePress - june 2012. ISBN: 978-989-8565-16-7
- Franco Tuveri, Manuela Angioni. 2012b. *A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs*. Proceedings of the 6th International Global Wordnet Conference Tribun EU pages 365-370 Christiane Fellbaum, Piek Vossen Global WordNet Association. ISBN: 978-80-263-0244-5
- Maurizio Viridis. 1978. *Fonetica del dialetto sardo campidanese*. Cagliari, Edizioni della Torre.
- Maurizio Viridis. 1988. *Sardisch: Areallinguistik (aree linguistiche)*. In: Lexikon der Romanistischen Linguistik, vol. IV, Italienisch, Korsisch, Sardisch. A cura di Günter Holtus, Michael Metzeltin, Christian Schmitt, Tübingen, Niemeyer, 1988, pp. 897-913.
- Maurizio Viridis. 1982. *Note sui dialetti dell'area arborense e la lingua del Condaghe di Santa Maria di Bonarcado*. In il Condaghe di Santa Maria di Bonarcado, riedizione del testo di Enrico Besta a cura di Maurizio Viridis, Oristano, S'Alvure.
- Maurizio Viridis. 2003a. *La lingua sarda fra le lingue neolatine: storia uso e problemi*, in: *La lingua e la cultura della Sardegna*. Rapporto del Convegno internazionale. La lingua e la cultura della Sardegna. Tokyo, 9-10 maggio 2003. (pp. 15-24). TOKYO: Waseda University (JAPAN).
- Maurizio Viridis. 2003b. *Tipologia e collocazione del sardo tra le lingue romanze*. In «IANUA. REVISTA PHILOLOGICA ROMANICA (on line)», 4.
- Piek Vossen (ed.). 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
- Max Leopold Wagner. 1960-64. *D.E.S. - Dizionario etimologico sardo, DES*. 3 Vol. Heidelberg, Winter.
- Max Leopold Wagner. 2008. *D.E.S. - Dizionario etimologico sardo*. 2 Vol. Ed. Giulio Paulis, Nuoro, Ilisso. ISBN 978-88-6202-030-5.

Extraction of Verbal Synsets and Relations for FarsNet

Fatemeh Khalghani

Faculty of Computer Science and
Engineering
Shahid Beheshti University
Tehran, Iran.
f.khalghani@gmail.com

Mehrnoush Shamsfard

Faculty of Computer Science and
Engineering
Shahid Beheshti University
Tehran, Iran.
m-shams@sbu.ac.ir

Abstract

WordNet or ontology development for resource-poor languages like Persian, requires composition of several strategies and employment of appropriate heuristics. Lexical and linguistic structured resources are limited for Persian and there is a lot of diversity and structural and syntagmatic complexities. This paper proposes a system for extraction of verbal synsets and relations to extend FarsNet (Persian WordNet). The proposed method extracts verbal words and concepts using noun and adjective words and synsets. It exploits the data from digital lexicon glossaries, which leads to the identification of 6890 proper verbal words and 2790 verbal synsets, with 91% and 67% precision respectively. The proposed system also extracts relations such as semantic roles of verbal arguments (instrument, location, agent, and patient) and also “related-to” (unlabeled) relations and co-occurrence among verbs and other concepts. For this purpose, a combination of linguistic approaches such as morphological analysis of words, semantic analysis, and use of key phrases and syntactic and semantic patterns, corpus-based approach, statistical techniques and co-occurrence analysis have been utilized. The presented strategy extracts 5600 proper relations between the existing concepts in FarsNet 2.0 with 76% precision.

1 Introduction

Semantic or conceptual relation extraction between concepts and appropriate relation labeling forms an important part of ontology learning and

ontology construction process that is widely used in information retrieval, question-answering systems, summarization, and word sense disambiguation (WSD) (Girju, 2008).

Learning and labeling of conceptual relations has been introduced as the most complex and challenging element in most of systems, especially in the construction of ontologies or WordNets (Sánchez & Moreno, 2008; Kavalec & Svátek, 2005). This problem can be divided into two separate parts of relationship extraction and labeling. The latter which tries to label an existing unlabeled relation between two concepts has been less addressed in previous studies.

