CLiC-it 2014

The First Italian Conference on Computational Linguistics

Proceedings of the Conference

December 9-10, 2014

Copyright ©2014 by Pisa University Press srl.

These proceedings were published by Pisa University Press.
The original papers are available at:
https://clic2014.fileli.unipi.it/it/atti
The papers are mirrored in the ACL Anthology.

ISBN 978-886741-472-7

Table of Contents

Preface
Roberto Basili, Alessandro Lenci and Bernardo Magnini

Creating a standard for evaluating distant supervision for relation extraction
Azad Abad and Alessandro Moschitti 1

Towards Compositional Tree Kernels
Paolo Annesi, Danilo Croce and Roberto Basili 7

Initial explorations in Kazakh to English statistical machine translation
Zhenisbek Assylbekov and Assulan Nurkas 12

Adapting linguistic tools for the analysis of Italian medical records
Giuseppe Attardi, Vittoria Cozza and Daniele Sartiano 17

Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado
Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi 23

Italian irony detection in Twitter: A first approach
Francesco Barbieri, Francesco Ronzano and Horacio Saggion 28

A retrieval model for automatic resolution of crossword puzzles in Italian language
Gianni Barlacchi, Massimo Nicosia and Alessandro Moschitti 33

Analysing word meaning over time by exploiting temporal Random Indexing
Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro 38

Combining distributional semantic models and sense distribution for effective Italian word sense disambiguation
Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro 43

A Lesk-inspired unsupervised algorithm for lexical choice from WordNet synsets
Valerio Basile 48

The Talmud System: a Collaborative web Application for the Translation of the Babylonian Talmud Into Italian
Andrea Bellandi, Davide Albanesi, Alessia Bellusci, Andrea Bozzi and Emiliano Giovannetti . 53

Towards a Decision Support System for Text Interpretation
Alessia Bellusci, Andrea Bellandi, Giulia Benotto, Amedeo Cappelli, Emiliano Giovannetti and Simone Marchi 58

An Italian dataset of textual entailment graphs for text exploration of customer interactions
Luisa Bentivogli and Bernardo Magnini 63

L'integrazione di informazioni contestuali e linguistiche nel riconoscimento automatico dell'ironia
Lorenzo Bernardini and Irina Prodanof 67

A generic tool for the automatic syllabification of Italian
Brigitte Bigi and Caterina Petrone 73

Errori di OCR e riconoscimento di entità nell'Archivio Storico de La Stampa
Andrea Bolioli, Eleonora Marchioni and Raffaella Ventaglio 78

Computer assisted annotation of themes and motifs in ancient Greek epigrams: First steps
Federico Boschetti, Riccardo Del Gratta and Marion Lamé 83

Defining an annotation scheme with a view to automatic text simplification
Dominique Brunato, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni 87

Senso Comune as a knowledge base of Italian language: the resource and its development
Tommaso Caselli, Isabella Chiari, Aldo Gangemi, Elisabetta Jezek, Alessandro Oltramari, Guido Vetere, Laure Vieu and Fabio Massimo Zanzotto 93

CorEA: Italian news corpus with emotions and agreement
Fabio Celli, Giuseppe Riccardi and Arindam Ghosh 98

Detecting attribution relations in speech: A corpus study
Alessandra Cervone, Silvia Pareti, Peter Bell, Irina Prodanof and Tommaso Caselli 103

Adattamento al progetto dei modelli di traduzione automatica nella traduzione assistita
Mauro Cettolo, Nicola Bertoldi and Marcello Federico 108

The new basic vocabulary of Italian as a linguistic resource
Isabella Chiari and Tullio De Mauro 113

Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi
Francesca Chiusaroli 117

Annotation of complex emotions in real-life dialogues: The case of empathy
Morena Danieli, Giuseppe Riccardi and Firoj Alam 122

Evaluating ImagAct-WordNet mapping for English and Italian through videos
Irene De Felice, Roberto Bartolini, Irene Russo, Valeria Quochi and Monica Monachini 128

CLaSSES: a new digital resource for Latin epigraphy
Irene De Felice, Margherita Donati and Giovanna Marotta 132

Online and Multitask Learning for Machine Translation Quality Estimation in Real-world Scenarios
José G. C. de Souza, Marco Turchi, Antonios Anastasopoulos and Matteo Negri 138

A Computational Approach to Poetic Structure, Rhythm and Rhyme
Rodolfo Delmonte 144

A reevaluation of dependency structure evaluation
Rodolfo Delmonte 151

Analisi Linguistica e stilostatistica - Uno studio predittivo sul campo
Rodolfo Delmonte 158

An adaptable morphological parser for agglutinative languages
Marina Ermolaeva 164

Distributed smoothed tree kernel
Lorenzo Ferrone and Fabio Massimo Zanzotto 169

Polysemy alternations extraction using the PAROLE SIMPLE CLIPS Italian lexicon
Francesca Frontini, Valeria Quochi and Monica Monachini 175

Rappresentazione dei concetti azionali attraverso prototipi e accordo nella categorizzazione dei verbi generali. Una validazione statistica.
Gloria Gagliardi 180

Correcting OCR errors for German in Fraktur font
Michel Généreux, Egon W. Stemle, Verena Lyding and Lionel Nicolas 186

Some issues on Italian to LIS automatic translation. The case of train announcements.
Carlo Geraci, Alessandro Mazzei and Marco Angster 191

ConParoleTue: crowdsourcing al servizio di un Dizionario delle Collocazioni Italiane per Apprendenti (Dici-A)
Andrea Gobbi and Stefania Spina 197

Making Latent SVMstruct practical for coreference resolution
Iryna Haponchyk and Alessandro Moschitti 203

Nominal coercion in space: Mass/count nouns and distributional semantics
Manuela Hürlimann, Raffaella Bernardi and Denis Paperno 208

Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language
Claudio Iacobini, Aurelio De Rosa and Giovanna Schirato 213

Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion
Elisabetta Jezek and Laure Vieu 219

Publishing PAROLE SIMPLE CLIPS as linguistic linked open data
Fahad Khan and Francesca Frontini 224

A preliminary comparison of state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank
Alberto Lavelli 229

SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations
Alessandro Lenci, Gianluca E. Lebani, Sara Castagnoli, Francesca Masini and Malvina Nissim 234

Più l'ascolto e più Mi piace! Social media e radio: uno studio preliminare del successo dei post
Eleonora Lisi, Emanuele Donati and Fabio Massimo Zanzotto 239

Estimating lexical resources Impact in text-to-text inference tasks
Simone Magnolini and Bernardo Magnini 244

Parting ways with the partitive view: a corpus-based account of the Italian particle "ne"
Alice Mariotti and Malvina Nissim 249

On the lexical coverage of some resources on Italian cooking recipes
Alessandro Mazzei 254

Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank
Anne-Lyse Minard, Alessandro Marchetti and Manuela Speranza 260

An English-Italian MWE dictionary
Johanna Monti 265

ALCIDE: An online platform for the analysis of language and content in a digital environment
Giovanni Moretti, Sara Tonelli, Stefano Menini and Rachele Sprugnoli 270

Inner speech, dialogue text and collaborative learning in virtual learning communities
Stefanos Nikiforos and Katia Lida Kermanidis 275

Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare
Maria Palmerini and Renata Savy 281

"Il Piave mormorava. . . ": Recognizing Locations and other Named Entities in Italian Texts on the Great War
Lucia Passaro and Alessandro Lenci 286

The importance of being sum. Network analysis of a Latin dependency treebank
Marco Passarotti 291

I-ChatbIT: An intelligent chatbot for the Italian Language
Arianna Pipitone, Vincenzo Cannella and Roberto Pirrone 296

Two-dimensional wordlikeness effects in lexical organisation
Vito Pirrelli, Claudia Marzi and Marcello Ferro 301

Toward disambiguating typed predicate-argument structures for Italian
Octavian Popescu, Ngoc Phuoc An Vo, Anna Feltracco, Elisabetta Jezek and Bernardo Magnini 306

Il corpus Speaky
Fabio Poroli, Massimiliano Todisco, Michele Cornacchia, Cristina Delogu, Andrea Paoloni and Mauro Falcone 311

Converting the parallel treebank ParTUT in Universal Stanford Dependencies
Manuela Sanguinetti and Cristina Bosco 316

Developing corpora and tools for sentiment analysis: the experience of the University of Turin group
Manuela Sanguinetti, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, Leonardo Allisio, Valeria Mussa and Cristina Bosco 322

Unsupervised antonym-synonym discrimination in vector space
Enrico Santus, Qin Lu, Alessandro Lenci and Chu-Ren Huang 328

Methods of textual archive preservation
Eva Sassolini, Sebastiana Cucurullo and Manuela Sassi 334

Combining unsupervised syntactic and semantic models of thematic fit
Asad Sayeed and Vera Demberg 339

Deep neural network adaptation for children's and adults' speech recognition
Romain Serizel and Diego Giuliani 344

An Italian corpus for aspect based sentiment analysis of movie reviews
Antonio Sorgente, Giuseppe Vettigli and Francesco Mele 349

Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione
Stefania Spina 354

Are quantum classifiers promising?
Fabio Tamburini 360

Geometric and statistical analysis of emotions and topics in corpora
Francesco Tarasconi and Vittorio Di Tomaso 365

Corpus ICoN: una raccolta di elaborati di italiano L2 prodotti in ambito universitario
Mirko Tavosanis 370

Coreference resolution for Italian: Assessing the impact of linguistic components
Olga Uryupina and Alessandro Moschitti 374

A context based model for Sentiment Analysis in Twitter for the Italian language
Andrea Vanzo, Giuseppe Castellucci, Danilo Croce and Roberto Basili 379

Semantic role annotation of instrument subjects
Rossella Varvara and Elisabetta Jezek 384

The Italian module for NooJ
Simonetta Vietri 389

We are glad to introduce CLiC-it 2014 (http://clic.humnet.unipi.it), the first edition of the Italian Conference on Computational Linguistics, a new event aiming to establish a reference forum for research on Computational Linguistics of the Italian community. CLiC-it covers all aspects of automatic language understanding, both written and spoken, and targets state-of-art theoretical results, experimental methodologies, technologies, as well as application perspectives, which may contribute to advance the field.

CLiC-it 2014 is held in Pisa on December 9-10 2014, and it is co-located with EVALITA-2014 (http://www.evalita.it), the fourth edition of the evaluation campaign of Natural Language Processing and Speech tools for Italian and with the XIII Symposium on Artificial Intelligence (Pisa, 10-12 December 2014, http://aiia2014.di.unipi.it/).

Pisa is a special place in the history of Italian Computational Linguistics. Here, Padre Roberto Busa carried out his pioneering research on automatic text processing in the late '60s with Antonio Zampolli, who then founded the Istituto di Linguistica Computazionale of CNR in Pisa, the first research center thoroughly devoted to Computational Linguistics and Natural Language Processing. The University of Pisa also hosted the first professorship in Computational Linguistics held by Antonio Zampolli until his death in 2003.

It is therefore highly symbolic that the Italian community on Computational Linguistics gathers for the first time in Pisa, there where its roots lie. Italian Computational Linguistics has come a long way. Research groups and centers are now spread nationwide and play an active role on the international scene. The large number of researchers that have decided to present their work to CLiC-it is the best proof of the maturity of our community, strongly committed to shape the future of Computational Linguistics.

The spirit of CLiC-it is inclusive. In the conviction that the complexity of language phenomena needs cross-disciplinary competences, CLiC-it intends to bring together researchers of related disciplines such as Computational Linguistics, Linguistics, Cognitive Science, Machine Learning, Computer Science, Knowledge Representation, Information Retrieval and Digital Humanities.

CLIC-it covers all aspects of automated language processing. Relevant topics for the conference include, but are not limited to, the following thematic areas:

- *Cognitive modeling of language processing and psycholinguistics.* Area chairs: Marco Baroni (University of Trento) and Vito Pirrelli (ILC-CNR, Pisa).
- *Digital Humanities*. Area chairs: Sara Tonelli (FBK, Trento) and Fabio Massimo Zanzotto (University of Rome Tor Vergata).
- *Information Extraction.* Area chairs: Maria Teresa Pazienza (University of Rome Tor Vergata) and Paola Velardi (University of Rome Sapienza)
- *Information Retrieval.* Area Chair: Fabrizio Sebastiani (ISTI-CNR, Pisa)
- *Linguistic Resources*. Area chairs: Elisabetta Jezek (University of Pavia) and Monica Monachini (ILC-CNR, Pisa)
- *Machine Translation.* Area chair: Marcello Federico (FBK, Trento)
- *Pragmatics and Creativity*. Area chairs: Rodolfo Delmonte (University of Venezia) and Malvina Nissim (University of Bologna)

- *Semantics and Knowledge Acquisition*. Area chairs: Gianni Semeraro (University of Bari) and Alessandro Moschitti (University of Trento)
- *Spoken Language Processing*. Area chairs: Franco Cutugno (University of Napoli Federico II) and Cinzia Avesani (ISTC-CNR, Padova)
- *Syntax and Parsing.* Area chairs: Giuseppe Attardi (University of Pisa) and Alessandro Mazzei (University of Torino)

We have received a total of 97 paper submissions, out of which 75 have been accepted to appear in the Conference proceedings, which are available online in a joint volume with Evalita 2014.

We are very proud of the scientific program of the conference: it includes two invited speakers, Eduard Hovy (Carnegie Mellon University) and John Nerbonne (University of Groningen), long and short oral presentations, as well as two poster sessions. We are also happy to assign best paper awards to young authors (PhD students and Postdocs) who appear as first author of their paper.

We would like to share the great success of CLiC-it 2014, the first edition of the Italian Conference on Computational Linguistics, with the whole Italian community. We thank the conference sponsors for their generous support:

We also thank the following organizations and institutions for endorsing CLiC-it:

- Università di Pisa
- Società Italiana di Glottologia (SIG)
- Associazione Italiana per l'Intelligenza Artificiale (AI*IA)
- Società di Linguistica Italiana (SLI)
- Associazione Italiana di Linguistica Applicata (AITLA)
- Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)
- Associazione Italiana Scienze della Voce (AISV)

Special thanks to Gianluca Lebani for his precious help to organize CLiC-it 2014 and prepare the conference proceedings.

Last but not least, we thank the area chairs and all the reviewers for their incredible work, the invited speakers for their contribution to make CLIC-it an international event, and all the persons involved in the organization of the conference in Pisa.

November 2014

CLiC-it 2014 CO-CHAIRS

Roberto Basili
Alessandro Lenci
Bernardo Magnini

10.12871/CLICIT201411

Creating a standard for evaluating Distant Supervision for Relation Extraction

Azad Abad[1] and **Alessandro Moschitti**[2,1]
[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Qatar Computing Research Institute
abad@disi.unitn.it, amoschitti@gmail.com

Abstract

English. This paper defines a standard for comparing relation extraction (RE) systems based on a Distant Supervision (DS). We integrate the well-known New York Time corpus with the more recent version of Freebase. Then, we define a simpler RE system based on DS, which exploits SVMs, tree kernels and a simple one-vs-all strategy. The resulting model can be used as a baseline for system comparison. We also study several example filtering techniques for improving the quality of the DS output.

Italiano. *Questo articolo definisce uno standard per comparare sistemi per l'estrazione di relazioni (ER) basati su Distant Supervision. In questo lavoro, integriamo il famoso corpus New York Time con la recente versione di Freebase. Quindi, definiamo in sistema di ER che usa DS basato su SVMs, tree kernels e la strategia uno-contro-tutti. Il modello risultante puó essere usato come baseline per la comparazione di sistemi. In aggiunta, studiamo diverse tecniche di filtraggio degli esempi prodotti dalla DS per migliorare la qualitá del suo output.*

1 Introduction

Relation Extraction (RE) is a well-known Natural Language Processing subarea, which aims at extracting relation types between two named entities from text. For instance, in the sentence:*"Alaska is a U.S. state situated in the North American continent."*, the identified relation type between two entity mentions can be denoted by a tuple r<e_1,e_2>$\in E \times E$, where the tuple name *r* is the relation type and e_1 and e_2 are the entities that participate in the relation.

$$\underbrace{Location/Contains}_{r} < \underbrace{Alaska}_{e1}, \underbrace{United\,States}_{e2} >$$

Currently, supervised learning approaches are widely used to train relation extractors. However, manually providing large-scale human-labeled training data is costly in terms of resources and time. Besides, (i) a small-size corpus can only contains few relation types and (ii) the resulting trained model is domain-dependent.

Distance Supervision (DS) is an alternative approach to overcome the problem of data annotation (Craven et al., 1999) as it can automatically generate training data by combining (i) a structured Knowledge Base (KB), e.g., Freebase[1] with a large-scale unlabeled corpus, C. The basic idea is: given a tuple r<e_1,e_2> contained in a referring KB, if both e_1 and e_2 appear in a sentence of C, that sentence is assumed to express the relation type r, i.e., it is considered a training sentence for r. For example, given the KB relation, `president(Obama,USA)`, the following sentence, *Obama has been elected in the USA presidential campaign*, can be used as a positive training example for `president(x,y)`.

However, DS suffers from two major drawbacks: first, in early studies, Mintz et al. (2009) assumed that two entity mentions cannot be in a relation with different relation types r_1 and r_2. In contrast, Hoffmann et al. (2011) showed that 18.3% of the entities in Freebase that also occur in the New York Times 2007 corpus (NYT) overlap with more than one relation type.

Second, although DS method has shown some promising results, its accuracy suffers from noisy training data caused by two types of problems (Hoffmann et al., 2011; Intxaurrondo et al., 2013; Riedel et al., 2010): (i) possible mismatch between the sentence semantics and the relation type mapped in it, e.g., the KB correct relation, `located_in(Renzi,Rome)`, cannot be mapped into the sentence, *Renzi does not love the Rome soccer team*; and (ii) coverage of the KB,

[1]http://www.freebase.com/

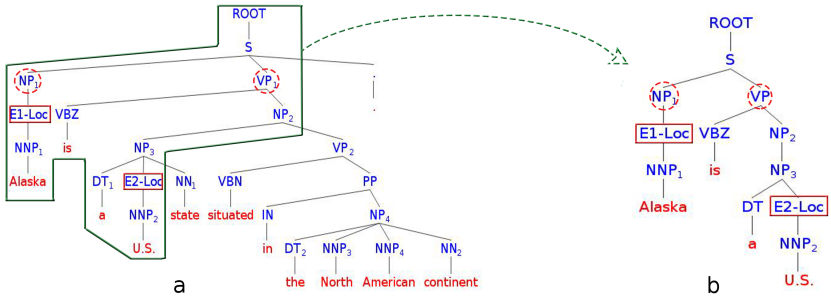


Figure 1: a) The constituent parse tree of the example sentence where "E1-Loc" denotes the source entity mentions and "E2-Loc" denotes the target entity. b) PT relation instance space of the sentence.

e.g., a sentence can express relations that are not in the KB (this generates false negatives).

Several approaches for selecting higher quality training sentences with DS have been studies but comparing such methods is difficult for the lack of well-defined benchmarks and models using DS.

In this paper, we aim at building a standard to compare models based on DS: first of all, we considered the most used corpus in DS, i.e., the combination of NYT and Freebase (NYT-FB).

Secondly, we mapped the Freebase entity IDs used in NYT-FB from the old version of 2007 to the newer Freebase 2014. Since entities changed, we asked an annotator to manually tag the entity mentions in the sentence. As the result, we created a new dataset usable as a stand-alone DS corpus, which we make available for research purposes.

Finally, all the few RE models experimented with NYT-FB in the past are based on a complex conditional random fields. This is necessary to encode the dependencies between the overlapping relations. Additionally, such models use very particular and sparse features, which make the replicability of the models and results complex, thus limiting the research progress in DS. Indeed, for comparing a new DS approach with the previous work using NYT-FB, the researcher is forced to re-implement a very complicated model and its sparse features. Therefore, we believe that simpler models can be very useful as (i) a much simpler re-implementation would enable model comparisons and (ii) it would be easier to verify if a DS method is better than another. In this perspective, our proposed approach is based on convolution tree kernels, which can easily exploit syntactic/semantic structures. This is an important aspect to favor replicability of our results.

Moreover, our method differers from previous state of the art on overlapping relations (Riedel et al., 2010) as we apply a modification of the simple one-vs-all strategy, instead of the complex graphical models. To make our approach competitive, we studied several parameters for optimizing SVMs and filtering out noisy negative training examples. Our extensive experiments show that our models achieve satisfactory results.

2 Related Work

Extracting relations from the text has become popular in IE community. In fully-supervised approach, all the instances are manually labeled by humans and it has been the most popular method so far (Zelenko et al., 2003; Culotta and Sorensen, 2004; Kambhatla, 2004). In semi-supervised approach, initially a small number of seed instances are manually annotated and used to extract the patterns from a big corpus (Agichtein and Gravano, 2000; Blum and Mitchell, 1998).

Distant Supervision (DS) has emerged to be a popular method for training semantic relation extractors. It was used for the first time in the biomedical domain (Craven et al., 1999) and the basic idea was to extract binary relations between protein and cell/tissues by using Yeast Protein Database (YPD) corpus. This method is getting more and more popular and different types of RE problems are being addressed (Bunescu and Mooney, 2007; Mintz et al., 2009; Riedel et al., 2010; Nguyen and Moschitti, 2011; Hoffmann et al., 2010; Riedel et al., 2013; Surdeanu et al., 2012; Hoffmann et al., 2011). Among others, tree kernels (TKs) have been widely used in supervised and weakly supervised setting and shown promising results. (Bunescu and Mooney, 2005; Nguyen et al., 2009; Nguyen and Moschitti, 2011; Bunescu and Mooney, 2005; Zelenko et al., 2003)

3 Basic RE using SVMs and TKs

Support Vector Machines (SVMs) are linear supervised binary classifiers that separate the class boundaries by constructing hyperplanes in a multidimensional space. They can also be used in non-separable linear space by applying kernel func-

tions. Tree kernels (TKs) (Collins et al., 2001) have been proved to achieve state-of-the-art in relation extraction (Zhang et al., 2006b). Different TKs have been proposed in the past (Moschitti, 2006). We modeled our RE system by using feature vectors along with syntactic/semantic trees (see (Zhang et al., 2006a; Nguyen et al., 2009)).

3.1 Feature Vectors

In our experiment, we used the features proposed by Mintz et al. (2009). It consists of two standard lexical and syntactic feature levels. Lexical/syntactic features extracted from a candidate sentence are decorated with different syntactic features such as: (i) Part of Speech (POS); (ii) the window of k words of the left and right of matched entities; (iii) the sequences of words between them; and (iv) finally, syntactic features extracted in terms of dependency patterns between entity pairs. The proposed features yield low-recall as they appear in conjunctive forms but at the same time they produce a high precision.

3.2 Tree Kernels for RE

We used the model proposed in (Zhang et al., 2006a). This, given two relation examples, R_1 and R_2, computes a composite kernel $K(R_1, R_2)$, which combines a tree kernel with a linear kernel. More formally:

$$K(R_1, R_2) = \alpha \vec{x_1} \cdot \vec{x_2} + (1 - \alpha) K_T(T_1, T_2),$$

where α is a coefficient that assigns more weight to the target kernel, $\vec{x_1}$ and $\vec{x_2}$ are feature vectors representing the two relations R_1 and R_2, respectively, and $K_T(T_1, T_2)$ is the tree kernel applied to the syntactic/semantic trees representing the two relations. T_i $(i = 1, 2)$ is the minimal subtree containing the shortest path between the two target entity mentions. Figure 1 shows a sentence tree (part a) and its associated tree (part b).

4 Experiments

Corpus. We trained our system on the NYT news wire corpus (Sandhaus, 2008). The original corpus includes 1.8 million articles written and published by the NYT between January 1987 and June 2007. We used the same subset of data as Riedel et al. (2010). The data set consists of two parts for training and the test, where the first part refers to the years 2005-2006 of the NYT whereas the second refer to the year 2007.

In the corpus provided by Riedel et al. (2010), instead of the entity mentions, their corresponding IDs in Freebase have been tagged (this because of previous copyright issues). The old version of Freebase 2007 is not available anymore and in many cases the IDs or entities have changed in Freebase 2014. So, it was not possible to combine NYT with the newer Freebase to apply DS. To deal with this problem, we mapped the old Freebase IDs with Freebase 2014 and, if the entities were not the same, we asked an annotator to manually tag the entity mentions in the sentence. As the result, we created a new dataset that is mapped with Freebase 2014 and it is usable as a stand-alone DS corpus, which we are making freely available[2]. Overall, we found 4,700 relations in the training set and 1,950 in the test set. The number of positive and negative examples is heavily imbalanced (1:134). So, we applied simple filtering to discard noisy negative examples from the training set.

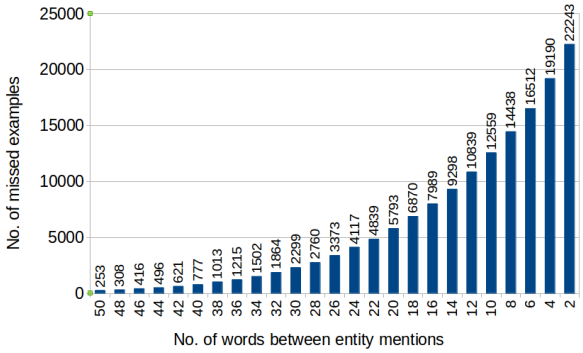


Figure 2: Recall of positive examples with respect to word distance between entity mentions.

4.1 Data Pre-processing

In the introduction, we pointed out that (i) some sentences containing the target entities may not semantically realize the target relation and (ii) other sentences express a correct relation not in the KB. We tackle such problems by applying sentence filtering and enriching the relations of previous KB.

Sentence Filtering. We used four levels of noise cleaning to remove potential incorrect sentences from the corpus. More specifically, we remove a sentence if:

- The distance between the two target entity mentions is more than k words (e.g., 26). We set the k threshold value equal to 10% of the total number of positive examples as shown in Figure 2.
- The number of tagged entities between the entity mentions are greater than a constant h (e.g.,10).
- None of the entity mentions in the sentence appeared in positive examples before, i.e., at least one of the entity in the negative example has to be

[2]http://goo.gl/M7I7fL

Relation Type	P%	R%	F1%
company/founders	66.7	11.4	19.5
location/contains	13.5	40.4	20.3
person/company	11.6	60.7	19.5
company/place_founded	20.0	6.7	10.0
person/place_lived	10	20.2	13.46

Table 1: Precision and recall of different relation types.

	P%	R%	F1%
Mintz++	31.28	15.43	20.67
Intxaurrondo et al.	29.79	17.48	22.03
Basic SVM	12.4	7.6	9.5
Our Model	11.3	23.0	15.1
Our Model + filtering	13.2	22.5	16.6

Table 2: Results for different models

in a relation with another entity (i.e., it has to be part of previously generated positive examples).

- The same entity pairs were in a relation in positive examples but with different relation type (Overlap Relation). For instance, in the mention *Edmonton, Alberta , one of six Canadian N.H.L. markets, is the smallest in the league.*, the entity mentions $<Edmonton, Alberta>$ are in relations with two relation types: *Province/Capital* and *Location/Contains*. Thus, to train Rel. 1, all the instances of Rel. 2 are removed and viceversa.

Enriching KB with new relations types. We analyzed the entity pairs in the sentences of our corpus with respect to the relations in Freebase 2007. We discovered that many pairs receive no-relation because they did not exist in Freebase 2007. This creates many false negative (FN) errors in the generation of training data. In the new release of Freebase many new relations are added, thus we could recover many of such FNs. However to keep the compatibility with the previous NYT-FB corpus, we simply discard such examples from the training set (instead of including them as new positive examples). We could match 1,131 new pairs, which are around 1.4% of the total number of the matched pairs in the training set. Overall, 3,373 mentions from the positive examples and 11,818 mentions from negative examples are discarded from the training set.

4.2 NLP Pipeline

Configurations. We use standard NLP tools in our pipeline: we parsed all the sentences using the Charniak parser (Charniak, 2000) and tagged the named entities with the Stanford NER toolkit (Finkel et al., 2005) into 4 classes (e.g. Person, Location, Organization and Other). We used SVM-Light-TK [3] for training our classifiers, and employed the one-vs-all strategy for multi-class classification but with some modifications to handle the overlap relations: instead of selecting the class with the highest score assigned by the classifier to sentences, we selected all the labels if the assigned scores are larger than a certain threshold (e.g., 0). Hence, the classifier can select more than one class for each example. We normalize both the tree kernel and the feature vectors.

Parameter Optimization. The SVM accuracy is highly influenced by selecting the suitable values for the cost-factor (option j) and trade-off (option c) parameters. As we mentioned, the dataset is very imbalance thus we tuned the j parameter to outweigh the positive example errors with respect to the negative examples during training. We used 30% of our training set as a development set to optimize the parameters. Then, the best combination of c and j values with the highest F-measure in the development set are used to train the classifier.

Evaluation. We compared our model with the two recent state-of-the-art algorithms such as: (1) Mintz++ (Surdeanu et al., 2012), which is an improved version of the original work by Mintz et al. (2009) and (2) Intxaurrondo et al. (2013). The results for different classes and the overall Micro-average F1 are shown in tables 1 and 2, respectively. Noted that, due to lack of space, only the performance of the most populated 5 classes out of 52 are reported. The results show that (i) our model improves the micro-average F1 of the basic RE implementation (basic SVM), i.e., by Zhang et al. (2006b), by more than 7 absolute percent points, i.e., 74% relative; and (ii) applying our simple filtering approach improves our model by 1.5% absolute points. However, our models are still outperformed by the state of the art: this is not critical considering that our aim is to build simpler baseline systems.

5 Conclusion

We have proposed a standard framework, simple RE models and an upgraded version of NYT-FB for more easily measuring the research progress in DS research. Our RE model is based on SVMs, can manage overlapping relations and exploit syntactic information and lexical features thanks to tree kernels. Additionally, we have shown that filtering techniques applied to DS data can discard noisy examples and significantly improve the RE accuracy.

[3] http://disi.unitn.it/moschitti/Tree-Kernel.htm

Acknowledgements

The research described in this paper has been partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant #288024: LIMOSINE – Linguistically Motivated Semantic aggregation engiNes.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Annual meeting-association for Computational Linguistics*, volume 45, page 576.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics.

Michael Collins, Nigel Duffy, et al. 2001. Convolution kernels for natural language. In *NIPS*, volume 2001, pages 625–632.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, and Daniel S Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 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*, pages 541–550. Association for Computational Linguistics.

Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. 2013. Removing noisy mentions for distant supervision. In *Proceedings of the 29th "Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural" (SEPLN 2013)*.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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*, pages 1003–1011. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning: ECML 2006*, pages 318–329. Springer.

Truc-Vien T Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 277–282. Association for Computational Linguistics.

Truc-Vien T Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1378–1387. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. philadelphia: Linguistic data consortium.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106.

Min Zhang, Jie Zhang, and Jian Su. 2006a. Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 288–295. Association for Computational Linguistics.

Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. 2006b. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 825–832. Association for Computational Linguistics.

10.12871/CLICIT201412

Towards Compositional Tree Kernels

Paolo Annesi, Danilo Croce and Roberto Basili
Department of Enterprise Engineering
University of Roma, Tor Vergata
Via del Politecnico 1, 00133 Roma, Italy
{annesi,croce,basili}@info.uniroma2.it

Abstract

English. Several textual inference tasks rely on kernel-based learning. In particular Tree Kernels (TKs) proved to be suitable to the modeling of syntactic and semantic similarity between linguistic instances. In order to generalize the meaning of linguistic phrases, Distributional Compositional Semantics (DCS) methods have been defined to compositionally combine the meaning of words in semantic spaces. However, TKs still do not account for compositionality. A novel kernel, i.e. the Compositional Tree Kernel, is presented integrating DCS operators in the TK estimation. The evaluation over Question Classification and Metaphor Detection shows the contribution of semantic compositions w.r.t. traditional TKs.

Italiano. *Sono numerosi i problemi di interpretazione del testo che beneficiano dall'applicazione di metodi di apprendimento automatico basato su funzioni kernel. In particolare, i Tree Kernel (TK) sono applicati alla modellazione di metriche di similaritá sintattica e semantica tra espressioni linguistiche. Allo scopo di generalizzare i significati legati a sintagmi complessi, i metodi di Distributional Compositional Semantics combinano algebricamente i vettori associati agli elementi lessicali costituenti. Ad oggi i modelli di TK non esprimono criteri di composizionalitá. In questo lavoro dimostriamo il beneficio di modelli di composizionalitá applicati ai TK, in problemi di Question Classification e Metaphor Detection.*

1 Introduction

Tree Kernels (TKs) (Collins and Duffy, 2001) are consolidated similarity functions used in NLP for their ability in capturing syntactic information directly from parse trees and used to solve complex tasks such as Question Answering (Moschitti et al., 2007) or Semantic Textual Similarity (Croce et al., 2012). The similarity between parse tree structures is defined in terms of all possible syntagmatic substructures. Recently, the Smoothed Partial Tree Kernel (SPTK) has been defined in (Croce et al., 2011): the semantic information of the lexical nodes in a parse tree enables a smoothed similarity between structures, which are partially similar and whose nodes can differ but are nevertheless related. Semantic similarity between words is evaluated in terms of vector similarity in a Distributional Semantic Space (Sahlgren, 2006; Turney and Pantel, 2010; Baroni and Lenci, 2010). Even if achieving higher performances w.r.t. traditional TKs, the main limitations of SPTK are that the discrimination between words is delegated only to the lexical nodes and semantic composition of words is not considered.

We investigate a kernel function that exploits semantic compositionality to measures the similarity between syntactic structures. In our perspective the semantic information should be emphasized by compositionally propagating lexical information over an entire parse tree, making explicit the head/modifier relationships between words. It enables the application of Distributional Compositional Semantic (DCS) metrics, that combine lexical representations by vector operator into the distributional space (Mitchell and Lapata, 2008; Erk and Pado, 2008; Zanzotto et al., 2010; Baroni and Lenci, 2010; Grefenstette and Sadrzadeh, 2011; Blacoe and Lapata, 2012; Annesi et al., 2012), within the TKs computation. The idea is to i) define a procedure to mark nodes of a parse tree that allows to spread lexical bigrams across the tree nodes ii) apply DCS smoothing metrics between such compositional nodes iii) enrich the SPTK formulation with compositional distributional seman-

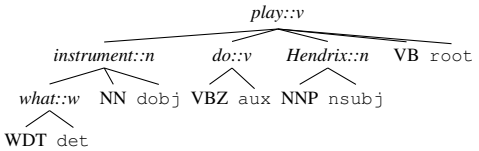


Figure 1: Lexical centered tree of the sentence "*What instrument does Hendrix play?*"

tics. The resulting model has been called Compositional Smoothed Partial Tree Kernel (CSPTK). The entire process of marking parse trees is described in Section 2. Therefore, in Section 3 the CSPTK is presented. Finally, in Section 4, the evaluations over Question Classification and Metaphor Detection tasks are shown.

2 Explicit compositions in Parse Trees

Compositional semantic constraints over a tree kernel computation can be applied when syntagms corresponding to nodes are made explicit. Given the question "*What instrument does Hendrix play?*" and its dependency structure, the corresponding syntactic structure is shown in Figure 1 in terms of a Lexically Centered Tree (LCT), as in (Croce et al., 2011). Nodes are partitioned into: **lexical** nodes in terms of non-terminals $\langle l_n ::pos_n \rangle$, such as *instrument::n*, where *l* is the lemma of the token and `pos` the part-of-speech; **syntactic** nodes, i.e. children of each lexical node which encodes a dependency function $d \in \mathcal{D}$ (e.g. PREP$_{OF}$) and the `pos`-tag of the parent (e.g. NN).

In order to introduce lexical compositionality to these syntactic representations, a mark-up process is introduced, enabling the compositional extension of the tree kernel. Each link between two non-terminal nodes in a LCT representation reflects a dependency relation d, encoded by the child of the lowest non-terminal node. For example, the dependency between the node *instrument::n* and its parent node *play::v* is of type `dobj`. Thus, semantic compositionality is introduced in terms of a head/modifier pair (h, m) over non-terminal nodes, where lexical head is always the upper node. Every non-terminal node is now marked as

$$\langle d_{h,m}, \langle l_h :: pos_h, l_m :: pos_m \rangle \rangle \quad (1)$$

Figure 2 shows a fully compositionally labeled tree, called Compositional Lexically Centered Tree (CLCT), for the sentence whose unlabeled version has been shown in Figure 1. Now nodes are partitioned so that: non-terminal nodes represent **compositional lexical pairs** (h, m) marked as in Equation 1: notice that the modifier is missing in the root node; **dependency functions** (`dobj`) and **POS-Tags** (VBZ) are encoded in the terminal nodes as in the original LCT; **lexical nodes**, e.g. *play::v*, are repeated as terminal nodes, in order to reduce data sparseness that may be introduced by considering only compositional compounds. A DCS model can be adopted, allowing to estimate an expressive similarity function between head-modifier pairs (h_1, m_1), (h_2, m_2) within the resulting kernel. In (Mitchell and Lapata, 2008) three general classes of compositional models have been defined: a linear *additive* model $\vec{p} = \mathbf{A}\vec{u} + \mathbf{B}\vec{v}$; a *multiplicative* model $\vec{p} = \mathbf{C}\vec{u}\vec{v}$ and the *dilation* model $\vec{p}_d = (\vec{u} \cdot \vec{u})\vec{v} + (\lambda - 1)(\vec{u} \cdot \vec{v})\vec{u}$. $\mathbf{A}$ and $\mathbf{B}$ are weight matrices; $\mathbf{C}$ is a weight tensor that project lexical vectors $\vec{u}$ and $\vec{v}$ onto the space of $\vec{p}$, i.e. the vector resulting from the composition; eventually, dilation is an asymmetric function where $\vec{u}$ can be used to dilate $\vec{v}$, and viceversa according with a dilation factor λ. Another compositional model adopted here is the so-called **Support Subspace**, proposed in (Annesi et al., 2012), which assumes that a composition is expressed by projecting vectors into subspaces. A projection reflects a selection function over the set of semantic features shared in the (h, m) compound. A subspace local to (h, m) can be found such that only the space dimensions specific to its meaning are selected. Support Subspaces seem very effective for simple syntactic structures by capturing bi-gram semantics, but they are not sensitive to complex linguistic structures.

3 The Compositional Smoothed Partial Tree Kernel

A Tree Kernel function is a function $TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2)$, where T_1 and T_2 are parse trees, while N_{T_1} and N_{T_2} are the sets of the T_1's and T_2's nodes. The Δ function recursively computes the amount of similarity between tree structures in terms of the similarity among substructures. The type of considered fragments determines the expressiveness of the kernel space and different tree kernels are characterized by different choices. In early models, e.g. (Collins and Duffy, 2001), lexical generalization has been neglected in the recursive matching, so that only exact matching between node labels was given a weight higher than 0. Lexical contribution was proposed by (Croce et al., 2011), in the so called Smoothed Partial Tree Kernel (SPTK). In SPTK, the TK extends

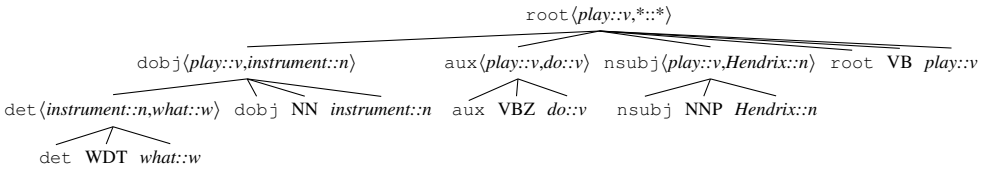


Figure 2: Compositional Lexically Centered Tree (CLCT) of the sentence "*What instrument does Hendrix play?*"

the similarity between tree structures allowing a smoothed function of node similarity σ. It allows to measure the similarity between syntactic tree structures, which are semantically related even when lexical nodes differ. This is achieved by the following formulation of the function Δ:

$\Delta_\sigma(n_1, n_2) = \mu\lambda\sigma(n_1, n_2)$, where n_1 and n_2 are leaves, else

$$\Delta_\sigma(n_1, n_2) = \mu\sigma(n_1, n_2)\Big(\lambda^2 + \sum_{\vec{I}_1, \vec{I}_2, l(\vec{I}_1)=l(\vec{I}_2)} \lambda^{d(\vec{I}_1)+d(\vec{I}_2)} \prod_{j=1}^{l(\vec{I}_1)} \Delta_\sigma(c_{n_1}(\vec{I}_{1j}), c_{n_2}(\vec{I}_{2j}))\Big) \quad (2)$$

In Eq. 2 , $\vec{I}_{1j}$ represents the sequence of subtrees, dominated by node n_1, that are shared with the children of n_2 (i.e. $\vec{I}_{2j}$) as all other non-matching substructures are neglected. The semantic similarity between nodes is measure by $\sigma(n_1, n_2)$.