Semantic analysis requires composition of various approaches like pattern-based and corpus-based techniques for languages such as Persian that lack accurate analytical instruments and structured sources and suitable tagged corpora. Thus, several lexical resources including syntactic verbal valency lexicon (Rasooli et al., 2011), comprehensive lexicon of synonyms and antonyms (Khodaparasti, 1997), online and digital lexicons such as Vajehyab browser¹, Wikipedia, Dadeqan dependency Treebank (Rasooli et al., 2013) and Wortschatz statistical corpus (Goldhahn et al., 2012) have been used in the presented strategy. In addition, Persian preprocessing tools such as Negar text editing tool, STeP-1 morphological analyzer (Shamsfard et al., 2010) and ParsiPardaz dependency parser (Sarabi et al., 2013), and a composition of linguistic, syntactic, and statistical approaches have been used for semantic relation extraction.

As verbs are the main core of sentences in many languages, extending the verbal part of wordnets may improve their efficiency and application in semantic analysis of texts.

¹ <http://www.vajehyab.com>

This paper focuses on extraction of verbs, verbal synsets and non-taxonomic relations in which at least one of the related terms is verb.

The given strategy in this paper for extraction of verbal concepts emphasizes on wide range of compound verbs and prefixes in Persian with highly metaphorical concepts, and in addition to implementation of verbal construction rules and paying attention to Arabic rhythms of words, starting from concepts of noun and adjective, it derives correspondent verbal concepts from several online lexicons with analyzing of entries and text of explanations and examples in thesauruses.

The rest of the paper is organized as following: section 2 present a brief introduction to FarsNet and its current situation, section 3 discusses related work, section 4 describes the proposed method including verb extraction, verbal synset composition and verbal relation extraction. Section 5 concludes the paper and suggests some further work.

2 FarsNet

FarsNet, the first Persian WordNet (Shamsfard et al., 2010) is a lexical database for Persian words. The first and second versions of FarsNet have been established in Natural Language Processing lab of Shahid Beheshti University at 2008 and 2010 respectively. FarsNet 3.0 which is currently under development is expected to have 100,000 lexical entries (currently about 87,000 are available). FarsNet like other wordnets is formed by a large set of lexical entries (words or phrases) organized in a network of synsets (a set of synonym terms). The edges of this network are semantic relations among synsets, including inner-POS and inter-POS ones. The relations defined between synsets in FarsNet include hypernym/hyponym, holonym/ meronym, antonym, domain, related-to, co-occurrence, cause, entails, salient-defining feature, potential-defining feature, unit and attribute; besides, some semantic roles as instrument, location, agent and patient. The report of variation trend of FarsNet versions and the existing semantic relations has been presented in (Shamsfard & Ghazanfari, 2016). Also, Table 1 displays statistics of words, synsets, and the relations between senses and synsets in various versions.

FarsNet version	Words	Word senses	Synsets	Synset relations	Sense relations
1.0	17842	24480	10012	6980	360
2.0	30222	36115	19398	36848	7043
2.5	33290	39735	20559	47761	19021
current	86747	98370	37959	91744	28739

Table 1: Statistics of words, synsets and relations of FarsNet

The strategies given in this paper have been adapted to extend and improve FarsNet 3.0 verbal synsets and relations.

3 Related Work

In the related field of automatic wordnet development, several efforts have been made. According to classification of Vossen (1998) for wordnet development approaches, two major approaches can be considered as merge and expansion. The merge approach is constructing a wordnet with independent use of target language resources and language specific properties and usually creating synsets from scratch, whereas the expansion method relies on existing wordnets (especially the English WordNet) and uses multilingual resources to translate words of existing synsets to target language and therefore preserves the source wordnet structure. However developing a wordnet by use of merge method is not always cost effective due to budget constraints and is more time-consuming than expansion method, it leads to a higher quality and extensive wordnet to be effectively used in certain and real NLP applications. Also a wordnet developed with merge approach will reserve the target language culture and region specific concepts and semantic relations and there is no need to deal with translation ambiguity, compared to expansion approach (Prabhu et al. , 2012).

Prabhu et al. (2012) use a hybrid approach of merge and expansion for developing IndoWordNet to benefit from the advantages of both.

Recently, word embedding models, especially Word2Vec (Mikolov et al., 2013), have been the focus of much research in NLP tasks. These models are widely used to calculate semantic relatedness of words and thus they can be applied in synset construction and semantic relation extraction subtasks of a wordnet development process.

The proposed work by Al Tarouti (2016) on Arabic wordnet and Mousavi and Faili (2017) on Persian wordnet use vectors created by Word2Vec to move towards a wordnet by an expansion method.

There are some other efforts to build a wordnet for the Persian language by either semiautomatic or automatic methods. Among semiautomatic ones we can mention (Bagherbeygi & Shamsfard, 2012), (Fadaei & Shamsfard, 2010)

and (Shamsfard, et al., 2010) which mainly use a merge method to build a Persian wordnet.

Among automatic methods we can mention (Dehkharghani & Shamsfard, 2011), and (Taghizadeh & Faili, 2016) which mainly use expansion methods and extract some mappings between Persian words and Princeton synsets. These systems do not build a wordnet but can be used to initiate building a wordnet. They are good in coverage and development time but not as well in precision of result.

Most of the research conducted on extraction and labeling of conceptual relation for Persian language (such as (Shamsfard & Barforoush, 2004) and (Fadaei & Shamsfard, 2010)) work on a limited predefined relations such as synonymy, hyper/hyponymy and holo/meronymy relations; and they lack favorable and needed efficiency for non-taxonomic relations and those relations corresponding to semantic roles (role relations).

Boudabous (2013) proposes a linguistic method based on morpho-lexical patterns to extract semantic relations in order to improve the Arabic WordNet (AWN) performance using Arabic Wikipedia articles as the input corpus.

The methods proposed by Shamsfard & Mousavi (2008) and Jafarinejad & Shamsfard (2012) carry out labeling thematic roles in sentence through rule-based approaches using shallow parsing of text; the mentioned conducted works lack favorable efficiency for extraction of conceptual relations among FarsNet high-level concepts and they don't have appropriate recall either. The work done by Zadeh Khosravi Forooshani & Rezaei Sharifabadi (2016) carries out semantic role labeling in Persian sentences by dependency parsing; that in comparison to works implemented with shallow-parsing, has higher accuracy and better efficiency; but it does not yet propose any solution for extraction of corresponding semantic relations among high level concepts of a wordnet.

The strategy suggested by Bagherbeygi & Shamsfard (2012) is one of the works conducted for automatic extraction of Persian verbal concepts in which FarsNet noun and adjective concepts are used for compound verbs extraction. It considers any combination of each noun/adjective and Persian light verbs as a compound verb candidate and then verifies correct words by checking up in Bijankhan Corpus and Arianpour Dictionary. Then it makes verbal synsets by a rule based method from noun and adjective synsets. Despite appropriate efficiency of this technique in derivation of phrasal verbs,

with respect to reliance of this method on combination of noun and adjective with light verbs and limitation of the used lexical sources, many prefixed and propositional verbs as well as more complex expressions and verbal phrases with metaphorical concepts are not identified.

4 The Proposed Method

4.1 Verbal Synset Extraction

The proposed method for verbal synset extraction uses the existing noun and adjective synsets and it is focused on the principle that automatic learning of concepts by starting from synsets instead of words, reduces processing and time costs for building synsets and extension of database.

The basic concept of the proposed approach is to consider the internal structure of the phrasal and prefixed verbs and verbal and phrasal terms with metaphorical concepts; this has led to expansion of verbs in Persian. The non-verbal parts of compounds are derived from the noun and adjective concepts and the appropriate verbal parts and prepositions and prefixes should be extracted.

To this end, we first consider each noun synset and for each noun in it, apply some rules to find a corresponding verbal concept based on its semantic category, grammatical structure and Arabic rhythm (for words with Arabic origin). Considering all semantic classes of nouns, we extracted the semantic classes for which verb extraction is possible. These classes are act, attribute, possession, motive, feeling, event, cognition, state, relation, and process.

Afterwards, based on structural rules of Persian and Arabic gerunds, the verbal and non-verbal parts are derived for the words for which these rules are applicable. Some of these rules are as following:

- Words with *Fe'Alat* rhythm such as *TebAbat* (طبابت: medicine), *VekAlat* (وکالت: attorneyship), and *KetAbat* (کتابت: writing) can be combined with the light verb *Kardan* (کردن: to do) to make a compound;
- Words with *Fa'Al* rhythm e.g. *KaffAsh* (کفاش: shoemaker), *AkkAs* (عکاس: photographer), and *NaqqAl* (نقال: narrator) can be used to make a phrasal verb by the rule *word+ ye+ kardan* (ی + کردن).
- Words with suffix *Gari* (گری ~) can be participate in verb construction with/without deletion of suffix before adding to *Kardan*, e.g. *Efsha Kardan* (افشا کردن: to disclose) from

EfshA+Gari (افشاگری: disclosure); and *Soda Kardan* (سودا کردن: to speculate) from *So-dA+Gari* (سوداگری: speculation).