One main limitation of SPTK is that σ does not consider compositional interaction between words. Given the phrases "*to play sport*" and "*to play instrument*", the SPTK relies only on a unique meaning for *play*, ignoring the compositional role of each modifier. Let us consider the application of the SPTK on the tree shown in Figure 2. When estimating the similarity with a tree derived from sentences such as "*What instrument does Hendrix play?*" or "*What sport does Bolt play?*", the kernel will estimate the similarity among all nodes. Then, the σ function in Equation 2 would not be able to exploit the different senses of the verb *play*, as a traditional DCS model would provide a unique vector representation.

The Compositional Smoothed Partial Tree Kernel (CSPTK) tries to overcome this limitation by measuring the similarity between constituency structures in which lexical compositionality have been made explicit. DCS operators are employed within the CSPTK computation. The core novelty of the CSPTK is the new estimation of σ as described in Algorithm 1. For the lexical nodes the kernel σ_{LEX} is applied, i.e. the cosine similarity between words sharing the same `pos`-tag. Moreover, the other non-lexical nodes contribute according to a strict matching policy: they provide full similarity only when the same `pos`, or dependency, is matched and 0 otherwise. The factor lw is here adopted to reduce the contribution of non-lexical nodes. The novel part of Algorithm 1 is introduced with the similarity computation over compositional nodes. In order to activate the similarity function between non-terminal nodes, they must have the same $d_{h,m}$. In this case a DCS metric can be applied between the involved (h, m) compounds: the lexical information related to pairs are checked and if their respective heads and modifiers share the corresponding POS, a compositional similarity function is applied.If a *modifier is missing*, e.g. the compounds are $(h_x, *)$ and (h_y, m_y), the virtual pair (h_x, h_x) and the pair (h_y, m_y) are used; if *both modifiers are missing*, e.g. the compounds are $(h_x, *)$ and $(h_y, *)$, the σ_{LEX}, i.e. the cosine similarity between word vectors, is adopted.

Algorithm 1 $\sigma_\tau(n_x, n_y, lw)$ Compositional estimation of the lexical contribution to semantic tree kernel

```
σ_τ ← 0,
if n_x = ⟨lex_x::pos⟩ and n_y = ⟨lex_y::pos⟩ then
    σ_τ ← σ_LEX(n_1, n_2)
end if
if (n_x = pos or n_x = dep) and n_x = n_y then
    σ_τ ← lw
end if
if n_x = ⟨d_{h,m}, ⟨li_x⟩⟩ and n_y = ⟨d_{h,m}, ⟨li_y⟩⟩ then
    /*Both modifiers are missing*/
    if li_x = ⟨h_x::pos⟩ and li_y = ⟨h_y::pos⟩ then
        σ_τ ← σ_COMP((h_x), (h_y)) = σ_LEX(n_x, n_y)
    end if
    /*One modifier is missing*/
    if li_x = ⟨h_x::pos_h⟩ and li_y = ⟨h_y::pos_h, m_y::pos_m⟩
    then
        σ_τ ← σ_COMP((h_x, h_x), (h_y, m_y))
    else
        /*General Case*/
        σ_τ ← σ_COMP((h_x, m_x), (h_y, m_y))
    end if
end if
return σ_τ
```

4 Experimental Evaluation

We evaluated CSPTK w.r.t. two inference tasks, i.e. Question Classification (QC) and Metaphor Detection (MI). Texts are processed with Stanford CoreNLP and compositional trees are generated as discussed in Section 2. The lexical similarity func-

tion is derived from a co-occurrence Word Space, acquired through the distributional analysis of the UkWaC corpus, as in (Croce et al., 2011).

CSPTK in Question Classification. In the QC task, the reference corpus is the UIUC dataset (Li and Roth, 2002), including 5,452 questions for training and 500 questions for test, organized in six coarse-grained classes. SVM training has been carried out over the UIUC by applying (i) the PTK and SPTK kernels over the LCT representation of the questions and (ii) the compositional tree kernels (CSPTKs), according to different compositional similarity metrics σ_{COMP}, to the CLCT representation. For learning our models, we used an extension of the SVM-LightTK software. Different compositional kernels are distinct according to the adopted compositionality metrics: *simple additive model* (Mitchell and Lapata, 2010), denoted by a "$^+$" superscript with $\alpha = \beta$; the *pointwise product operator*, denoted by a "$^{\cdot}$" superscript; the *dilation operator* model, denoted by a ^{d} superscript with $\lambda = 1$; the *support subspace* model of (Annesi et al., 2012), denoted by ^{SS}.

Kernel	Accuracy	Std. Dev.
BoW	86.3%	±0.3%
PTK_{LCT}	90.3%	±1.8%
SPTK_{LCT}	92.2%	±0.6%
CSPTK^{+}_{CLCT}	**95.6%**	±0.6%
$\text{CSPTK}^{\cdot}_{CLCT}$	94.6%	±0.5%
CSPTK^{d}_{CLCT}	94.2%	±0.4%
CSPTK^{ss}_{CLCT}	93.3%	±0.7%

Table 1: Results in the Question Classification task

In Table 1 the accuracy achieved by the different systems is reported as the percentage of sentences correctly assigned to the proper question class. As a baseline, a simple bag-of-word model (i.e. BoW) is also computed: it represents questions as binary word vectors and it results in a kernel measuring the lexical overlap. The introduction of lexical semantic information in tree kernel operators, such as in SPTK vs. PTK, is beneficial thus confirming the outcomes of (Croce et al., 2011). $CSPTKs$ seem to make an effective use of the lexical semantic smoothing as they all outperform the non-compositional counterparts. In particular CSPTK^{+}_{CLCT} outperforms all the other compositional operators. Eventually, the error reduction ranges between 12% and 42%.

CSPTK for Metaphor Detection. For the second experiment we choose the annotated Metaphor corpus by (Hovy et al., 2013). The task consists to classify the target words use as literal or metaphorical. The dataset consists of 3,872 sentences divided into training, development, and test sets, using a 80-10-10 split. In Table 2, the accuracy achieved by the different systems is reported. The complexity of the task is confirmed by the low inter annotator agreement achieved over the dataset, i.e. 0.57. As detecting metaphor depends on the deep interaction among words, it seems reasonable that the models using only syntactic information (i.e. PTK) or distributional words in isolation (i.e. BoW) or both (i.e. SPTK) achieve poor performances. The method proposed in (Srivastava et al., 2013) confirms the impact of a proper semantic generalization of the training material. It reaches the SoA by applying a walk-based graph kernel that generalizes the notion of tree kernel as a general framework for word-similarity, and incorporates distributed representations in a flexible way. In our test the syntactic information together with the compositional smoothing, activated by the compositional nodes of the CSPTK, make also an effective use of the lexical semantic smoothing and outperform all the non-compositional counterparts, achieving an accuracy of 75.3%. Even though CSPTK does not outperform (Srivastava et al., 2013), it represents a completely automatic method, largely applicable to different tasks.

Kernel	Accuracy
BoW	71.3%
PTK_{LCT}	71.6%
SPTK_{LCT}	71.0%
CSPTK^{+}_{CLCT}	72.4%
CSPTK^{ss}_{CLCT}	**75.3%**
(Srivastava and Hovy, 2013)	**76.0%**

Table 2: Results in the Metaphor Detection task

5 Conclusions

In this paper, a novel kernel function has been proposed in order to exploit Distributional Compositional operators within Tree Kernels. The proposed approach propagates lexical semantic information over an entire tree, by building a Compositionally labeled Tree. The resulting Compositional Smoothed Partial Tree Kernel measures the semantic similarity between complex linguistic structures by applying metrics sensible to distributional compositional semantics. Empirical results in the Question Classification and Metaphor Detection tasks demonstrate the positive contribution of compositional information for the generalization capability within the proposed kernel.

References

P. Annesi, V. Storch, and R. Basili. 2012. Space projections as distributional models for semantic composition. In *In Proceedings of CICLing 2012*, volume 7181 of *Lecture Notes in Computer Science*, pages 323–335. Springer.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 546–556, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Collins and N. Duffy. 2001. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 625–632.

D. Croce, A. Moschitti, and R. Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of EMNLP*, Edinburgh, Scotland, UK.

D. Croce, P. Annesi, V. Storch, and R. Basili. 2012. Unitor: Combining semantic text similarity functions through sv regression. In **SEM 2012*, pages 597–602, Montréal, Canada, 7-8 June.

K. Erk and S. Pado. 2008. A structured vector space model for word meaning in context. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. ACL.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *EMNLP*, pages 1394–1404. ACL.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of ACL '02*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *In Proceedings of ACL/HLT 2008*, pages 236–244.

J. Mitchell and M Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *ACL*. The Association for Computer Linguistics.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

Shashank Srivastava and Dirk Hovy, 2013. *A Walk-based Semantically Enriched Tree Kernel Over Distributed Word Representations*, pages 1411–1416. Association for Computational Linguistics.

Shashank Srivastava, Dirk Hovy, and Eduard H. Hovy. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In *EMNLP*, pages 1411–1416.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1263–1271, Stroudsburg, PA, USA. Association for Computational Linguistics.

10.12871/CLICIT201413

Initial Explorations in Kazakh to English Statistical Machine Translation

Zhenisbek Assylbekov, Assulan Nurkas
School of Science and Technology
Nazarbayev University
53 Kabanbay batyr ave., Astana, Kazakhstan
{zhassylbekov, anurkas}@nu.edu.kz

Abstract

English. This paper presents preliminary results of developing a statistical machine translation system from Kazakh to English. Starting with a baseline model trained on 1.3K and then on 20K aligned sentences, we tried to cope with the complex morphology of Kazakh by applying different schemes of morphological word segmentation to the training and test data. Morphological segmentation appears to benefit our system: our best segmentation scheme achieved a 28% reduction of out-of-vocabulary rate and 2.7 point BLEU improvement above the baseline.

Italiano. *Questo articolo presenta dei risultati preliminari relativi allo sviluppo di un sistema di traduzione automatica statistica dal Kazaco all'Inglese. Partendo da un modello di base, addestrato su 1.3K e 20K coppie di frasi, proviamo a gestire la complessa morfologia del Kazaco utilizzando diversi schemi di segmentazione morfologica delle parole sui dati di addestramento e di valutazione. La segmentazione morfologica sembra apportare benefici al nostro sistema: il nostro migliore schema di segmentazione ottiene una riduzione del 28% del "Out-of-Vocabulary Rate" ed un miglioramento di 2.7 punti della misura "BLEU" rispetto al sistema di base.*

1 Introduction

The availability of considerable amounts of parallel texts in Kazakh and English has motivated us to apply statistical machine translation (SMT) paradigm for building a Kazakh-to-English machine translation system using publicly available data and open-source tools. The main ideas of SMT were introduced by researchers at IBM's Thomas J. Watson Research Center (Brown et al., 1993). This paradigm implies that translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. We show how one can compile a Kazakh-English parallel corpus from publicly available resources in Section 2.

It is well known that challenges arise in statistical machine translation when we deal with languages with complex morphology, e.g. Kazakh. However recently there were attempts to tackle such challenges for similar languages by morphological pre-processing of the source text (Bisazza and Federico, 2009; Habash and Sadat, 2006; Mermer, 2010). We apply morphological preprocessing techniques to Kazakh side of our corpus and show how they improve translation performance in Sections 5 and 6.

2 Corpus preparation

In order to build an SMT system for any languages one needs to obtain a substantial amount of parallel texts in those languages.

2.1 Small corpus

First we decided to mine a parallel corpus from e-mail messages circulated within one of Kazakhstani organizations with a considerable amount of international staff. In that organization e-mail messages that are addressed to all employees are usually written in three languages: Kazakh, English and Russian. But sometimes they are written solely in English. To identify among all messages only those that contained at least Kazakh and English parts we examined several such e-mails, and we found out that most of them had 'Dear', 'Құрметті' and 'Уважаемые' as beginnings of English, Kazakh and Russian parts respectively as in the example below:

Dear Library Patrons, Please see the …
Құрметті оқырмандар, Қосымшадан …
Уважаемые читатели, Пожалуйста, …

Statistical analysis showed that at 0.9 confidence level a simple heuristic method that classified an e-mail message as trilingual if it contained the words 'Dear', 'Құрметті' and 'Уважаемые' would get not less than 77% of such e-mails.

Out of 1,609 e-mails addressed to all employees that were dumped in April 2014 from one of the company workers' mailbox, we could get 636 trilingual messages. In order to extract Kazakh and English parts from each text chunk we assumed that the Kazakh part began with 'Құрметті', the English part began with 'Dear' and the Russian part began with 'Уважаемые' as in the example above. There are better approaches to detect languages in a multilingual document, e.g. Compact Language Detector (`https://code.google.com/p/cld2/`) or `langid.py` (Lui and Baldwin, 2012), and we are going to use them in our future work.

We trained the *Punkt* sentence splitter from NLTK (Loper and Bird, 2002) on Kazakh side of the corpus and used it along with the pre-trained model for English to perform sentence segmentation for each e-mail message. Then sentence alignment for each pair of e-mails was performed using *hunalign* (Varga et al., 2005). After removing all repeating sentences we obtained 1,303 parallel sentences. We sampled 100 sentence pairs for tuning and 100 sentence pairs for testing purposes.

2.2 Larger corpus

A larger corpus was mined from the official site of the President of the Republic of Kazakhstan located at `http://akorda.kz`. Text extraction from HTML was performed through a Perl-script that used HTML::TreeBuilder module from CPAN. After sentence splitting and sentence alignment we obtained 22,180 parallel sentences. Unfortunately, there were misalignments and sometimes Russian sentences found their way into Kazakh side of the corpus. This happened because the President of Kazakhstan sometimes gave bilingual speeches in Kazakh and Russian and the Russian parts were not translated. We sampled 2,200 sentence pairs from the larger corpus, and 242 of them turned out to be misaligned. So, it seems that approximately $242/2200 = 11\%$ of all sentence pairs are "bad" and the data is subject to further cleaning. We used the "good" 1,958 sentence pairs out of 2,200 for tuning and testing purposes.

3 Kazakh morphology and MT

Kazakh is an agglutinative language, which means that words are formed by joining suffixes to the stem. A Kazakh word can thus correspond to English phrases of various length as shown in Table 1.

дос	friend
дос**тар**	friend**s**
достар**ым**	**my** friends
достар**ымыз**	**our** friends
достарымыз**да**	**at** our friends
достарымызда**мыз**	**we are** at our friends

Table 1: Example of Kazakh suffixation

The effect of rich morphology can be observed in our corpora. Table 2 provides the vocabulary sizes, type-token ratios (TTR) and out-of-vocabulary (OOV) rates of Kazakh and English sides of larger corpus.

	English	Kazakh
Vocabulary size	18,170	35,984
Type-token ratio	3.8%	9.8%
OOV rate	1.9%	5.0%

Table 2: Vocabulary sizes, TTR and test set OOV rates

It is easy to see that rich morphology leads to sparse data problems for SMT that make translation of rare or unseen word forms difficult. That is why we need to use morphological segmentation to reduce data sparseness.

4 Related work

Few small-sized (0.2K–1.3K sentences) and one medium-sized (69.8K sentences) parallel corpora for Kazakh-English pair are available within the OPUS project (Tiedemann, 2012). We were not aware of these resources at the beginning of our research, and therefore we decided to compile our own corpora.

Rule-based approach and preliminary ideas on statistical approach for Kazakh-to-English machine translation were discussed by Tukeyev et al. (2011). Sundetova et al. (2013) presented

structural transfer rules for English-to-Kazakh machine translation system based on Apertium platform (Forcada et al., 2011).

To our knowledge, this is the first paper on the application of SMT methods and morphological segmentation to Kazakh language. However preprocessing of morphologically-rich languages was considered previously in several works: for the Arabic-to-English task Habash and Sadat (2006) presented morphological preprocessing schemes; for the Turkish-to-English direction Bisazza and Federico (2009) developed morphological segmentation schemes and Mermer (2010) presented unsupervised search for the optimal segmentation. In our work we implemented four schemes suggested by Bisazza and Federico (2009), and developed three new schemes for verbs and gerunds.

5 Morphological segmentation schemes

5.1 Preprocessing technique

We performed morphological analysis for our corpora using an open-source finite-state morphological transducer *apertium-kaz* (Washington et al., 2014). It is based on Helsinki Finite-State Toolkit and is available within the Apertium project (Forcada et al., 2011). The analysis was carried out by calling `lt-proc` command of the Lttoolbox (Ortis-Rojas et al., 2005). Since more than one analysis was possible, disambiguation was performed through a Constrained Grammar rules (Karlsson et al., 1995) by calling the `cg-proc` command, which decreased ambiguity from 2.4 to 1.4 analyses per form (see an example of disambiguation in Table 3). In cases when ambiguity still remained we used the first analysis from the output of `cg-proc`.

‘in 2009 , we started the construction works .’		
2009 <u>*жылы*</u> *біз құрылысты бастадық .*		
	жылы⟨adj⟩	‘warm’
	жылы⟨adj⟩⟨advl⟩	‘warmly’
→	**жыл⟨n⟩⟨px3sp⟩⟨nom⟩**	**‘year’**
	жылы⟨adj⟩⟨subst⟩⟨nom⟩	‘warmth’

Table 3: Morphological disambiguation of a Kazakh word in context.

Consequently, each surface form is changed to one of its lexical forms. Now simple regular expressions can be used to describe different segmentation rules on lexical forms.

5.2 Segmentation schemes

Below we present segmentation schemes which are combinations of splitting and removal of tags from the analyzed lexical forms. Segmentation rules MS2–MS11 were suggested by Bisazza and Federico (2009).

MS2. Dative, ablative, locative and instrumental cases are split off from words, since they often align with the English prepositions ‘to’, ‘from’, ‘in’ and ‘with/by’, respectively. The remaining case tags – nominative, accusative and genitive – are removed from the words because they are not expected to have English counterparts.

MS6. After treating case tags we split off from nouns the possessive tags of all persons except the 3rd singular ⟨*px3sp*⟩, which is removed.

MS7. This rule splits off copula from words, in addition to MS6’s rules.

MS11. This rule splits off person suffixes from finite verb forms and copula, in addition to MS7’s rules.

MS11a. This rule removes person suffixes from finite verb forms, in addition to MS7’s rules.

MS12. In addition to MS11a’s rules this rule splits off dative, ablative, locative and instrumental cases from gerunds that are derived from verbs in active form. The remaining case tags – nominative, accusative and genitive – are removed.

MS13. In addition to MS12’s rules this rule splits off from gerunds the possessive tags of all persons except the 3rd singular ⟨*px3sp*⟩, which is removed.

The Kazakh side of our corpora was preprocessed by the aforementioned segmentation schemes. After that angle brackets ‘⟨⟩’ around tags were replaced by plus sign ‘+’ at the beginnings of tags for compatibility with SMT toolkit Moses (Koehn et al., 2007). The benefit of segmentation for word alignment in Kazakh-to-English direction is shown in Figure 1.

6 Experiments

6.1 Baseline

The open-source SMT toolkit Moses (Koehn et al., 2007) was used to build the baseline system. Phrase pairs were extracted from symmetrized word alignments generated by GIZA++ (Och and Ney, 2003). The decoder features a statistical log-linear model including a phrase-based translation model, a 5-gram language model, a lexicalized dis-

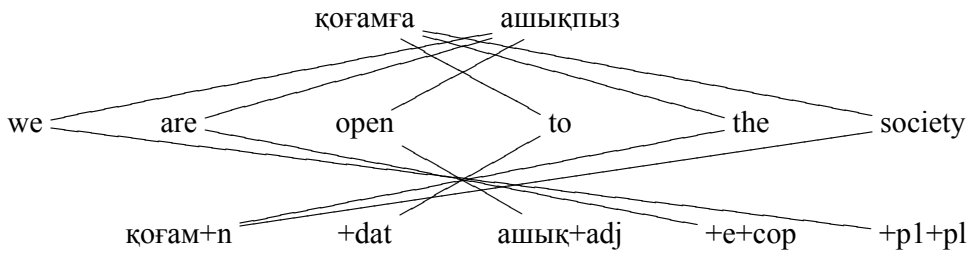


Figure 1: Word alignment before (up) and after (down) morphological segmentation MS11.

tortion model and word and phrase penalties. Distortion limit is 6 by default.

The weights of the log-linear combination were optimized by means of a minimum error rate training procedure (Och, 2003) run on tuning sets mentioned in section 2. Evaluation was performed on test sets.

6.2 Morphological segmentation

The impact of morphological segmentation on training corpus dictionary size and the test set OOV rate is shown in Table 4. One can see that better segmentation schemes lower the vocabulary size and OOV rate.

Scheme	Small corpus		Larger corpus	
	Vocab.	OOV	Vocab.	OOV
baseline	6,143	19.4	35,984	5.0
MS2	5,754	16.0	31,532	4.1
MS6	5,404	15.0	29,430	3.9
MS7	5,393	15.0	29,270	3.9
MS11	5,368	14.1	28,928	3.7
MS11a	5,362	14.8	28,923	3.8
MS12	5,283	14.5	28,079	3.7
MS13	5,241	14.3	27,792	3.6

Table 4: Effect of preprocessing on Kazakh side's training corpus vocabulary size and test set OOV rate.

6.3 Distortion limit

Since the number of words in each sentence has grown on average after segmentation, it seems reasonable to increase the distortion limit (DL) consequently. Thus, we allowed the distortion to be unlimited.

Table 5 shows how morphological preprocessing and unlimited distortion affects translation performance. In each system the same preprocessing was applied to the training, tuning and test data. Each system was run with limited and unlimited distortion but the set of weights for both cases was optimized with the default DL equal to 6.

Scheme	small corpus		larger corpus	
	DL=6	DL=∞	DL=6	DL=∞
baseline	17.69	17.32	22.75	23.70
MS2	18.50	**18.54**	23.77	25.23
MS6	17.29	17.32	23.77	25.06
MS7	17.63	17.43	23.90	25.41
MS11	14.95	15.13	23.62	25.21
MS11a	18.03	17.97	23.95	25.30
MS12	17.80	17.84	23.82	25.18
MS13	**18.74**	18.49	**24.05**	**25.46**

Table 5: BLEU scores.

7 Discussion and Future Work

The experiments have shown that a selective morphological segmentation improves the performance of an SMT system. One can see that in contrast to Bisazza and Federico's results (2009), in our case MS11 downgrades the translation performance. One of the reasons for this might be that Bisazza and Federico considered translation of spoken language in which sentences were shorter on average than in our corpora.

In this work we mainly focused on nominal suffixation. In our future work we are planning to: increase the dictionary of morphological transducer – currently it covers 93.3% of our larger corpus; improve morphological disambiguation using e.g. perceptron algorithm (Sak et al., 2007); develop more segmentation rules for verbs and other parts of speech; mine more mono- and bilingual data using official websites of Kazakhstan's public authorities.

Acknowledgments

We would like to thank Dr. Francis Morton Tyers and Jonathan North Washington for their constant attention to this work. We are grateful to the reviewers for their useful comments and careful readings. This research was financially supported by the grant of the Corporate Fund "Fund of Social Development".

References

Arianna Bisazza and Marcello Federico. 2009. Morphological Pre-Processing for Turkish to English Statistical Machine Translation. *Proceedings of IWSLT 2009*, 129–135.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics,* 19(2):263–311.

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.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. *Proceedings of the Human Language Technology Conference of the NAACL 2006*, 49–-52.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila, eds. 1995. *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. 2007. Moses: Open source toolkit for statistical machine translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Demo and Poster Sessions,* Prague, Czech Republic, 177–180.

Steven Bird. 2006. Nltk: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions,* 69–72. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Demo Session, Jeju, Republic of Korea.

Coşkun Mermer and Ahmet Afşin Akin. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. *Proceedings of the ACL 2010 Student Research Workshop*, 31–36. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models, *Computational Linguistics,* 29(1):19–-51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* 160–-167.

Sergio Ortiz Rojas, Mikel L. Forcada, and Gema Ramírez Sánchez. 2005. Construccion y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento del Lenguaje Natural*, 35:51-–57.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of Turkish text with perceptron algorithm. *Computational Linguistics and Intelligent Text Processing,* 107–118. Springer Berlin Heidelberg.

A. Sundetova, M. L. Forcada, A. Shormakova, A. Aitkulova. 2013. Structural transfer rules for English-to-Kazakh machine translation in the free/open-source platform Apertium. *Компьютерная обработка тюркских языков. Первая международная конференция: Труды.* Астана: ЕНУ им. Л. Н. Гумилева, 2013. – с. 317–326.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218

U. A. Tukeyev, A. K. Melby, Zh. M. Zhumanov. 2011. Models and algorithms of translation of the Kazakh language sentences into English language with use of link grammar and the statistical approach. *Proceedings of the IV Congress of the Turkic World Mathematical Society*, 1(3):474.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy. 2005. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, 590–596.

Jonathan N. Washington, Ilnar Salimzyanov and Francis Tyers. 2014. Finite-state morphological transducers for three Kypchak languages. *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, 3378–3385.

10.12871/CLICIT201414

Adapting Linguistic Tools for the Analysis of Italian Medical Records

Giuseppe Attardi
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
attardi@di.unipi.it

Vittoria Cozza
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
cozza@di.unipi.it

Daniele Sartiano
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
sartiano@di.unipi.it

Abstract

English. We address the problem of recognition of medical entities in clinical records written in Italian. We report on experiments performed on medical data in English provided in the shared tasks at CLEF-ER 2013 and SemEval 2014. This allowed us to refine Named Entity recognition techniques to deal with the specifics of medical and clinical language in particular. We present two approaches for transferring the techniques to Italian. One solution relies on the creation of an Italian corpus of annotated clinical records and the other on adapting existing linguistic tools to the medical domain.

Italiano. *Questo lavoro affronta il problema del riconoscimento di entità mediche in referti medici in lingua italiana. Riferiamo su degli esperimenti svolti su testi medici in inglese forniti nei task di CLEF-ER 2013 e SemEval 2014. Questi ci hanno consentito di raffinare le tecniche di Named Entity recognition per trattare le specificità del linguaggio medico e in particolare quello dei referti clinici. Presentiamo due approcci al trasferimento di queste tecniche all'italiano. Una soluzione consiste nella creazione di un corpus di referti medici in italiano annotato con entità mediche e l'altro nell'adattare strumenti tradizionali per l'analisi linguistica al dominio medico.*

1 Introduction

One of the objectives of the *RIS* project (RIS 2014) is to develop tools and techniques to help identifying patients at risk of evolving their disease into a chronic condition. The study relies on a sample of patient data consisting of both medical test reports and clinical records. We are interested in verifying whether text analytics, i.e. information extracted from natural language texts, can supplement or improve information extracted from the more structured data available in the medical test records.

Clinical records are expressed as plain text in natural language and contain mentions of diseases or symptoms affecting a patient, whose accurate identification is crucial for any further text mining process.

Our task in the project is to provide a set of NLP tools for extracting automatically information from medical reports in Italian. We are facing the double challenge of adapting NLP tools to the medical domain and of handling documents in a language (Italian) for which there are few available linguistic resources.

Our approach to information extraction exploits both supervised machine-learning tools, which require annotated training corpora, and unsupervised deep learning techniques, in order to leverage unlabeled data.

For dealing with the lack of annotated Italian resources for the bio-medical domain, we attempted to create a silver corpus with a semi-automatic approach that uses both machine translation and dictionary based techniques. The corpus will be validated through crowdsourcing.

2 Medical Training Corpus

Currently Italian corpora annotated with mentions of medical terms are not easily available. Hence we decided to create a corpus of Italian medical reports (IMR), annotated with medical mentions and to make it available on demand.

The corpus consists of 10,000 sentences, extracted from a collection of 23,695 clinical records of various types, including discharge summaries, diagnoses, and medical test reports.

The annotation process consists in two steps: creating a silver corpus using automated tools

and then turning the corpus into a gold one by manually correcting the annotations.

For building the silver corpus we used:

- a NER trained over a silver English resource translated to Italian;
- a dictionary-based entity recognition approach.

For converting the silver corpus into a gold one, validation by medical experts is required. We organized a crowdsourcing campaign, for which we are recruiting volunteers to whose we will assign micro annotation tasks. Special care will be taken to collected answers reliability.

2.1 Translation based approach

The CLEF-ER 2013 challenge (Rebholz-Schuhmann et al., 2010) aimed at the identification of mentions in bio-medical texts in various languages, starting from an annotated resource in English, and at assigning to them a concept unique identifier (CUI) from the UMLS thesaurus (Bodenreider, 2004). UMLS combines several multilingual medical resources, including Italian terminology from MedDRA Italian (MDRITA15_1) and MESH Italian (MSHITA2013), bridged through their CUI's to their English counterparts.

The organizers provided a silver standard corpus (SSC) in English, consisting of 364,005 sentences extracted from the EMEA corpus, which had been automatically annotated by combining the outputs of several Named Entity taggers (Rebholz-Schuhmann et al., 2010).

In (Attardi et al., 2013) we proposed a solution for annotating Spanish bio-medical texts, starting from the SSC English resource. Our approach combined techniques of machine translation and NER and consists of the following steps:

1. phrase-based statistical machine translation is applied to the SSC in order to obtain a corresponding annotated corpus in the target language. A mapping between mentions in the original and the corresponding ones in the translation is preserved, so that the CUIs from the original can be transferred to the translation. This produces a Bronze Standard Corpus (BSC) in the target language. A dictionary of entities is also created, which associates to each pair (entity text, semantic group) the corresponding CUIs that appeared in the SSC.
2. the BSC is used to train a model for a Named Entity tagger, capable of assigning semantic groups to mentions.
3. the model built at step 2) is used for tagging entities in sentences in the target language.
4. the annotated document is enriched by adding CUIs to each entity, looking up the pair (entity, group) in the dictionary of CUIs, of step 1.

For machine translation we trained Moses (Koehn, 2007) using a biomedical parallel corpus consisting of EMEA, Medline and the Spanish Wikipedia for the language model.

In task A of the challenge, on mention identification, our submission achieved the best score for the categories disease, anatomical part, live being and drugs, with scores ranging between 91.5% and 97.4% (Rebholz-Schuhmann et al., 2013). In task B on CUI identification, the scores were however much lower.

As NE tagger, we used the Tanl NER (Attardi et al., 2009), a generic sequence tagger based on a Maximum Entropy Markov Model, that uses a rich feature set, customizable by providing a feature model. Such kinds of taggers perform quite well on newswire documents, where capitalization features are quite helpful in identifying people, organization and locations. With a proper feature model we were able to achieve satisfactory results also for medical domain.

Adapting the CLEF-ER approach to Italian required repeating the translation step with an en-it parallel corpus, consisting of EMEA and UMLS for the medical domain and (Europarl, 2014; JRC-Acquis, 2014) for more general domains.

A NE tagger for Italian was the trained on the translated silver corpus.

Due to a lack of annotated Italian medical texts, we couldn't perform validation on the resulting tagger. Manual inspection confirms the hypothesis that accuracy is similar to the Spanish version, given that the major difference in the process is the translation corpus and that Spanish and Italian are similar languages.

2.2 Dictionary based approach

Since the terminology for entities in medical records is fairly restricted, another method for identifying mentions in the IMR corpus is to use a dictionary. We produced an Italian medical thesaurus by merging:

- over 70,000 definitions of treatments and diagnosis from the ICD-9-CM terminology;
- about 22,000 definitions from the SnoMed semantic group "Symptoms and Signs, Disease and Anatomical part" in the UMLS;
- over 2,600 active ingredients and drugs from the "Lista dei Farmaci" (AIFA, 2014).

We identified mentions in the IMR corpus using two techniques: *n*-gram based and parser based.

The IMR text is preprocessed first with the Tanl pipeline, performing sentence splitting, tokenization, POS tagging and lemma extraction.

To ease matching, text is normalized by lowercasing each word.

The *n*-gram based technique tries matching each *n*-gram (with *n* between 1 and 10) in the corpus with entries in the thesaurus in two ways: with lemma match and with approximate match. Approximate matching involves removing prepositions, punctuations and articles from both *n*-grams and entries and performing an exact match.

The parse based matching enables to deal also with some kinds of discontiguous mentions, i.e. entity mentions interleaved with modifiers, e.g. adjectives, verb or adverbs. The matching process involves parsing each sentence in the IMR with the DeSR parser (Attardi et al., 2009), selecting noun phrases corresponding to subtrees whose root is a noun and consisting of certain patterns of nouns, adjectives and prepositions, and finally searching these noun phrases in the thesaurus.

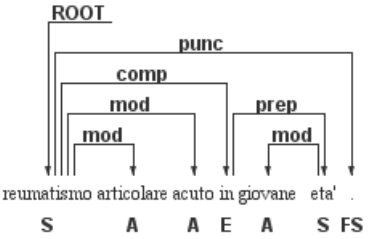


Figure 1: A parsed sentence from the IMR.

Figure 1 shows a sample sentence and its parse tree from which the following noun phrases are identified: `reumatismo acuto`, `reumatismo articolare`, `reumatismo in giovane età`, `giovane età`. Among these, `reumatismo acuto` is recognized as an ICD9 disease.

Overall, we were able to identify over 100,000 entities in the IMR corpus by means of the dictionary approach. The process is expected to guarantee high precision: manual inspection of 4,000 sentences, detected a 96% precision.

Besides recognized entities in the thesaurus, we annotated temporal expressions by a version of HeidelTime (Strotgen and Gertz, 2013) adapted to Italian (Attardi and Baronti, 2014).

3 NE Recognition

We split the analysis of medical records into two steps: recognition of mentions and assignment of unique identifiers (CUI) to those mentions.

The training set consisted of 199 notes with 5,816 annotations, the development set of 99 notes with 5,351 annotations.

For the first step we explored using or adapting traditional NER techniques. We performed experiments in the context of the *SemEval 2014 task 7, Analysis of Clinical Text,* where we could try these techniques using suitable training and test data, even though in English rather than Italian (Attardi et al., 2014). We tested several NER tools: the Tanl NER, the Stanford NER (CRF_NER, 2014) and a Deep Learning NER (NLPNET, 2014), which we developed based on the SENNA architecture (SENNA, 2011). While the first two taggers rely on a rich feature sets and supervised learning, the Deep Learning tagger uses almost no features and relies on word embeddings, learned through an unsupervised process from unannotated texts, along the approach by Collobert et al. 2011.

We created an unannotated corpus (UC) combining 100,000 terms from the English Wikipedia and 30,000 additional terms from a subset of unannotated medical texts from the MIMIC corpus (Moody and Marks, 1996). The word embeddings for the UC are computed by training a deep learning architecture initialized to the values provided by Al-Rfou' et al. (2013) for the English Wikipedia and to random values for the medical terms.

All taggers use dictionary features. We created a dictionary of disease terms (about 22,000 terms from the "Disease or Syndrome" semantic type of UMLS) excluding the most frequent words from Wikipedia.

The Tanl NER could be customized with additional cluster features, extracted from a small window of input tokens. The clusters of UC terms were calculated using the following algorithms:

- DBSCAN (Ester et al., 1996) as implemented in scikit-learn (SCIKIT, 2014). Applied to the word embeddings it produced a set of 572 clusters.
- Continuous Vector Representation of Words (Mikolov et al., 2013), using the word2vec library (WORD2VEC, 2014) with several settings.

We obtained the best accuracy with word2vec using a set of 2,000 clusters.

3.1 Conversion to IOB format

Before applying NE tagging, we had to convert the medical records into the IOB format used by

most NE taggers. The conversion is not straightforward since clinical reports contain discontiguous and overlapping mentions. For example, in:

```
Abdomen is soft, nontender, non-
distended, negative bruits
```

there are two mentions: `Abdomen nontender` and `Abdomen bruits`.

The IOB format does not allow either discontinuity or overlaps. We tested two conversions: one by replicating a sentence, each version having a single mention from a set of overlapping ones. The second approach consisted in using additional tags for disjoint and shared mentions (Tang et al., 2013): DISO for contiguous mentions; DDISO for disjoint entity words that are not shared by multiple mentions; HDISO for the head word that belongs to more than one disjoint mentions.

We tested the accuracy of various NE taggers on the SemEval development set. The results are reported in Table 1. Results marked with *discont* were obtained with the additional tags for discontiguous and overlapping mentions.

NER	Precision	Recall	F- score
Tanl	80.41	65.08	71.94
Tanl+dbscan	**80.43**	64.48	71.58
Tanl+word2vec	79.70	**67.44**	**73.06**
Nlpnet	80.29	62.51	70.29
Stanford	80.30	64.89	71.78
CRFsuite	79.69	61.97	69.72
Tanl discont	78.57	65.35	71.35
Nlpnet discont	77.37	63.76	69.61
Stanford discont	80.21	62.79	70.44

Table 1: Accuracy on the development set.

3.2 Semeval 2014 NER for clinical text

The task 7 of SemEval 2014 allowed us to test NE tagging techniques on medical records and to adapt them to the task. Peculiarly, only one class of entities, namely diseases, is present in the corpus.

We dealt with overlapping mentions by converting the annotations. Discontiguous mentions were dealt in two steps: the first step identifies contiguous portions of a mention with a traditional sequence labeler; then separate portions of mentions are combined into a full mention with guidance from a Maximum Entropy classifier (Berger et al., 1996), trained to recognize which pairs belong to the same mention. The training set consists of all pairs of terms within a document annotated as disorders. A positive instance is created if the terms belong to the same mention, a negative one otherwise.

The classifier was trained using a binned distance feature and dictionary features, extracted for each pair of words in the training set.

For mapping entities to CUIs we applied fuzzy matching (Fraser, 2011) between the extracted mentions and the textual description of entities present in a set of UMLS disorders. The CUI from the match with highest score is chosen.

Our submission reached a comparable accuracy to the best ones based on a single system approach (Pradhan et al., 2014), with an F-score of 0.65 for Task A and 0.83 for Task A relaxed. For Task B and Task B relaxed the accuracies were 0.46 and 0.70 respectively. Better results were achieved by submissions that used an ensemble of taggers.

We also attempted combinations of the outputs from the Tanl NER (with word2vec cluster features), Nlpnet NER and Stanford NER in several ways. The best results were obtained by a simple one voting approach, taking the union of all annotations. The results of the evaluation, for both the multiple copies and discount annotation style, are shown below:

NER	Precision	Recall	F- score
Agreement multiple	73.96	**73.68**	**73.82**
Agreement discont	**81.69**	65.85	72.92

Table 2: Accuracy of NER system combination.