- From words with suffix *Gi* (گی ~) proper verbs can be made by a set of rules. For example *Kooftan* or *Koofteh Shodan* (کوفته شدن: to concuss) from *Kooftegi* (کوفتگی: concussion) and *RAnandegi Kardan* (رانندگی: to drive) from *RAnandegi* (رانندگی: driving).
- From the combined words whose structure ends to (present lemma + ی (i)) the corresponding verbs can be obtained by substituting the (present lemma + i) with its corresponding gerund form. *Taj GozAshtan* (تاج گذاشتن: to crown) from *TAjgozAri* (تاج گذاری: crowning), and *Tasmim Gereftan* (تصمیم گیری: to decide) from *Tasmimgiri* (تصمیم گیری: decision making).

Some of noun words which are very numerous do not follow any certain rule; but they have participation in structure of phrasal verbs as verbal part(s). For example, making verb of *Habs Keshidan* (حبس کشیدن: to imprison) from noun of *Habs* (حبس: prison) and *Shak DAShtan* (شک داشتن: having suspicion) from *Shak* (شک: suspicion) and *Be Haghighat Peyvasthan* (به حقیقت پیوستن: to come true) from word of *Haghighat* (حقیقت: truth); extracting a rule for these cases is not easy and they can be validated through analysis on lexicons and the related corpora. For this purpose, the suitable verbal part can be obtained for each noun non-verbal part of a compound verb automatically by searching for the word(s) in entries and body of the group of digital lexicons including Khodaparasti Glossary, Moein Thesaurus, Dehkhoda Dictionary, Amid Thesaurus, and Glossary of Refined Words and by benefitting from Vajehyab Dictionary Browser.

In the next step, the verbs obtained from each noun in a synset, are considered as candidates for making a synset; moreover, for each verb in the synset, its synonyms can be extracted from available lexical sources to participate in the synset. After completion of the verbal synset an unlabeled relation (related-to) will be held between the original noun synset and the derived verbal synset.

Finally after completion of automatic phases, with respect to error possibility, expert supervision for synset verification would be necessary. The possible errors might comprise non prevalence (obsolescence) of the generated verb by means of grammatical rules, and or non-idiomaticness of the verb obtained by surveying

in glossaries of the day. It is also possible that the obtained verb might be very specific and rarely used. For example, Persian verb *Derakhshandegi Kardan* (درخشندگی کردن) (to do brighten) that has been derived by means of grammatical rules is not correct, or verb of *KhAb Dookhtan* (خواب دوختن) (To sew sleep) that is found in glossaries is not used today. The other error is forming a synsets with words with similar structure and non-verbal part but different meanings. For example Persian phrasal verbs such as *KhAb Raftan* (خواب رفتن: to go asleep), *KhAb Beh KhAb Raftan* (خواب به خواب رفتن: to die in asleep), and *KhAb Didan* (خواب دیدن: to see or have night dream) that are all derived from Persian term *KhAb* (خواب: sleep) each one has a separate meaning. With respect to these errors, we conclude that building verbal synsets from noun synsets is not 100% automatically feasible and expert supervision and analysis would be inevitable; nevertheless, the approach used might be highly efficient in automatic extraction of new and synonymous terms and the data obtained might be efficient in reducing processing size and time spent for building synsets.

4.2 Non-Taxonomic Relation Extraction

The proposed method for extracting non-taxonomic relations employs lexical sources, various tools and combination of different linguistic, syntactic and statistical methods to improve efficiency and to increase precision and recall. This system primarily extracts a pair of concepts in semantic relation in concept pair extraction subsystem. The type and label of some of the relations can be identified during the concept pair extraction phase. For others, the labeling is postponed to the next phase and just adds the pair as “related-to” into the candidate set. These unlabeled relations will go through the labeling subsystem to determine their labels. These subsystems and their algorithms are discussed in this section.

• The concept pair extraction subsystem

To extract concept pairs with a semantic relation we applied three methods:

In the first method, all synsets with words containing any derivational form of a verb, Arabic rhythms, and keywords denoting a semantic role (location, instrument, agent, and patient) have been extracted from FarsNet. Then for each of the above words their corresponding verb (e.g. with the same stem) is extracted. The word and its corresponding verb make a concept pair to be

used as the input of further morphological and semantic analysis.