4 Conclusions

We presented a series of experiments on biomedical texts from both medical literature and clinical records, in multiple languages, that helped us to refine the techniques of NE recognition and to adapt them to Italian. We explored supervised techniques as well as unsupervised ones, in the form of word embeddings or word clusters. We also developed a Deep Learning NE tagger that exploits embeddings. The best results were achieved by using a MEMM sequence labeler using clusters as features improved in an ensemble combination with other NE taggers.

As an further contribution of our work, we produced, by exploiting semi-automated techniques, an Italian corpus of medical records, annotated with mentions of medical terms.

Acknowledgements

Partial support for this work was provided by project RIS (POR RIS of the Regione Toscana, CUP n° 6408.30122011.026000160).

References

AIFA open data. 2014. Retrieved from: http://www.agenziafarmaco.gov.it/it/content/dati-sulle-liste-dei-farmaci-open-data

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In Proc. of Conference on Computational Natural Language Learning, CoNLL 2013, pp. 183-192, Sofia, Bulgaria.

Giuseppe Attardi et al., 2009. Tanl (Text Analytics and Natural Language Processing). SemaWiki project: http://medialab.di.unipi.it/wiki/SemaWiki

Giuseppe Attardi, et al. 2009. The Tanl Named Entity Recognizer at Evalita 2009. In *Proc. of Workshop Evalita'09* - Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, ISBN 978-88-903581-1-1.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proc. of Workshop Evalita'09* - Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, ISBN 978-88-903581-1-1.

Giuseppe Attardi, Andrea. Buzzelli, Daniele Sartiano. 2013. Machine Translation for Entity Recognition across Languages in Biomedical Documents. *Proc. of CLEF-ER 2013 Workshop*, September 23-26, Valencia, Spain.

Giuseppe Attardi, Vittoria Cozza, Daniele Sartiano. 2014. UniPi: Recognition of Mentions of Disorders in Clinical Text. *Proc. of the 8th International Workshop on Semantic Evaluation*. SemEval 2014, pp. 754–760

Giuseppe Attardi, Luca Baronti. 2014. Experiments in Identification of Temporal Expressions in Evalita 2014. *Proc. of Evalita 2014*.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.* 22, 1 (March 1996), 39-71.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, vol. 32, no. supplement 1, D267–D270.

Ronan Collobert et al. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, pp. 2461–2505.

CRF-NER. 2014. Retrieved from: http://nlp.stanford.edu/software/CRF-NER.shtml

EUROPARL. European Parliament Proceedings Parallel Corpus 1996-2011. 2014. http://www.statmt.org/europarl/

Jenny Rose Finkel, Trond Grenager and Christopher D. Manning 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of the 43nd Annual Meeting of the ACL*, 2005, pp. 363–370.

Neil Fraser. 2011. Diff, Match and Patch libraries for Plain Text.

JRC-Acquis Multilingual Parallel Corpus, Version 2.2. 2014. Retrieved from: http://optima.jrc.it/Acquis/index_2.2.html

Philipp Koehn, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL* on Interactive Poster and Demonstration Sessions. ACL.

MDRITA15_1. 2012. Medical Dictionary for Regulatory Activities Terminology (MedDRA) Version 15.1, Italian Edition; MedDRA MSSO; September, 2012.

MSHITA2013. 2013. Italian translation of Medical Subject Headings. Istituto Superiore di Sanità, Settore Documentazione. Roma, Italy.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*.

George B. Moody and Roger G. Mark. 1996. A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring. *Computers in Cardiology* 23:657–660.

NLPNET. 2014. Retrieved from https://github.com/attardi/nlpnet/

Sameer Pradhan, et al. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. *Proc. of the 8th International Workshop on Semantic Evaluation* (SemEval 2014), August 2014, Dublin, Ireland, pp. 5462.

Dietrich Rebholz-Schuhmann et al. 2010. The CALBC Silver Standard Corpus for Biomedical Named Entities - A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. In *Proc. of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta, Malta.

Dietrich Rebholz-Schuhmann, et al. 2013. Entity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview. *Lecture Notes in Computer Science*, Vol. 8138, 353-367

RIS: Ricerca e innovazione nella sanità. 2014. POR RIS of the Regione Toscana. homepage: http://progetto-ris.it/

SCIKIT. 2014 Retrieved from http://scikit-learn.org/

SENNA. Semantic/syntactic Extraction using a Neural Network Architecture. 2011. Retrieved from http://ml.nec-labs.com/senna/

Jannik Strotgen and Michael Gertz. Multilingual and Cross-domain Temporal Tagging. 2013. Language Resources and Evaluation, June 2013, Volume 47, Issue 2, pp 269-298.

Buzhou Tang et al. 2013. Recognizing and Encoding Discorder Concepts in Clinical Text using Machine Learning and Vector Space Model. *Workshop of ShARe/CLEF eHealth Evaluation Lab 2013*.

Buzhou Tang, et al. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed Research International*, Volume 2014, Article ID 240403.

Jorg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov et al., eds.: *Recent Advances in Natural Language Processing*. Volume V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, pp. 237–248.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL'03 Shared Task: Language-Independent Named Entity Recognition. In: *Proc. of CoNLL '0*3, Edmonton, Canada, 142–147.

WORD2VEC. 2014 Retrieved from http://code.google.com/p/word2vec/

10.12871/CLICIT201415

Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado

Alessia Barbagli•, Pietro Lucisano•,
Felice Dell'Orletta◇, Simonetta Montemagni◇, Giulia Venturi◇
•Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"
alessia.barbagli@gmail.com, pietro.lucisano@uniroma1.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

Abstract

Italiano. L'ultimo decennio ha visto l'affermarsi a livello internazionale dell'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento. Questo contributo, che si colloca all'interno di una ricerca più ampia di pedagogia sperimentale, riporta i primi e promettenti risultati di uno studio finalizzato al monitoraggio dell'evoluzione del processo di apprendimento della lingua italiana condotto a partire dalle produzione scritte degli studenti con strumenti di annotazione linguistica automatica e di estrazione di conoscenza.

English. *Over the last ten years, the use of language technologies was successfully extended to the study of learning processes. The paper reports the first results of a study, which is part of a broader experimental pedagogy project, aimed at monitoring the evolution of the learning process of the Italian language based on a corpus of written productions by students and exploiting automatic linguistic annotation and knowledge extraction tools.*

1 Introduzione

L'uso di tecnologie del linguaggio per lo studio dei processi di apprendimento e, in termini più applicativi, di costruzione dei cosiddetti *Intelligent Computer–Assisted Language Learning systems (ICALL)* è sempre più al centro di ricerche interdisciplinari che mirano a mettere in luce come metodi e strumenti di annotazione linguistica automatica e di estrazione della conoscenza siano oggi maturi per essere usati anche nel contesto educativo e scolastico. A livello internazionale, ciò è dimostrato dal successo del *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, arrivato nel 2014 alla sua nona edizione[1].

Il presente contributo si pone in questo contesto di ricerca, riportando i primi risultati di uno studio tuttora in corso, finalizzato a descrivere, con strumenti di carattere quantitativo e qualitativo, l'evoluzione delle abilità di scrittura, sia a livello del contenuto testuale sia delle competenze linguistiche, dalla prima alla seconda classe della scuola secondaria di primo grado. Si tratta di un lavoro esplorativo finalizzato a costruire un modello di analisi empirica in grado di consentire l'osservazione dei processi e dei prodotti della didattica della produzione scritta. Il carattere innovativo di questa ricerca nel panorama nazionale e internazionale si colloca a vari livelli.

Sul versante metodologico, la ricerca qui delineata rappresenta il primo studio finalizzato al monitoraggio dell'evoluzione del processo di apprendimento linguistico della lingua italiana condotto a partire dalle produzione scritte degli studenti e con strumenti di annotazione linguistica automatica e di estrazione di conoscenza. L'utilizzo di tecnologie del linguaggio per il monitoraggio dell'evoluzione della competenza linguistica di apprendenti affonda le radici in un filone di studi avviato a livello internazionale nell'ultimo decennio e all'interno del quale analisi linguistiche generate da strumenti di trattamento automatico del linguaggio sono usate, ad esempio, per: monitorare lo sviluppo della sintassi nel linguaggio infantile (Sagae et al., 2005; Lu, 2007); identificare deficit cognitivi attraverso misure di complessità sintattica (Roark et al., 2007) o di associazione semantica (Rouhizadeh et al., 2013); monitorare la capacità di lettura come componente centrale della competenza linguistica (Schwarm e Ostendorf, 2005; Petersen e Ostendorf, 2009). Prendendo le mosse da questo filone di ricerca, Dell'Orletta e Montemagni (2012) e Dell'Orletta et al. (2011) hanno di-

[1] http://www.cs.rochester.edu/∼tetreaul/acl-bea9.html

mostrato all'interno di due studi di fattibilità che le tecnologie linguistico–computazionali possono giocare un ruolo centrale nella valutazione della competenza linguistica italiana di studenti in ambito scolastico e nel tracciarne l'evoluzione attraverso il tempo. Questo contributo rappresenta uno sviluppo originale e innovativo di questa linea di ricerca, in quanto la metodologia di monitoraggio linguistico proposta è utilizzata all'interno di uno studio più ampio di pedagogia sperimentale, basato su un corpus significativo di produzioni scritte di studenti e finalizzato a rintracciare l'evoluzione delle competenze in una prospettiva diacronica e/o socio–culturale.

L'oggetto delle analisi rappresenta un altro elemento di novità: è stato scelto il primo biennio della scuola secondaria di primo grado come ambito scolastico da analizzare perché poco indagato dalle ricerche empiriche e poiché poche sono state sino ad oggi le indagini che hanno verificato l'effettiva pratica didattica derivata dalle indicazioni previste dai programmi ministeriali relativi a questo ciclo scolastico, a partire dal 1979 fino alle Indicazioni Nazionali del 2012.

2 Il contesto e i dati della ricerca

Il contesto di riferimento è rappresentato dalla ricerca IEA IPS (*Association for the Evaluation of Educational Achievement, Indagine sulla Produzione Scritta*) che agli inizi degli anni '80 ha coinvolto quattordici paesi di tutto il mondo (tra cui l'Italia) in un'indagine sull'insegnamento e sull'apprendimento della produzione scritta nella scuola. I risultati dell'indagine sono riportati in Purves (1992), e per l'Italia in Lucisano (1988) e Lucisano e Benvenuto (1991).

Lo studio più ampio, tuttora in corso, in cui il presente contributo si colloca si basa sull'ipotesi che nei due anni presi in esame si realizzino dei cambiamenti rilevanti nelle modalità di approccio alla scrittura degli studenti e nella loro produzione scritta, e che tali cambiamenti siano dovuti allo stimolo di un insegnamento più formale. Si ritiene che tali cambiamenti possono essere verificati osservando le variazioni che risultano dai prodotti dell'attività di scrittura scolastica.

La ricerca è stata organizzata individuando tre tipi di variabili: di sfondo (es. background familiare, territoriale, personale), di processo (es. misura di abilità linguistiche degli studenti) e di prodotto (es. misure sui testi degli studenti).

Abbiamo preso come riferimento un campione di giudizio composto da studenti di sette diverse scuole secondarie di primo grado di Roma; la scelta delle scuole è avvenuta basandosi sul presupposto che esista una relazione tra l'area territoriale in cui è collocata la scuola e l'ambiente socio–culturale di riferimento. Sono state individuate due aree territoriali: il centro storico e la periferia rappresentativi rispettivamente di un ambiente socio–culturale medio–alto e medio–basso. Per ogni scuola è stata individuata una classe, per un totale di 77 studenti in centro e 79 in periferia. Per ogni studente, sono state raccolte due tipologie di produzioni scritte: le tracce assegnate dai docenti nei due anni scolastici e due prove comuni relative alla percezione dell'insegnamento della scrittura, svolte dalle classi al termine del primo e del secondo anno[2]. È stato così possibile raccogliere un corpus di 1.352 testi che sono stati digitalizzati per le successive fasi di analisi. Per entrambe le tipologie di produzioni, l'analisi ha riguardato sia il contenuto sia la struttura linguistica sottostante. In quanto segue, ci focalizzeremo sull'analisi delle prove comuni sull'insegnamento della scrittura.

3 Analisi delle produzioni scritte

Il corpus di produzioni scritte, una volta digitalizzato, è stato arricchito automaticamente con annotazione morfo–sintattica e sintattica. A tal fine è stata utilizzata una piattaforma consolidata di metodi e strumenti per il trattamento automatico dell'italiano sviluppati congiuntamente dall'ILC–CNR e dall'Università di Pisa[3]. Per quanto riguarda l'annotazione morfo–sintattica, lo strumento utilizzato è descritto in Dell'Orletta (2009); sul versante dell'analisi a dipendenze, abbiamo utilizzato DeSR (Attardi et al., 2009).

Il testo arricchito con informazione linguistica ("linguisticamente annotato") costituisce il punto di partenza per le analisi successive, riconducibili a due filoni principali finalizzati rispettivamente all'identificazione dei contenuti più salienti e alla definizione del profilo delle competenze linguistiche di chi lo ha prodotto. L'accuratezza dell'annotazione, seppur decrescente attraverso i diversi livelli, è sempre più che accettabile da permettere la tracciabilità nel testo di una vasta tipologia di

[2] Le tracce somministrate derivano dalla Prova 9 della Ricerca IEA–IPS (Lucisano, 1984; Corda Costa e Visalberghi, 1995) che consiste in una lettera di consigli indirizzata a un coetaneo su come scrivere un tema.

[3] http://linguistic-annotation-tool.italianlp.it/

tratti riguardanti diversi livelli di descrizione linguistica, che possono essere sfruttati nei compiti di monitoraggio linguistico (Montemagni, 2013).

Partendo dall'assunto di base che i termini costituiscono la rappresentazione linguistica dei concetti più salienti di una collezione documentale e per questo motivo il compito di estrazione terminologica costituisce il primo e fondamentale passo verso l'accesso al suo contenuto, il corpus delle prove comuni morfo–sintatticamente annotato è stato sottoposto ad un processo di estrazione terminologica finalizzato all'identificazione e all'estrazione delle unità lessicali monorematiche e polirematiche rappresentative del contenuto. A tal fine è stato utilizzato $T2K^2$ (Text–to–Knowledge)[4], una piattaforma web finalizzata all'acquisizione di informazione semantico–lessicale da corpora di dominio (Dell'Orletta et al., 2014). Il monitoraggio delle competenze e abilità linguistiche degli apprendenti, che rappresenta un ambito inesplorato di applicazione di tali tecnologie, ha riguardato sia il livello lessicale sia aspetti della struttura linguistica (in particolare, morfo–sintattica e sintattica). A questo scopo è stato usato MONITOR–IT, lo strumento che, implementando la strategia di monitoraggio descritta in Montemagni (2013), analizza la distribuzione di un'ampia gamma di caratteristiche linguistiche (di base, lessicali, morfo–sintattiche e sintattiche) rintracciate automaticamente in un corpus a partire dall'output dei diversi livelli di annotazione linguistica (Dell'Orletta et al., 2013): la Tabella 1 riporta una selezione dei tratti più significativi utilizzati.

4 Primi risultati

I risultati riguardano il corpus delle prove comuni di scrittura somministrate nel primo e secondo anno per un totale di 240 testi. L'analisi di ciascuna collezione è stata condotta sia in rapporto al contenuto che alla struttura linguistica in relazione ad un'ampia gamma di tratti linguistici.

4.1 Analisi del contenuto

La Tabella 2 riporta i primi 15 termini estratti da $T2K^2$ a partire dalle prove comuni del primo e del secondo anno, ordinati per rilevanza. Tra i termini più salienti emersi dall'analisi delle prove del primo anno si segnalano *paura dei compiti, paura dei lavori di scrittura* così come *difficoltà nei compiti, esperienza in quinta*, che rivelano una tipologia di consigli appartenente alla sfera psico–emotiva. Nel secondo anno, i termini più significativi estratti dal testo fanno riferimento a consigli che riguardano aspetti più 'tecnici' come *uso di parole, pertinenza alla traccia, uso dei verbi*.

È interessante qui far notare come tali risultati siano in linea con la codifica manuale del contenuto condotta sulla base della griglia predisposta dalla ricerca IEA (Fabi e Pavan De Gregorio, 1988; Asquini, 1993; Asquini et al., 1993), che divide i consigli contenuti nei testi in sei macro aree (Contenuto, Organizzazione, Stile e registro, Presentazione, Procedimento, Tattica). Analizzando i risultati ottenuti sono evidenti i cambiamenti avvenuti tra il primo e il secondo anno. Nel primo anno la maggior parte dei consigli dati riflettono la didattica della scuola primaria e pertengono all'area della tattica; anche i consigli relativi al procedimento, sono focalizzati sulla sfera del comportamento e della realtà psico–emotiva. Nel secondo anno l'attenzione si sposta verso gli aspetti più prettamente linguistici come il contenuto e la presentazione (che comprende ortografia, calligrafia e grammatica), riflettendo la didattica della scuola secondaria di primo grado. La differenza appare ancora più significativa nel confronto tra i consigli più frequenti delle prove dei due anni. I consigli che hanno registrato le maggiori frequenze nelle prove del primo anno riguardavano esclusivamente l'aspetto psico–emotivo e il comportamento (es. *Aspetta un po', rifletti prima di scrivere*, *Leggi/scrivi molto*, *Non avere paura*) mentre nelle prove del secondo anno tra i dieci consigli più frequenti (es. *Scrivi con calligrafia ordinata*, *Usa una corretta ortografia*, *Attieniti all'argomento; solo i punti pertinenti*) non compare nessun consiglio di tattica.

4.2 Analisi della struttura linguistica

Il monitoraggio comparativo tra le caratteristiche linguistiche rintracciate nel corpus di prove comuni realizzate nel primo e nel secondo anno è stato condotto con l'obiettivo di tracciare l'evoluzione delle abilità linguistiche degli studenti nei due anni. Dall'ANOVA delle prove comuni risulta che esistono differenze significative tra primo e secondo anno a tutti i livelli di analisi linguistica considerati. Ad esempio, rispetto alle catteristiche 'di base' risulta che la variazione del *numero medio di token per frase* nelle due prove dei due anni

[4] http://www.italianlp.it/demo/t2k-text-to-knowledge/

Catteristiche di base
Lunghezza media dei periodi e delle parole
Catteristiche lessicali
Percentuale di lemmi appartenenti al *Vocabolario di Base (VdB)* del *Grande dizionario italiano dell'uso* (De Mauro, 2000)
Distribuzione dei lemmi rispetto ai repertori di uso (Fondamentale, Alto uso, Alta disponibilità)
Type/Token Ratio (TTR) rispetto ai primi 100 e 200 tokens
Catteristiche morfo–sintattiche
Distribuzione delle categorie morfo–sintattiche
Densità lessicale
Distribuzione dei verbi rispetto al modo, tempo e persona
Catteristiche sintattiche
Distribuzione dei vari tipi di relazioni di dipendenza
Arità verbale
Struttura dell'albero sintattico analizzato (es. altezza media dell'intero albero, lunghezza media delle relazioni di dipendenza)
Uso della subordinazione (es. distribuzione di frasi principali vs. subordinate, lunghezza media di sequenze di subordinate)
Modificazione nominale (es. lunghezza media dei complementi preposizionali dipendenti in sequenza da un nome)

Tabella 1: Selezione delle caratteristiche linguistiche più salienti oggetto di monitoraggio linguistico.

Prova I anno			**Prova II anno**		
DomainRel.	Termine	Freq.	DomainRel.	Termine	Freq.
1	compiti di scrittura	26	1	errori di ortografia	15
2	maestra di italiano	21	2	professoressa di italiano	10
3	lavori di scrittura	17	3	uso di parole	9
4	compiti in classe	30	4	tema in classe	7
5	errori di ortografia	11	5	compiti in classe	9
6	paura dei compiti	9	6	pertinenza alla traccia	4
7	compiti in classe d'italiano	7	7	professoressa di lettere	4
8	anno di elementari	7	8	tema	369
9	classe d'italiano	7	9	voti a tema	3
10	compiti di italiano	7	10	temi a piacere	3
11	maestra	405	11	contenuto del tema	3
12	compiti per casa	6	12	errori di distrazione	3
13	esperienze in quinta	4	13	professoressa	131
14	maestra delle elementari	4	14	frasi	80
15	maestra di matematica	4	15	traccia	81

Tabella 2: Un estratto dell'estrazione terminologica automatica dalle prove comuni del I e II anno.

è significativa. Mentre le prove scritte nel primo anno contengono frasi lunghe in media 23,82 token, la lunghezza media delle frasi delle prove del secondo anno è pari a 20,71 token. Significativa è anche la variazione nell'uso di voci riconducibili al *VdB*, che diminuisce dall'83% del vocabolario nelle prove del primo anno al 79% nel secondo anno, così come i valori di *TTR* (rispetto ai primi 100 tokens), che aumentano passando dallo 0,66 allo 0,69. In entrambi i casi, tali mutamenti possono essere visti come conseguenza di un arricchimento lessicale. Per quanto riguarda il livello morfo–sintattico, sono soprattutto le caratteristiche che catturano l'uso dei tempi e dei modi verbali a essere particolarmente significative. A livello del monitoraggio sintattico, è ad esempio l'uso del *complemento oggetto in posizione pre– o post–verbale* a variare significativamente. Se nelle prove del primo anno il 19% dei complementi oggetto è in posizione pre–verbale, nel secondo anno la percentuale diminuisce passando al 13%; mentre nel primo anno i complementi oggetti post–verbali sono l'81% e aumentano passando all'87% nel secondo anno. Nelle prove del secondo anno si osserva dunque un maggiore rispetto dell'ordinamento canonico soggetto–verbo–oggetto, più vicino alle norme dello scritto che del parlato.

Sebbene i risultati ottenuti siano ancora preliminari rispetto al più ampio contesto della ricerca, crediamo mostrino chiaramente le potenzialità dell'incontro tra linguistica computazionale ed educativa, aprendo nuove prospettive di ricerca. Le linee di attività in corso includono l'analisi della correlazione tra le evidenze acquisite attraverso il monitoraggio linguistico e le variabili di processo e di sfondo così come lo studio dell'evoluzione delle abilità linguistiche del singolo studente.

References

G. Asquini, G. De Martino, L. Menna. 1993. Analisi della prova 9. In AA.VV *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise*, IRRSAE MOLISE, Campobasso, Lampo, pp. 77–100.

G. Asquini. 1993. Prova 9 lettera di consigli. In AA.VV *La produzione scritta nel biennio superiore. Ricerca nelle scuole superiori del Molise*, IRRSAE MOLISE, Campobasso, Lampo, pp. 67–75.

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

M. Corda Costa and A. Visalberghi. 1995. *Misurare e valutare le competenze linguistiche. Guida scientifico-pratica per gli insegnanti*. Firenze, ed. La Nuova Italia.

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia.

F. Dell'Orletta, S. Montemagni, E.M. Vecchi, and G. Venturi. 2011. Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. In G.C. Bruno, I. Caruso, M. Sanna, I. Vellecco (a cura di) *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, pp. 319–336, Milano, McGraw-Hill.

F. Dell'Orletta F. and S. Montemagni. 2012. Tecnologie linguistico–computazionali per la valutazione delle competenze linguistiche in ambito scolastico. In *Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010)*, 27-29 settembre, Viterbo.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2013. Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose. *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, 7–11 September, Hissar, Bulgaria, pp. 189–197.

F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni. 2014. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2062–2070, 26-31 May, Reykjavik, Iceland.

T. De Mauro. 2000. *Grande dizionario italiano dell'uso (GRADIT)*. Torino, UTET.

A Fabi and G. Pavan De Gregorio. 1988. La prova 9: risultati di una ricerca sui contenuti in una prova di consigli sulla scrittura. In *Ricerca educativa*, 5, pp. 2–3.

X. Lu. 2007. Automatic measurement of syntactic complexity in child language acquisition. In *International Journal of Corpus Linguistics*, 14(1), pp. 3–28.

P. Lucisano. 1984. L'indagine IEA sulla produzione scritta. In *Ricerca educativa*, Numero 5.

P. Lucisano. 1988. La ricerca IEA sulla produzione scritta. In *Ricerca educativa*, 2–3, pp. 3–13.

P. Lucisano e G. Benvenuto. 1991. Insegnare a scrivere: dalla parte degli insegnanti. In *Scuola e Città*, 6, pp. 265–279.

S. Montemagni. 2013. Tecnologie linguistico–computazionali e monitoraggio della lingua italiana. In *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, Anno XLII, Numero 1, pp. 145–172.

S.E. Petersen e M. Ostendorf. 2009. A machine learning approach to reading level assessment. In *Computer Speech and Language (23)*, pp. 89–106.

A. Purvues. 1992. *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries vol 6*. Oxford, Pergamon.

B. Roark, M. Mitchell, K. Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pp. 1–8.

M. Rouhizadeh, E. Prud'hommeaux, B. Roark, J. van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 709–714.

K. Sagae, A. Lavie, B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 197–204.

S.E. Schwarm e M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05)*, pp. 523–530.

10.12871/CLICIT201416

Italian Irony Detection in Twitter: a First Approach*

Francesco Barbieri, Francesco Ronzano, Horacio Saggion
Universitat Pompeu Fabra, Barcelona, Spain
name.surname@upf.edu

Abstract

English. Irony is a linguistic device used to say something but meaning something else. The distinctive trait of ironic utterances is the opposition of literal and intended meaning. This characteristic makes the automatic recognition of irony a challenging task for current systems. In this paper we present and evaluate the first automated system targeted to detect irony in Italian Tweets, introducing and exploiting a set of linguistic features useful for this task.

Italian. *L'ironia è una figura retorica mediante la quale si vuole conferire a una espressione un significato differente da quello letterale. Il riconoscimento automatico dell'ironia è reso difficile dalla sua principale caratteristica: il contrasto tra significato inteso e significato letterale. In questo studio proponiamo e valutiamo il primo sistema per il riconoscimento automatico di Tweets ironici in italiano.*

1 Introduction

Sentiment Analysis is the interpretation of attitudes and opinions of subjects on certain topics. With the growth of social networks, Sentiment Analysis has become fundamental for customer reviews, opinion mining, and natural language user interfaces (Yasavur et al., 2014). During the last decade the number of investigations dealing with sentiment analysis has considerably increased, targeting most of the time English language. Comparatively and to the best of our knowledge there are only few works for the Italian language.

*The research described in this paper is partially funded by the Spanish fellowship RYC-2009-04291, the SKATER-TALN_UPF project (TIN2012-38584-C06-03), and the EU project Dr. Inventor (n. 611383).

In this paper we explore an important sentiment analysis problem: *irony detection*. Irony is a linguistic device used to say something when meaning something else (Quintilien and Butler, 1953). Dealing with figurative languages is one of the biggest challanges to correctly determine the polarity of a text: analysing phrases where literal and indented meaning are not the same, is hard for humans, hence even harder for machines. Moreover, systems able of detect irony can benefit also other A.I. areas like Human Computer Interaction.

Approaches to detect irony have been already proposed for English, Portuguese and Dutch texts (see Section 2). Some of these systems used words, or word-patterns as irony detection features (Davidov et al., 2010; González-Ibáñez et al., 2011; Reyes et al., 2013; Buschmeier et al., 2014). Other approaches, like Barbieri and Saggion (2014a), exploited lexical and semantic features of single words like their frequency in reference corpora or the number of associated synsets. Relying on the latter method, in this paper we present the first system for automatic detection of irony in Italian Tweets. In particular, we investigate the effectiveness of Decision Trees in classifying Tweets as ironic or not ironic, showing that classification performances increase by considering lexical and semantic features of single words instead of pure bag-of-words (BOW) approaches. To train our system, we exploited as ironic examples the Tweets from the account of a famous collective blog named Spinoza and as not ironic examples the Tweets retrieved from the timelines of seven popular Italian newspapers.

2 Related Work

The standard definition of irony is "saying the opposite of what you mean" (Quintilien and Butler, 1953). Grice (1975) believes that irony is a rhetorical figure that violates the maxim of quality, while Giora (1995) says that irony can be any

form of negation with no negation markers. Wilson and Sperber (2002) defined irony as echoic utterance that shows a negative aspect of someone's else opinion. Utsumi (2000) and Veale and Hao (2010a) stated that irony is a form of pretence that is violated.

Irony has been approached computationally by Veale and Hao (2010b) who proposed an algorithm for separating ironic from non-ironic similes in English, detecting common terms used in this ironic comparison. Reyes et al. (2013) proposed a model to detect irony in English Tweets, pointing out that skipgrams which capture word sequences that contain (or skip over) arbitrary gaps, are the most informative features. Barbieri and Saggion (2014a) and Barbieri and Saggion (2014b) designed a model that avoided the use of the words (or pattern of words) as the use of single words or word-patterns as features. They focused on the lexical and semantic information that characterises each word in an Tweet, like its frequency in different corpora, its length, the number of associated synsets, etc. The system of Buschmeier et al. (2014) included features proposed in previous systems and gave for the first time a baseline for the irony detection problem in English (best F1-measure obtained was 0.74). Little research has been carried out on irony detection in languages other than English. Carvalho et al. (2009) and de Freitas et al. (2014) dealt with irony in Portuguese newspapers. Liebrecht et al. (2013) designed a model to detect irony in Dutch Tweets.

Gianti et al. (2012) collected and annotate a set of ironic examples from a common collective Italian blog. This corpus is also used in Bosco et al. (2013) for the study of sentiment analysis and opinion mining in Italian.

3 Data and Text Processing

The corpus[1] we used is composed of 25,450 Tweets: 12.5% are ironic and 87.5% non-ironic. The set of ironic examples (3,185) is an aggregation of the posts from the Twitter accounts "spinozait" and "LiveSpinoza". Spinoza is an Italian collective blog that includes posts of sharp satire on politics (the posts are suggested by the community and a group of volunteers filter the content to be published). Spinoza is a very popular blog and there is a collective agreement on the irony of its posts (Bosco et al., 2013). The non-ironic examples (22,295) are Tweets retrieved from Twitter accounts of the seven most popular Italian daily newspapers, including "Corriere della Sera", "Gazzetta dello Sport", "Il Messaggero", "Repubblica", "Il Resto del Carlino", "Il Sole 24 Ore", and "La Stampa". Almost the totality of these posts do not contain irony, they only describe news. We decided to consider newspaper Tweets as negative items for two reasons. Firstly because Spinoza Tweets are about politics and news, thus they deal with topics related to the same domain of Italian daily newspapers. Secondly, because the style of Spinoza Tweets is similar to the style typical of newspapers. Hence Spinoza and newspapers posts have have similar content, similar style, but different intentions.

In order to process the text and build our model we used freely available tools. We used the tokenizer, POS tagger and UKB Word Sense Disambiguation algorithm provided by Freeling (Carreras et al., 2004). We also exploited the Italian WordNet 1.6[2] to get synsets and synonyms, and the sentiment lexicon Sentix[3] (Basile and Nissim, 2013) derived from SentiWordnet (Esuli and Sebastiani, 2006). We used on the CoLFIS Corpus of Written Italian[4] to obtain the usage frequency of a word in written Italian.

4 Method

This section describes two systems: both exploit Decision Trees to classify Tweets as ironic or not. The first system (Section 4.1) is the irony detection approach we propose that relies on lexical and semantic features characterising each word of a Tweet. The second system (Section 4.2) exploits words occurrences (BOW approach) as features useful to train a Decision Tree. The latter system is used as a reference (baseline) to evaluate our irony detection approach.

4.1 Irony Detection Model

Our model for irony detection includes five types of features: Frequency, Synonyms, Ambiguity, Part of Speech, and Sentiments. We included in our model a subset of the features proposed by Barbieri and Saggion (2014a), describing implicit characteristics of each word in a Tweet. We do

[1] The reader can find the list of the Tweet IDs at http://sempub.taln.upf.edu/tw/clicit2014/

[2] http://multiwordnet.fbk.eu/english/home.php

[3] http://www.let.rug.nl/basile/twita/sentix.php

[4] http://linguistica.sns.it/CoLFIS/Home_eng.htm

not consider features such as punctuation, emoticons or number of characters of the Tweet. The proposed features aim to detect two aspects of Tweets that we consider particularly relevant to detect irony: the style used (e.g. register used, frequent or rare words, positive or negative words, etc.) and the unexpectedness in the use of words (Lucariello, 1994) i.e. the presence of "out of context" words (the *gap* feature, see below).

4.1.1 Frequency

We retrieved from the CoLFIS Corpus, the frequency of the word of each Tweet. Thus, we derive three types of Frequency features: *rarest word frequency* (frequency of the most rare word included in the Tweet), *frequency mean* (the arithmetic average of all the frequency of the words in the Tweet) and *frequency gap* (the difference between the two previous features). These features are computed for all the words of each Tweet. We also computed these features by considering only Nouns, Verbs, Adjectives, and Adverbs.

4.1.2 Synonyms

Irony conveys two messages to the audience at the same time, the literal and the intended message (Veale, 2004). We consider the frequencies (in CoLFIS Corpus) of the synonyms of each word in the Tweet, as retrieved from WordNet. Then we compute: the *greatest / lowest number of synonyms* with frequency higher than the one present in the Tweet, the *mean number of synonyms* with frequency greater / lower than the frequency of the related word present in the Tweet. We determine also the greatest / lowest number of synonyms and the mean number of synonyms of the word with frequency greater / lower than the one present in the the Tweet (*gap* feature). We also computed these features separately, considering each of the four POS as before

4.1.3 Ambiguity

Ambiguity plays an important role in irony: a word with more than one meaning can be used to say two (or more) things at the same time. To model the ambiguity of the terms in the Tweets we use the WordNet synsets associated to each word. Our hypothesis is that if a term has many meanings (synsets) it is more likely to be used in an ambiguous way. For each Tweet we calculate the *maximum number of synsets* associated to a single word, the *synset number mean* of all the words, and the *synset gap* that is the difference between the two previous features. We determine the value of these features considering all the words of a Tweet and as well as including only Nouns, Verbs, Adjectives or Adverbs.

4.1.4 Part Of Speech

The features included in the Part Of Speech group are designed to capture the style of the Tweets. The features of this group are eight and each of them counts the number of occurrences of words characterised by a certain POS. The eight POS considered are *Verbs*, *Nouns*, *Adjectives*, *Adverbs*, *Interjections*, *Determiners*, *Pronouns*, and *Adpositions*.

4.1.5 Sentiments

The sentiments of the words in ironic Tweets are important for two reasons: to detect the *sentiment* style (e.g. if ironic Tweets contain mainly positive or negative terms) and to capture unexpectedness created by a negative word in a positive context and viceversa. Relying on Sentix (see Section 3) we compute the *number of positive/negative words*, the *sum of the intensities of the positive/negative words*, the *mean of intensities of positive/negative words*, the *greatest positive/negative score*, the *gap between greatest positive/negative and positive/negative mean*. Then, as before we compute these features for each of the POSs Noun, Verb, Adjetive, and Adverbs.

4.2 Bag Of Word Baseline

Our baseline model is a Decision Tree trained on features represented by the occurrence of the 200 most frequent words in the training set (we calculate the frequent words in each experiment, see Section 5). We only considered words of the message itself, removing expressions such as the name of the newspapers and common patterns like "Continue to read [link]" or "See the Video Gallery on [link]" often present in specific Twitter accounts.

5 Experiments and Results

We obtained from our initial corpus two kinds of datasets: the ironic dataset (that includes all the Tweets from the two Spinoza accounts) and the non-ironic dataset (that is composed by the newspaper Tweets). We choose to classify tweets by a Decision Tree algorithm coupled with the SubsetEvaluation feature selection approach. For our

experiments we used Weka (Witten and Frank, 2005). We train our classifier in a dataset composed of 80% of the Tweets of the ironic dataset and 80% of the Tweets of the non-ironic dataset. The performance of the trained model are tested on a set of Tweets that includes the remaining portions of both ironic and non ironic datasets (20% of each dataset). Examples in the train and test sets are chosen randomly, to avoid correlation of Tweets close in time that are likely to be on the same topic. In addition we run a 10-cross validation using a balanced binary dataset (irony VS one negative topic). We carried out two experiments using the above framework (train/test and 10-cross validation):

1 - We consider as positive examples the ironic Tweets from Spinoza, and as negative examples each Tweet of the seven newspapers (this experiment is performed seven times, as we compare irony with each newspaper).

2 - We consider as positive example the ironic Tweets from Spinoza as before, while the negative dataset includes Tweets from all the seven newspaper (each newspaper contributes with a number of Tweets equal to 455).

We run the two experiments using both our feature set for irony detection (Section 4.1) and the BOW baseline features (Section 4.2). The results are reported in Table 1, organised in Precision, Recall and F-Measure.

6 Discussion

Our system always outperforms the BOW baseline. In Experiment 1 (irony versus each newspaper) our model outperforms the BOW approach by at least 4 points (F1). In Experiment 2 (irony versus a composition of all the newspapers) the results of BOW are still worse, six points less, and not due by chance (according to the McNemar's statistical test, 0.01 significance level). Moreover, in Experiment 2 the BOW baseline obtains its worst result, suggesting that this approach models the style of a specific newspaper rather than the ironic Tweets. On the other hand our system seems to better adapt to this situation indicating that it is less influenced by the non-ironic examples (a good characteristic as in a realistic case the non-ironic examples are unknown and of any type). The best features (information gain 0.20/0.15) are number of verbs and synset related features (Ambiguity, Section 4.1.3).

		Test Set			10-Folds		
	Data	**P**	**R**	**F**	**P**	**R**	**F**
Bag Of Words	Corr	.74	.68	.71	.72	.69	.70
	Gazz	.67	.70	.69	.71	.70	.70
	Mess	.71	.66	.68	.71	.67	.69
	Repu	.72	.68	.70	.70	.67	.69
	Rest	.77	.70	.73	.76	.72	.74
	Sol24	.71	.71	.71	.70	.70	.70
	Stam	.73	.66	.64	.70	.64	.66
	MIX	.69	.62	.65	.70	.61	.65
Our Model	Corr	.77	.76	.76	.78	.73	.75
	Gazz	.77	.76	.76	.75	.75	.75
	Mess	.73	.72	.72	.71	.70	.70
	Repu	.80	.75	.77	.73	.73	.73
	Rest	.87	.77	.82	.80	.78	.79
	Sol24	.76	.79	.78	.74	.72	.73
	Stam	.74	.75	.75	.74	.73	.72
	MIX	**.75**	**.76**	**.76**	.72	.70	.71

Table 1: Precision, Recall and F-Measure of each run of Experiment 1 and Experiment 2 ("MIX")

7 Conclusion and Future Work

In this study we evaluate a novel system to detect irony in Italian, focusing on Tweets. We tackle this problem as binary classification, where the ironic examples are posts of the Twitter account Spinoza and the non-ironic examples are Tweets from seven popular Italian newspapers. We evaluated the effectiveness of Decision Trees with different feature sets to carry out this classification task. Our system only focuses on characteristics on lexical and semantic information that characterises each word, rather than the words themselves as features. The performance of the system is good if compared to our baseline (BOW) considering only word occurrences as features, since we obtain an F1 improvement of 0.11. This result shows the suitability of our approach to detect ironic Italian Tweets. However, there is space to enrich and tune the model as this is only a first approach. It is possible to both improve the model with new features (for example related to punctuation or language models) and evaluate the system on new and extended corpora of Italian Tweets as they become available. Another issue we faced is the lack of accurate evaluations of features performance considering distinct classifiers / algorithms for irony detection.

References

Francesco Barbieri and Horacio Saggion. 2014a. Modelling Irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, Gothenburg, Sweden, April. Association for Computational Linguistics.

Francesco Barbieri and Horacio Saggion. 2014b. Modelling Irony in Twitter, Features Analysis and Evaluation. In *Language Resources and Evaluation conference, LREC*, Reykjavik, Iceland, May.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis and opinion mining: the case of irony and senti-tut. *Intelligent Systems, IEEE*.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews.

Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *LREC*.

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

Larissa A de Freitas, Aline A Vanin, Denise N Hogetop, Marco N Bochernitsan, and Renata Vieira. 2014. Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 628–633. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of Language Resources and Evaluation Conference*, volume 6, pages 417–422.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals, Istanbul, Turkey*, pages 1–7.

Rachel Giora. 1995. On irony and negation. *Discourse processes*, 19(2):239–264.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying Sarcasm in Twitter: A Closer Look. In *ACL (Short Papers)*, pages 581–586. Citeseer.

H Paul Grice. 1975. Logic and conversation. *1975*, pages 41–58.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013*, page 29.

Joan Lucariello. 1994. Situational irony: A concept of events gone awry. *Journal of Experimental Psychology: General*, 123(2):129.

Quintilien and Harold Edgeworth Butler. 1953. *The Institutio Oratoria of Quintilian. With an English Translation by HE Butler*. W. Heinemann.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, pages 1–30.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Tony Veale and Yanfen Hao. 2010a. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650.

Tony Veale and Yanfen Hao. 2010b. Detecting Ironic Intent in Creative Comparisons. In *ECAI*, volume 215, pages 765–770.

Tony Veale. 2004. The challenge of creative information retrieval. In *Computational Linguistics and Intelligent Text Processing*, pages 457–467. Springer.

Deirdre Wilson and Dan Sperber. 2002. Relevance theory. *Handbook of pragmatics*.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Ugan Yasavur, Jorge Travieso, Christine Lisetti, and Naphtali Rishe. 2014. Sentiment analysis using dependency trees and named-entities. In *The Twenty-Seventh International Flairs Conference*.

10.12871/CLICIT201417

A Retrieval Model for Automatic Resolution of Crossword Puzzles in Italian Language

Gianni Barlacchi[1] and **Massimo Nicosia**[2,1] and **Alessandro Moschitti**[2,1]
[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Qatar Computing Research Institute
{gianni.barlacchi, m.nicosia, amoschitti}@gmail.com

Abstract

English. In this paper we study methods for improving the quality of automatic extraction of answer candidates for an extremely challenging task: the automatic resolution of crossword puzzles for Italian language. Many automatic crossword puzzle solvers are based on database system accessing previously resolved crossword puzzles. Our approach consists in querying the database (DB) with a search engine and converting its output into a probability score, which combines in a single scoring model, i.e., a logistic regression model, both the search engine score and statistical similarity features. This improved retrieval model greatly impacts the resolution accuracy of crossword puzzles.

Italiano. *In questo lavoro abbiamo studiato metodi per migliorare la qualità dell'estrazione automatica di risposte da utilizzare nella risoluzione di cruciverba in lingua italiana. Molti risolutori automatici utilizzano database di definizioni e risposte provenienti da cruciverba risolti in precedenza. Il nostro approccio consiste nell'applicare tecniche di Information Retrieval alla base di dati, accedendo a questa per mezzo di un motore di ricerca. Gli score associati ai risultati sono combinati con altre misure di similiarità in un singolo modello di regressione logistica, che li convertite in probabilità. Il risultante modello é in grado di individuare con piú affidabilità definizioni simili e migliora significativamente l'accuratezza nella risoluzione dei cruciverba.*

1 Introduction

Crossword Puzzles (CPs) are probably one of the most popular language game. Automatic CP solvers have been mainly targeted by the artificial intelligence (AI) community, who has mostly focused on AI techniques for filling the puzzle grid, given a set of answer candidates for each clue. The basic idea is to optimize the overall probability of correctly filling the entire grid by exploiting the likelihood of each candidate answer, fulfilling at the same time the grid constraints. After several failures in approaching the human expert performance, it has become clear that designing more accurate solvers would not have provided a winning system. In contrast, the Precision and Recall of the answer candidates are obviously a key factor: very high values for these performance measures would enable the solver to quickly find the correct solution.

Similarly to the Jeopardy! challenge case (Ferrucci et al., 2010), the solution relies on Question Answering (QA) research. However, although some CP clues are rather similar to standard questions, there are some specific differences: (i) clues can be in interrogative form or not, e.g., ≪Capitale d'Italia: *Roma*≫; (ii) they can contain riddles or be deliberately ambiguous and misleading (e.g., ≪Se fugge sono guai: *gas*≫); (iii) the exact length of the answer keyword is known in advance; and (vi) the confidence in the answers is an extremely important input for the CP solver.

There have been many attempts to build automatic CP solving systems. Their goal is to outperform human players in solving crosswords more accurately and in less time. Proverb (Littman et al., 2002) was the first system for the automatic resolution of CPs. It includes several modules for generating lists of candidate answers. These lists are merged and used to solve a Probabilistic-Constraint Satisfaction Problem. Proverb relies on a very large crossword database as well as several expert modules, each of them mainly based on domain-specific databases (e.g., movies, writers and geography). WebCrow (Ernandes et al.,

2005) is based on Proverb. In addition to its predecessor, WebCrow carries out basic linguistic analysis such as Part-Of-Speech tagging and lemmatization. It takes advantage of semantic relations contained in WordNet, dictionaries and gazetteers. Its Web module is constituted by a search engine, which can retrieve text snippets or documents related to the clue. WebCrow uses a WA* algorithm (Pohl, 1970) for Probabilistic-Constraint Satisfaction Problems, adapted for CP resolution. To the best of our knowledge, the state-of-the-art system for automatic CP solving is Dr. Fill (Ginsberg, 2011). It targets the crossword filling task with a Weighted-Constraint Satisfaction Problem. Constraint violations are weighted and can be tolerated. It heavily relies on huge databases of clues.

All of these systems queries the DB of previously solved CP clues using standard techniques, e.g., SQL Full-Text query. The DB is a very rich and important knowledge base. In order to improve the quality of the automatic extraction of answer candidate lists from DB, we provide for the Italian language a completely novel solution, by substituting the DB and the SQL function with a search engine for retrieving clues similar to the target one. In particular, we define a reranking function for the retrieved clues based on a logistic regression model (LRM), which combines the search engine score with other similarity features. To carry out our study, we created a clue similarity dataset for the Italian language. This dataset constitutes an interesting resource that we made available to the research community[1].

2 WebCrow Architecture

We compare our methods with one of the best systems for automatic CP resolution, WebCrow (Ernandes et al., 2005). It was kindly made available by the authors. The solving process is divided in two phases: in the first phase, the coordinator module forwards the clues of an input CP to a set of modules for the generation of several candidate answer lists. Each module returns a list of possible solutions for each clue. Such individual clue lists are then merged by a specific *Merger* component, which uses list confidence values and the probabilities of correctness of each candidate in the lists. Eventually, a single list of candidate-probability pairs is generated for each input clue. During the second phase WebCrow fills the crossword grid

[1] http://ikernels-portal.disi.unitn.it/projects/webcrow/

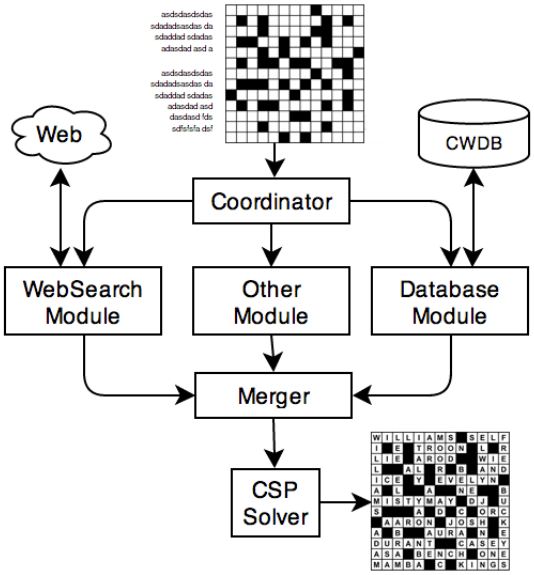


Figure 1: Overview of WebCrow's architecture.

by solving a constraint-satisfaction problem. WebCrow selects a single answer from each candidate merged list, trying to satisfy the imposed constraints. The goal of this phase is to find an admissible solution maximizing the number of correctly inserted words. In this paper, we focus on the DB module, and we describe it here.

Knowledge about previous CPs is essential for solving new ones. Indeed, clues often repeat in different CPs, thus the availability of a large DB of clue-answer pairs allows for easily finding the answers to previously used clues. To exploit the database of clue-answer pairs, WebCrow uses three different modules:

CWDB-EXACT, which simply checks for an exact match between the target clue and those in the DB. The score of the match is computed using the number of occurrences of the matched clue.

CWDB-PARTIAL, which employs MySQL's partial matching, query expansion and positional term distances to compute clue-similarity scores, along with the Full-Text search functions.

CWDB-DICTIO, which simply returns the full list of words of correct length, ranked by their number of occurrences in the initial list.

We compare our method with the CWDB-PARTIAL module. We improved it by applying a different retrieval function and using a linear model for scoring each possible answer.

3 Clue Retrieval from Database

This work is inspired by our earlier paper on learning to rank models for the automatic resolution of crossword puzzles for English language (Barlac-

Rank	Clue	Answer	Score
1	L'ente dei petroli	eni	8.835
2	Un colosso del petrolio	eni	8.835
3	Il petrolio americano	oil	8.835
4	Il petrolio della Mobil	oil	8.835
5	Il petrolio della Shell	oil	8.835

Table 1: Clue ranking for the query: *Il petrolio BP: oil*

chi et al., 2014). In that work, we showed that learning to rank models based on relational syntactic structures defined between the clues and the similar clue candidates can improve the retrieval of clues from a database of solved crossword puzzles. We cannot yet use our learning to rank model for the Italian language as we are implementing the needed syntactic/semantic parsers for such language. However, we have integrated the same search engine based on BM25 for Italian. Then, the completely new contribution is the use of supervised LRM to convert the Lucene scores into probabilities.

3.1 Clue Similarity for Italian language

WebCrow creates answer lists by retrieving clues from the DB of previously solved crosswords. As described before, it simply uses the classical SQL operator and full-text search. We verified the hypothesis that a search engine could achieve better results and we opted for indexing the DB of clues and their answers. We used the Open Source search engine Lucene (McCandless et al., 2010), its state-of-the-art BM25 retrieval model and the provided Italian Analyzer for processing the query. The analyzer performs basic operations, such as stemming and tokenization, over the input text. However, although this alone improved the quality of the retrieved clue list, a post-processing step is necessary for weighting the answer candidates appearing multiple times in the list. For example, Table 1 shows the first five clues, retrieved for a query originated by the clue: ≪Il petrolio BP: *oil*≫ (literally: *The petroleum BP*). Three answers out of five are correct, but they are not ranked before the others in the list. The Lucene scores of repeated candidates are not probabilities, thus their sum is typically not meaningful, i.e., it does not produce aggregated scores comparable between different answer candidates. For this reason, a LRM converts the Lucene score associated with each word into a probability. This way, we can sum the probabilities of the same answer candidates in the list and then normalize them considering the size of the list. We apply the following formula to obtain a single final score for each different answer candidate:

$$Score(G) = \sum_{c \in G} \frac{P^{LR}(y = 1|\vec{x}_c)}{n}$$

where c is the answer candidate, G is the set of answers matching exactly with c and n is the size of the answer candidate list. $\vec{x}_c$ is the feature vector associated with $c \in G$, $y \in \{0, 1\}$ is the binary class label ($y = 1$ when c is the correct answer). The conditional probability computed with the linear model is the following:

$$P^{LR}(y = 1|c) = \frac{1}{1 + e^{-y\vec{w}^T\vec{x}_c}}$$

where $\vec{w} \in \mathbb{R}^n$ is a weight vector (Yu et al., 2011).

In order to capture the distribution of the Lucene scores over the answer candidates list, we used the following simple features.

Lucene scores. These features are useful to characterize the distribution of the BM25 scores over the list. They include: the BM25 score of the target candidate and the maximum and minimum BM25 scores of the entire list. In particular, the last two features give the model information about the Lucene score range.

Rank. For each candidate answer c we include the rank $r \in [1, n]$ provided by the search engine Lucene. n is the size of the answer candidate list.

Clue distance. It quantifies how dissimilar the input clue and the retrieved clue are. This formula is mainly based on the well known Levenshtein distance.

For building the training set, we used a set of clues to query the search engine. We obtained candidates from the indexed clues and we marked them using the available ground truth. Clues sharing the same answer of the query clue are positive examples. During testing, clues are again used as search queries and the retrieved clue lists are classified.

4 Experiments

In this section, we present the results of our model. Our referring database of Italian clues is composed by 46,270 unique pairs of clue-answer, which belong to three different crossword editors.

4.1 Experimental Setup

For training the classifier we used Scikit-learn LRM implementation (Pedregosa et al., 2011) with default parameters. To measure the impact of the rerankers as well as the baselines, we used well known metrics for assessing the accuracy of

Model	MRR	AvgRec	REC@1	REC@5	REC@10
WebCrow	73.00	78.13	64.93	83.51	86.11
BM25	77.27	86.30	65.75	93.40	100.00
BM25+LRM	**81.20**	**88.94**	**71.12**	**95.70**	**100.00**

Table 2: Clue retrieval

QA and retrieval systems, i.e.: Recall at rank 1 (R@1, 5 and 10), Mean Reciprocal Rank (MRR), the average Recall (AvgRec). R@k is the percentage of questions with a correct answer ranked at the first position. MRR is computed as follows: $MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank(q)}$, where *rank(q)* is the position of the first correct answer in the candidate list. AvgRec and all the measures are evaluated on the first 10 retrieved clues.

4.2 Similar clue retrieval

We created the training and test sets using the clues contained in the database. The database of clues can be indexed for retrieving similar clues. The training set contains 10,000 clues whose answer may be found in the first ten position. With the same approach, we created a test set containing 1,000 clues that (i) are not in the training set and (ii) have at least an answer in the first ten position. We used the search engine to retrieve the clues in both training and test dataset creation.

We experimented with two simple different models: (i) BM25 and (ii) BM25 + LRM. However, since WebCrow includes a database module, in Tab. 2, we have an extra row indicating its accuracy evaluated using the CWDB-PARTIAL module. We note that in the BM25 model the list is ranked using the Lucene score while, in the BM25 + LRM the list is ranked using the probability score as described in the previous section. The result derived from the test set show that:
(i) BM25 is very accurate, i.e., an MRR of 77.27%. It improves on WebCrow about 4.5 absolute percent points, demonstrating the superiority of an IR approach over DB methods.
(ii) LRM achieves higher MRR, up to 4 absolute percent points of improvement over BM25 and thus about 8.5 points more than WebCrow.
(iii) Finally, the relative improvement on REC@1 is up to 9.5% (6.19% absolute). This high result is promising in the light of improving WebCrow for the end-to-end task of solving complete CPs.

4.3 Impact on WebCrow

In these experiments, we used our retrieval model for similar clues (BM25+LRM) using 5 complete CPs (for a total of 397 clues) created for a past Italian competition, organized by the authors of WebCrow. This way, we could measure the impact of our model on the complete task carried out by WebCrow. More specifically, we give our list of answers to WebCrow in place of the list that would have been extracted by the CWDB module . It should be noted that to evaluate the impact of our list, we disabled the WebCrow access to other lists, e.g., dictionaries. This means that the absolute resolution accuracy of WebCrow using our and its own lists can be higher (see (Ernandes et al., 2008) for more details).

Model	MRR	REC@1	REC@5	REC@10
WebCrow	30.89	27.63	35.17	36.14
Our Model	**34.41**	**29.36**	**36.92**	**38.93**

Table 3: Performance on the word list candidates averaged over the clues of 5 entire CPs

The first result that we derived is the accuracy of the answer list produced from the new data, i.e., constituted by the 5 entire CPs. The results are reported in Tab. 3. We note that the improvement of our model is lower than before as a non-negligible percentage of clues are not solved using the clue DB. Additionally, when we computed the accuracy in solving the complete CPs, we noted a small improvement: this happens because BM25 does not retrieve enough correct candidates for our specific test set constituted of five entire crossword puzzles.

5 Conclusions

In this paper, we improve the answer extraction from DB for automatic CP resolution. We combined the state-of-the-art BM25 retrieval model and an LRM by converting the BM25 score into a probability score for each answer candidate. For our study and to test our methods, we created a corpora for clue similarity containing clues in Italian. We improve on the lists generated by WebCrow by 8.5 absolute percent points in MRR. However, the end-to-end CP resolution test does not show a large improvement as the percentage of retrieved clues is not high enough.

Acknowledgments

We would like to thank Marco Gori and Marco Ernandes for making available WebCrow. The research described in this paper has been partially supported by the EU FP7 grant #288024: LIMOSINE – Linguistically Motivated Semantic aggregation engiNes. Many thanks to the anonymous reviewers for their valuable work.

References

Gianni Barlacchi, Massimo Nicosia, and Alessandro Moschitti. 2014. Learning to rank answer candidates for automatic resolution of crossword puzzles. *CoNLL-2014*.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2005. Webcrow: A web-based system for crossword solving. In *In Proc. of AAAI '05*, pages 1412–1417. Menlo Park, Calif., AAAI Press.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2008. A web-based agent challenges human experts on crosswords. *AI Magazine*, 29(1).

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79.

Matthew L. Ginsberg. 2011. Dr.fill: Crosswords and an implemented solver for singly weighted csps. *J. Artif. Int. Res.*, 42(1):851–886, September.

Michael L. Littman, Greg A. Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134(1,Äì2):23 – 55.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ira Pohl. 1970. Heuristic search viewed as path finding in a graph. *Artificial Intelligence*, 1(3–4):193 – 204.

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.*, 85(1-2):41–75, October.

10.12871/CLICIT201418

Analysing Word Meaning over Time by Exploiting Temporal Random Indexing

Pierpaolo Basile and **Annalina Caputo** and **Giovanni Semeraro**
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.lastname}@uniba.it

Abstract

English. This paper proposes an approach to the construction of *WordSpace*s which takes into account temporal information. The proposed method is able to build a geometrical space considering several periods of time. This methodology enables the analysis of the time evolution of the meaning of a word. Exploiting this approach, we build a framework, called Temporal Random Indexing (*TRI*) that provides all the necessary tools for building *WordSpace*s and performing such linguistic analysis. We propose some examples of usage of our tool by analysing word meanings in two corpora: a collection of Italian books and English scientific papers about computational linguistics.

Italiano. *In questo lavoro proponiamo un approccio per la costruzione di WordSpaces che tengano conto di informazioni temporali. Il metodo proposto costruisce degli spazi geometrici considerando diversi intervalli temporali. Questa metodologia permette di studiare l'evoluzione nel tempo del significato delle parole. Utilizzando questo approccio abbiamo costruito uno strumento, chiamato Temporal Random Indexing (TRI), che permette la costruzione dei WordSpaces e fornisce degli strumenti per l'analisi linguistica. Nell'articolo proponiamo alcuni esempi di utilizzo del nostro tool analizzando i significati delle parole in due corpus: uno relativo a libri nella lingua italiana, l'altro relativo ad articoli scientifici in lingua inglese nell'ambito della linguistica computazionale.*

1 Introduction

The analysis of word-usage statistics over huge corpora has become a common technique in many corpus linguistics tasks, which benefit from the growth rate of available digital text and computational power. Better known as Distributional Semantic Models (DSM), such methods are an easy way for building geometrical spaces of concepts, also known as *Semantic* (or *Word*) *Spaces*, by skimming through huge corpora of text in order to learn the context of usage of words. In the resulting space, semantic relatedness/similarity between two words is expressed by the closeness between word-points. Thus, the semantic similarity can be computed as the cosine of the angle between the two vectors that represent the words. DSM can be built using different techniques. One common approach is the Latent Semantic Analysis (Landauer and Dumais, 1997), which is based on the Singular Value Decomposition of the word co-occurrence matrix. However, many other methods that try to take into account the word order (Jones and Mewhort, 2007) or predications (Cohen et al., 2010) have been proposed. Recursive Neural Network (RNN) methodology (Mikolov et al., 2010) and its variant proposed in the *word2vect* framework (Mikolov et al., 2013) based on the continuous bag-of-words and skip-gram model take a complete new perspective. However, most of these techniques build such *SemanticSpaces* taking a *snapshot* of the word co-occurrences over the linguistic corpus. This makes the study of semantic changes during different periods of time difficult to be dealt with.

In this paper we show how one of such DSM techniques, called Random Indexing (RI) (Sahlgren, 2005; Sahlgren, 2006), can be easily extended to allow the analysis of semantic changes of words over time. The ultimate aim is to provide a tool which enables to understand how words

change their meanings within a document corpus as a function of time. We choose RI for two main reasons: 1) the method is incremental and requires few computational resources while still retaining good performance; 2) the methodology for building the space can be easily expanded to integrate temporal information. Indeed, the disadvantage of classical DSM approaches is that *WordSpace*s built on different corpus are not comparable: it is always possible to compare similarities in terms of neighbourhood words or to combine vectors by geometrical operators, such as the tensor product, but these techniques do not allow a direct comparison of vectors belonging to two different spaces. Our approach based on RI is able to build a *WordSpace* on different time periods and makes all these spaces comparable to each another, actually enabling the analysis of word meaning changes over time by simple vector operations in *WordSpace*s.

The paper is structured as follows: Section 2 provides details about the adopted methodology and the implementation of our framework. Some examples of the potentiality of our framework are reported in Section 3. Lastly, Section 4 closes the paper.

2 Methodology

We aim at taking into account temporal information in a DSM approach, which consists in representing words as points in a *WordSpace*, where two words are similar if represented by points close to each other. Hence, this *Temporal WordSpace* will be suitable for analysing how word meanings change over time. Under this light, RI has the advantages of being very simple, since it is based on an incremental approach, and is easily adaptable to the *temporal* analysis needs. The *WordSpace* is built taking into account words co-occurrences, according to the distributional hypothesis (Harris, 1968) which states that words sharing the same linguistic contexts are related in meaning. In our case the linguistic context is defined as the words that co-occur with the *temporal* word, i.e. the word under the temporal analysis. The idea behind RI has its origin in Kanerva work (Kanerva, 1988) about the Sparse Distributed Memory. RI assigns a context vector to each unit; in our case, each word represents a context. The context vector is generated as a high-dimensional random vector with a high number of zero elements and a few number of elements equal to 1 or -1 randomly distributed over the vector dimensions. Vectors built using this approach generate a nearly orthogonal space. During the incremental step, a vector is assigned to each temporal element as the sum of the context vectors representing the context in which the temporal element is observed.The mathematical insight behind the RI is the projection of a high-dimensional space on a lower dimensional one using a random matrix; this kind of projection does not compromise distance metrics (Dasgupta and Gupta, 1999).

Formally, given a corpus C of n documents, and a vocabulary V of m words extracted form C, we perform two steps: 1) assigning a context vector c_i to each word in V; 2) generating for each word w_i a semantic vector sv_i computed as the sum of all the context vectors assigned to the words co-occurring with w_i. The context is the set of m words that precede and follow w_i. The second step can be defined by the equation:

$$sv_i = \sum_{d \in C} \sum_{-m<i<+m} c_i \qquad (1)$$

After these two steps, we obtain a set of semantic vectors assigned to each word in V representing our *WordSpace*.

2.1 Temporal Random Indexing

The classical RI does not take into account temporal information, but it can be easily adapted to the methodology proposed in (Jurgens and Stevens, 2009) for our purposes. In particular, we need to add a metadata containing information about the year in which the document was written, to each document in C. Then, Temporal RI can build several *WordSpace*s T_k for different time periods, with these spaces being comparable to each other. This means that a vector in the *WordSpace* T_1 can be compared with vectors in the space T_2. The first step in the classical RI is unchanged in Temporal RI, and represents the strength of our approach: the use of the same context vectors for all the spaces makes them comparable. The second step is similar to the one proposed for RI but it takes into account the temporal period. Let T_k be a period that ranges between years $y_{k_{start}}$ and $y_{k_{end}}$, where $y_{k_{start}} < y_{k_{end}}$; then, for building the *WordSpace* T_k we consider only the documents d_k written during T_k.

$$sv_{i_{T_k}} = \sum_{d_k \in C} \sum_{-m<i<+m} c_i \qquad (2)$$

Using this approach we can build a *WordSpace* for each time period over a corpus C tagged with information about the publication year.

2.2 The TRI System

We build a system, called TRI, able to perform Temporal RI using a corpus of documents with temporal information. TRI provides a set of features: 1) to build a *WordSpace* for each year, provided that a corpus of documents with temporal information is available; 2) to merge *WordSpace*s that belong to a particular time period (the new *WordSpace* can be saved on disk or stored in memory for further analysis); 3) to load a *WordSpace* and fetch vectors; 4) to combine and sum vectors; 5) to retrieve similar vectors using the cosine similarity; 6) to extract the neighbourhood of a word or compare neighbourhoods in different spaces for the temporal analysis of a word meaning. All these features can be combined to perform linguistic analysis using a simple shell. Section 3 describes some examples. The TRI system is developed in JAVA and is available on-line[1] under the GNU v.3 license.

3 Evaluation

The goal of this section is to show the usage of the proposed framework for analysing the changes of word meaning over time. Moreover, such analysis supports the detection of linguistics events that emerge in specific time intervals related to social or cultural phenomena. To perform our analysis we need a corpus of documents tagged with time metadata. Then, using our framework, we can build a *WordSpace* for each year. We study the semantics related to a word by analysing the nearest words in the *WordSpace*. For example, we can analyse how the meaning of word has changed in an interval spanning several periods of time. Given two time period intervals and a word w, we can build two *WordSpace*s (T_1 and T_2) by summing the *WordSpace*s assigned to the years that belong to each time period interval. Then using the cosine similarity, we can rank and select the nearest words of w in the two *WordSpace*s, and measure how the semantics of w is changed. Due to the fact that TRI makes *WordSpace*s comparable, we can extract the vectors assigned to w in T_1 and in T_2, and compute the cosine similarity between them. The similarity shows how the semantic of w is changed over time; a similarity equals to 1 means that the word w holds the same semantics. We adopt this last approach to detect words that mostly changed their semantics over time and analyse if this change is related to a particular social or cultural phenomenon. To perform this kind of analysis we need to compute the divergence of semantics for each word in the vocabulary.

Gutenberg Dataset. The first collection consists of Italian books with publication year by the Project Gutenberg[2] made available in text format. The total number of collected books is 349 ranging from year 1810 to year 1922. All the books are processed using our tool TRI creating a *WordSpace* for each available year in the dataset. For our analysis we create two macro temporal periods, before 1900 (T_{pre900}) and after 1900 ($T_{post900}$). The space T_{pre900} contains information about the period 1800-1899, while the space $T_{post900}$ contains information about all the documents in the corpus. As a first example, we analyse how the neighbourhood of the word *patria* (*homeland*) changes in T_{pre900} and $T_{post900}$. Table 1 shows the ten most similar words to *patria* in the two time periods; differences between them are reported in bold. Some words *(legge, virtù, onore)*[3] related to fascism propaganda occur in $T_{post900}$, while in T_{pre900} we can observe some concepts *(nostra, causa, italia)*[4] probably more related to independence movements in Italy. As an example, analysing word meaning evolution over time, we observed that the word *cinematografo* (*cinema*) clearly changes its semantics: the similarity of the word *cinematrografo* in the two spaces

Table 1: Neighbourhood of *patria* (*homeland*).

T_{pre900}	$T_{post900}$
libertà	libertà
opera	gloria
pari	giustizia
comune	comune
gloria	**legge**
nostra	pari
causa	**virtù**
italia	**onore**
giustizia	opera
guerra	**popolo**

[1] https://github.com/pippokill/tri

[2] http://www.gutenberg.org/

[3] In English: *(law/order, virtue, honour)*.

[4] In English: *(our, reason, Italy)*.

Table 2: Neighbourhoods of *semantics* across several decades.

1960-1969	1970-1979	1980-1989	1990-1999	2000-2010	2010-2014
linguistics	**natural**	syntax	syntax	syntax	syntax
theory	linguistic	natural	theory	theory	theory
semantic	semantic	**general**	interpretation	interpretation	interpretation
syntactic	**theory**	theory	general	description	description
natural	syntax	semantic	linguistic	**meaning**	complex
linguistic	language	syntactic	description	linguistic	meaning
distributional	processing	linguistic	**complex**	logical	linguistic
process	syntactic	**interpretation**	natural	complex	logical
computational	description	**model**	representation	representation	structures
syntax	**analysis**	**description**	**logical**	**structures**	representation

is very low, about 0.40. To understand this change we analysed the neighbourhood in the two spaces and we noticed that the word *sonoro* (*sound*) is strongly related to *cinematografo* in $T_{post900}$. This phenomenon can be ascribed to the sound introduction after 1900.

ANN Dataset. The ACL Anthology Network Dataset (Radev et al., 2013)[5] contains 21,212 papers published by the Association of Computational Linguistic network, with all metadata (authors, year of publication and venue). We split the dataset in decades (1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2010, 2010-2014), and for each decade we build a different *WordSpace* with TIR. Each space is the sum of *WordSpace*s belonging to all the previous decades plus the one under consideration. In this way we model the whole word history and not only the semantics related to a specific time period. Similarly to the Gutenberg Dataset, we first analyse the neighbourhood of a specific word, in this case *semantics*, and then we run an analysis to identify words that have mostly changed during the time. Table 2 reports in bold, for each decade, the new words that entered in the neighbourhood of *semantics*. The word *distributional* is strongly correlated to *semantics* in the decade 1960-1969, while it disappears in the following decades. Interestingly, the word *meaning* popped up only in the decade 2000-2010, while *syntax* and *syntactic* have always been present.

Regarding the word meaning variation over time, it is peculiar the case of the word *bioscience*. Its similarity in two different time periods, before 1990 and the latest decade, is only 0.22. Analysing its neighbourhood, we can observe that before 1990 *bioscience* is related to words such as *extraterrestrial* and *extrasolar*, nowadays the same word is related to *medline*, *bionlp*, *molecular* and *biomedi*. Another interesting case is the word *unsupervised*, which was related to *observe*, *partition*, *selective*, *performing*, before 1990; while nowadays has correlation of *supervised*, *disambiguation*, *technique*, *probabilistic*, *algorithms*, *statistical*. Finally, the word *logic* changes also its semantics after 1980. From 1979 to now, its difference in similarity is quite low (about 0.60), while after 1980 the similarity increases and always overcomes the 0.90. This phenomenon can be better understood if we look at the words *reasoning* and *inference*, which have started to be related to the word *logic* only after 1980.

4 Conclusions

We propose a method for building *WordSpace*s taking into account information about time. In a *WordSpace*, words are represented as mathematical points and the similarity is computed according to their closeness. The proposed framework, called TRI, is able to build several *WordSpace*s in different time periods and to compare vectors across the spaces to understand how the meaning of a word has changed over time. We reported some examples of our framework, which show the potential of our system in capturing word usage changes over time.

Acknowledgements

This work fullfils the research objectives of the project PON 01 00850 ASK-Health (Advanced System for the interpretation and sharing of knowledge in health care).

[5] Available on line: http://clair.eecs.umich.edu/aan/

References

Trevor Cohen, Dominique Widdows, Roger W. Schvaneveldt, and Thomas C. Rindflesch. 2010. Logical Leaps and Quantum Connectives: Forging Paths through Predication Space. In *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pages 11–13.

Sanjoy Dasgupta and Anupam Gupta. 1999. An elementary proof of the Johnson-Lindenstrauss lemma. Technical report, Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA.

Zellig S. Harris. 1968. *Mathematical Structures of Language*. New York: Interscience.

Michael N. Jones and Douglas J. K. Mewhort. 2007. Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, 114(1):1–37.

David Jurgens and Keith Stevens. 2009. Event Detection in Blogs using Temporal Random Indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16. Association for Computational Linguistics.

Pentti Kanerva. 1988. *Sparse Distributed Memory*. MIT Press.

Thomas K. Landauer and Susan T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, 104(2):211–240.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent Neural Network based Language Model. In *INTERSPEECH*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL Anthology Network Corpus. *Language Resources and Evaluation*, pages 1–26.

Magnus Sahlgren. 2005. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics.

10.12871/CLICIT201419

Combining Distributional Semantic Models and Sense Distribution for Effective Italian Word Sense Disambiguation

Pierpaolo Basile and **Annalina Caputo** and **Giovanni Semeraro**
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.surname}@uniba.it

Abstract

English. Distributional semantics approaches have proven their ability to enhance the performance of overlap-based Word Sense Disambiguation algorithms. This paper shows the application of such a technique to the Italian language, by analysing the usage of two different Distributional Semantic Models built upon ItWaC and Wikipedia corpora, in conjunction with two different functions for leveraging the sense distributions. Results of the experimental evaluation show that the proposed method outperforms both the most frequent sense baseline and other state-of-the-art systems.

Italiano. *Gli approcci di semantica distribuzionale hanno dimostrato la loro capacità nel migliorare le prestazioni degli algoritmi di Word Sense Disambiguation basati sulla sovrapposizione di parole. Questo lavoro descrive l'applicazione di questa tipologia di tecniche alla lingua italiana, analizzando l'utilizzo di due diversi Modelli di Semantica Distribuzionale costruiti sui corpora ItWaC e Wikipedia, in combinazione con due diverse funzioni che sfruttano le distribuzioni dei significati. I risultati della valutazione sperimentale mostrano la capacità di questo metodo di superare le prestazioni sia della baseline rappresentata dal senso più comune che di altri sistemi a stato dell'arte.*

1 Introduction

Given two words to disambiguate, Lesk (1986) algorithm selects those senses which maximise the overlap between their definitions (i.e. glosses), then resulting in a pairwise comparison between all the involved glosses. Since its original formulation, several variations of this algorithm have been proposed in an attempt of reducing its complexity, like the *simplified* Lesk (Kilgarriff and Rosenzweig, 2000; Vasilescu et al., 2004), or maximizing the chance of overlap, like in the *adapted* version (Banerjee and Pedersen, 2002). One of the limitations of Lesk approach relies on the exact match between words in the sense definitions. Semantic similarity, rather than word overlap, has been proposed as a method to overcome such a limitation. Earlier approaches were based on the notion of semantic relatedness (Patwardhan et al., 2003) and tried to exploit the relationships between synsets in the WordNet graph. More recently, Distributional Semantic Models (DSM) have stood up as a way for computing such semantic similarity. DSM allow the representation of concepts in a geometrical space through word vectors. This kind of representation captures the semantic relatedness that occurs between words in paradigmatic relations, and enables the computation of semantic similarity between whole sentences. Broadening the definition of semantic relatedness, Patwardhan and Pedersen (2006) took into account WordNet contexts: a gloss vector is built for each word sense using its definition and those of related synsets in WordNet. A distributional thesaurus is used for the expansion of both glosses and the context in Miller et al. (2012), where the overlap is computed as in the original Lesk algorithm. More recently, Basile et al. (2014) proposed a variation of Lesk algorithm based on both the simplified and the adapted version. This method combines the enhanced overlap, given by the definitions of related synsets, with the reduced number of matching that are limited to the contextual words in the simplified version. The evaluation was conducted on the SemEval-2013 Multilingual Word Sense Disambiguation task (Navigli et al., 2013), and involved the use of BabelNet

(Navigli and Ponzetto, 2012) as sense inventory. While performance for the English task was above the other task participants, the same behaviour was not reported for the Italian language.

This paper proposes a deeper investigation of the algorithm described in Basile et al. (2014) for the Italian language. We analyse the effect on the disambiguation performance of the use of two different corpora for building the distributional space. Moreover, we introduce a new sense distribution function (SDfreq), based on synset frequency, and compare its capability in boosting the distributional Lesk algorithm with respect to the one proposed in Basile et al. (2014).

The rest of the paper is structured as follows: Section 2 provides details about the *Distributional* Lesk algorithm and DSM, and defines the two above mentioned sense distribution functions exploited in this work. The evaluation, along with details about the two corpora and how the DSM are built, is presented in Section 3, which is followed by some conclusions about the presented results.

2 Distributional Lesk Algorithm

The distributional Lesk algorithm (Basile et al., 2014) is based on the simplified version (Vasilescu et al., 2004) of the original method. Let $w_1, w_2, ...w_n$ be a sequence of words, the algorithm disambiguates each target word w_i by computing the semantic similarity between the glosses of the senses associated to the target word and its context. This similarity is computed by representing in a DSM both the gloss and the context as the sum of the words they are composed of; then this similarity takes into account the co-occurrence evidences previously collected through a corpus of documents. The corpus plays a key role since the richer it is the higher is the probability that each word is fully represented in all its contexts of use. Finally, the correct sense for a word is selected by choosing the one whose gloss maximizes the semantic similarity. Despite the use of a *SemanticSpace* for computing the similarity, still the sense description can be too short for a meaningful comparison with the word context. Following this observation, we adopted an approach inspired by the adapted Lesk (Banerjee and Pedersen, 2002), and we decided to enrich the gloss of the sense with those of related meanings, duly weighted to reflect their distances with respect to the original sense. As sense inventory we choose BabelNet 1.1, a huge multilingual semantic network which comprises both WordNet and Wikipedia. The algorithm consists of the steps described as follows.

Building the glosses. We retrieve the set $S_i = \{s_{i1}, s_{i2}, ..., s_{ik}\}$ of senses associated to w_i by firstly looking up to the WordNet portion of BabelNet, then if no sense is found we seek for senses from Wikipedia, since probably the word is a named entity. This strategy was selected after tuning our system. For each sense s_{ij}, the algorithm builds the extended gloss representation g^*_{ij} by adding to the original gloss g_{ij} the glosses of related meaning retrieved through the BabelNet function "getRelatedMap", with the exception of "antonym" senses. Each word in g^*_{ij} is weighted by a function inversely proportional to the distance d between s_{ij} and the related glosses where the word occurs. Moreover, in order to emphasize more discriminative words among the different senses, we introduce in the weight a variation of the inverse document frequency (*idf*) for retrieval that we named inverse gloss frequency (*igf*). The *igf* for a word w_k occurring gf^*_k times in the set of extended glosses for all the senses in S_i, the sense inventory of w_i, is computed as $IGF_k = 1 + log_2 \frac{|S_i|}{gf^*_k}$. The final weight for the word w_k appearing h times in the extended gloss g^*_{ij} is given by:

$$weight(w_k, g^*_{ij}) = h \times IGF_k \times \frac{1}{1+d} \quad (1)$$

Building the context. The context C for the word w_i is represented by all the words that occur in the text.

Building the vector representations. The context C and each extended gloss g^*_{ij} are represented as vectors in the *SemanticSpace* built through the DSM described in Subsection 2.1.

Sense ranking. The algorithm computes the cosine similarity between the vector representation of each extended gloss g^*_{ij} and that of the context C. Then, the cosine similarity is linearly combined with a function which takes into account the usage of the meaning in the language. In this paper we investigate the two functions described in Subsection 2.2. The output of this step is a ranked list of synsets. The sense with the highest similarity is selected.

2.1 Distributional Semantics

Distributional Semantic Models are a means for representing concepts through vectors in *Semantic* (or *Word*) *Spaces*. Building the *SemanticSpace* only requires the analysis of big amounts of text data in order to collect evidence about word usage in the language in a complete unsupervised method. These methods rely on the construction of a word-to-word matrix M, which reflects the paradigmatic relations between words that share the same contexts, e.g. between words that can be used interchangeably. In this space, the vector proximity expresses the semantic similarity between words, traditionally computed as the cosine of the angle between the two word-vectors. Moreover, the concept of *semantic similarity* can be extended to whole sentences via the vector addition ($+$) operator. A sentence can always be represented as the sum of the word vectors it is composed of. Then, vector addition can be exploited to represent both the extended gloss and the target word context in order to asses their similarity.