For instance, between concept pair of *DastgAhe Tasfiyeh HavA* (air refinement system: دستگاه تصفیه هوا) and *Tasfieh Kardan* (to refine: تصفیه کردن) there is an “instrument” relation; and between the concept of *PanAhgAh* (shelter: پناهگاه) and verbs *PanAh DAdan* (to shelter: پناه دادن) and *PanAh Bordan* (to take refuge: پناه بردن) there are “location” relations; and in concept pair of *NAzer* (supervisor: ناظر) and *NezArat Kardan* (to supervise: نظارت کردن) there is an “agent” relation. All of these relations can be extracted by morphological analysis according to derivational affixes or Arabic patterns (rhythms).

Lexico-semantic analysis of synset glosses is another technique to extract related concepts. In this method a group of lexico-semantic patterns and key phrases correspondent to each of the semantic roles has been utilized for semantic analysis of glosses. After using a verb detection module to detect simple and compound verbs in the gloss, some patterns are used to extract the relation between the synset and the detected verb. For example, the synset of *Rahbar* (leader: رهبر), *Rahnama* (guide: راهنما) and *SarjonbAn* (mentor: سرجنبان) is defined as “*someone who leads and commands*”. Applying the agent patterns on this gloss lead to extraction of an “agent” relation between the synset and the verbs “*Hedayat Kardan* (to guide: هدایت کردن) and *FarmAn DAdan* (to command: فرمان دادن)”. As another example the synset of “*HammAm* (bathroom: حمام) and *GarmAbeh* (bathhouse: گرمابه)” is defined as “*a location that is built for washing body*”. Using a location pattern leads to extracting a “location” relationship between the synset of bathroom and the synset of wash (شستشو کردن).

The other approach for extraction of a concept pair participant in semantic relation is to consider all of the existing verbal synsets (concepts) in FarsNet as the first input and obtaining the second selected concept by means of the following statistical approach. The input of the statistical module is the set of all words (with all of their written forms) Then using Wortschatz statistical corpus for each verb, its co-occurrent nouns are derived and sorted according to their frequency. This way the most frequent co-occurrent nouns to each verb are extracted. But we need a synset as a member of concept pair not a word. Thus we extract all the synsets which include the co-occurrent noun as a candidate and at the next steps employ a Word Sense Disambiguation

(WSD) module to determine the suitable sense (synset).

• Semantic relation labeling subsystem

In the previous steps some concept pairs (a pair of two synsets with a relation among them) were extracted and some of their relations were labeled during the extraction process. In this step we are going to extract more labeled relations or label some remained unlabeled ones. In order to enhance precision and recall in the system we employed several aforesaid sources. In this step we first find dependents (synonyms, hyper/hyponyms and instances) for the input concept pair. Then we label the relations between the concept pairs and their dependents. These two steps will be discussed in more details in the following.

- Finding dependents for input concept pair

In order to derive dependent for each of input concepts, we have used various sources including FarsNet synsets, Khodaparasti lexicon, and also redirect pages in Wikipedia to find synonymies and FarsNet taxonomic relations, and Wikipedia categories and subcategories to achieve hierarchical relations as father and child concepts for any concept.

We execute a shallow preprocessing on the given dependents to improve system efficiency including text normalization and unifying various word forms, omission of inflectional affixes, refinement of additional descriptors and finding of NP head especially for Wikipedia categories.

After determination of dependents, the labels are acquired for semantic relations among input concept pair and pair of dependent concepts by various techniques. The used approaches include morphological analysis, employing syntactic patterns, and adjustment of these patterns for identifying semantic roles which are discussed in the following.

- Morphological analysis module

We have utilized STeP-1 stemmer and morphological analyzer as the main tool in this module. This module tries to find stems and derivational affixes for any input term. We have prepared anaffix lexicon-and a rich set of morpho-patterns that covers various types of derivational affixes for combining with noun, adjective and verb stems.

Likewise, we also utilize a group of pattern or templates (rhythms) in Arabic language from which many words have been made in Persian. These rhythms include construction patterns for

gerund, noun of place, nominative noun, past participle, and noun of exaggeration. For instance for active participles of *NAzer* (supervisor: ناظر) or *TAjer* (merchant: تاجر), these gerunds are derived *NezArat* (supervision: نظارت) and *TejArat* (trade: تجارت) and they refer to “agent” semantic role. The noun of exaggeration also usually refers to a job. For example, the label for relation among *KhayyAt* (tailor: خیاط) and *KhayyAti Kardan* (to sew: خیاطی کردن) is also an agent.