2.2 Sense Distribution

We analyse two functions to compute the probability assigned to each synset. The first one has already been proposed in the original version of the distributional Lesk algorithm (Basile et al., 2014), the second one is based on synset frequency. It is important to point out that many synsets in BabelNet refer to named entities that do not occur in WordNet. In order to compute the probability of these synsets using a synset-tagged corpus we try to map them to WordNet and select the WordNet synset with the maximum probability. If no WordNet synset is provided, we assign a uniform probability to the synset.

Distribution based on conditional probability (SDprob). We define the probability $p(s_{ij}|w_i)$ that takes into account the sense distribution of s_{ij} given the word w_i. The sense distribution is computed as the number of times the word w_i is tagged with the sense s_{ij} in a sense-tagged corpus. Zero probabilities are avoided by introducing an additive (Laplace) smoothing. The probability is computed as follows:

$$p(s_{ij}|w_i) = \frac{t(w_i, s_{ij}) + 1}{\#w_i + |S_i|} \quad (2)$$

where $t(w_i, s_{ij})$ is the number of times the word w_i is tagged with the sense s_{ij}.

Distribution based on frequency (SDfreq). We compute the probability $p(s_{ij})$ of a meaning s_{ij} in a tagged corpus. The frequency is computed by taking into account all the occurrences of the whole set of meanings assigned to the word w_i. Given S_i, the set of the k possible meanings of w_i, the frequency of each s_{ij} in S_i is computed as:

$$p(s_{ij}) = \frac{t(s_{ij}) + 1}{\sum_{k=1}^{l}(t(s_{ik})) + |S_i|} \quad (3)$$

where $t(s_{ij})$ are the occurrences of s_{ij} in the tagged corpus.

3 Evaluation

The evaluation is performed using the dataset provided by the organizers of the Multilingual WSD (Task-12) of SemEval-2013 (Navigli et al., 2013), a traditional WSD all-words experiment where BabelNet is used as sense inventory. Our evaluation aims at: 1) analysing the algorithm performance changes in function of both the two synset distribution functions and the corpus used to built the DSM; 2) comparing our system with respect to the other task participants for the Italian language.

System Setup. Our algorithm[1] is developed in JAVA and exploits the BabelNet API 1.1.1[2]. We adopt the standard Lucene analyzer to tokenize both glosses and the context. The *SemanticSpace*s for the two corpora are built using proprietary code derived from (Widdows and Ferraro, 2008) which relies on two Lucene indexes, denoted as ItWaC and Wiki, containing documents form ItWaC Corpus (Baroni et al., 2009) and the Wikipedia dump for Italian, respectively. For each corpus, the co-occurrence matrix M contains information about the top 100,000 most frequent words. Co-occurrences are computed by taking into account a window of 5 words. M is built by using Random Indexing and by setting a reduced dimension equal to 400 and the seed to 10. Sense distribution functions are computed over MultiSemCor (Bentivogli and Pianta, 2005), a parallel (English/Italian) sense labelled corpus of SemCor. Since BabelNet Italian glosses are taken from MultiWordNet, which does not contain glosses for all the synsets, we replaced each missing gloss with the other synonym words that belong to the

[1] Available on line: `https://github.com/pippokill/lesk-wsd-dsm`

[2] Available on line: `http://lcl.uniroma1.it/babelnet/download.jsp`

Table 1: Comparison between DSMs with different Sense Distribution functions.

Run	*DSM*	*SenseDistr.*	*P*	*R*	*F*	*A*
MFS	-	-	0.572	0.572	0.572	-
ItWaC	ItWaC	-	0.614	0.613	0.613	99.73%
Wiki	Wiki	-	0.596	0.594	0.595	99.73%
ItWaCprob	ItWaC	SDprob	0.732	0.730	0.731	99.73%
ItWaCfreq	ItWaC	SDfreq	0.718	0.716	0.717	99.73%
Wikiprob	Wiki	SDprob	0.703	0.700	0.701	99.73%
Wikifreq	Wiki	SDfreq	0.700	0.698	0.699	99.73%

synset. The gloss term scoring function is always applied, since it provides better results. The synset distance d used to expand the gloss is fixed to 1 (the experiments with a distance d set to 2 did not result in any improvement). The sense distribution is linearly combined with the cosine similarity score through a coefficient set to 0.5. Using only sense distribution to select a sense is somehow similar to the most frequent sense (MFS) technique, i.e. the algorithm always assigns the most probable meaning. The MFS reported in Table 1 and Table 2 is the one computed by the task organizers in order to make results comparable. Evaluation is performed in terms of F measure.

Results of the Evaluation. Table 1 shows the results obtained by the distributional Lesk algorithm on the Italian language by exploiting different corpora and sense distribution functions. It is well known that the MFS approach obtains very good performance and it is hard to be outperformed, especially by unsupervised approaches. However, all the proposed systems are able to outperform the MFS, even those configurations that do not make use of sense distribution (ItWaC and Wiki). With respect to DSM, ItWaC corpus consistently provides better results (ItWaC vs. Wiki, ItWaCprob vs. Wikiprob, and ItWaCfreq vs. Wikifreq). By analysing the sense distribution functions, the best overall result is obtained when the SDprob function is exploited (ItWaCprob vs. ItWaCfreq), while there are no differences between SDprob and SDfreq in the DSM built on Wikipedia (Wikiprob vs. Wikifreq).

Table 2 compares the two systems built on the ItWaC corpus, with and without the sense distribution (SDprob), to the other task participants (UMCCDLSI2, DAEBAK!, GETALPBN) (Navigli et al., 2013). Moreover, we report the results of Babelfy (Moro et al., 2014) and UKB (Agirre et al., 2010), which hitherto have given the best performance on this dataset. While the system without sense distribution (ItWaC) is over the baseline but still below many task participants, the run which exploits the sense distribution (ItWaCprob) always outperforms the other systems.

Table 2: Comparison with other systems.

System	*F*
ItWaCprob	0.731
UKB	0.673
Babelfy	0.666
UMCC-DLSI-2	0.658
ItWaC	0.613
DAEBAK	0.613
MFS	*0.572*
GETALP-BN	0.528

4 Conclusions and Future Work

This paper proposed an analysis for the Italian language of an enhanced version of Lesk algorithm, which replaces the word overlap with distributional similarity. We analysed two DSM built over the ItWaC and Wikipedia corpus along with two sense distribution functions (SDprob and SDfreq). The sense distribution functions were computed over MultiSemCor, in order to avoid missing references between Italian and English synsets. The combination of the ItWaC-based DSM with the SDprob function resulted in the best overall result for the Italian portion of SemEval Task-12 dataset.

Acknowledgements

This work fulfils the research objectives of the project "VINCENTE - A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" (PON 02 00563 3470993) funded by the Italian Ministry of University and Research (MIUR).

References

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based Word Sense Disambiguation of Biomedical Documents. *Bioinformatics*, 26(22):2889–2896, November.

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 136–145. Springer Berlin Heidelberg.

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.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34(1-2):15–48.

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.

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, Mumbai, India, December. The COLING 2012 Organizing Committee.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.

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.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, April.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'03, pages 241–257, Berlin, Heidelberg. Springer-Verlag.

Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 633–636.

Dominic Widdows and Kathleen Ferraro. 2008. Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

10.12871/CLICIT2014110

A Lesk-inspired Unsupervised Algorithm for Lexical Choice from WordNet Synsets

Valerio Basile
University of Groningen, The Netherlands
v.basile@rug.nl

Abstract

English. The generation of text from abstract meaning representations involves, among other tasks, the production of lexical items for the concepts to realize. Using WordNet as a foundational ontology, we exploit its internal network structure to predict the best lemmas for a given synset without the need for annotated data. Experiments based on re-generation and automatic evaluation show that our novel algorithm is more effective than a straightforward frequency-based approach.

Italiano. *La generazione di testo a partire da rappresentazioni astratte comporta, tra l'altro, la produzione di materiale lessicale per i concetti da generare. Usando WordNet come ontologia fondazionale, ne sfruttiamo la struttura interna per individuare il lemma più adatto per un dato synset, senza ricorrere a dati annotati. Esperimenti basati su ri-generazione e valutazione automatica mostrano che il nostro algoritmo è più efficace di un approccio diretto basato sulle frequenze.*

1 Introduction

Many linguists argue that true synonyms don't exist (Bloomfield, 1933; Bolinger, 1968). Yet, words with similar meanings do exist and they play an important role in language technology where lexical resources such as WordNet (Fellbaum, 1998) employ *synsets*, sets of synonyms that cluster words with the same or similar meaning. It would be wrong to think that any member of a synset would be an equally good candidate for every application. Consider for instance the synset {food, nutrient}, a concept whose gloss in WordNet is "any substance that can be metabolized by an animal to give energy and build tissue". In (1), this needs to be realized as "food", but in (2) as "nutrient".

1. It said the loss was significant in a region where fishing provides a vital source of **food**|**nutrient**.
2. The Kind-hearted Physician administered a stimulant, a tonic, and a **food**|**nutrient**, and went away.

A straightforward solution based on n-gram models or grammatical constraint ("a food" is ungrammatical in the example above) is not always applicable, since it would be necessary to generate the complete sentence first, to exploit such features. This problem of lexical choice is what we want to solve in this paper. In a way it can be regarded as the reverse of WordNet-based Word Sense Disambiguation, where instead of determining the right synset for a certain word in a given context, the problem is to decide which word of a synset is the best choice in a given context.

Lexical choice is a key task in the larger framework of Natural Language Generation, where an ideal model has to produce varied, natural-sounding utterances. In particular, generation from purely semantic structures, carrying little to no syntactic or lexical information, needs solutions that do not depend on pre-made choices of words to express generic concepts. The input to a lexical choice component in this context is some abstract representation of meaning that may specify to different extent the linguistic features that the expected output should have.

WordNet synsets are good candidate representations of word meanings, as WordNet could be seen as a dictionary, where each synset has its own definition in written English. WordNet synsets are also well suited for lexical choice, because they consist in actual sets of lemmas, considered to be synonyms of each other in specific contexts. Thus, the problem presented here is restricted to

the *choice of lemmas from WordNet synsets*.

Despite its importance, the task of lexical choice problem is not broadly considered by the NLG community, one of the reasons being that it is hard to evaluate. Information retrieval techniques fail to capture not-so-wrong cases, i.e. when a system produces a different lemma from the gold standard but still appropriate to the context.

In this paper we present an unsupervised method to produce lemmas from WordNet synsets, inspired by the literature on WSD and applicable to every abstract meaning representation that provides links from concepts to WordNet synsets.

2 Related Work

Stede (1993) already noticed the need to exploit semantic context, when investigating the criteria for lexical choice in NLG. Other systems try to solve the lexical choice problem by considering situational aspects of the communication process such as pragmatics (Hovy, 1987), argumentative intent (Elhadad, 1991) or the degree of salience of semantic elements (Wanner and Bateman, 1990).

A whole line of research in NLG is focused on domain-specific or domain-independent generation from ontologies. Few works have underlined the benefits of a general concept hierarchy, such as the Upper Model (Bateman, 1997) or the MIAKT ontology (Bontcheva and Wilks, 2004), to serve as pivot for different application-oriented systems. Bouayad-Agha et al. (2012) employ a layered framework where an upper ontology is used together with a domain and a communication ontology for the purpose of robust NLG.

WordNet can be seen as an upper ontology in itself, where the synsets are concepts and the hypernym/hyponym relation is akin to generalization/specialization. However, to our knowledge, WordNet has not been used so far as supporting ontology for generation, even though there exists work on the usefulness of such resource for NLG-related tasks such as domain adaptation and paraphrasing (Jing, 1998).

3 The Ksel Algorithm

The Lesk algorithm (Lesk, 1986) is a classic solution to the Word Sense Disambiguation problem that, despite its simple scheme, achieves surprisingly good results by only relying on an external knowledge source, e.g. a dictionary. Inspired by the Lesk approach to WSD, and by the symmetrical relation between WSD and our present problem, we devised an algorithm that exploits semantic similarity between candidate lemmas of a synset and its semantic context. We call this algorithm *Ksel*. Lesk computes the relatedness between the candidate senses for a lemma and the linguistic context as a function of all the words in the synsets' definitions and the context itself – in the simplest case the function is computed by considering just word overlap. Similarly, Ksel computes a score for the candidates lemmas as a function of all the synsets they belong to and the semantic context. Just as not every word in a synset gloss is relevant to the linguistic context, not every synset of a lemma will be related to the semantic context, but carefully choosing the aggregation function will weed out the unwanted elements. The intuition is that in most cases the synsets of a word in WordNet are related to each other, just as Lesk's original algorithm for WSD leverages the fact that the words in a sense definition are often semantically related.

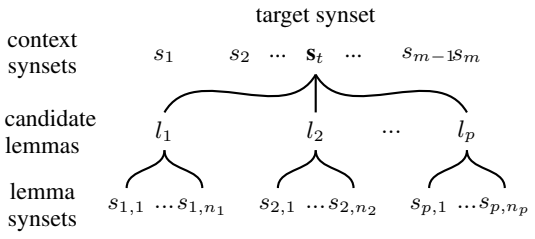


Figure 1: Elements of the Ksel algorithm.

Referring to Figure 1, the task at hand is that of choosing the right lemma l among the candidates $l_1, l_2, ..., l_p$ for the target synset s_t. The other synsets given in input form the context $C = s_1, ..., s_m, s_i \neq s_t$. We define the similarity between a lemma and a generic synset as a function of the similarities of all the synsets to which the lemma belongs and the synset under consideration:

$$s_{LS}(l_j, s_i) = f_1(sim(s_1, s_{j,k}) : 1 \leq k \leq n_j) \quad (1)$$

Using the lemma-synset similarity, we define the relatedness of a lemma to the semantic context as a function of the similarities of the lemma itself with the context synsets:

$$s_{LC}(l_j, C) = f_2(s_{LC}(l_j, s_i) : s_i \in C, 1 \leq i \leq m) \quad (2)$$

Three functions are still not specified in the definitions above – they are actually parameters of

the algorithm. f_1 and f_2 are aggregation functions over a set of similarity scores, that is, they take a set of real numbers, typically limited to the $[-1, 1]$ interval, and return a value in the same interval. sim is a similarity measure between WordNet synsets, like one of the many that have been proposed in literature – see Budanitsky and Hirst (2006) for a survey and an evaluation of WordNet-based similarity measures.

The target lemma, according to the Ksel algorithm, is the one that maximizes the measure in 2:

$$l_t = \arg\max_j s_{LC}(l_j, C) \tag{3}$$

To better clarify how Ksel works, here is an example of lexical choice between two candidate lemmas given a semantic context. The example is based on the sense-annotated sentence "The Kind-hearted Physician administered a stimulant, a tonic, and a food|nutrient, and went away.". The context C is the set of the synsets representing the meaning of the nouns "stimulant" (c_1 = {stimulant, stimulant drug, excitant}), "tonic" (c_2 = {tonic, restorative}) and "physician" (c_3 = {doctor, doc, physician, MD, Dr., medico}). The target synset is {food, nutrient}, for which the algorithm has to decide which lemma to generate between *food* and *nutrient*. *food* occurs in three synsets, while *nutrient* occurs in two:

- $s_{1,1}$: {food, nutrient}
- $s_{1,2}$: {food, solid_food}
- $s_{1,3}$: {food, food_for_thought, intellectual_nourishment}
- $s_{2,1}$: {food, nutrient}
- $s_{2,2}$: {nutrient}

For the sake of the example we will use the basic WordNet path similarity measure, that is, the inverse of the length of the shortest path between two synsets in the WordNet hierarchy. For each synset of *food*, we compute the mean of its path similarity with all the context synsets, and we take the average of the scores. This way, we have an aggregate measure of the semantic relatedness between a lemma (i.e. all of its possible synsets) and the semantic context under consideration. Then we repeat the process with *nutrient*, and finally choose the lemma with the highest aggregate similarity score. The whole process and the intermediate results are summarized in Table 1. Since .152 is greater than .117, the algorithm picks *nutrient* as the best candidate for this semantic context. Even if, for instance, $sim(s_{1,2}, c_1)$ were higher than 0.200, the aggregation mechanism would have averaged out the effect on the final choice of lemma.

Table 1: Running Ksel to select the best lemma between *food* and *nutrient* in a context composed of the three synsets c_1, c_2 and c_3.

lemma	**synset**	similarity to			**average**
		c_1	c_2	c_3	
food	$s_{1,1}$	.200	.166	.090	.152
food	$s_{1,2}$	.142	.125	.090	.119
food	$s_{1,3}$	.090	.083	.071	.081
lemma-context similarity (average):					**.117**
nutrient	$s_{2,1}$	.200	.166	.090	.152
nutrient	$s_{2,2}$	.200	.166	.090	.152
lemma-context similarity (average):					**.152**

4 Experiments

We conducted a few tests to investigate which parameters have influence over the performance of the Ksel algorithm. We took 1,000 documents out of the Groningen Meaning Bank (Basile et al., 2012), a semantically annotated corpus of English in which the word senses are encoded as WordNet synsets. The GMB is automatically annotated, partly corrected by experts and via crowdsourcing, and provides for each document an integrated semantic representation in the form of a Discourse Representation Structure (Kamp and Reyle, 1993), i.e. logical formulas consisting of predicates over discourse referents and relations between them. In the GMB, concepts are linked to WordNet synsets.

Our experiment consists of generating a lemma for each concept of a DRS, comparing it to the gold standard lemma, and computing the average precision and recall over the set of documents.

The Ksel algorithm, as described in Section 3, has three parameters functions. For the two aggregating functions, we experimented with mean, median and maximum. For the WordNet similarity measures between synsets, we took advantage of the Python NLTK library[1] that provides implementation for six different measures on WordNet 3.0 data:

- Path similarity, based on the shortest path that connects the synsets in the hypernym/hypnoym taxonomy.
- Leakcock & Chodorow's measure, which takes into account the maximum depth of the taxonomy tree (Leacock and Chodorow, 1998).
- Wu & Palmer's measure, where the distances are computed between the target synsets and

[1] http://www.nltk.org/

Table 2: Comparison of the performance of the Ksel algorithm with two baselines.

Method	Accuracy
Random	0.552
Most Frequent Lemma	0.748
Ksel (median, median, RES)	0.776

their most specific common ancestor (Wu and Palmer, 1994).

- Three methods based in Information Content: Resnik's measure (Resnik, 1995), Jiang's measure (Jiang and Conrath, 1997) and Lin's measure (Lin, 1998).

In the case of WSD, a typical baseline consists of taking the most frequent sense of the target word. The Most Frequent Sense baseline in WSD works very well, due to the highly skewed distribution of word senses. We investigate if the intuition behind the MFS baseline is applicable to the the lexical choice problem by reversing its mechanics, that is, the baseline looks at the frequency distribution of the target synset's lemmas in the data and selects the one that occurs more often.

We ran our implementation of Ksel on the GMB dataset with the goal of finding the best combination of parameters. Three alternatives for the aggregation functions and six different similarity measures result in 54 possible combination of parameters. For each possibility, we computed the accuracy relative to the gold standard lemmas in the data set corresponding to the concepts and found that the best choice of parameters is the median for both aggregation functions and the Resnik's measure for synset similarity.

Next we compared Ksel (with best-performing parameters) to a baseline that selects one uniformly random lemma among the set of synonyms, and the Most Frequent Lemma baseline described earlier. The results of the experiment, presented in Table 2, show how Ksel significantly outperform the MFL baseline. The accuracy of Ksel using Resnik's similarity measure with other aggregation functions range between 0.578.and 0.760.

5 Discussion

The aggregation functions play a big role in ruling out irrelevant senses from the picture, for instance the third sense of *food* in the example in Section 3 has very low similarity to the semantic context. As said earlier, the intuition is that the intra-relatedness of different synsets associated with the same words is generally high, with only few exceptions.

One case where the Ksel algorithm cannot be applied is when a synset is made of two or more monosemous words. In this case, a choice must be made that cannot be informed by semantic similarity, for example a random choice – this has been the strategy in this work. However, in our dataset only about 5% of all the synsets belong to this particular class.

WordNet synsets usually provide good quality synonyms for English lemmas. However, this is not always the case, for instance in some cases there are lemmas (or sequences of lemmas) that are not frequent in common language. As an example, the first synset of the English noun *month* is made of the two lemmas *month* and *calendar_month*. The latter occurs very seldom outside specific domains but Ksel produced it in 177 out of 181 cases in our experiment. Cases like this result in awkward realizations such as "Authorities blame Azahari bin Husin for orchestrating last *calendar month*'s attacks in Bali." (example from the test set). Fortunately, only a very small number of synsets are affected by this phenomenon.

Finally, it must be noted that Ksel is a totally unsupervised algorithm that requires only an external lexical knowledge base such as WordNet. This is not the case for other methods, including the MFL baseline.

6 Conclusion and Future Work

In this paper we presented an unsupervised algorithm for lexical choice from WordNet synsets called Ksel that exploits the WordNet hierarchy of hypernyms/hyponyms to produce the most appropriate lemma for a given synset. Ksel performs better than an already high baseline based on the frequency of lemmas in an annotated corpus.

The future direction of this work is at least twofold. On the one hand, being based purely on a lexical resource, the Ksel approach lends itself nicely to be applied to different languages by leveraging multi-lingual resources like BabelNet (Navigli and Ponzetto, 2012). On the other hand, we want to exploit existing annotated corpora such as the GMB to solve the lexical choice problem in a supervised fashion, that is, ranking candidate lemmas based on features of the semantic structure, in the same track of our previous work on generation from work-aligned logical forms (Basile and Bos, 2013).

References

Valerio Basile and Johan Bos. 2013. Aligning formal meaning representations with surface strings for wide-coverage text generation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 1–9, Sofia, Bulgaria, August. Association for Computational Linguistics.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Joost Venhuizen. 2012. Developing a large semantically annotated corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(1):15–55.

L. Bloomfield. 1933. *Language*. University of Chicago Press.

Dwight Bolinger. 1968. Entailment and the meaning of structures. *Glossa*, 2(2):119–127.

Kalina Bontcheva and Yorick Wilks. 2004. Automatic report generation from ontologies: The miakt approach. In *In Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems*, pages 324–335.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, Marco Rospocher, Horacio Saggion, Luciano Serafini, and Leo Wanner. 2012. From ontology to nl: Generation of multilingual user-oriented environmental reports. In Gosse Bouma, Ashwin Ittoo, Elisabeth Mtais, and Hans Wortmann, editors, *Natural Language Processing and Information Systems*, volume 7337 of *Lecture Notes in Computer Science*, pages 216–221. Springer Berlin Heidelberg.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Michael Elhadad. 1991. Generating adjectives to express the speaker's argumentative intent. In *In Proceedings of the 9th Annual Conference on Artificial Intelligence. AAAI*, pages 98–104.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689 – 719.

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

Hongyan Jing. 1998. Usage of wordnet in natural language generation. In *Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98) workshop on Usage of WordNet in Natural Language Processing Systems*, pages 128–134.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

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.

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.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Manfred Stede. 1993. Lexical choice criteria in language generation. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*, EACL '93, pages 454–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leo Wanner and John A. Bateman. 1990. A collocational based approach to salience-sensitive lexical selection.

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.

10.12871/CLICIT2014111

The Talmud System: a Collaborative web Application for the Translation of the Babylonian Talmud Into Italian

Andrea Bellandi, Davide Albanesi,
Alessia Bellusci, Andrea Bozzi, Emiliano Giovannetti
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. In this paper we introduce the Talmud System, a collaborative web application for the translation of the Babylonian Talmud into Italian. The system we are developing in the context of the "Progetto Traduzione del Talmud Babilonese" has been designed to improve the experience of collaborative translation using Computer-Assisted Translation technologies and providing a rich environment for the creation of comments and the annotation of text on a linguistic and semantic basis.

Italiano. *In questo articolo presentiamo il Sistema Talmud, un'applicazione web collaborativa per la traduzione del Talmud babilonese in italiano. Il sistema, che stiamo sviluppando nel contesto del "Progetto Traduzione del Talmud Babilonese", stato progettato per migliorare l'esperienza di traduzione collaborativa utilizzando tecnologie di Computer-Assisted Translation e fornendo un ambiente ricco per la creazione di commenti e l'annotazione del testo su base linguistica e semantica.*

1 Introduction

Alongside the Bible, the Babylonian Talmud (BT) is the Jewish text that has mostly influenced Jewish life and thought over the last two millennia. The BT corresponds to the effort of late antique scholars (*Amoraim*) to provide an exegesis of the *Mishnah*, an earlier rabbinic legal compilation, divided in six "orders" (*sedarim*) corresponding to different categories of Jewish law, with a total of 63 tractates (*massekhtaot*). Although following the inner structure of the *Mishnah*, the BT discusses only 37 tractates, with a total of 2711 double sided folia in the printed edition (Vilna, XIX century). The BT is a comprehensive literary creation, which went through an intricate process of oral and written transmission, was expanded in every generations before its final redaction, and has been the object of explanatory commentaries and reflexions from the Medieval Era onwards. In its long history of formulation, interpretation, transmission and study, the BT reflects inner developments within the Jewish tradition as well as the interactions between Judaism and the cultures with which the Jews came into contact (Strack and Stemberger, 1996). In the past decades, online resources for studying Rabbinic literature have considerably increased and several digital collections of Talmudic texts and manuscripts are nowadays available (Lerner, 2010). Particularly, scholars as well as a larger public of users can benefit from several new computing technologies applied to the research and the study of the BT, such as (i.) HTML (Segal, 2006), (ii.) optical character recognition, (iii.) three-dimensional computer graphics (Small, 1999), (iv.) text encoding, text and data mining (v.) image recognition (Wolf et al., 2011(a); Wolf et al., 2011(b); Shweka et al., 2013), and (vi.) computer-supported learning environments (Klamma et al., 2005; Klamma et al., 2002). In the context of the "Progetto Traduzione del Talmud Babilonese", the Institute for Computational Linguistics of the Italian National Research Council (ILC-CNR) is in charge of developing a collaborative Java-EE web application for the translation of the BT into Italian by a team of translators. The Talmud System (TS) already includes Computer-Assisted Translation (CAT), Knowledge Engineering and Digital Philology tools, and, in future versions, will include Natural Language Processing tools for Hebrew/Aramaic, each of which will be outlined in

detail in the next Sections.

2 Description of the System

The general architecture of the TS is represented in Figure 1. Each system component implements specific functionalities targeted at different types of users. Translators and revisors are assisted in the translation process by CAT technologies, including indexers and a Translation Memory (TM); philologists and linguists are enabled to insert notes, comments, semantic annotations and bibliographical references; domain experts are allowed to structure relevant terms into glossaries, and, possibly, into domain ontologies; researchers and scholars can carry out complex searches both on a linguistic and semantic basis; editors are enabled to produce the printed edition of the translation of the BT in an easier manner, by arranging translations and notes in standard formats for desktop publishing software. In what follows, we briefly outline the TS main components and the progress state of their development.

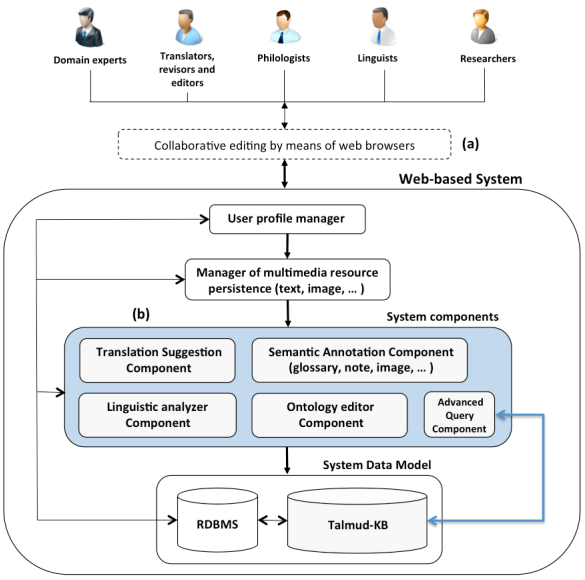


Figure 1: The Talmud System's architecture. (a) Collaborative editing - (b) Component based structure.

2.1 Translation Suggestion Component

We chose to adopt a Translation Memory (TM) based approach due to the literary style of the BT. Composed in a dialogical form and characterized by formulaic language, the BT presents several standard expressions. Furthermore, as an exegetical text, the BT contains innumerable quotation from the Bible, the *Mishnah*, other tannaitic sources and even from amoraitic statements discussed in other passages of the BT itself.

To the best of our knowledge, our implementation mainly contemplates aspects related to the specific needs of the translators community working on the BT in a collaborative environment, that the main non commercial CAT tools (OpenTM, OmegaT, Olanto, Transolution) and commercial ones (Dèjà Vu, Trados, Wordfast, Multitrans, Star Transit) do not take suitably into account (see (Bellandi et al., 2014(b)) for details). These specific requisites can be generalized to other complex ancient texts, where the emphasis of the translation work shall concern the quality instead of the translation pace. Exhibiting exceptionally concise sentences, which remain often unclear even to expert Talmudists, the BT cannot be treated and translated as a modern text. It is worth considering the Matecat Project[1], where the authors combine CAT and machine translation (MT) technologies, providing both suggestions by MT which are consistent with respect to the whole text, and methods for the automatic self-correction of MT making use of the implicit feedback of the user. The lack of linguistically annotated resources, and large collections of parallel texts regarding the languages present in the BT, prevented us to consider any statistical MT toolkit. We implemented a TM enabling translators to re-elaborate the plain and literal translation of the text and integrate it with explicative additions. The TM is organized at the segment level. A segment is a portion of original text having an arbitrary length. We formally defined the translation memory $M_{BT} = \{(s_i, T_i, A_i, c_i)\}$ with i ranging 1 to n, as a set of n tuples, where each tuple is defined by:

- s_i, the source segment;
- $T_i = \{t_i^1, \ldots, t_i^k\}$, the set of translations of s_i with $k \geq 1$, where each t_i^j has its literal part $\overline{t}_i^j$, and its contextual information $\widetilde{t}_i^j$, with $1 \leq j \leq k$;
- $A_i = \{a_i^1, \ldots, a_i^k\}$, the set of translators id of each translation of s_i in T_i with $k \geq 1$;
- c_i, the context of s_i referring to the tractate which belongs to;

Each segment's translation is obtained by differentiating the "literal" translation (using the bold

[1] http://www.matecat.com/matecat/the-project/

style) from explicative additions, i.e. "contextual information". Segments exhibiting the same literal part may convey different contextual information. By the term "context", we refer to the tractate to which the source segment belongs. The translation environment we created allows to acquire the segment to be translated, to query the TM, and to suggest the Italian translations related to the most similar strings. Since the BT does not exhibit a linguistic continuity, thus preventing an automatic splitting into sentences, we opted for a manual segmentation. Each translator selects a certain source segment to translate from a specific window of the system's GUI, which contains the specific tractate of the BT. This process may have a positive outcome: translators, being forced to manually detect the segments, could acquire a deeper awareness of the text they are about to translate. Clearly, the manual segmentation implies the engagement of the translators in a deep cognitive process aimed to establish the exact borders of a segment. The thorough reflection of the segmentation affects deeply also the final translation, by orienting the content and nature of the TM. So far, we could not include neither grammatical nor syntactic information in the similarity search algorithm (see, Section 2.4). Thus, we adopted similarity measures based on edit distance, by considering that two source segments are more similar when exhibiting the same terms in the same order. The novelty of this approach consists in the way we rank suggestions with the same value, based on external information, stored as metadata inside the TM, i.e. (i.) authors of translations and (ii.) the context (the tractate of reference). These informations are highly valuable, enabling (i.) translators to evaluate the reliability of the suggested translations according to the scientific authority of their authors, and (ii.) revisors to pervain to a more coherent, homogeneous and fluent translation. Since each suggested translation can be shown with or without its contextual information, each translator is enabled to approve and choose the literal translation, editing only the contextual information. Thus, our system relieves human translators from routine work, but always enabling them to control and orientate the translation process. Such a system is particularly useful for a complex ancient text such as the BT, which demands the linguistic and scholarly input of human users. Finally, Figure 2 shows the TM performance in terms of redundancy rate, roughly estimated by conducting a jackknife experiment (Wu, 1986). Redundancy curves are drawn by considering the ranking of the similarity function. The percentage of source segments found both verbatim and fuzzy in the memory appears to grow logarithmically with time (and consequently with the size of the memory).

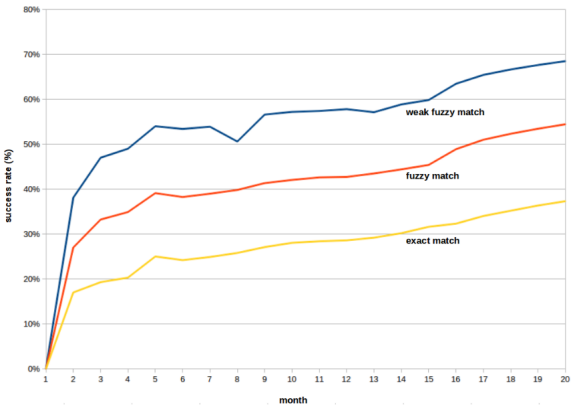


Figure 2: Redundancy of the translation memory in function of time.

2.2 Knowledge Engineering

Dealing with ethics, jurisprudence, liturgy, ritual, philosophy, trade, medicine, astronomy, magic and so much more, the BT represents the most important legal source for Orthodox Judaism and the foundation for all the successive developments of *Halakhah* (legal knowledge) and *Aggadah* (narrative knowledge). By means of an annotation module, translators can then semantically annotate arbitrary portions of text on the basis of the above fields. To date, the annotation process exploits an initial set of five predefined semantic classes: people's names, animals, plants, idiomatic expressions (e.g., *the Master said*), concepts (e.g., *Terumah*). This functionality allows the creation of specialized glossaries and, when fully implemented, the automatization of the annotation process. Furthermore, it enables experts specialized in the various Talmudic subjects to annotate, in a collaborative environment, relevant and technical terms and, eventually, structure them in a Talmudic Knowledge base (Talmud-KB, in Figure 1), using a formal knowledge representation language. To face the plurality of opinions, which generally originates in a collegial environment when assigning semantic labels, especially in the context of translation, the TS is fitted to enable domain

experts to represent uncertain knowledge through "weighted" relations, according to their scientific confidence (Bellandi and Turini, 2012; Danev et al., 2006; Ding et al., 2005).

2.3 Digital Philology

The system also responds to the specific needs of philological work and specialized analyses of the text, allowing to insert annotations at various levels of granularity. The parts of the Italian translation that appear in bold, for example, correspond to literal translations, while those in plain are explicative additions, i.e phrases added to make concepts expressed in Hebrew/Aramaic understandable to an Italian reader. Other annotations of greater granularity include: i) the addition of (explanatory) notes by translators and revision notes by revisors, ii) semantic annotations based on predefined types (see 2.2) designed to offer greater philological precision to the analysis and indexing of the text and for the construction of glossaries. A further element designed to perform more in-depth analysis of the translated text is provided by a dedicated component to introduce, in a standardized way, partially precompiled bibliograpghic references (e.g. for biblical citations to be completed with chapter and verse numbers) and names of Rabbis.

2.4 Language Analysis

Within the BT, we distinguish: (i.) quotations of portions from the *Mishnah*, (ii.) long amoraic discussions of mishnaic passages aimed at clarifying the positions and lexicon adopted by the *Tannaim*, and (iii.) external tannaitic material not incorporated in the canonical *Mishnah*. The content and philological depth of the BT implies an elevated degree of linguistic richness. In its extant form, the BT attests to (i.) different linguistic stages of Hebrew (Biblical Hebrew, Mishnaic Hebrew, Amoraic Hebrew), (ii.) different variants of Jewish Aramaic (Babylonian Aramaic and Palestinian Aramaic), and (iii.) several loanwords from Akkadian, ancient Greek, Latin, Pahlavi, Syriac and Arabic. To date, there are no available Natural Language Processing (NLP) tools suitable for processing ancient North-western Semitic languages, such as the different Aramaic idioms attested to in the BT, and for detecting the historical variants of Hebrew language as used in the Talmudic text. Several computational studies have been recently carried out on Modern Semitic Languages, including Modern Hebrew, and two high quality NLP tools are implemented for this language (Itai, 2006; HebMorph, 2010). Nevertheless, Modern Hebrew has been through a process of artificial revitalization from the end of the XIX century and does not correspond to the idioms recurring in the BT, even not to Biblical Hebrew or Mishnaic Hebrew. For this dissimilarity between the new and the ancient Hebrew languages, the existing NLP tools for Hebrew are highly unfit for processing the BT. In its multifaceted form, the "language" of the BT is unique and attested to only in few other writings. In addition, only few scholars have a full knowledge of the linguistic peculiarities of the BT and even fewer experts in Talmudic Studies are interested in collaborating to the creation of computational technologies for this textual corpus. These two main reasons have prevented, so far, the development of NLP tools for the BT, which would require a huge and very difficult effort probably not entirely justified by the subsequent use of the new technologies developed. The only attempts in these direction have been conducted within the Responsa Project on rabbinic texts, including the BT, and the Search And Mining Tools with Linguistic Analysis (SAMTLA[2]) on the corpus of Aramaic Magic Texts from Late Antiquity (AMTLA), some of which are written in Jewish Babylonian Aramaic, the dialect characterizing the BT. In the future phases of our project, we aim to develop some language resources for processing the linguistic and dialectic variants attested to in the BT.

3 Conclusion

We here introduced the Talmud System, a collaborative web application for the translation of the Babylonian Talmud into Italian integrating technologies belonging to the areas of (i.) Computer-Assisted Translation, (ii.) Digital Philology, (iii.) Knowledge Engineering and (iv.) Natural Language Processing. Through the enhancement of the already integrated components (i., ii., iii.) and the inclusion of new ones (iv.) the TS will allow, in addition to the improvement of the quality and pace of the translation, to provide a multi-layered navigation (linguistic, philological and semantic) of the translated text (Bellandi et al., 2014(c)).

[2] http://samtla.dcs.bbk.ac.uk/

4 Acknowledgements

This work has been conducted in the context of the research project TALMUD and the scientific partnership between S.c.a r.l. "Progetto Traduzione del Talmud Babilonese" and ILC-CNR and on the basis of the regulations stated in the "Protocollo d'Intesa" (memorandum of understanding) between the Italian Presidency of the Council of Ministers, the Italian Ministry of Education, Universities and Research, the Union of Italian Jewish Communities, the Italian Rabbinical College and the Italian National Research Council (21/01/2011).

References

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti, Enrico Carniani. 2014(a). Content Elicitation: Towards a New Paradigm for the Analysis and Interpretation of Text. Mohamed H. Hamza, ed., In *Proceedings of the IASTED International Conference on Informatics*, pp. 507-532.

Andrea Bellandi, Franco Turini. 2012. Mining Bayesian Networks Out of Ontologies. *Journal of Intelligent Information Systems*, 38(2):507-532.

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti. 2014(b). Computer Assisted Translation of Ancient Texts: the Babylonian Talmud Case Study. In *Proceedings of the 11th International Natural Language Processing and Cognitive Systems*.

Andrea Bellandi, Alessia Bellusci, Amedeo Cappelli, and Emiliano Giovannetti. 2014(c). Graphic Visualization in Literary Text Interpretation. In *Proceedings of the 18th International Conference on Information Visualisation*. Paris, France. July 15-18.

Boris Danev, Ann Devitt, Katarina Matusikovai. 2006. *Constructing Bayesian Networks Automatically using Ontologies*. Second Workshop on Formal Ontologies Meets Industry.

Zhongli Ding, Yun Peng, Rong Pan. 2005. *BayesOWL: Uncertainty Modeling in Semantic Web Ontologies*. Soft Computing in Ontologies and Semantic Web Springer-Verlag.

HebMorph - Morphological Analyser and Disambiguator for Hebrew Language. 2010. http://code972.com/hebmorph.