Each word, after morphological analysis is examined for inclusion of an entry of the affix lexicon or obeying of Persian morpho-patterns or Arabic templates (rhythms) and if it is composed of one of meaningful derivational affixes, proportional to the semantic role, a semantic label would be attached to it. Then all the words in a concept pair and are compared with each other. If they have any common infinitive stem we may be able to extract new relations among them. For example consider that the words *ArAyeshgar* (barber-hair dresser: آرایشگر) and *ArAyeshgAh* (barber shop: آرایشگاه) appear in a concept pair. As they have the common infinitive stem *ArAyesh kardan* (hair dressing: آرایش کردن) and the first is the agent and the second is the location of this act, we can include that there is a “location” relation held in this concept pair. We have employed verbs valency lexicon in order to find the dependent stems of an infinitive.

Whereas STeP-1 stemmer does not analyze compound nouns and verbs, thus we have improved function of stemmer for morphological analysis of compound words. For example in the initial stemmer, some terms like *DAneshAmooz* (student: دانش آموز), *Ashpaz* (cook: آشپز), and *GolkAr* (gardener: گلکار) are identified as single noun or adjective words; while it will be very useful for labeling their corresponding relations, if they are analyzed into constructional terms with saving all constituent stems. We have utilized glossary of verbs to solve this problem and we check ending of noun or adjective compound words with present stems. If the compound word passes the check, we save the stems of both terms as stem of the given word and create a semantic label of “agent” for infinitive of the present stem. For example, label of relation among concept pair *GolkAr- KAshtan* (gardener-to plant: گلکار- کاشتن) would be “agent”.

Similarly, input concepts that are noun phrases are analyzed in this module in terms of presence of keywords correspondent to role relations. For example, many categories are expressed in Wikipedia pages by descriptors e.g. “*VasAyeI- Va-*

sileh- abzAr- abzArAlAt- LavAzem- TajhizAt (devices- means- tools- apparatuses- equipment: وسایل-وسيله-ابزار-ابزار آلات-لوازم- (تجهيزات) and or most of concepts in FarsNet include descriptors e.g. “*MakAn- Mahal- Zamin- Mo’asseseh- OtAgh- EdAreh- Sherkat* (place-location-land-institute-chamber-department-company: مکان-محل-موسسه-اتاق-اداره-شرکت)”. Therefore, proposing an approach for morphological analysis on them increases system recall. To this end, we save any word, including one of the given descriptors with correspondent semantic label e.g. instrument and location and stem of the term after descriptor. In order to achieve its semantic relations with the other input concepts we act similar as above-said process. For example, “instrument” will be assumed as label for relation of concept pair of *TajhizAt SAKhtemAni- SAKhtan* (constructional equipment- to build: تجهیزات- ساختن).

- Syntactic analysis module and dependency analysis

Dependency treebanks include a group of sentences which have been analyzed according to dependency command, and generally verb of sentence is selected as root and origin and the relation of other words of the sentence with each other and the verb would be characterized. These corpora are considered as rich sources for finding deep syntactic patterns and the resulted processing would be highly accurate; though frequency of occurrence and recall in them is not that much high.

The studied concept pair is analyzed in terms of nature (being noun or verb) after entering into this module; for this purpose we employ stemmer and also utilize lexicon of verbs to identify the compound verbs. Then, we survey corpus to find sentences including both of them. Whereas the concept may occur in corpus in singular or plural form, or other inflection such as a noun preceding an unknown Persian article (*Ya-e-Nakareh*: ی نکره) and also Dadegan dependency treebank comprises of root of words in sentence, therefore, input concepts are compared with the specified roots in corpus as well.

By finding dependency of noun on verb and application of some rules and conditions and adjustment of semantic patterns to syntactic patterns, we label these relations for semantic role of noun to input verb. For instance, if the dependency of concept-to-verb relation is of subjective and the given verb is of active voice the label of conceptual relation or semantic role will correspond to agent, and if the verb is of passive

voice the label will be of patient type. For example, in this sentence: “*Flags were hoisted as symbol of lament*”, the concept of “flag” will have role of “patient” for the concept of “to hoist”. Likewise, the additional composition including infinitive is examined with left and right neighbors; for example, the label of hidden relation in additional composition *FAsh Kardan RAz* (to disclose secret: فاش کردن راز) will denote “patient”.