Alon Itai. 2006. Towards a Research Infrastructure for Language Resources. In *Proceedings of the Language Resources and Evaluation Conference*.

Ralf Klamma, Marc Spaniol, Matthias Jarke. 2005. MECCA: Hypermedia Capturing of Collaborative Scientific Discourses about Movies. *Informing Science Journal*, 8:3-38.

Ralf Klamma, Elisabeth Hollender, Matthias Jarke, Petra Moog, Volker Wulf. 2002. Vigils in a Wilderness of Knowledge: Metadata in Learning Environments. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 519-524.

Heidi Lerner. 2010. Online Resources for Talmud Research, Study and Teaching. *Association for Jewish Studies*, pp. 46-47.

MateCat Project. A CAT Tool for Your Business. Simple. Web-Based, 2012. http://www.matecat.com/matecat/the-project/.

Eliezer Segal. A Page from the Babylonian Talmud. http://www.ucalgary.ca/ elsegal/TalmudPage.html.

Eliezer Segal. 2006. Digital Discipleship: Using the Internet for the Teaching of Jewish Thought. H. Kreisel, ed., *Study and Knowledge in Jewish Thought*, pp. 359-373.

Roni Shweka, Yaacov Choueka, Lior Wolf, Nachum Dershowitz. 2013. Automatic Extraction of Catalog Data from Digital Images of Historical Manuscripts. *Literary and Linguistic Computing*, pp. 315-330.

David L. Small. 1999. *Rethinking the Book, unpubl. PhD Dissertation*. Massachusetts Institute of Technology, http://www.davidsmall.com/portfolio/talmud-project/.

H. L. Strack, G. Stemberger. 1996. *Introduction to Talmud and Midrash* . tr. and ed. by M. Bockmuehl, pp. 190-225.

Lior Wolf, Liza Potikha, Nachum Dershowitz, Roni Shweka, Yaacov Choueka. 2011(a). Computerized Palaeography: Tools for Historical Manuscripts. In *Proceedings 18th IEEE International Conference on Image Processing*, pp. 3545-3548.

Lior Wolf, Lior Litwak, Nachum Dershowitz, Roni Shweka, Yaacov Choueka. 2011(b). Active Clustering of Document Fragments using Information Derived from Both Images and Catalogs. In *Proceedings IEEE International Conference on Computer Vision*, pp. 1661-1667.

Chien-Fu Jeff Wu. 1986. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261-1295.

10.12871/CLICIT2014112

Towards a Decision Support System for Text Interpretation

Alessia Bellusci, Andrea Bellandi, Giulia Benotto,
Amedeo Cappelli, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. This article illustrates the first steps towards the implementation of a Decision Support System aimed to recreate a research environment for scholars and provide them with computational tools to assist in the processing and interpretation of texts. While outlining the general characteristics of the system, the paper presents a minimal set of user requirements and provides a possible use case on Dante's Inferno.

Italiano. *Questo articolo illustra i primi passi verso la realizzazione di un Sistema di Supporto alle Decisioni volto a ricreare un ambiente di ricerca per gli studiosi e assisterli, anche mediante strumenti computazionali, nell'elaborazione e nell'interpretazione di testi. Oltre a delineare le caratteristiche generali del sistema, l'articolo presenta una serie minima di requisiti utente e fornisce un possibile caso d'uso sull'Inferno di Dante.*

1 Introduction

A text represents a multifaceted object, resulting from the intersection of different expressive layers (graphemic, phonetic, syntactic, lexico-semantic, ontological, etc.). A text is always created by a writer with a specific attempt to outline a certain subject in a particular way. Even when it is not a literary creation, a given text follows its writer's specific intention and is written in a distinct form. The text creator's intention is not always self-evident and, even when it is, a written piece might convey very different meanings proportionally to the various readers analysing it. Texts can be seen, in fact, as communication media between writers and readers. Regardless of the epistemological theory about where meaning emerges in the reader-text relationship (Objectivism, Constructivism, Subjectivism), a text needs a reader as much as a writer to be expressive (Chandler, 1995). The reader goes beyond the explicit information given in the text, by making certain inferences and evaluations, according to his/her background, experience, knowledge and purpose. Therefore, interpretation depends on both the nature of the given text and the reader/interpreter; it can be understood as the goal, the process and the outcome of the analytic activity conducted by a certain reader on a given text under specific circumstances. Interpretation corresponds to the different – virtually infinite – mental frameworks and cognitive mechanisms activated in a certain reader/interpreter when examining a given text. The nature of the interpretation of a given text can be philological, historical, psychological, etc.; a psychological interpretation can be Freudian, Jungian, etc... Furthermore, the different categories of literary criticism and the various interpretative approaches might be very much blurred and intertwined, i.e. an historical interpretation might involve philological, anthropological, political and religious analyses.

While scholars are generally aware of their mental process of selection and categorization when reading/interpreting a text and, thus, can re-adjust their interpretative approach while they operate, an automatic system has often proved unfit for qualitative analysis due to the complexity of text meaning and text interpretation (Harnad, 1990). Nevertheless, a few semi-automatic systems for qualitative interpretation have been proposed in the last decades. The most outstanding of them is ATLAS.ti, a commercial system for qualitative analysis of unstructured data, which has been applied in the early nineties to text interpretation (Muhr, 1991). ATLAS.ti, however, appears too general to respond to the articulated needs

of a scholar studying a text, lacking of advanced text analysis tools and automatic knowledge extraction features. The University of Southampton and Birkbeck University are currently working on a commercial project, SAMTLA[1], aimed to create a language-agnostic research environment for studying textual corpora with the aid of computational technologies. In the past, concerning the interpretation of literary texts, the introduction of text annotation approaches and the adoption of high-level markup languages allowed to go beyond the typical use of concordances (DeVuyst, 1990; Sutherland, 1990; Sperberg-Mc Queen and Burnard, 1994). In this context, several works have been proposed for the study of Dante's Commedia. One of the first works involved the definition of a meta representation of the text of the Inferno and the construction of an ontology formalizing a portion of Dante's Commedia's world (Cappelli et al., 2002). Data mining procedures able to conceptually query the aforementioned resources have also been implemented (Baglioni et al., 2004). Among the other works on Dante we cite The World of Dante (Parker, 2001), Digital Dante of the Columbia University (LeLoup and Ponterio, 2006) and the Princeton Dante Project (Hollander, 2013). A "multidimensional" social network of characters, places and events of Dante's Inferno have been constructed to make evident the innermost structure of the text (Cappelli et al., 2011) by leveraging on the expressive power of graph representations of data (Newman, 2003; Newman et al., 2006; Easley and Kleinberg, 2010; Meirelles, 2013). A touch table approach to Dante's Inferno, based on the same social network representation, has been also implemented (Bordin et al., 2013). More recently, a semantic network of Dante's works has been developed alongside a RDF representation of the knowledge embedded in them (Tavoni et al., 2014). Other works involving text interpretation and graph representations have been carried out on other literary texts, such as Alice in Wonderland (Agarwal et al., 2012) and Promessi Sposi (Bolioli et al., 2013).

As discussed by semiologists, linguists and literary scholars (Eco, 1979; Todorov, 1973; Segre, 1985; Roque, 2012) the interpretation of a text may require a complex structuring and interrelation of the information belonging to its different expressive layers.

[1] http://samtla.dcs.bbk.ac.uk/

The Decision Support System (DSS) we here introduce aims to assist scholars in their research projects, by providing them with semi-automatic tools specifically developed to support the interpretation of texts at different and combined layers. We chose to start from the analysis of literary texts to be able to face the most challenging aspects related to text interpretation. This work is the third of a series describing the progressive development of the general approach: for the others refer to (Bellandi et al., 2013; Bellandi et al., 2014). In what follows, we describe the general characteristics of the DSS we plan to develop accompanied by a minimal set of user requirements (2.), we present a possible scenario, in which the system can be applied (3.), and we provide some conclusive notes (4.).

2 Towards a Decision Support System for Text Interpretation

In this section, we present our vision of a DSS (Shim et al., 2002) specifically aimed to recreate a *research environment* for scholars and provide them with computational tools developed to assist data elaboration and content interpretation of texts. Theoretically, each automatic act operated by a computational system on a given text can be seen as an interpretative act. Yet, in our view, users shall remain the main decision-makers within their interpretative process, while the system and the integrated tools we aim to create shall function only as instruments enabling users to achieve their research goals in a clearer and easier manner. In the computational metaphor, our DSS would represent the writing desk and library of the historian or the laboratory and microscope of the biologist.

Within the system, users shall be able to carry out a *research project* based on one or more *textual sources* from the beginning through its end, whether the project is the analysis of medical records, the interpretation of a literary work, the production of a critical edition of a given text, or the historical analysis of textual material. Similarly, our system shall assist the creation of text interpretations either for personal purposes (student exercise, amateur research) or for scientific productions (article, monograph, critical edition). Although conceived for the use of a single scholar, the system shall enable users also to selectively share their results in a collaborative space. With the aid of our DSS, users shall be able to consult, search and analyze

a text dynamically and according to their specific interest. The system shall enable to conduct the study of a given text on several and different *layers*, each of which is already implicit in the text and explicated by the interpretative activity of the reader/scholar through specific tools and visual solutions provided by the system.

2.1 Minimal User Requirements

In order to define a minimal set of user requirements we first introduce the following key terms: *textual source*, *layer*, *element*, *relation* and *network*. As ***textual source*** we intend every object presenting at least one grapheme, which has been either digitized or scanned as image and uploaded into the system (i.e., page from a digitized literary book, image of an inscribed pottery, image of a folium from a manuscript, transcription of a manuscript). The term *source* can refer to (i.) a textual corpus (i.e., Dante's writings), (ii.) a specific section/unit/book of the given corpus (i.e., Inferno), and (iii.) a passage from a specific book of a given corpus (i.e., XVI Canto of Inferno). A ***layer*** is a specific set of features embedded in a given *textual source*, which can be explicated by users through analysis and annotation tools. Each *source* exhibits, at least, a graphemic layer (grapheme/s on a given writing surface) and may include an unlimited number of *layers*, according to the user's research interest. Some basic *layers* (i.e., graphemic, phonetic, terminological, ontological) are already provided by the DSS, while others (arbitrary layers) can be defined by users (e.g., dialogical layer, anthropological layer). An ***element*** is an atomic unit forming a *layer*, i.e. a grapheme of the graphemic layer, a phoneme of the phonetic layer, a term of the terminological layer, or a concept of the ontological layer; an *element* can be visualized as a node of a network in the interface of the DSS. A ***relation*** is a link between two or more *elements*, intra and inter-layer; a relation can be visualized as an arc of a *network* in the interface of the DSS. Finally, a ***network*** is a set of *elements* and the *relations* among them visualized as a graph.

We have grouped the minimal requirements we identified for the development of our DSS in four main categories. To the first group, **(A.) Upload and Source Management**, belong the following requirements: **(1.)** creation of a new *research project*; **(2.)** management of a variety of different *research projects* for each user; **(3.)** upload of the relevant *sources* for a specific *project*; **(4.)** running of OCR on the scanned source, when dealing with images of manuscripts or material objects; **(5.)** sharing of selected sources with selected users; **(6.)** execution of catalographic searches. To the second group, **(B.) Layers**, belong: **(1.)** use of predefined basic *layers* **(2.)** definition of arbitrary *layers*; **(3.)** use of (manual and automatic) tools for the elicitation of the elements of a specific *layer*; **(4.)** addition of notes (footnotes, endnotes, general notes, philological, linguistic, ...) and comments of different types to a specific *element*. To the third category, **(C) Research and Comparison**: **(1.)** execution of searches on the selected textual sources within one or more *layers*; **(2.)** execution of searches with boolean and regular expressions; **(3.)** execution of manual and semi-automatic comparisons between two or more *sources*, also on different *layers*, by presenting them together on the screen; **(4.)** highlighting of the differences between two or more *sources* selected for the comparison; **(5.)** highlighting of features shared by two or more *sources* selected for the comparison; **(6.)** visualization of the results of each specific search and comparison in structured lists. Finally, for the fourth category, **(D) Construction of Networks**, we identified the following requirements: **(1.)** manual or, when possible, automatic construction of a *network*, realized by defining *relations* among *elements* belonging to the same *layer* or different *layers*; **(2.)** editing of an automatically generated *network*.

3 A Possible Use Case on Dante's Inferno

Here, we present a possible use case on Dante's Inferno, a highly complex and rich writing, which gathers a great amount of information, thus requiring very different scholarly skills to be fully understood and analysed. Particularly, our use case studies the dialogues of Guelfi and Ghibellini, two rival Florentine political factions. Although in our vision the DSS would enable users to annotate chunks of text as dialogues and to define the *text ontology* (Bellandi et al., 2013) including the characters of the *al di là*, we chose to exploit an existing XML-encoded advanced representation of Inferno (Cappelli et al., 2011).

An analysis of this type can be articulated in a series of steps, each one bringing to the construction of a portion of the *network* (requirement

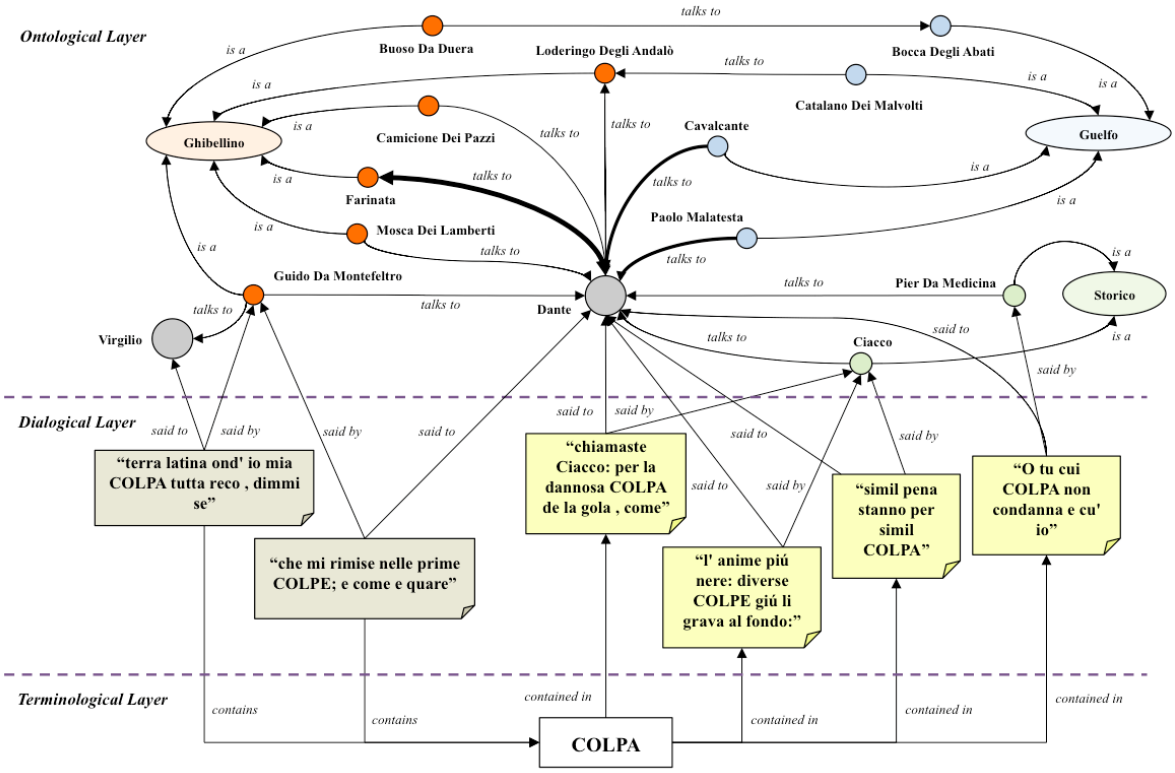


Figura 1: Example of network

D.1), of Figure 1. The first step involves the *ontological layer* (requirement B.1): the user would build the upper part of the network by introducing the relation *talks to* (the thickness of the relative arc representing the number of dialogical interactions) among the elements *Guelfo*, *Ghibellino*, *Dante*, and *Virgilio*. The obtained network shows that the only interactions between the two factions are those of *Buoso Da Duera* who talks to *Bocca degli Abati*, and *Catalano Dei Malvolti* who talks to *Loderingo Degli Andalò*. Furthermore, *Guido Da Montefeltro* is the only *Ghibellino* who talks to both *Dante* and *Virgilio*. The user could then be interested in analysing his dialogues (the two added on the left part of the network as elements of the dialogical layer), by using a terminology extractor, bringing to the elicitation of the elements (terms) constituting the terminological layer (requirement B.3). The user could select the term *colpa* ("guilt" in English) since being present in both dialogues and add it to the network. In the final part of this example the user could verify if the term *colpa* appears in other dialogues. To do this the user would search the pattern "colp[ae]" (representing the singular and plural forms of the lemma *colpa*) inside the elements of the dialogical layer (requirement C.2). As a result, the network would be populated with four more dialogues, showing that only *Ciacco* and *Pier Da Medicina* talk to *Dante* using the term *colpa*. These two characters are not politically characterized, being classified, in the ontology, as "Storico" (historical character).

4 Conclusions

In this work, we presented our vision of a Decision Support System for the analysis and interpretation of texts. In addition to outlining the general characteristics of the system, we illustrated a case study on Dante's Inferno showing how the study of a text can involve elements belonging to three different layers (ontological, dialogical and terminological) thus allowing to take into account, in an innovative way, both textual and contextual elements.

The next steps will consist in the extension of the user requirements and the design of the main components of the system. We plan to start with the basic features allowing a user to create a project and upload documents and then provide the minimal text processing tools necessary for the definition and management of (at least) the graphemic layer.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 88-96, Montréal, Canada. June 8.

Miriam Baglioni, Mirco Nanni, and Emiliano Giovannetti. 2004. Mining literary texts by using domain ontologies. In *Proceedings of the Workshop on Knowledge Discovery and Ontologies (KDO-2004).* Pisa, Italy. September 20-24.

Andrea Bellandi, Alessia Bellusci, Emiliano Giovannetti, and Enrico Carniani. 2013. Content Elicitation: Towards a New Paradigm for the Analysis and Interpretation of Text. In *Proceedings of the IASTED International Conference on Informatics.* Innsbruck, Austria. February 17-19.

Andrea Bellandi, Alessia Bellusci, Amedeo Cappelli, and Emiliano Giovannetti. 2014. Graphic Visualization in Literary Text Interpretation. In *Proceedings of the 18th International Conference on Information Visualisation.* Paris, France. July 15-18.

Andrea Bolioli, Matteo Casu, Maurizio Lana, and Renato Roda. 2013. Exploring the Betrothed Lovers. *OASIcs-OpenAccess Series in Informatics*, 32:30–35.

Silvia Bordin, Massimo Zancanaro, and Antonell De Angeli. 2013. Touching Dante: A Proximity-based Paradigm for Tabletop Browsing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, 27:1-10. Trento, Italy. September 16-20.

Amedeo Cappelli, Maria Novella Catarsi, Patrizia Michelassi, Lorenzo Moretti, Miriam Baglioni, Franco Turini, and Mirko Tavoni. 2002. Knowledge Mining and Discovery for Searching in Literary Texts. In *Proceedings of LREC 2002.* Las Palmas, Canary Islands, Spain. 29-31 May.

Amedeo Cappelli, Michele Coscia, Fosca Giannotti, Dino Pedreschi, and Salvo Rinzivillo. 2011. The social network of Dante's Inferno. *Leonardo*, 44(3):246–247.

Daniel Chandler. 1995. *The Act of Writing: A Media Theory Approach* . Aberystwyth, pp. 4-8.

Jan De Vuyst. 1990. Knowledge representation for text interpretation. *Literary and linguistic computing* 5(4): 296–302.

David Easley, and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Umberto Eco. 1979. *Lector in Fabula*. Bompiani, Milano.

Stevan Harnad. 1990. Against Computational Hermeneutics. *Social Epistemology*, 4:167–172.

Robert Hollander. 2013. The Princeton Dante Project. *Humanist Studies and the Digital Age* 3(1):53-59. http://etcweb.princeton.edu/dante/index.html

Jean W. LeLoup, and Robert Ponterio. 2006. Dante: Digital and on the Web. *Language Learning & Technology* 10(1): 3–8. http://dante.ilt.columbia.edu

Isabel Meirelles. 2013. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers.

Thomas Muhr. 1991. ATLAS.ti - A Prototype for the Support of Text Interpretation. *Qualitative Sociology* 14:349–371. Human Science Press, New York.

Mark E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*. 45:167–256.

Mark E. J. Newman, Albert László Barabási, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton University Press.

Deborah Parker. 2001. The World of Dante: a hypermedia archive for the study of the inferno. *Literary and linguistic computing* 16(3): 287–297. http://www.worldofdante.org/about.html

Antonio Roque. 2012. Towards a computational approach to literary text analysis. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 97–104, Montréal, Canada. June 8.

Cesare Segre, 1985. Testo letterario, interpretazione, storia: linee concettuali e categorie critiche. In *Alberto Asor Rosa: Letteratura italiana* 21–140. Einaudi, Torino.

Jung P. Shim, Merril Warkentin, James F. Courtney, Daniel J. Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision support systems*, 33(2):111–126. Elsevier.

Michael Sperberg-McQueen, and Lou Burnard. 1994. *Guidelines for electronic text encoding and interchange*, 1. Chicago and Oxford: Text Encoding Initiative.

Kathryn Sutherland. 1990. A Guide Through the Labyrinth: Dickens's Little Dorrit as Hypertext. *Literary and Linguistic Computing*. 5(4):305–309.

Mirko Tavoni, Paola Andriani, Valentina Bartalesi, Elvira Locuratolo, Carlo Meghini, and Loredana Versienti. 2014. Towards a semantic network of Dante's works and their contextual knowledge. In *Proceedings of The Digital Humanities 2014 conference*. Lausanne, Switzerland. July 7-12.

Tzevetan Todorov. 1973. Postscriptum. *In R. Jakobson, Questions de poétique* 485–504. Editions du Seuil, Paris.

10.12871/CLICIT2014113

An Italian Dataset of Textual Entailment Graphs for Text Exploration of Customer Interactions

Luisa Bentivogli and Bernardo Magnini
FBK, Trento, Italy
bentivo,magnini@fbk.eu

Abstract

English. This paper reports on the construction of a dataset of *textual entailment graphs* for Italian, derived from a corpus of real customer interactions. Textual entailment graphs capture relevant semantic relations among text fragments, including equivalence and entailment, and are proposed as an informative and compact representation for a variety of text exploration applications.

Italiano. *Questo lavoro riporta la costruzione di un dataset di grafi di implicazione testuale per la lingua italiana, derivati da un corpus di interazioni reali tra cliente e call centre. I grafi di implicazione testuale catturano relazioni semantiche significative tra porzioni di testi, incluse equivalenze e implicazioni, e sono proposti come un formato di rappresentazione informativo e compatto per applicazioni di esplorazione di contenuti testuali.*

1 Introduction

Given the large production and availability of textual data in several contexts, there is an increasing need for representations of such data that are able at the same time to convey the relevant information contained in the data and to allow compact and efficient text exploration. As an example, customer interaction analytics requires tools that allow for a fine-grained analysis of the customers' messages (*e.g.* complaining about a particular aspect of a particular service or product) and, at the same time, allow to speed up the search process, which commonly involves a huge amount of interactions, on different channels (*e.g.* telephone calls, emails, posts on social media), and in different languages.

A relevant proposal in this direction has been the definition of *textual entailment graphs* (Berant et al., 2010), where graph nodes represent predicates (*e.g.* $marry(x, y)$), and edges represent the entailment relations between pairs of predicates. This recent research line in Computational Linguistics capitalizes on results obtained in the last ten years in the field of *Recognizing Textual Entailment* (Dagan et al., 2009), where a successful series of shared tasks have been organized to show and evaluate the ability of systems to draw text-to-text semantic inferences.

In this paper we present a linguistic resource consisting of a collection of textual entailment graphs derived from real customer interactions in Italian social fora, which is our motivating scenario. We extend the earlier, predicate-based, variant of entailment graphs to capture entailment relations among more complex text fragments. The resource is meant to be used both for training and evaluating systems that can automatically build entailment graphs from a stream of customer interactions. Then, entailment graphs are used to browse large amount of interactions by call center managers, who can efficiently monitor the main reasons for customers' calls. We present the methodology for the creation of the dataset as well as statistics about the collected data.

This work has been carried out in the context of the EXCITEMENT project[1], in which a large European consortium aims at developing a shared software infrastructure for textual inferences, *i.e.* the EXCITEMENT Open Platform[2] (Padó et al., 2014; Magnini et al., 2014), and at experimenting new technology (*i.e.* entailment graphs) for customer interaction analytics.

[1] excitement-project.fbk.eu
[2] http://hltfbk.github.io/Excitement-Open-Platform/

2 Textual Entailment Graphs

Textual Entailment is defined as a directional relationship between two text fragments - T, the entailing text and H, the entailed text - so that *T entails H if, typically, a human reading T would infer that H is most likely true* (Dagan et al., 2006). While Recognizing Textual Entailment (RTE) datasets are typically composed of independent T-H pairs, manually annotated with "entailment" or "non entailment" judgments (see (Bentivogli et al., Forthcoming) for a survey of the various RTE datasets), the text exploration scenario we are addressing calls for a representation where entailment pairs are highly interconnected. We model such relations using *Textual Entailment Graphs*, where each node is a textual proposition (*e.g.* a predicate with arguments and modifiers), and each edge indicates a directional entailment relation.

An example of textual entailment graph is presented in Figure 1, where the node *"chi ha la chiavetta non riesce a connettersi"* entails *"non riesco a navigare con la chiavetta"*. Entailment judgments in this context are established under an existential interpretation: if there is a situation where someone *"non riesce a connettersi"*, then it is true (*i.e.* it is entailed) that, under appropriate meaning interpretation of the sentences, a situation exists in which someone *"non riesce a navigare"*. In the entailment graph, mutually entailing nodes (corresponding to pharaphrases) are represented unified in the same node, as in the case of *"chi ha la chiavetta non riesce a connettersi"*, *"la mia chiavetta non si connette"*, *"non riesco a collegarmi con la chiavetta"* in Figure 1. The graph representation also allows to derive implicit relations among nodes. For instance, since the entailment relation is transitive, the graph in Figure 1 allows to infer that *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* entails *"non riesco a navigare con la chiavetta"*. In addition, the lack of a path in the graph represents non-entailment relations, as for instance the fact that *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* does not entail *"da domenica non riesco a navigare con la chiavetta"*, because we can not establish a temporal relation between *"dal giorno 20/4"* and *"da domenica"*.

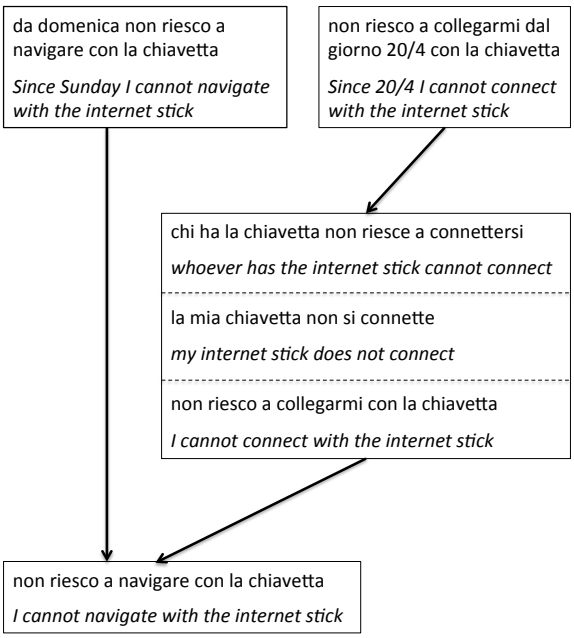


Figure 1: Portion of textual entailment graph.

3 Dataset Creation

The entailment graph creation process starts from customer interactions collected for a given topic and is composed of two main phases: (*i*) for each interaction all the relevant text fragments are extracted and the corresponding fragment graphs are created; (*ii*) all the individual fragment graphs are merged into the final entailment graph. The complete workflow of the dataset creation process is shown in Figure 2.

The starting interactions are posts taken from the official webpage of a mobile service provider in a social network, and contain reasons for dissatisfaction concerning the provider. The texts are anonymized to eliminate any reference to both the provider and the customers writing the posts.

As Figure 2 shows, the process alternates manual and automatic steps. In step 1, for each interaction the relevant text fragments are manually identified. A fragment is defined as a content unit that conveys one complete statement related to the topic (*i.e.* one reason for dissatisfaction). In our example, *"da domenica non riesco a navigare con la chiavetta"*, *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"*, *"la mia chiavetta non si connette"* are all fragments extracted from different interactions. Fragments are then generalized in order to increase the probability of recognizing entailing texts in the collection and provide a richer hierarchical structure to the entailment graph. Such generalization is performed automatically after *grammatical modifiers* of the fragments, *i.e.* tokens which can be removed

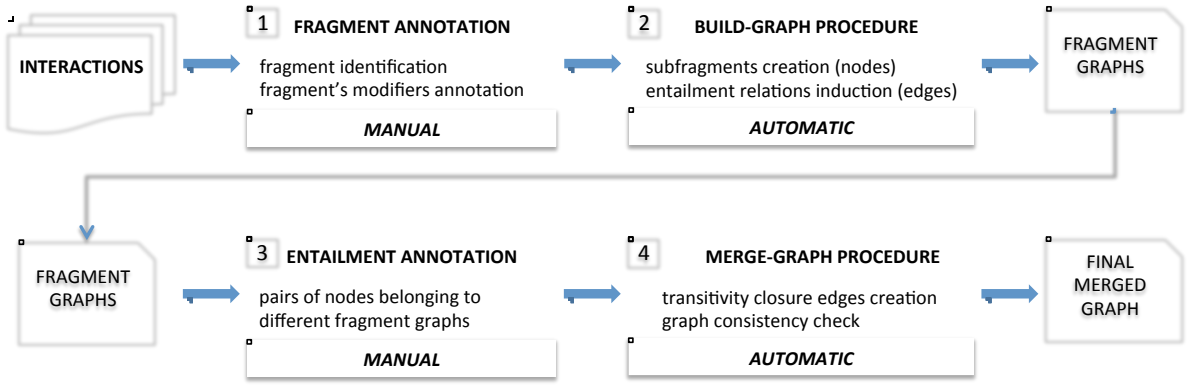


Figure 2: Entailment graph creation process.

from a fragment without affecting its comprehension, are manually specified. For example, *"da domenica"* and *"dal giorno 20/4"* are annotated as modifiers of respectively the first and second fragment above. This first manual annotation phase was carried out with the CAT Tool (Lenzi et al., 2012).[3] In step 2, given a fragment and its annotated modifiers, the corresponding subfragments are automatically created by incrementally removing its modifiers until no modifiers are left. In addition, entailment relations are automatically induced following the principle that a more specific text (*i.e.* containing more modifiers) entails a more generic one (*i.e.* containing less modifiers). As a result, an entailment graph of the corresponding fragment - Fragment Graph - is constructed, where the nodes are the fragment and its subfragments, and the edges are the entailment relations between them. In our example, for the fragment *"da domenica non riesco a navigare con la chiavetta"*, the more general subfragment *"non riesco a navigare con la chiavetta"* is automatically created as well as the entailment relation from the entailing fragment to the entailed subfragment.

To obtain the final textual entailment graph, individual fragment graphs are merged by finding all the entailment relations between their nodes. In order to minimize the number of node pairs to be manually annotated in step 3, two strategies were adopted prior to annotation, one manual and one automatic. First, clustering of fragment graphs was manually performed according to the specific topic (*i.e.* reason for dissatisfaction) expressed by the fragments. The assumption behind this strategy is that there are no entailment relations between fragment graphs belonging to different clusters (*i.e.* dealing with different reasons for dissatisfaction). As an example, two different clusters were created for fragment graphs expressing dissatisfaction about "Telefoni smartphone e cellulari" and "Consolle". The merging phase is then performed cluster by cluster, and one final merged entailment graph for each cluster is created. Second, an algorithm aimed at skipping unnecessary manual annotations is integrated in the manual annotation interface. The interface presents to annotators all the pairwise comparisons between minimal subfragments (*i.e.* texts with no modifiers). If there is no entailment relation, then all the other pairwise comparisons between the other nodes of the fragments are automatically annotated as "no entailment". If an entailment relation is annotated between minimal subfragments, then also their respective ancestors are paired and proposed for manual annotation. In our example, *"non riesco a collegarmi con la chiavetta"* is annotated as entailing *"non riesco a navigare con la chiavetta"*. Due to this entailment relation, also *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* and *"da domenica non riesco a navigare con la chiavetta"* are paired and presented for annotation, which in this case is a negative entailment judgment. Also mutual entailment can be annotated, as for *"non riesco a collegarmi con la chiavetta"*, *'chi ha la chiavetta non riesce a connettersi"*, and *"la mia chiavetta non si connette"*.

Once step 3 has been completed, in the final automatic step 4 the individual fragment graphs are merged, transitive closure edges are added, and a consistency check aimed at ensuring that there are

[3]The tool is freely available at `https://dh.fbk.eu/resources/cat-content-annotation-tool`.

Clusters	Interactions	Fragment Graphs	Total Nodes	Total Edges	Intra-Fragment Edges	Inter-Fragment Edges
19	294	344	760	2316	733	1583

Table 1: Composition of the dataset.

no transitivity violations is carried out.

As a result of fragment graph merging, a textual entailment graph over the input fragments is constructed.

Statistics about the composition of the dataset created according to the described procedure are presented in Table 1. The final dataset contains 19 consistent textual entailment graphs, one for each of the clusters into which the fragment graphs were subdivided. The table also shows the number of original interactions and the fragment graphs derived from them (step 1 of the process), and the total number of nodes and edges composing the 19 final entailment graphs resulting from the merging of fragment graphs (step 4 of the process). Finally, the total number of edges contained in the final graphs is further subdivided into intra-fragment and inter-fragment edges. Intra-fragment edges denote edges connecting the nodes within fragment graphs, *i.e.* edges generated during fragment graph construction. Inter-fragment edges are edges generated during the merge phase.

The dataset is released for research purposes under a Creative Commons Attribution-NonCommercial-ShareAlike license, and will be available at the EXCITEMENT project website by the end of the project (31/12/2014). The release will also contain information about Inter-Annotator Agreement, which is being currently calculated for the two manual annotation phases carried out during dataset creation, namely (*i*) the identification of modifiers within text fragments, which is necessary to build the fragment graphs (step 1 of the process), and (*ii*) the annotation of entailment relations between statements (nodes) belonging to different fragment graphs, which is required to merge the fragment graphs (step 3).

4 Conclusion

We have presented a new linguistic resource for Italian, based on textual entailment graphs derived from real customer interactions. We see a twofold role of this resource: *(i)* on one side it provides empirical evidences of the important role of semantic relations and provides insights for new developments of the textual entailment framework; *(ii)* on the other side, a corpus of textual entailment graphs is crucial for the realization and evaluation of automatic systems that can build entailment graphs for concrete application scenarios.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923).

References

Luisa Bentivogli, Ido Dagan, and Bernardo Magnini. Forthcoming. The recognizing textual entailment challenges datasets. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global learning of focused entailment graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1220–1229, Uppsala, Sweden.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering*, 15(4):i–xvii.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics, Demo papers*.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*. doi:10.1017/S1351324913000351.

10.12871/CLICIT2014114

L'integrazione di informazioni contestuali e linguistiche nel riconoscimento automatico dell'ironia

Lorenzo Bernardini, Irina Prodanof
Dept. of Humanities, University of Pavia
lorenzo.bernardini01@ateneopv.it; irina.prodanof@unipv.it

Abstract

Italiano. L'ironia verbale è una figura retorica altamente complessa che appartiene al livello pragmatico del linguaggio. Finora, tuttavia, tutti i tentativi in campo computazionale volti al riconoscimento automatico dell'ironia si sono limitati a ricercare indizi linguistici che potessero segnalarne la presenza senza considerare fattori pragmatici e contestuali. In questo lavoro si è cercato di valutare la possibilità di integrare semplici fattori contestuali computabili con informazioni di tipo linguistico al fine di migliorare l'efficacia dei sistemi di riconoscimento automatico di ironia nei commenti dei lettori di quotidiani online.

English. *Verbal irony is a highly complex figure of speech which belongs to the pragmatic level of language. So far, however, all the computational approaches aimed at automatically recognizing irony have only searched for linguistic cues which could signal the presence of irony without taking into account pragmatic and contextual factors. In this work we have considered the possibility of integrating simple contextual and computable factors with linguistic information in order to improve the performance of irony detection systems in the comments of readers of web newspapers.*

1 Introduzione

L'ironia verbale è una figura retorica molto complessa che si colloca al livello pragmatico del linguaggio. Per quanto un ironista possa servirsi di elementi fonologici, prosodici, morfologici, lessicali, sintattici e semantici per produrre ironia, quest'ultima non è una proprietà interna all'enunciato stesso e non è determinata dalle sue caratteristiche formali.

L'ironia è piuttosto un fenomeno interpretativo legato alle aspettative che un ascoltatore sviluppa riguardo alle intenzioni dell'autore di un enunciato prodotto in uno specifico contesto a partire da uno sterminato insieme di informazioni enciclopediche e contestuali.

2 Potenziali applicazioni e stato dell'arte

Il riconoscimento automatico di ironia porterebbe benefici nel campo della Sentiment Analysis. Non individuare l'intenzione ironica dell'autore di un enunciato può portare infatti a gravi fraintendimenti riguardo alle sue opinioni. Poiché l'uso dell'ironia è pervasivo e quantitativamente non trascurabile in molti contesti online, un solido sistema di Sentiment Analysis dovrebbe considerare i problemi legati all'uso dell'ironia e sviluppare metodi per il suo riconoscimento automatico.

Il riconoscimento automatico di ironia è tuttavia ancora a uno stadio pionieristico. Ad oggi tutti gli approcci computazionali al riconoscimento di ironia hanno cercato esclusivamente di individuare elementi linguistici interni al testo (grafici, semantici e lessicali) che potessero indicarne la presenza. Carvalho et al. (2009) hanno proposto una lista di indizi espliciti per individuare l'ironia nei commenti dei lettori di un quotidiano online in lingua portoghese (virgolette intorno ad aggettivi o nomi positivi, espressioni/simboli che indicano una risata nella scrittura (e.g. *lol*), interiezioni usate per esprimere sentimenti positivi (e.g. *viva*) e punteggiatura marcata). Reyes et al. (2013) hanno invece costruito un complesso modello considerando 11 parametri linguistici posti su 4 dimensioni (*signatures*, *unexpectedness*, *style*, *emotional scenarios*) per rilevare l'ironia nei *tweets*.

Per quanto questi e altri lavori abbiano ottenuto

risultati parzialmente soddisfacenti, tuttavia nessuno di essi ha mai preso in considerazione gli elementi contestuali che costituiscono il cuore del fenomeno dell'ironia, limitandosi a considerare esclusivamente il materiale linguistico alla ricerca di caratteristiche formali che possano segnalarne la presenza. Tuttavia molto spesso l'informazione linguistica da sola non è sufficiente a stabilire se un enunciato è ironico o meno, tanto che lo stesso materiale linguistico, se proveniente da fonti o contesti differenti, può essere interpretato come ironico o meno.

Il contesto gioca dunque un ruolo fondamentale. Ma esistono fattori contestuali computabili che possano rivelarsi efficaci per riconoscere automaticamente l'ironia in testi online?

3 Analisi di una comunità online

Per individuare quali fattori contestuali possano rivelarsi computabili si è deciso di svolgere un'analisi empirica[1] riguardo all'uso dell'ironia all'interno di una comunità virtuale. L'analisi è stata realizzata sui commenti lasciati in calce agli articoli dai lettori abituali de *Il Fatto Quotidiano* online. L'analisi dei commenti si è svolta in tre fasi distinte.

Prima fase. Innanzitutto si è cercato di cogliere la significatività in termini quantitativi dei commenti ironici rispetto al totale dei commenti, verificando se i commenti ironici presentassero indizi linguistici espliciti della presenza di ironia.

Sono stati analizzati 20 commenti casuali a 6 articoli di giornale. Per ogni commento si è verificato:

- se l'autore fosse un commentatore abituale (oltre 500 commenti sul sito).
- se il commento fosse ironico o meno.
- se fossero presenti gli indizi espliciti proposti da Carvalho et al. (2009).

I numeri complessivi di quest'analisi empirica non possono certo costituire un campione statistico valido, ma sembrano tuttavia suggerire alcune tendenze degne di nota. Innanzitutto si è rilevato come l'ironia fosse presente in un numero considerevole di commenti. Almeno un quarto dei commenti a ciascun articolo presentava (almeno in parte) ironia, per un totale di 39 commenti su 120 (15 i casi di dubbia attribuzione). Questo dato suggerirebbe che l'ironia non debba essere trascurata come un fenomeno marginale in Sentiment Analysis.

In secondo luogo si è osservato come la maggior parte dei commentatori fossero abituali (85 su 120). Questo fatto è un presupposto fondamentale per avvalorare l'assunto che essi possano conoscersi tra loro e formare così una comunità.

Per quanto riguarda invece l'uso di indizi espliciti si è notato come essi comparissero soltanto in una minoranza, tuttavia significativa, dei commenti ironici (16/39). Ciò mostra che, servendosi solamente di questi indizi linguistici, molti dei commenti ironici non potrebbero assolutamente essere individuati. Per giunta la stessa serie di indizi appariva in quantità non trascurabile anche in commenti non ironici, seppure in proporzione minore (11/66). Pertanto l'uso di questi indizi linguistici, pur se inefficaci in molti casi di ironia e talvolta addirittura fuorvianti, non deve essere accantonato, vista comunque la maggior probabilità che essi siano presenti in un commento ironico rispetto ad uno non ironico.

Seconda fase. Nella seconda fase di analisi si è indagata l'attitudine ironica complessiva di una serie di commentatori abituali. Sono stati analizzati gli ultimi 25 commenti postati da 22 commentatori abituali scelti casualmente. Per ognuno di essi sono stati raccolti i seguenti dati:

- numero di commenti ironici tra gli ultimi 25 postati dall'utente.
- numero di commenti, tra quelli ironici sopra individuati, che presentassero gli indizi proposti da Carvalho et al. (2009).

In questa fase di analisi si è osservato come sembri possibile individuare un ristretto gruppo di commentatori che si contraddistingue dagli altri poiché pare manifestare un'attitudine spiccatamente ironica (oltre 80% dei commenti sono interamente ironici). La maggioranza degli utenti sembra invece affidarsi più sporadicamente all'uso di ironia e spesso l'uso è limitato a parti ridotte del testo, perlopiù all'inizio o nel finale del commento. Si è notato inoltre come gli indizi linguistici siano pressoché assenti nell'ironia dei commentatori "spiccatamente" ironici, mentre sono decisamente più frequenti nei commenti ironici degli altri utenti. Volendo deliberatamente semplificare, sembrano delinearsi due tendenze di comportamento particolarmente significative:

[1] Per una discussione più approfondita e maggiori dettagli sul lavoro di analisi si rimanda a Bernardini (2014). L'intero corpus è disponibile su richiesta.

- Commentatori quasi sempre ironici in quasi tutti i commenti e nell'interezza del commento. La loro ironia è raramente "marcata" da indizi linguistici espliciti.
- Commentatori poco ironici. L'ironia è usata più di rado e spesso solo nella frase iniziale o conclusiva. La loro ironia presenta più frequentemente indizi linguistici espliciti.

Terza fase. La terza fase di analisi si è concentrata sul comportamento dei commentatori più ironici della comunità. Inizialmente sono state osservate le interazioni avvenute tra i quattro utenti più ironici individuati nella seconda fase di analisi e gli altri commentatori nell'arco di 3 mesi.

Attraverso quest'analisi sono emerse due considerazioni interessanti. In primo luogo si è visto come gli utenti abituali riconoscano e siano spesso consapevoli dello stile "spiccatamente ironico" di alcuni utenti. In secondo luogo è emerso come gli utenti "spiccatamente ironici" tendano a rispondere con più frequenza a commenti ironici che a commenti non ironici rispetto agli altri utenti.

Un altro aspetto peculiare dei commentatori spiccatamente ironici riguarda le loro scelte di *nickname* e *avatar*. Questo non é particolarmente sorprendente poiché il nome e l'immagine scelta rappresentano la *faccia* che un utente vuole mostrare di sé. Pertanto un utente che desideri apparire molto ironico tenderà plausibilmente a registrarsi con una *faccia* che risulti simpatica. Si è perciò cercato di verificare empiricamente se *nickname* e *avatar* divertenti potessero risultare indicativi per individuare commentatori spiccatamente ironici. Sono stati individuati manualmente 13 utenti con una *faccia* riconducibile a contesti divertenti. Di questi, 8 manifestavano un'attitudine spiccatamente ironica. In particolare tra i 7 utenti con una *faccia* riconducibile a contesti più specificamente satirici, 6 di essi apparivano spiccatamente ironici. Sembrerebbe pertanto che la scelta di usare un *nickname* e/o un *avatar* riconducibili a contesti divertenti, o tanto meglio satirici, possa davvero essere un ottimo indicatore per individuare quegli utenti che esibiscono una attitudine molto ironica all'interno della comunità.

4 Verso un modello integrato di informazioni linguistiche e contestuali

Le informazioni contestuali che un ascoltatore può utilizzare per attribuire ironia a un enunciato sono soggettive e potenzialmente infinite. Al momento è però impensabile ipotizzare una macchina che possa contenere un insieme infinito e indefinito di credenze e conoscenze dal quale, in qualche modo, riconoscere l'ironia in un testo.

L'obiettivo deve essere dunque cercare di individuare ed estrarre alcune semplici informazioni contestuali computabili per poterle poi integrare con l'informazione linguistica. Il modello qui proposto integrerebbe informazioni di tre tipologie.

4.1 Informazioni relative all'attitudine ironica dell'enunciato

Dall'analisi dell'attitudine ironica della comunità online (Sez. 2, Seconda Fase) emergevano tendenze di comportamento differenti tra i commentatori quasi sempre ironici e gli altri utenti. In questo modello la prima operazione consisterebbe dunque nel riconoscere l'attitudine ironica generale di un commentatore per avviare una differente trattazione del materiale linguistico. Per fare ciò si dovrebbe innanzitutto considerare la *faccia* esibita dall'utente. Attraverso ricerche automatiche per parole o immagini sarebbe semplice collegare i *nickname* o gli *avatar* a contesti satirici e ipotizzare con buona probabilità che il commentatore sia spiccatamente ironico. Inoltre, poiché i commentatori molto ironici tendono con più probabilità a rispondere ad altri commenti ironici, si potrebbe estrarne manualmente una breve lista per considerare i commenti ai quali essi hanno replicato come più probabilmente ironici. Se un alto numero di commenti di uno stesso utente fosse poi commentato da più commentatori di questa lista, tale utente sarebbe inserito a sua volta nella lista (soprattutto in caso di *faccia* riconducibile a contesti divertenti o satirici).

4.2 Informazioni relative alla rete di rapporti nella comunità

In secondo luogo bisognerebbe analizzare come i commenti siano stati recepiti dalla comunità. Nelle risposte a un commento possono infatti trovarsi chiari indizi che tale commento sia stato recepito come ironico. Gli indizi da ricercare nelle risposte a un commento sarebbero:

- Espressioni che indicano una risata, soprattutto nell'incipit o isolati senza ulteriore testo.
- Parole appartenenti al campo semantico di "ironia".

Inoltre, poiché le possibilità di interpretare correttamente un enunciato ironico sono incrementate

dalla familiarità esistente tra coloro che interagiscono, il rilievo degli indizi di ironia presenti nelle risposte a un commento potrebbe essere ponderato in proporzione al numero totale di interazioni sul sito tra il commentatore originale e colui che risponde.

4.3 Informazioni di natura linguistica

Il ricorso ad alcune informazioni contestuali non escluderebbe comunque l'uso di elementi linguistici nel processo di individuazione automatica di ironia. I sistemi ad oggi implementati hanno ottenuto risultati parzialmente soddisfacenti e sarebbe errato ignorare l'aiuto che potrebbe derivarne. Gli indizi proposti da Carvalho et al. (2009) sembrano soprattutto convincenti poiché facilmente rilevabili e pensati appositamente per un identico contesto. Oltre a questi indizi, anche i puntini di sospensione e i giochi di parole si sono rivelati dei buoni segnalatori della presenza di ironia durante l'analisi empirica.

4.4 Modello integrato

Le serie di informazioni contestuali e linguistiche sarebbero integrate in un unico sistema di identificazione automatica di ironia.

Innanzitutto si isolerebbero gli utenti che esibiscono un'attitudine ironica molto spiccata all'interno della comunità. L'appartenenza a questo gruppo condizionerebbe la successiva trattazione del materiale linguistico sia interno al commento sia nelle sue risposte (v. Appendice A). Se l'utente dovesse appartenere al gruppo dei commentatori spiccatamente ironici:

- il suo commento sarà considerato molto probabilmente ironico a prescindere da indizi linguistici presenti nel commento stesso e nelle risposte.
- la presenza di indizi linguistici nel commento ne determinerebbe la classificazione come ironico. L'assenza sarebbe ininfluente.
- la presenza di indizi linguistici nel commento ne determinerebbe la classificazione come ironico. L'assenza influirebbe solo quando, in numerose risposte, nessuna li presentasse.
- se classificati come ironici, i commenti sarebbero considerati ironici nella loro interezza.

Se invece l'utente non dovesse appartenere al gruppo dei commentatori spiccatamente ironici:

- il suo commento sarà considerato ironico solo in presenza di indizi linguistici nel commento stesso o nelle sue risposte.
- la presenza di indizi linguistici nel commento inciderebbe positivamente sulla sua classificazione come ironico. L'assenza inciderebbe negativamente.
- la presenza di indizi linguistici nelle risposte inciderebbe positivamente sulla sua classificazione come ironico. L'assenza inciderebbe negativamente.
- se i commenti fossero classificati come ironici a partire da indizi linguistici interni al commento stesso, saranno ritenute ironiche solo le frasi contenenti tali indizi, soprattutto al principio o alla fine del commento.

5 Conclusioni

In questo lavoro è stata presentata la possibilità di usare informazioni contestuali per identificare automaticamente l'ironia nei commenti dei lettori abituali di giornali online. A tal fine è stato proposto un possibile approccio computazionale che identifichi i commentatori maggiormente ironici di una comunità, suggerendo un trattamento differente del materiale linguistico tra essi e gli altri commentatori. L'integrazione di informazioni contestuali e informazioni linguistiche potrebbe influire positivamente sull'efficacia dei sistemi di riconoscimento automatico di ironia, che avrebbero una funzione importante nel campo della Sentiment Analysis.

Al momento stiamo ampliando la ricerca valutando l'influenza di informazioni come il tipo di quotidiano, l'argomento della notizia e la lunghezza del commento su un corpus di commenti più ampio e costruito su più quotidiani.

Ovviamente un'integrazione di informazioni contestuali così basilari non risolverebbe completamente il problema di come identificare automaticamente l'ironia in testi online.

Tuttavia questo lavoro riflette la ferma convinzione che questi progressivi tentativi di integrare informazioni contestuali semplici e computabili con l'informazione linguistica siano oggi la migliore strada da percorrere per tentare di affrontare automaticamente fenomeni di natura pragmatica così complessa e sfaccettata come l'ironia.

Ringraziamenti

Vorremmo ringraziare Caterina Mauri, Silvia Pareti, Diego Frassinelli e Bonnie Webber. Un ringraziamento va inoltre ad Alessandra Cervone.

Bibliografia

Katharina Barbe. 1995. *Irony in context*. John Benjamins Publishing Co. Amsterdam.

Lorenzo Bernardini. 2014. Un nuovo parametro per l'individuazione automatica dell'ironia: la fonte dell'enunciato. Tesi della Magistrale in Linguistica Teorica e Applicata. Università di Pavia, Pavia.

Penelope Brown and Steven Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press. Cambridge.

Paula Carvalho, Luís Sarmento, Mário J. Silva and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. ACM. 53–56.

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology General*. 121–126.

Sigmund Freud. 1948. *Der Witz und seine Beziehung zum Unbewussten: Gesammelte Werke, 6*. Imago Publishing Co. London.

Rachel Giora. 1995. On irony and negation. *Discourse processes*. 19(2): 239–265.

Erving Goffman. 1967. *Interaction ritual: essays on face-to-face behavior*. Aldine Publishing Co. Chicago.

Herbert P Grice. 1978. Logic and conversation. *Syntax and semantics: Vol. 9. Pragamatics*. Academic. New York.

Norman Knox. 1972. On the classification of ironies. *Modern Philology*. 70: 53–62.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*. 22(2): 126–142.

Walter Nash. 1985. *The language of humour: Style and technique in comic discourse*. Longman. Londra.

Antonio Reyes, Paolo Rosso and Davide Buscaldi. 2009. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*. 18(4): 311–331.

Antonio Reyes and Paolo Rosso. 2011. Mining subjective knowledge from customer reviews: A specific case of irony detection. *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*. Association for Computational Linguistics. 118–124.

Antonio Reyes, Paolo Rosso and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*. 47: 239–268.

Antonio Sarmento, Paolo Rosso and Tony Veale. 2013. Automatic creation of a reference corpus for political opinion mining in user-generated content. *Proceedings of the st international CIKM workshop on topic-sentiment analysis for mass opinion*. ACM. 29–36.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Blackwell. Oxford.

Wolf-Dieter Stempel. 1976. Ironie als Sprechhandlung. In Wolfgang Preisendanz and Rainer Warning. *Das Komische*. Fink. Munich. 205–237.

Akira Utsumi. 1996. A unified theory of irony and its computational formalization. *Proceedings of the 16th conference on computational linguistics*. Association for Computational Linguistics. Morristown, NJ. 962–967.

Tony Veale and Yanfen Hao. 2010. Detecting irony in creative comparisons. *Proceedings of 19th European conference on artificial intelligence-ECAI 2010*. IOS Press. Amsterdam. 765–770.

Cynthia M. Whissell. 1989. The dictionary of affect in language. In Robert Plutchik and Henry Kellerman (Eds.). *Emotion: Theory, Research and Experience*. Academic Press. New York. 113–131.

Appendici

A

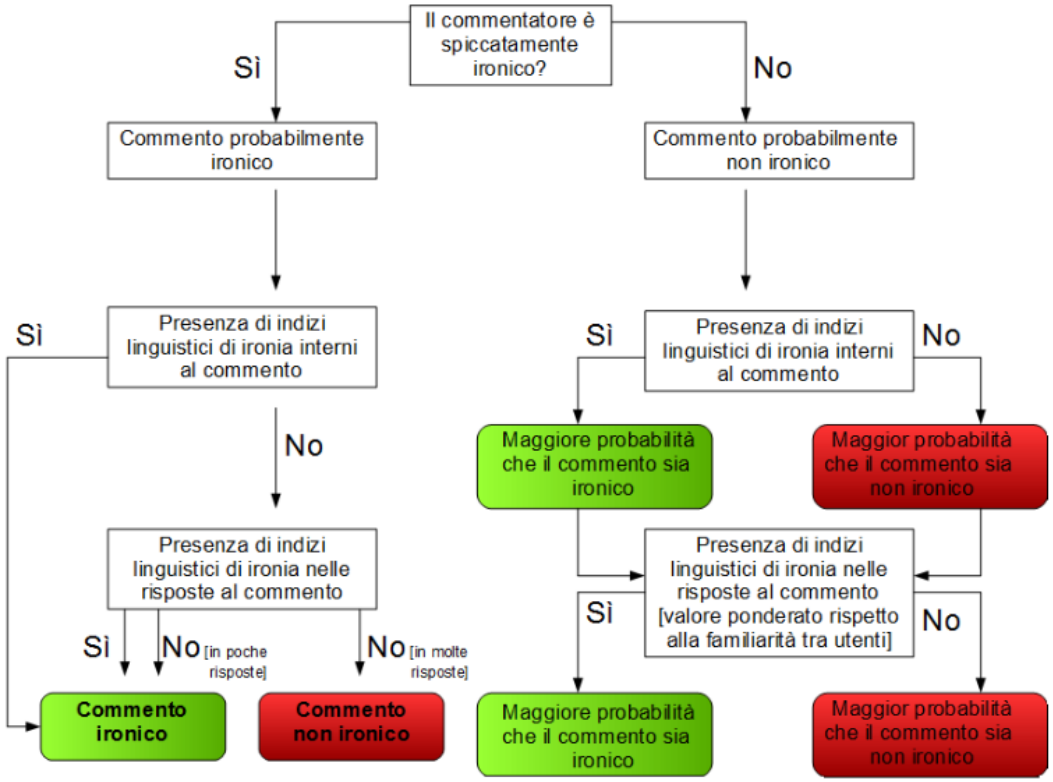


Figura 1: La figura 1 rappresenta il modello integrato di informazioni linguistiche e contestuali per il riconoscimento automatico di ironia. La classificazione dell'utente come spiccatamente ironico implicherebbe una differente trattazione del materiale linguistico contenuto nel commento stesso e nelle sue risposte. Per i commentatori ironici sarebbe sufficiente la presenza di un qualsiasi indizio nel commento o nelle sue risposte affinché esso sia classificato come interamente ironico, mentre l'unica condizione perché sia classificato come non ironico sarebbe la presenza di più risposte senza alcun indizio di ironia. Per gli altri commentatori invece la classificazione del commento come ironico dipenderebbe maggiormente dalla presenza di indizi sia nel commento che nelle sue risposte (ponderandone la rilevanza in base alla familiarità tra gli utenti). Inoltre i commenti classificati come ironici a partire da indizi presenti nel commento stesso saranno considerati ironici limitatamente alle porzioni di testo nelle quali tali indizi sono presenti.

10.12871/CLICIT2014115

A generic tool for the automatic syllabification of Italian

Brigitte Bigi, Caterina Petrone
Laboratoire Parole et Langage, CNRS, Aix Marseille Université
5 avenue Pasteur, 13100 Aix-en-Provence, France
{brigitte.bigi,caterina.petrone}@lpl-aix.fr

Abstract

English. This paper presents a rule-based automatic syllabification for Italian. Differently from previously proposed syllabifiers, our approach is more user-friendly since the Python algorithm includes both a Command-Line User and a Graphical User interfaces. Moreover, phonemes, classes and rules are listed in an external configuration file of the tool which can be easily modified by any user. Syllabification performance is consistent with manual annotation. This algorithm is included in SPPAS, a software for automatic speech segmentation, and distributed under the terms of the GPL license.

Italiano. *Questo articolo presenta una procedura di sillabificazione automatica per l'italiano basata su regole. Diversamente da altri sillabificatori, la nostra procedura è più facile da usare perchè l'algoritmo, compilato in Python, include un'interfaccia a linea di comando e un'interfaccia grafica. Inoltre i fonemi, le classi e le regole sono elencate in un file di configurazione esterno che può essere facilmente modificato. I risultati della sillabificazione automatica sono congruenti con quelli ottenuti dalle annotazioni a mano. L'algoritmo è incluso in SPPAS, un software per la segmentazione automatica del parlato distribuito secondo le condizioni di licenza GPL.*

1 Introduction

This paper presents an approach to automatic detection of syllable boundaries for Italian speech. This syllabifier makes use of the phonetized text.

The syllable is credited as a linguistic unit conditioning both segmental (e.g., consonant or vowel lengthening) and prosodic phonology (e.g., tune-text association, rhythmical alternations) and its automatic annotation represent a valuable tool for quantitative analyses of large speech data sets. While the phonological structure of the syllable is similar across different languages, phonological and phonotactic rules of syllabification are language-specific. Automatic approaches to syllable detection have thus to incorporate such constraints to precisely locate syllable boundaries.

The question then arises of how to obtain an acceptable syllabification for a particular language and for a specific corpus (a list of words, a written text or an oral corpus of more or less casual speech). In the state-of-the-art, the syllabification can be made directly from a text file as in (Cioni, 1997), or directly from the speech signal as in (Petrillo and Cutugno, 2003).

There are two broad approaches to the problem of the automatic syllabification: a rule-based approach and a data-driven approach. The rule-based method effectively embodies some theoretical position regarding the syllable, whereas the data-driven paradigm tries to infer new syllabifications from examples syllabified by human experts. In (Adsett et al., 2009), three rule-based automatic systems and two data-driven automatic systems (Syllabification by Analogy and the Look-Up Procedure) are compared to syllabify a lexicon.

Indeed, (Cioni, 1997) proposed an algorithm for the syllabification of written texts in Italian, by syllabifying words directly from a text. It is an algorithm of deterministic type and it is based upon the use of recursion and of binary tree in order to detect the boundaries of the syllables within each word. The outcome of the algorithm is the production of the so-called canonical syllabification (the stream of syllabified words).

On the other side, (Petrillo and Cutugno, 2003)

presented an algorithm for speech syllabification directly using the audio signal for both English and Italian. The algorithm is based on the detection of the most relevant energy maxima, using two different energy calculations: the former from the original signal, the latter from a low-pass filtered version. This method allows to perform the syllabification with the audio signal only, so without any lexical information.

More recently, (Iacoponi and Savy, 2011) developed a complete rule-based syllabifier for Italian (named Sylli) that works on phonemic texts. The rules are then based on phonological principles. The system is composed of two transducers (one for the input and one for the output), the syllabification algorithm and the mapping list (i.e., the vocabulary). The two transducers convert the two-dimensional linear input to a three-dimensional phonological form that is necessary for the processing in the phonological module and then sends the phonological form back into a linear string for output printing. The system achieved good performances compared to a manual syllabification: more than 0.98.5% (syllabification of spoken words). This system is distributed as a package written in C language and must be compiled; the program is an interactive test program that is used in command-line mode. After the program reads in the phone set definition and syllable structure parameters, it loops asking for the user to type in a phonetic transcription, calculating syllable boundaries for it, and then displaying them. When the user types in a null string, the cycling stops and execution ends. Finally, there are two main limitations: this tool is only dedicated to computer scientists, and it does not support time-aligned input data.

With respect to these already existing approaches and/or systems, the novel aspect of the work reported in this paper is as follows:

- to propose *a generic and easy-to-use tool* to identify syllabic segments from phonemes;
- to propose *a generic algorithm*, then a set of rules for the particular context of Italian spontaneous speech.

In this context, "generic" means that the phone set, the classes and the rules are easily changeable; and "easy-to-use" means that the system can be used by any user.

2 Method description

In the current study, we report on the adaptation of a rule-based system for automatic syllabification of phonemes' strings of the size greater than a graphic word. The system was initially developed for French (Bigi et al., 2010) and here adapted on Italian since there are currently no freely available system that can be used either by computer scientists and linguists.

The problem we deal with is the automatic syllabification of a phoneme sequences. The proposed phoneme-to-syllable segmentation system is based on 2 main principles:

1. a syllable contains a vowel, and only one;
2. a pause is a syllable boundary.

These two principles focus the problem on the task of finding a syllabic boundary between two vowels in each Inter-Pausal Unit (IPU), as described in Figure 1.

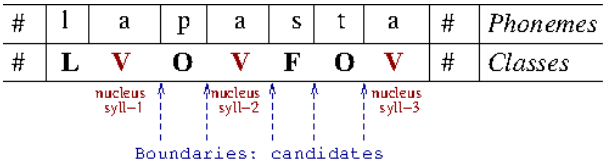


Figure 1: Main principles ("#" means a pause)

As in the initial system for French, we group phonemes into classes and establish language-specific rules dealing with these classes. The identification of relevant classes is then very important. The following classes were used:

V - Vowels: a e i o u O E oe ae
G - Glides: j w
L - Liquids: l L r
O - Occlusives: p t k b d g
F - Fricatives: s S f z tS ts v dz dZ
N - Nasals: m nf ng

Uppercase bold-letters indicate the abbreviations used for classes throughout this paper. The letter **C** is also used to mention one of G, L, O, N or F.

The system firstly check if the observed sequence of classes corresponds to an exception. If not, the general rules are applied (see Table 1).

The exception rules are:

- (Consonant + Glide) can't be segmented
- (Consonant + Liquid) can't be segmented
- (Consonant + Liquid + Glide) can't be segmented

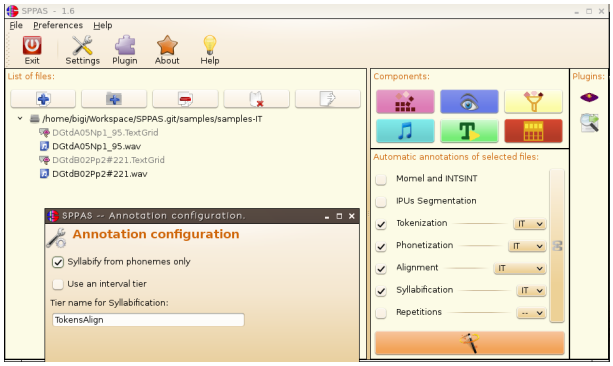


Figure 2: SPPAS: Graphical User Interface with syllabification options

	Observed sequence	Segmentation rule
1	VV	V.V
2	VCV	V.CV
3	VCCV	VC.CV
4	VCCCV	VC.CCV
5	VCCCCV	VC.CCCV
6	VCCCCCV	VCC.CCCV

Table 1: General Rules (V and C are phonological vowels/consonant respectively)

Notice that the rules we propose follow usual phonological statements for most of the spoken corpus. Our aim is not to propose a true set of syllabification rules for Italian, but to provide an acceptable syllabification for the most part of spoken corpora. We do not suggest that our solutions are the only ones particularly for syllables with complex structures as long as they are fairly uncommon in a given specific corpus. This is the reason why the tool implementing these rules was developed to be as generic as possible: any user can change either the phone set or the rules.

Finally, in the system described in (Bigi et al., 2010), the syllabification is performed between 2 silences (as defined in the main principles). From this system, we added the possibility to perform the syllabification between any kind of boundaries. In such case, a ”reference tier” is given by the user to the system. Table 2 shows an example when the time-aligned tokens are used as reference tier.

Of course, the reference tier can contain any type of annotation (we used tokens in the example, but prosodic contours, syntactic segments, etc. can be used if this annotation is available).

segment type	sentence	phonemes	syllables
sentence	la pasta	/lapasta/	la.pas.ta
	la stella	/lastela/	las.te.la
token	la.pasta	/la/ /pasta/	la.pas.ta
	la.stella	/la/ /stela/	la.ste.la

Table 2: Syllabification into segments, without changing the rules.

3 Implementing in a tool

The system proposed in this paper is included in SPPAS (Bigi, 2012), a tool distributed under the terms of the GNU Public License[1]. It is implemented using the programming language Python 2.7. Among other functions, SPPAS proposes an automatic speech segmentation at the phone and token levels for French, English, Spanish, Italian, Chinese, Taiwanese and Japanese. Moreover, the proposed software fulfills the specifications listed in (Dipper et al., 2004): it is a linguistic tool, free of charge, ready and easy to use, it runs on any platform and it is easy to install, the maintenance is guaranteed (at least until 2016), and it is XML-based. To download it, use the URL:

`http://www.lpl-aix.fr/~bigi/sppas/`

The current version (i.e. 1.6) allows to import data from Praat, Elan, Transcriber or from CSV files. The output can be one of ”xra” (native file format), ”TextGrid”, ”eaf” or ”csv”. The time-aligned phonemes (produced by SPPAS from the speech audio file and the orthographic transcription) are used as input to the syllabifier to produce 3 tiers with time-aligned syllables, classes and structures (as shown in Figure 3). A dictionary can be syllabified by using the same program, by ”simulating” time-alignments, and exporting the result in CSV format.

A simple ASCII text file that the user can change as needed contains the phoneset and the rules for the syllabification process.

4 Evaluation

All testing material was taken from CLIPS (Savy and Cutugno, 2009), distributed during the Evalita 2011 evaluation campaign. This corpus is made of about 15 map-task dialogues recorded by couples of speakers exhibiting a wide variety of Italian variants. Dialogues length ranges from 7/8 min-

[1]See: `http://www.gnu.org/licenses/gpl-3.0.en.html` for details

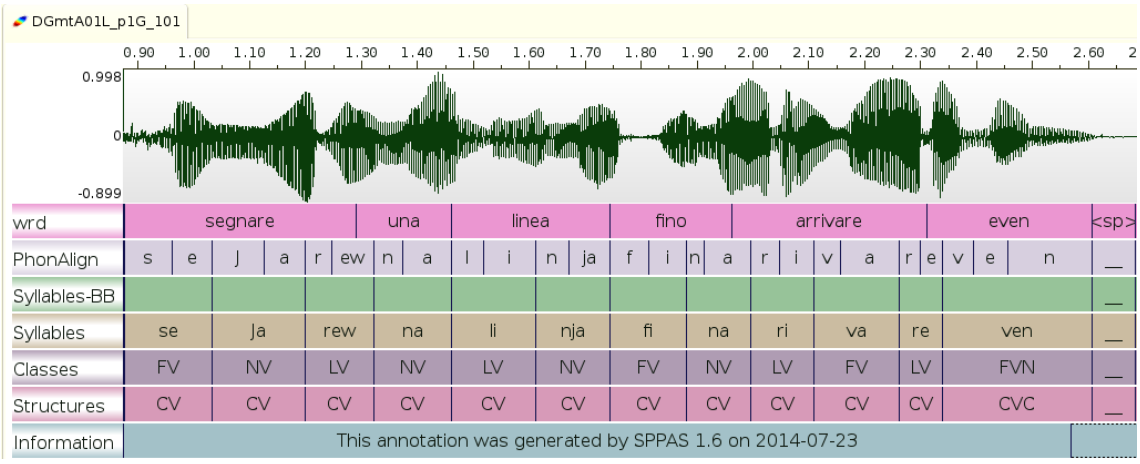


Figure 3: System output example on spoken Italian. The "Syllables-BB" tier (green) was manually annotated. Two assimilation phenomena can be observed in "fino arrivare even", which is phonetized as /finarivareven/ and that impact on the syllabification.

utes to 15/20 minutes, including word segmentation and phonetic segmentation. The test corpus was automatically extracted from these data with the following criteria: 1/ last 2 utterances of each speaker in each dialogue and 2/ all utterances containing from 100 to 106 phonemes. From such data, we kept only files containing more than 5 words, which represents about 10 minutes of spoken speech, and 1935 syllable boundaries have to be fixed. Notice that we have not corrected the transcription of phonemes for which we have not agree upon with the transcribers (as in Figure 3 for /ew/ in /seJarewna/).

The authors (one French - BB, one Italian - CP) manually syllabified the corpus and the resulting syllables were then compared with automatic syllabification obtained from the same corpus. In both cases, the syllabification was done by submitting the time-aligned phonemic representations of the sentences. One run was performed by using the basic system (phonemes only), and not by segmenting into intervals (see Figure 2 for both options). The agreement rates are:

- CP & BB: 99.12%
- CP & SPPAS-basic: 97.13%
- BB & SPPAS-basic: 97.80%

As the automatic system is using the phonemes only, it is important to notice that a part of the errors are due to the segmentation of words starting by 's' followed by a plosive (see Table 2). Unfortunately, by using the tokens as a reference tier for the segmentation, the results decrease to 96.1% (compared to BB). This is due to the large number of reductions and asimilations of spontaneous speech. However, we can create a tier with boundaries at pauses and specific boundaries before the /s/ for all words starting by /s/+plosive. The syllabification between such segments can then be used *to improve results to 98.2% (compared to CP) or 98.9% (compared to BB)*.

The results show that the program syllabification is very close to those made by human experts. Then, syllabification in Italian can be mostly predicted algorithmically, even when accounting for minor boundary segmentation phenomena found in speech.

5 Conclusion

The paper presented a new feature of the SPPAS tool that lets the user provide syllabification rules and perform automatic segmentation by means of a well-designed graphical user interface. The system is mainly dedicated to linguists that would like to design and test their own set of rules. A manual verification of the output of the program confirmed the accuracy of the proposed set of rules for syllabification of dialogues. Furthermore, the rules or the list of phonemes can be easily modified by any user. Possible uses of the program include speech corpus syllabification, dictionary syllabification, and quantitative syllable analysis.

References

Connie R Adsett, Yannick Marchand, et al. 2009. Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of italian. *Computer Speech & Language*, 23(4):444–463.

B. Bigi, C. Meunier, I. Nesterenko, and R. Bertrand. 2010. Automatic detection of syllable boundaries in spontaneous speech. In *Language Resource and Evaluation Conference*, pages 3285–3292, La Valetta (Malta).

B. Bigi. 2012. The SPPAS participation to Evalita 2011. In *Working Notes of EVALITA 2011, ISSN: 2240-5186*, Roma (Italy).

L. Cioni. 1997. An algorithm for the syllabification of written Italian. pages 22–24, Santiago de Cuba.

S. Dipper, M. Götze, and M. Stede. 2004. Simple annotation tools for complex annotation tasks: an evaluation. In *Proc. of the LREC Workshop on XML-based richly annotated corpora*, pages 54–62.

L. Iacoponi and R. Savy. 2011. Sylli: Automatic phonological syllabification for Italian. In *Proc. of INTERSPEECH*, pages 641–644, Florence (Italy).

M. Petrillo and F. Cutugno. 2003. A syllable segmentation algorithm for English and Italian. In *Proc. of INTERSPEECH*, Geneva (Switzerland).

R. Savy and F. Cutugno. 2009. CLIPS. diatopic, diamesic and diaphasic variations in spoken Italian. In *Proc. of the 5th Corpus Linguistics Conference*, Liverpool (England).

10.12871/CLICIT2014116

Errori di OCR e riconoscimento di entità nell'Archivio Storico de La Stampa

Andrea Bolioli
CELI Torino
abolioli@celi.it

Eleonora Marchioni
CELI Torino
marchioni@celi.it

Raffaella Ventaglio
CELI Torino
ventaglio@celi.it

Abstract

Italiano. In questo articolo presentiamo il progetto di riconoscimento delle menzioni di entità effettuato per l'Archivio Storico de La Stampa e una breve analisi degli errori di OCR incontrati nei documenti. L'annotazione automatica è stata effettuata su circa 5 milioni di articoli, nelle edizioni dal 1910 al 2005.

English. *In this paper we present the project of named entity recognition (NER) carried out on the documents of the historical archive of La Stampa and we show a short analysis of the OCR errors we had to deal with. We automatically annotated the authors of the articles, mentions of persons, geographical entities and organizations in approximately 5 million newspaper articles ranging from 1910 to 2005.*

1 Introduzione

In questo articolo descriveremo sinteticamente il progetto di annotazione automatica di menzioni di entità effettuato sui documenti dell'Archivio Storico de La Stampa, cioè il riconoscimento automatico delle menzioni di persone, entitá geografiche ed organizzazioni (le "named entities") effettuato su circa 5 milioni di articoli del quotidiano, seguito al progetto più ampio di digitalizzazione dell'Archivio Storico. [1]

Anche se il progetto risale ad alcuni anni fa (2011), pensiamo che possa essere d'interesse in quanto molti dei problemi affrontati e alcune delle metodologie utilizzate sono ancora attuali, a causa della maggiore disponibilità di vasti archivi storici di testi in formati digitali con errori di OCR. Si è trattato del primo progetto di digitalizzazione dell'intero archivio storico di un quotidiano italiano, e uno dei primi progetti internazionali di annotazione automatica di un intero archivio. Nel 2008 il New York Times aveva rilasciato un corpus annotato contenente circa 1,8 milioni di articoli dal 1987 al 2007 (New York Times Annotated Corpus, 2008), in cui erano state annotate manualmente persone, organizzazioni, luoghi e altre informazioni rilevanti utilizzando vocabolari controllati.

L'Archivio Storico de La Stampa comprende complessivamente 1.761.000 pagine digitalizzate, per un totale di oltre 12 milioni di articoli, di diverse pubblicazioni (La Stampa, Stampa Sera, Tuttolibri, Tuttoscienze, ecc.), dal 1867 al 2005. Il riconoscimento automatico di entità si è limitato agli articoli della testata La Stampa successivi al 1910, identificati come tali dalla presenza di un titolo, cioè a circa 4.800.000 documenti.

L'annotazione delle menzioni negli articoli consente di effettuare analisi sulla co-occorrenza tra entità e altri dati linguistici, sui loro andamenti temporali, e la generazione di infografiche, che non possiamo approfondire in questo articolo. Nella figura 1 mostriamo solamente come esempio il grafico delle persone più citate negli articoli del giornale nel corso dei decenni.

Nel resto dell'articolo presentiamo brevemente una analisi degli errori di OCR presenti nelle trascrizioni, prima di descrivere le procedure adottate per il riconoscimento automatico delle menzioni e i risultati ottenuti.

2 Analisi degli errori di OCR

Le tecniche di OCR (Optical Character Recognition) per il riconoscimento e la trascrizione auto-

[1] Come si legge nel sito web dell'Archivio Storico (www.archiviolastampa.it), "Il progetto di digitalizzazione dell'Archivio Storico La Stampa è stato realizzato dal Comitato per la Biblioteca dell'Informazione Giornalistica (CBDIG) promosso dalla Regione Piemonte, la Compagnia di San Paolo, la Fondazione CRT e l'editrice La Stampa, con l'obiettivo di creare una banca dati online destinata alla consultazione pubblica e accessibile gratuitamente."

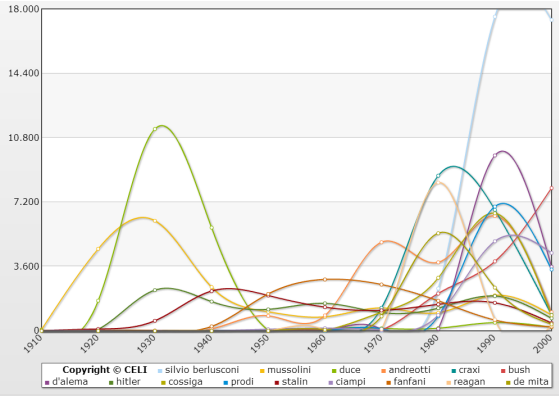


Figure 1: Grafico delle persone più citate

matica del testo comportano di per sé un margine di errore ”fisiologico”. Quando il documento cartaceo è di ottima qualità (buona carta e ottima qualità di stampa), le tecniche di OCR più moderne possono arrivare a percentuali di accuratezza molto alte. Nel caso di un archivio storico, dove la qualità dei documenti originali è ben lungi dall'essere ottimale, la quantità di errori di OCR aumenta notevolmente, soprattutto per i documenti più vecchi o in peggiore stato di conservazione, come attestato ad es. in (Holley, 2009) e nelle pubblicazioni del progetto europeo (IMPACT, 2010)).

Gli errori di OCR incidono sulle possibilità e sulla qualità delle elaborazioni successive sul testo. Una ricerca ”full-text” sui testi degli articoli non sarà ad esempio in grado di trovare le parole che contengono errori, oppure il testo sarà talvolta travisato, quando l'errore dà origine a una parola esistente ma diversa da quella originale. Allo stesso modo, il riconoscimento automatico delle menzioni di entità dovrà confrontarsi con questa situazione problematica.