In order to find supplementary relations and to increase precision and efficiency of labeling system, if a noun is related to a preposition with a verb, that preposition is also used for semantic analysis and identifying of label of relation. For instance, Persian prepositions like *BA-Dar-Az-Tavasot-Bevasileh* (with- in- from- to- via- by: با- در- از- به- توسط- به وسیله) can represent various semantic roles, for example label of role relation for concept pair ‘*Goldoozi Kardan-Charkh-e-KhayyAti*’ (needlework- sewing machine: گلدوزی کردن- چرخ خیاطی) with respect to the presence of preposition *Ba* (by: با), through participation with them and using of semantic category and semantic analysis of the gloss for concept of “*sewing machine*” would be determined as “instrument”.

The other technique which has been designed in this module to determine semantic relation among input concept pair comprises of using ParsiPardaz tool for dependency analysis of example sentences of any synset in Synset table of FarsNet database. After dependency analysis of these sentences, we act as what was mentioned above and determine label of relation by adjustment of syntactic and semantic patterns.

- **Word sense disambiguation module**

Finally, after identifying and labeling conceptual relations among a concept pair, it is necessary to adapt a method for selecting the best and most appropriate synset for ambiguous words. To this end, a method has been designed that preserves recall and efficiency of the system while having reasonable precision. In this technique, we primarily select the appropriate synset among candidates according to their semantic categories and its relation to the label of the identified conceptual relation; for instance, if we embed word *Cinema* (سینما) in a “location” relation we expect that its corresponding synset has *location* in its semantic category.

In the next step, we apply a Lesk-like algorithm for WSD. To find the most appropriate synset for a polysemous word or for a new synset to be merged with, we compare the word (or words in the new synset) with the words in the

gloss and example of candidate synsets after omitting the stopwords; the synset with more common words is more appropriate.

The precision of this method is low when the candidate synsets have just one word or if the candidate synsets are semantically similar and so there is textual similarity between their glosses and examples. In these cases human supervision is needed to resolve the ambiguity. For instance, there are several synsets semantically close together for these words *NaghAsh* (painter: نقاش) and *Rang Kardan* (to paint: رنگ کردن) or words of *BANk* (bank: بانک) and *Poul* (money: پول) that makes difficult automatic recognition of the most appropriate synset. therefore presence of these commonalities in glosses and examples of all of them makes automatic recognition of the most appropriate synset difficult.

5 Results and Conclusion

This paper discusses the application of various automatic linguistic, syntactic and statistical methods on various resources to extend FarsNet by a merge method. The proposed method not only has reasonable precision and coverage, but also covers culture and language specific concepts and relations which cannot be captured by expansion methods. It can either extend the existing verbal synsets by a new verb or create a new synset for new and specific verbs of Persian lexicons with metaphorical meanings

This strategy significantly increases recall and the number of verbs and extracted semantic relations. Although it is applied to Persian, it can be used for extracting and labeling semantic relations in other languages as well.

The experimental results show that the proposed verb-extraction method, extracts 6890 correct verbs - regardless of polysemy and number of senses for each word and add them to FarsNet 2.0—that already had 7820 verbs. The synset extraction method added 2790 verbal synsets to 3670 verbal synsets existing in FarsNet 2.0. The synsets need manual judgment and semantic disambiguation of senses by lexicographers. Table 2 demonstrates the results of the proposed method for verbal word and synset extraction. The results show that the hybrid method (using structural rules plus digital lexicons) significantly increases both the number of extracted verbs and their precision; however using lexicons decreases the precision of results while increasing the number of correctly extracted synsets. This happens due

to polysemous words with different meanings in a synset.

Verb extraction approach	No. of correctly extracted verbs	Precision for verb Ex.	No. of Correctly extracted synsets	Precision for syn-set Ex.
Applying structural rules	750	79%	396	76%
Applying structural rules and digital lexicons	6890	91%	2790	67%

Table 2: Number of correct words and synsets and precision of the proposed method for verb and synset extraction

The given results for automatic extraction of non-taxonomic relations contain 5600 correct relations among existing synsets in FarsNet 2.0; with accuracy rate of 76%. FarsNet 2.0 had 1040 semantic relations (excluding hyper/hypo-nymy, domain, and holo/mero-nymy) before applying the proposed strategy which formed only about 2.8% of the relations in FarsNet 2.0. This rate reached to 15.7% after implementation of the suggested method. Thus, the proposed automatic method has efficiently contributed to improve the number of non-taxonomic relations corresponding to thematic roles and co-occurrence relations and reduced size of manual processing for relation extraction. The presented strategy still leads to extraction of further and more accurate conceptual relations by increase in number of synsets and words and examples for each of the concepts by extension of FarsNet.

References

- Al Tarouti, F. a. (2016). Enhancing Automatic Wordnet Construction Using Word Embeddings. *In Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*.
- Bagherbeygi, S., & Shamsfard, M. (2012). Corpus based Semi-Automatic Extraction of Persian Compound Verbs and their Relations. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, (pp. 2863-2867). Istanbul.
- Boudabous, M. M. (2013). Arabic wordnet semantic relations enrichment through morpho-lexical patterns. *In Proceeding of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSA)*.
- Dehkharghani, R., & Shamsfard, M. (2011). *Bilingual Ontology Mapping*. Germany: LAMBERT Publisher.
- Fadaei, H., & Shamsfard, M. (2010). Extracting conceptual relations from persian resources. *Seventh International Conference on Information Technology: New Generations (ITNG)*, (pp. 244-248).
- Girju, R. (2008). Semantic relation extraction and its applications. *ESSLLI*.
- Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *LREC*, (pp. 759-765).
- Hwang, C. L., & Yoon, K. (1981). *Multiple Attributes Decision Making Methods and Applications*. Berlin: Springer.
- Jafarinejad, F., & Shamsfard, M. (2012, March). Extracting Generalized Semantic Roles from Corpus. *IJCSI International Journal of Computer Science Issues*, 9(2).
- Kavalec, M., & Svátek, V. (2005). A study on automated relation labelling in ontology learning. (P. Buitelaar, P. Cimiano, & B. Magnini, Eds.) *Ontology learning from text: Methods, evaluation and applications*, 44-58.
- Khodaparasti, F. (1997). *A Comprehensive Dictionary of Persian Synonyms and Antonyms*. Shiraz: Daneshnameye Fars.
- Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *hlt-Naacl*, 13, 746-751.
- Mousavi, Z. a. (2017). Persian Wordnet Construction using Supervised Learning. *arXiv preprint arXiv:1704.03223*.
- Prabhu, V., Desai, S., Redkar, H., Prabhugaonkar, N., Nagvenkar, A., & Karmali, R. (2012). An efficient database design for IndoWordNet development using hybrid approach. *In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language*, (pp. 229-236). Mumbai, India.
- Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013). Development of a Persian Syntactic Dependency Treebank. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. Atlanta, USA.
- Rasooli, M. S., Moloodi, A., Kouhestani, M., & Bidgoli, B. M. (2011). A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, (pp. 227-231). Poznań, Poland.
- Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

- Sánchez, D., & Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data and Knowledge Engineering*, 64(3), 600–623.
- Sarabi, Z., Mahyar, H., & Farhoodi, M. (2013). ParsiPardaz: Persian Language Processing Toolkit. *Computer and Knowledge Engineering (ICCKE)* (pp. 73-79). IEEE.
- Shamsfard, M., & Barforoush, A. A. (2004). Learning Ontologies from Natural Language Texts. *Int. J. Hum.-Comput. Stud.*, 60(1), 17-63.
- Shamsfard, M., & Ghazanfari, Y. (2016). Augmenting FarsNet with New Relations and Structures for verbs. *8th Global Wordnet Conference*. Bucharest.
- Shamsfard, M., & Mousavi, M. (2008). Thematic role extraction using shallow parsing. *International Journal of Computational Intelligence*, 4(2), 126-132.
- Shamsfard, M., Hesabi, A., Fadaei, H., Mansoori, N., Famian, A., Bagherbeigi, S., et al. (2010). Semi automatic development of FarsNet; The persian WordNet. *5th Global WordNet Conference (GWA2010)*. Mumbai, India.
- Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. *LREC*.
- Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-resource Languages Using Cross-lingual WSD. *J. Artif. Int. Res.*, 56(1), 61-87.
- Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD. *J. Artif. Int. Res.*, 56(1), 61-87.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic*. Springer.
- Zadeh Khosravi Forooshani, P., & Rezaei Sharifabadi, M. (2016). Automatic semantic role labeling of Persian sentences by the aid of dependency Treebank. pp. 27-38.