2.1 Annotazione manuale e tipi di errore di OCR

Una misura affidabile dell'accuratezza dell'OCR si basa sul confronto tra il risultato dell' OCR e la trascrizione manuale corretta. Abbiamo quindi svolto un lavoro di annotazione manuale durante il quale dei linguisti hanno esaminato un campione dell'archivio (”corpus di valutazione”), individuando e classificando le anomalie di riconoscimento del testo.

Forniamo qui una sintesi della tipologia degli errori:

- Segmentazione degli articoli (”segmentation”): l'errata interpretazione degli elementi grafici della pagina (linee, titoli, cambiamento del corpo del carattere) può portare ad una errata segmentazione degli articoli, con diversi effetti possibili: un articolo risulta diviso in più parti, oppure diversi articoli vengono interpretati come uno solo, oppure un articolo risulta composto da porzioni di testo provenienti in realtà da articoli diversi.

- Segmentazione delle parole (”wordSep”): errori nell'interpretazione delle spaziature tra caratteri, parole o righe, che danno origine ad errori di segmentazione, ad esempio ”parlamenta re” (parlamentare), ”doc um e nto” (documento)

- Sillabazione (”hyphenation”): parole che vanno a capo nel testo originale (normalmente con trattino) vengono interpretate come parole separate, o come un'unica parola con trattino infisso. Ad esempio ”disprezzare” , ”prin-cipio”, ”re- latore”. Nei casi in cui neppure il trattino viene interpretato correttamente, la parola viene spezzata in due parti, es. ”seces sionista”.

- Riconoscimento dei caratteri alfabetici (”charError”): difetti nella qualità di stampa, macchie sulla carta, pieghe, graffi sul microfilm, ecc. possono portare ad una errata interpretazione dei caratteri. Ad esempio ” Kffégati” al posto di ”delegati”, ”coiitr'amnvirag'iao” anziché ”contrammiraglio”, ”cattchlico” per ”cattolico”. Un caso particolare è rappresentato dalla confusione tra lettere e numeri, ad esempio ”c0n” invece di ”con”, ecc.

- Sequenza delle parole (”wordSequence”): talvolta l'individuazione delle righe di testo operata dall'OCR può commettere errori dando origine a un testo dove le righe sono frammentate e mescolate impropriamente.

- Interpretazione di elementi grafici (”graphics”): linee, disegni, immagini possono essere interpretate dall'OCR come testo, dando origine a sequenze di caratteri errate.

- Punteggiatura (”punct”): l'errata interpretazione della punteggiatura può portare all'introduzione di segni di punteggiatura inesistenti, o all'assenza di segni necessari.

Spesso accade che punti, virgole, apici appaiono in posti sbagliati, ad es. ”sconqua.ssate””

Altri errori di OCR rilevanti per l'analisi automatica del testo riguardano l'interpretazione di maiuscole e minuscole, (ad es. ”DOSti” anziché ”posti”) e il significato delle parole: gli errori nel riconoscimento dei caratteri alfabetici possono dare origine a parole di senso compiuto che possono essere corrette soltanto considerando il contesto in cui occorre la parola (ad esempio ”casa” per ”cosa”, ”Baciale” al posto di ”sociale”).

2.2 Risultati dell'analisi degli errori di OCR

L'annotazione manuale degli errori di OCR è stata effettuata utilizzando una piattaforma web sviluppata ad hoc per consentire una annotazione collaborativa e veloce. Oltre ad annotare l'errore, il linguista annotava anche la possibile correzione. Sono stati annotati a mano 894 articoli di prima pagina del periodo 1930-2005, secondo le modalità descritte nel paragrafo precedente. Gli errori annotati sono complessivamente 16.842. I più frequenti sono gli errori di tipo ”charError”, cioè errori nell'interpretazione dei caratteri di una parola; seguiti dagli errori di tipo ”hyphenationSeparate”, cioè casi in cui una parola che andava a capo è stata interpretata come due parole distinte, con o senza trattino infisso.

A titolo di esempio elenchiamo alcuni degli errori più frequenti nelle edizioni dei decenni '90 e 2000, rilevanti per la NER: l'articolo ”una” è trascritto come ”ima”; la sequenza di caratteri ”li” viene riconosciuta come ”h” (ad es. l'articolo ”gli” é spesso scritto: ”gh”, ”pohtica” = ”politica”, ”poh”=”poli”); ”o” si trova scritto come ”0”; la lettera ”c” è interpretata come ”e” (es: ”dc” diventa ”de”, ”pci” diventa ”pei”).

Una analisi sistematica degli errori indotti da OCR direttamente sui nomi propri, abbastanza frequenti e variegati, sarebbe sicuramente interessante e non banale. Tra le annotazioni automatiche di persone, ad es., sono emerse le menzioni ”dustin hoffmann”, ”dustin hoflman”, ”dustin hoftman”, ”dustin holfman”, ”dustin hollman”, ”dustin hotfman”, ”dustin hotlman”, che potrebbero riferirsi all'attore americano.

Il post-processing dei documenti consentirebbe la correzione di alcuni degli errori risultati dall'OCR, utilizzando diverse tecniche, tra le quali:

- utilizzo di risorse linguistiche e semantiche, come dizionari ad alta copertura, risorse del semantic web come DBpedia, pattern sintattici;
- utilizzo di modelli statistici creati con apprendimento automatico;
- correzione manuale da parte degli utenti in modalità crowdsourcing, realizzata ad es. nel British Newspaper Archive [2] e nell'archivio dei Digitised newspapers della National Library of Australia [3]

3 Il riconoscimento delle menzioni di entità

All'analisi degli errori di OCR è seguito l'arricchimento semantico dei documenti tramite il riconoscimento automatico delle entità nominate (o ”Named Entity Recognition”, NER), cioè le persone, i luoghi e le organizzazioni menzionate negli articoli. Oltre alle persone citate nei testi, abbiamo annotato automaticamente gli autori degli articoli, per aggiungere un metadato utile ma, inaspettatamente, non banale da riconoscere.

Per effettuare il riconoscimento delle entità abbiamo utilizzato un metodo misto di apprendimento automatico e regole linguistiche, cioè abbiamo applicato in cascata un classificatore automatico SVM (Support Vector Machine) e un annotatore a regole (pattern linguistici). L'apprendimento automatico ha ovviamente richiesto una fase di annotazione manuale per creare il training set e il test set, utilizzato per valutare l'accuratezza.

3.1 Annotazione manuale e automatica

La vastità dell'archivio, sia come numero di documenti che come copertura temporale (5 milioni di articoli dal 1910 al 2005) e la varietà dei documenti (tutti gli articoli del giornale, dalla politica allo sport, dalla cultura alla cronaca, da inizio novecento al 2005) hanno posto problemi di scelta del corpus di articoli da annotare a mano per creare il data set di sviluppo.[4]

Abbiamo effettuato l'annotazione manuale delle menzioni di entità su un corpus di circa

[2] http://www.britishnewspaperarchive.co.uk

[3] http://trove.nla.gov.au

[4] La selezione del corpus di sviluppo è stato un processo molto articolato che non possiamo descrivere in dettaglio nel presente articolo.

1800 articoli, selezionati prevalentemente dalle prime pagine, dal 1910 al 2005 (per un totale di circa 582.000 token). Nell'annotazione manuale abbiamo seguito, per quanto possibile, le linee guida dell'I-CAB, Italian Content Annotation Bank, utilizzato a partire da Evalita 2007 nel task di Named Entity Recognition su articoli di giornale in italiano [5], contenente circa 525 articoli (I-CAB Evalita, 2007) del giornale L'Adige dell'anno 2004.

Il riconoscimento automatico di entità in testi storici che presentano errori di OCR è un tema affrontato in letteratura, ad es. in (Packer, 2010) e più recentemente in (Rodriquez, 2012), che riceverà probabilmente maggiore attenzione nei prossimi anni, grazie alla maggiore diffusione di archivi storici in formati digitali. Non disponendo di una soluzione sicura per questo problema, né di studi specifici per l'italiano, abbiamo deciso di utilizzare una metodologia di NER affidabile ed efficiente, cioè quella descritta in (Pianta, 2007), e utilizzata nel sistema che aveva dato i risultati migliori in Evalita 2007. Nelle edizioni successive di Evalita (2009 e 2011) le percentuali di accuratezza non sono migliorate in modo significativo e sono, viceversa, peggiorate nel task di NER da trascrizioni di notizie radiofoniche (Evalita, 2011), che contengono errori di trascrizione.

L'analisi linguistica di pre-processing del testo è stata effettuata con la pipeline UIMA (Apache UIMA, 2009) di CELI (annotatori UIMA in cascata per tokenizzazione, sentence splitting, analisi morfologica, disambiguazione, uso di gazetteers, ecc).

Il componente SVM utilizzato per il training e la creazione del modello è YamCha, Yet Another Multipurpose CHunk Annotator ((Kudo, 2001)). Per l'analisi automatica dei 5 milioni di testi, abbiamo integrato nella pipeline UIMA un componente di classificazione delle NE (cioè un annotatore di NE) che utilizzava il modello SVM creato. Dopo l'annotazione automatica con SVM, veniva applicato un componente a regole (che usava pattern linguistici), indispensabile per migliorare la correttezza delle annotazioni sia in casi particolari, come il riconoscimento degli autori, sia in altri casi rilevanti che non erano stati inclusi nel corpus di training.

[5] http://www.evalita.it/2007/tasks/ner

3.2 Risultati della NER

Forniamo qui sinteticamente alcuni dati sui risultati ottenuti dall'annotazione automatica dell'Archivio. Nella tabella seguente mostriamo il numero di named entities estratte da 4.800.000 articoli (solo le named entities che occorrono più di 10 volte) :

Tipo di NE	Num di NE	Num di documenti
PER	113.397	1.586.089
GPE	10.276	1.693.496
ORG	6.535	1.203.345
Autori	1.027	350.732

Nella tabella seguente mostriamo le misure di accuratezza (precision e recall), ottenute sul corpus di testing di 500 documenti.

Tipo di NE	Precision %	Recall %
PER	80.19	78.61
GPE	87.82	82.54
ORG	75.47	50.49
Autori	91.87	47.58

Tra le entità "standard", quelle di tipo ORG si sono dimostrate le più difficili da annotare automaticamente, come era prevedibile. Sorprendentemente invece è stato difficile il riconoscimento automatico degli autori, a causa degli errori di segmentazione degli articoli, dell'uso di sigle, della posizione variabile a inizio articolo o alla fine, e della mancanza, a volte, di punteggiatura nelle porzioni di testo rilevanti.

Conclusioni

In questo breve articolo abbiamo accennato alcune delle metodologie e delle problematiche del progetto di annotazione automatica di 5 milioni di articoli dell'Archivio Storico de La Stampa. Abbiamo segnalato alcune difficoltà legate alla presenza considerevole di errori di OCR e alla vastità e varietà dell'archivio (l'intero archivo va dal 1867 dal 2005). Queste problematiche potrebbero essere affrontate positivamente utilizzando informazioni e metodologie che non abbiamo potuto sperimentare in questo progetto, come ad es. il crowdsourcing.

Acknowledgments

Ringraziamo Francesco Cerchio, Vittorio Di Tomaso, Dario Gonella, Gianna Cernuschi, Roberto Franchini e gli altri colleghi che hanno contribuito alla realizzazione del progetto.

References

Anderson, N., A. Conteh and N. Fitzgerald. 2010. *IMPACT: Building Capacity in Mass Digitisation*. Presentation at the IS&T Archiving conference (1-4 June, The Hague, The Netherlands)

The Apache Software Foundation. 2009. *Apache UIMA Specifications (Unstructured Information Management Architecture)* http://uima.apache.org/uima-specification.html The Apache Software Foundation

Rose Holley. 2009. *How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs* D-Lib Magazine.

Taku Kudo e Yuji Matsumoto. 2001. *Chunking with Support Vector Machines*. Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies Pages 1-8

Valentina Bartalesi Lenzi, Manuela Speranza, Rachele Sprugnoli 2013. *Named Entity Recognition on Transcribed Broadcast News at EVALITA 2011*. In Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Rome, Italy, January 24-25, 2012. Springer 2013 Lecture Notes in Computer Science

Packer, T. L., J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi, and L. S. Jensen. 2010. *Extracting person names from diverse and noisy OCR text*. Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM.

Pianta, E., Zanoli, R. 2007. *Exploiting SVM for Italian Named Entity Recognition*. In Intelligenza Artificiale, Special Issue on NLP Tools for Italian, IV-2.

K.J. Rodriquez and Michael Bryant and T. Blanke and M. Luszczynska 2012. *Comparison of Named Entity Recognition tools for raw OCR text*. KONVENS Proceedings 2012 (LThist 2012 workshop).

Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium, Philadelphia.

Manuela Speranza 2007. *EVALITA 2007: The Named Entity Recognition Task*. In Proceedings of EVALITA 2007, Workshop held in conjunction with AI*IA, Rome, 10 September 2007. Intelligenza artificiale, 4-2.

10.12871/CLICIT2014117

Computer Assisted Annotation of Themes and Motifs in Ancient Greek Epigrams: First Steps

Federico Boschetti **Riccardo Del Gratta** **Marion Lamé**
ILC-CNR
Via Moruzzi, 1
56124 Pisa - ITALY
{firstname.lastname}@ilc.cnr.it

Abstract

English. This paper aims at illustrating some tools to assist the manual annotation of themes and motifs in literary and epigraphic epigrams for the PRIN 2010/2011 Memorata Poetis Project.

Italiano. *Questo contributo ha lo scopo di illustrare alcuni strumenti per assistere l'annotazione manuale di temi e motivi in epigrammi letterari ed epigrafici, nell'ambito del progetto PRIN 2010/2011 Memorata Poetis.*

1 Overview

The Memorata Poetis Project is a national funded project (PRIN 2010/2011), led by Professor Paolo Mastandrea, "Ca' Foscari" University of Venice, in continuity with the Musisque Deoque Project (Mastandrea and Spinazzè, 2011). It aims at the study of the intertextuality between epigraphic and literary epigrams in Greek, Latin, Arabic and Italian languages. Some of those epigrams are translated in more languages. Currently the access to the website (http://memoratapoetis.it) is restricted to the project workgroups but the access will be public before the end of the project, i.e. February 2016.

To understand the specific goal of this work in progress, a broader presentation of the project is necessary. Epigrams are short poems and follow specific schemes, contents and structures. Those short poems are transmitted both by epigraphs and by manuscripts, with interesting relations between the different traditions: an epigram can have been copied from stone to parchment, losing its original function and contextualization or, on the contrary, a literary epigram can have been adapted to a new epigraphic situation. As inscription, epigrams are a communication device inserted in a cultural construct. They are part of an information system and this implies, in addition to texts and their linguistics aspects: writings, contexts and iconotextual relationships. This holistic *and systemic* construction creates meanings: in Antiquity and in Middle-Ages, for instance, epigrams, as inscriptions, were often epitaphs.

Intertextuality also takes into account this relation between images of the context and the epigrams. For instance, "fountain" is a redundant motive in epigrams. An epigram that refers to divinities of water could be inscribed on a fountain. Such epigraphic situation participates to the global meaning. It helps to study the original audience and the transmission of epigrams. The reuse of themes and motifs illustrates how authors work and may influence other authors. From epigraphs to modern edition of epigrams, intertextuality draws the movement of languages and concepts across the history of epigrams.

Here is an example of a poetic English translation of a Theocritus' epigram:

> XV. [For a Tripod Erected by Damoteles to Bacchus] The precentor Damoteles, Bacchus, exalts / Your tripod, and, sweetest of deities, you. / He was champion of men, if his boyhood had faults; / And he ever loved honour and seemliness too.

(transl. by Calverly, 1892, https://archive.org/details/Theocritus/TranslatedIntoEnglishVerseByC.s.Calverley)

Effectively, European cultures enjoyed epigrams since the Antiquity, copied them, translated them, and epigrams became a genre that philology studies ardently. This intercultural process transforms epigrams and, at the same time, tries to keep their essence identifiable in those themes and

motifs. Naturally, those themes and motifs, such as "braveness", "pain", "love" or more concretely "rose", "shield", "bee" are reflecting the concepts in use in several different languages. The Memorata Poetis Project tries to capture metrical, lexical and semantic relations among the document of this heterogeneous multilingual corpus.

The study of intertextuality is important to understand the transmission of knowledge from author to author, from epoch to epoch, or from civilization to civilization. Even if the mechanisms of the transmission are not explicit, traces can be found through allusions, or thematic similarities. If the same themes are expressed through the same motif(s), probably there is a relation between the civilizations, which express this concept in a literary form, independently by the language in which it is expressed. For instance, the concept of the shortness of life and the necessity to enjoy this short time is expressed both in Greek and Latin literature:

Anthologia Graeca 11, 56 Πῖνε καὶ εὐφραίνου. τί γὰρ αὔριον ἢ τί τὸ μέλλον, / οὐδεὶς γινώσκει. (transl.: Drink and be happy. Nobody knows how will be tomorrow or the future.)

Catullus, *carmina*, 5 Viuamus, mea Lesbia, atque amemus / ... / Nobis cum semel occidit breuis lux, / Nox est perpetua una dormienda. (transl.: Let us live and love, my Lesbia [...] when our short light has set, we have to sleep a never ending night.)

Whereas other units are working on Greek, Latin, and Italian texts, the ILC-CNR unit of the project currently has in charge the semantic annotation of a small part of the Greek and of all the Arabic texts and it is developing computational tools to assist the manual annotation, in order to suggest the most suitable tags that identify themes and motifs. The semantic annotation of literary and historical texts in collaborative environments is a relevant topic in the age of the Semantic Web. At least two approaches are possible: a top-down approach, in which an ontology or a predefined taxonomy is used for the annotation, and a bottom-up approach, in which the text can be annotated with unstructured tags that will be organized in a second stage of the work. By combining these approaches, it is possible to collect more evidence to establish agreement on the annotated texts.

2 Manual Annotation

The distinction between theme and motif is a challenging theoretical question that this large scale intertextual work aims at studying in depth when the critical mass of annotations will be reached. Up to now, among heterogeneous and opposite discussions (see Lefèvre, 2006), the position shared by the Memorata Poetis working groups is that a theme is composed of motifs, even if the taxonomy adopted for the annotation does not reflect a neat distinction between these complex and interrelated concepts. The taxonomy of the themes and motifs has been established by an expert of ancient Greek literature, Gian Carlo Scarpa ("Ca' Foscari" University of Venice), and an expert of Latin literature, Paola Paolucci (University of Perugia). The items of the taxonomy are close to one hundred. The number varies due to the periodic revisions, coordinated by the central unit, according to the proposals of the other operative units.

Despite the large number of items for the classification, the taxonomy has only three levels of depth, for instance:

```
Res > Alimenta et potiones > Vinum
(Things > Food and drinks > Wine)
```

The repertory of themes and motifs is based on the study of the indices of "notabilia" in authoritative editions of Greek and Latin poetic collections of the last five centuries. Thus, the taxonomy adopted is grounded in a long tradition of studies, which organizes the themes in spheres of pertinence, such as the semantic spheres of plants, animals, human beings and gods. Hower, its hierarchical structure prevent transversal relationships, such as the relation between the body parts of human beings and the body parts of animals, which obviously do not share the same kind of body parts. An ontology-driven organisation of such themes and motifs should enrich the expressivity of the description.

3 Granularity Issues

Manual annotations are performed at the level of the entire epigram, when the annotator is aware that the theme (or the motif) interests the entire document, or at the level of a single verse, if the annotator identifies the line interested by a specific theme or motif. It is not possible to annotate a single word, because the annotation of single words is slower and the citation practices in the domain of

classical philology, related to the identification of themes and motifs, usually require the indication of the verse (or verse sequences). The automated tools that we are developing requires a finer granularity, at word level. Even if the manual annotation is performed at the granularity of verse, the individuation of the highest correlations between peculiar words and themes or motifs that they contribute to express, is a useful exploratory strategy. For this reason it is necessary to lemmatize the texts and to calculate the correlation between a specific word and a specific theme or motif. Relevant associations are used to rank the epigrams not yet manually annotated, but candidate to contain the pertinent theme or motif. In addition, those levels of granularity illustrate the complementarity between top-down and bottom-up approaches. A granularity based on words refers to a semantics defined by the strict content of the text and not on the prototype that the annotator has in mind. For instance, in one case the concept of flower is extensionally defined by the occurrences in the actual texts, through the names of specific flowers, such as "rose" and "violet". In the second case, the annotator could associate to a sequence of words that never contain the term "flower" a projection of his or her prototypical idea of flower.

4 Lemmatization

The lemmatization of the epigrams has been performed using *Morpheus*, the morphological analyzer for Ancient Greek developed at the Perseus Project (Crane, 1991). Multiple suggestions related to lemma and pos are scored according to the probabilities calculated on the Ancient Greek Treebank, `http://nlp.perseus.tufts.edu/syntax/treebank/greek.html`.

5 Identification of words highly associated to specific themes and motifs

Currently more than 10,000 manual annotations have been performed by the collaborators to the project. The annotated verses have been tokenized by the ILC-CNR unit and tokens have been lemmatized. By evaluating the correlation between lemma and theme or motif, the system that we are developing is able to suggest the most suitable tags. A couple of examples can clarify:

Bákchos is the name of the god of wine, and he is highly correlated to the following themes and motifs:

`Bacchus (Bacchus),`
`Crapula (pleasure),`
`Vinum (wine, as a drink),`
`Vinum curis remedium`
`(wine, as a solution to cares)`

sakós, which means *shield*, is highly associated to the motif `Instrumenta belli (war instruments)`

The association between synsets and themes is derived from the association between words and themes. Each verse has been lemmatized and the correlation of each lemma with the leading theme associated to the verse is evaluated calculating the log likelihood ratio, which involves as parameters the frequency of the pairs lemma - theme under observation; the frequency of the pairs lemma - different theme; the frequency of the pairs different lemma - theme under observation and, finally, the frequency of the pairs other lemma - other theme. From the pairs with the highest scores, lemmas are extracted and searched on Ancient Greek WordNet, in order to identify relevant synonyms.

6 The Ancient Greek WordNet

Even if a word has never been previously annotated by hand, it can be associated to a specific theme or motif, recurring to the growing Ancient Greek WordNet (Bizzoni et al., 2014). As shown in Fig. 1, the Ancient Greek WordNet is connected to Latin, Italian, Arabic, English and Croatian WordNets.

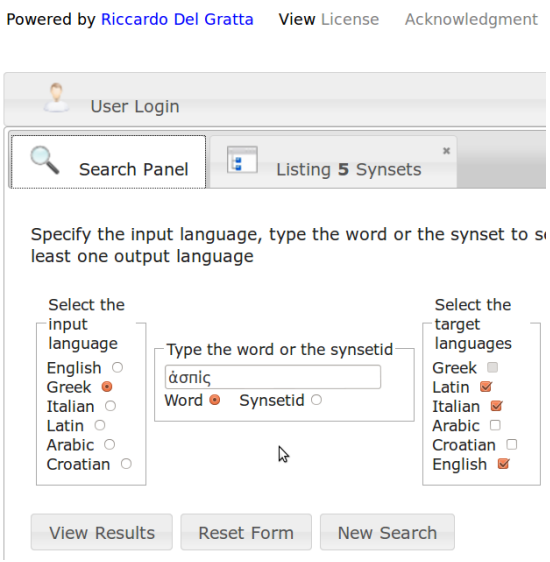


Figure 1: Ancient Greek WordNet GUI.

Searching for the aforementioned word *sákos*, it provides the following synset, composed both by frequent (such as *aspís*) and rare words (such as *prothorákion*).

LSJ Entry	LSJ Translation
sákos	shield
aspideîon	shield
aspís	shield
proballós	shield
párme	buckler
prothorákion	shield

Table 1: Greek-English pairs.

7 Improving the Coverage of Ancient Greek WordNet

The coverage of Ancient Greek WordNet (27% of the lexicon) is still poor, because the associations between Greek and English, the pivot language, has been performed extracting pairs from the LSJ bilingual dictionary and only precise matchings with an English word (or phrase) in Princeton WordNet create an association.

In order to improve the coverage, multiword English definitions have been parsed with the Stanford parser, and the identified head is assumed to be a word belonging to a hypernymic synset. An example:

word: boágrion
English translation: a shield of wild bull's hide
head: shield
part of speach: NN
number of words: 6
syntactic structure:

```
(ROOT(NP(NP(DT)(NN))
      (PP(IN)
          (NP(NP(JJ)
              (NN)(POS))(NN)))))
```

8 Conclusion

In conclusion, we have presented a work in progress related to the lexico-semantic instruments under development at the ILC-CNR to assist the annotators that collaborate to the Memorata Poetis Project.

Acknowledgments

We acknowledge Luigi Tessarolo, "Ca' Foscari" University of Venice, for the data related to the current status of annotation of the Memorata Poetis database and Marina Caputo, University of Perugia, for information related to the procedures adopted for manual annotation.

References

Luisa Bentivogli and Emanuele Pianta. 2004. *Extending WordNet with Syntagmatic Information*. In Proceedings of the Second Global WordNet Conference, Brno, Czech Republic, January 20-23: 47-53.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini and Gregory Crane. 2014. *The Making of Ancient Greek WordNet*. LREC 2014: 1140-1147

Gregory Crane. 1991. Generating and Parsing Classical Greek. Literary and Linguistic Computing, 6(4): 243-245.

Matteo Lefèvre. 2006. *Per un profilo storico della critica tematica*. In C. Spila, ed., *Temi e letture*, Roma, 11-29.

Paolo Mastandrea and Linda Spinazzè. 2011. Nuovi archivi e mezzi d'analisi per i testi poetici. I lavori del progetto Musisque Deoque. Amsterdam.

Pavese Carlo Odo. 1997. I temi e i motivi della lirica corale ellenica. Pisa.

Gunawan Pranata and Erick Pranata. 2010. *Acquisition of Hypernymy-Hyponymy Relation between Nouns for WordNet Building.* In Proceedings of the 2010 International Conference on Asian Language Processing (IALP '10). IEEE Computer Society, Washington, DC, USA: 114-117.

Di Donato Francesca and Morbidoni Christian and Fonda Simone and Piccioli Alessio and Grassi Marco and Nucci Michele. 2013. *Semantic Annotation with Pundit: A Case Study and a Practical Demonstration* In Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities, DH-CASE '13, 16:1–16:4

10.12871/CLICIT2014118

Defining an annotation scheme with a view to automatic text simplification

Dominique Brunato Felice Dell'Orletta Giulia Venturi Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR)
ItaliaNLP Lab – *www.italianlp.it*
Via G. Moruzzi,1 – Pisa (Italy)
{name.surname}@ilc.cnr.it

Abstract

English. This paper presents the preliminary steps of ongoing research in the field of automatic text simplification. In line with current approaches, we propose here a new annotation scheme specifically conceived to identify the typologies of changes an original sentence undergoes when it is manually simplified. Such a scheme has been tested on a parallel corpus available for Italian, which we have first aligned at sentence level and then annotated with simplification rules.

Italiano. *In questo contributo presentiamo i primi passi delle ricerche attuali sulla semplificazione automatica del testo. In linea con gli approcci più recenti, proponiamo qui un nuovo schema di annotazione teso specificamente a identificare le tipologie di cambiamenti che una frase originale subisce quando viene semplificata manualmente. Questo schema è stato testato su un corpus parallelo disponibile per l'italiano, che abbiamo precedentemente allineato a livello di frase e successivamente annotato con le regole di semplificazione.*

1 Introduction

Automatic Text Simplification (ATS) as a field of research in NLP is receiving growing attention over the last few years due to the implications it has for both machine- and human-oriented tasks. For what concerns the former, ATS has been employed as a pre-processing step, which provides an input that is easier to be analyzed by NLP modules, so that to improve the efficiency of, e.g., parsing, machine translation and information extraction. For what concerns the latter, ATS can also play a crucial role in educational and assistive technologies; e.g., it is used for the creation of texts adapted to the needs of particular readers, like children (De Belder and Moens, 2010), L2 learners (Petersen and Ostendorf, 2007), people with low literacy skills (Aluìsio et al., 2008), cognitive disabilities (Bott and Saggion, 2014) or language impairments, such as aphasia (Carroll et al., 1998) or deafness (Inui et al., 2003).

From the methodological point of view, while the first attempts were mainly developed on a set of predefined rules based on linguistic intuitions (Chandrasekar et al., 1996; Siddharthan, 2002), current ones are much more prone to adopt data-driven approaches. Within the latter paradigm, the availability of monolingual parallel corpora (i.e. corpora of authentic texts and their manually simplified versions) turned out to be a necessary prerequisite, as they allow for investigating the actual editing operations human experts perform on a text in the attempt to make it more comprehensible for their target readership. This is the case of Brouwers et al. (2014) for French; Bott and Saggion (2014) for Spanish; Klerke and Søgaard (2012) for Danish and Caseli et al. (2009) for Brazilian Portuguese. To our knowledge, only a parallel corpus exists for Italian which was developed within the EU project Terence, aimed at the creation of suitable reading materials for poor comprehenders (both hearing and deaf, aged 7-11)[1]. An excerpt of this corpus was used for testing purposes by Barlacchi and Tonelli (2013), who devised the first rule-based system for ATS in Italian focusing on a limited set of linguistic structures.

The approach proposed in this paper is inspired to the recent work of Bott and Saggion (2014) for Spanish and differs from the work of Barlacchi and Tonelli (2013) since it aims at learning from a parallel corpus the variety of text adaptations that characterize manual simplification. In particular, we focus on the design and development of a new annotation scheme for the Italian language intended to cover a wide set of linguistic phenomena implied in text simplification.

[1] More details can be found in the project website: http://www.terenceproject.eu/

2 Corpus alignment

The Terence corpus is a collection of 32 authentic texts and their manually simplified counterpart, all covering short novels for children. The simplification was carried out in a cumulative fashion with the aim of improving the comprehension of the original text at three different levels: global coherence, local cohesion and lexicon/syntax.

Given its highly structured approach and the clearly focused target, we believe the Terence corpus represents a very useful resource to investigate the manual simplification process with a view to its computational treatment. In particular, we proceeded as follows. First, we selected the outcomes of the last two levels of simplification (i.e. local cohesion and lexicon/syntax) which were considered respectively as the original and the simplified version of the corpus. This choice was motivated by the need of tackling only those textual simplification aspects with a counterpart at the linguistic structure level. We then hand-aligned the resulting 1036 original sentences to the 1060 simplified ones. The alignment results (table 1) provide some insights into the typology of human editing operations. As we can see, in 90% of the cases a 1:1 alignment is reported; 39 original sentences (3.75%) have a correspondence 1:2, thus suggesting an occurred split; 2 original sentences have undergone a three-fold split (0.19%), i.e. they correspond to three sentences in the simplified version; 15 pairs of original sentences have been merged into a single one (2.88%). Finally, the percentage of misaligned sentences is 1% (7 sentences were completely deleted after the simplification, whereas 4 novel ones have been introduced in the simplified corpus).

	1:1	**1:2**	**1:3**	**2:1**	**1:0**	**0:1**
N°sentences	958	39	2	30	7	4
%	92.1	3.75	0.19	2.88	0.67	0.38

Table 1: Corpus alignment results

3 Simplification annotation scheme

For the specific concerns of our study, we have defined the following annotation scheme, covering six macro-categories: split, merge, reordering, insert, delete and transformation. For some of them, a more specific subclass has been introduced, while for others (e.g. reordering) we are providing a finer internal distinction and a qualitative analysis focused on some selected constructs. Such a two-leveled structure has been similarly proposed by Bott and Saggion (2014) and we believe it is highly flexible and reusable, i.e. functional to capture similarities and variations across paired corpora from diverse domains and for different categories of readers. In table 2 we report the typology of rules covered by the annotation scheme. For each rule we also provide the frequency distribution within the Terence corpus.

Simplification Annotation Scheme		
Classes	***Sub-classes***	***Freq. %***
Split		1.75
Merge		0.57
Reordering		8.65
Insert	Verb	4.93
	Subject	1.79
	Other	12.03
Delete	Verb	2.04
	Subject	0.49
	Other	19.45
Transformation	Lexical Substitution	40.01
	Anaphoric replacement	0.61
	Noun_to_Verb	1.59
	Verb_to_Noun (nominalization)	0.61
	Verbal Voice	0.53
	Verbal Features	4.93

Table 2: Simplification annotation scheme

Split: it is the most investigated operation in ATS, for both human- and machine-oriented applications. Typically, a split affects coordinate clauses (introduced by coordinate conjunctions, colons or semicolons), subordinate clauses (e.g., non-restrictive relative clauses), appositive and adverbial phrases. Nevertheless, we do not expect that each sentence of this kind undergoes a split, as the human expert may prefer not to detach two clauses, for instance when a subordinate clause provides the necessary background information to understand the matrix clause. In (1) we give an example of split from the corpus[2].

(1) O: *Mamma Gorilla sembrava completamente distrutta per le cure che dava al suo vivace cucciolетto Tito,* ***che stava giocando vicino alle grosse sbarre di acciaio che circondavano il recinto***.

[2] In all the examples of aligned sentences from the corpus, O stands for original and S for simplified.

S: *Mamma Gorilla sembrava proprio distrutta per le cure che dava al suo vivace cuccioletto Tito.* ***Tito stava giocando vicino alle grosse sbarre di acciaio che erano intorno alla loro area***.

Merge: it has to be intended as the reverse of split, i.e. the operation by which two (or more) original sentences are joined into a unique simplified sentence. Such a kind of transformation is less likely to be adopted, as it creates semantically denser sentences, more difficult to process (Kintsh and Keenan, 1973). Yet, to some extent (see the alignment results), this is a choice the expert can make (ex. 2) and it can be interesting to verify whether the sentences susceptible to be merged display any regular pattern of linguistic features that can be automatically captured.

(2) O: *Clara pensò che fosse uno dei cigni. Ma poi si rese conto che stava urlando!*

S: *In un primo momento, Clara pensò che fosse uno dei cigni, ma poi sentì urlare!*

Reordering: this tag marks rearrangements of words between the original sentence and its simplified counterpart (3). Clearly, changing the position of the elements in a sentence is not an isolated event but it depends upon modifications at lexicon or syntax; e.g., replacing an object clitic pronoun (which is preverbal with finite verbs in Italian) with its full lexical antecedent[3] yields the unmarked order SVO, associated with easier comprehension and earlier acquisition (Slobin and Bever, 1982). Conversely, the author of the simplified text may sometimes prefer a non-canonical order, when s/he believes, e.g., that it allows the reader to keep the focus stable over two or more sentences.

(3) O: *Il passante gli spiegò che, per arrivare al bidone,* ***doveva contare ben 5 bidoni a partire dal semaforo***.
S: *Il signore spiegò a Ugolino che* ***doveva contare 5 bidoni a partire dal semaforo****, per arrivare al bidone della carta*.

Insert: the process of simplification may even result in a longer sentence, because of the insertion of words or phrases that provide supportive information to the original sentence. Despite the cognitive literature suggests to reduce the inference load of a text, especially with less skilled or low-knowledge readers (Ozuru et al., 2009), it is difficult to predict what the author of a simple text will actually add to the sentence to make it clearer. It can happen that the sentence is elliptical, i.e. syntactically compressed, and the difficulty depends on the ability to retrieve the missing arguments, which are then made explicit as a result of the simplification. Our annotation scheme has introduced two more specific tags to mark insertions: one for verbs and one for subject. The latter signals the transformation of a covert subject in a lexical noun phrase[4].

(4) O: *Essendo da poco andata in pensione dal suo lavoro, disse che le mancavano i suoi studenti* […]

S: *Essendo da poco andata in pensione dal suo lavoro* ***come insegnante****, disse che le mancavano i suoi studenti* […]

Delete: a text should be made easier by eliminating redundant information. As for the *insert* tag, also deletion is largely unpredictable, although we can imagine that simplified sentences would contain less adjunct phrases (e.g. adverbs or adjectives) than the authentic ones. Such occurrences have been marked with the underspecified *delete* rule (ex. 5); two more restricted tags, *delete_verb* and *delete_subj*, have been introduced to signal, respectively, the deletion of a verb and of an overt subject (made implicit and recoverable through verb agreement morphology).

(5) O: ***Sembrava veramente che*** *il fiume stesse per straripare*.
S: *Il fiume stava per straripare*.

Transformation: under this label we have included six main typologies of transformations that a sentence may be subject to, in order to become more comprehensible for the intended reader. Such modifications can affect the lexical, morpho-syntactic and syntactic levels of sentence representation, also giving rise to overlapping phenomena. Our annotation scheme has intended to cover the following phenomena:

- *Lexical substitution*: that is when a word (or a multi-word expression) is replaced with another (or more than one), which is usually a more common synonym or a less specific term. Given the relevance of lexical changes in text simplification, which is also confirmed by our results, previous works have proposed feasible ways to automatize lexical simplification, e.g. by relying on electronic resources, such as WordNet (De Belder et al., 2010) or word frequency lists (Drndarevic et al., 2012). In our annotation scheme this rule has been conceived to be quite generic, as synonyms or hypernyms replacements do not

[3] This is also a case of anaphora resolution, for which a dedicated tag has been conceived.

[4] The covert/overt realization of the subject is an option available in null-subject languages like Italian.

cover all the strategies an author can adopt to reduce the vocabulary burden of a text. A finer characterization will be part of a qualitative analysis.

(6) O: *Il **passante** gli spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni a partire dal semaforo*.

S: *Il **signore** spiegò a Ugolino che doveva contare 5 bidoni a partire dal semaforo, per arrivare al bidone della carta*.

- *Anaphoric replacement*: the substitution of a referent pronoun with its full lexical antecedent (a definite noun phrase or a proper noun);

(7) O: *Il passante **gli** spiegò che, per arrivare al bidone, doveva contare ben 5 bidoni* […].

S: *Il signore spiegò **a Ugolino** che doveva contare 5 bidoni a partire dal semaforo*[…]

- *Noun_to_verb*: when a nominalization or a support verb construction is replaced with a simple verb.

(8) O: *Il giorno **della partenza**, i bambini salutarono i loro genitori durante la colazione.*

*S: Il giorno **in cui i genitori partirono**, i bambini li salutarono durante la colazione.*

- *Verb_to_noun*: to mark the presence of a nominalization or of a support verb construction instead of an original simple verb.

(9) O: *Benedetto era molto arrabbiato e voleva **vendicare** sua sorella*.

S: *Benedetto era molto arrabbiato e voleva **ottenere vendetta** per sua sorella*.

- V*erbal voice*: to signal the transformation of a passive sentence into an active (ex. 10) or vice versa. In our corpus we found only one application of the latter; this finding was expected since passive sentences represent an instance of non-canonical order: they are acquired later by typically developing children (Maratsos, 1974, Bever, 1970; for Italian, Cipriani et al., 1993; Ciccarelli, 1998) and have been reported as problematic for atypical populations, e.g. deaf children (Volpato, 2010). Yet, the "passivization" rule may still be productive in other typologies of texts, where it can happen that the author of the simplification prefers not only to keep, but even to insert, a passive, in order to avoid more unusual syntactic constructs in Italian (such as impersonal sentences). This is also in line with what Bott and Saggion (2014) observed for passives in Spanish text simplification.

(10) O: *Solo il papà di Luisa, "Crispino mangia cracker" era dispiaciuto, perché **era stato battuto da Tonio Battaglia***.

S: *Solo il papà di Luisa era triste, perché **Tonio Battaglia lo aveva battuto***.

- V*erbal features:* Italian is a language with a rich inflectional paradigm and changes affecting verbal features (mood, tense, aspect) have proven useful in discriminating between easy- and difficult-to-read texts in readability assessment task (Dell'Orletta et al., 2011). The easy-to-read texts examined there were also written by experts in text simplification, but their target were adults with limited cognitive skills or a low literacy level. Poor comprehenders also find it difficult to properly master verbal inflectional morphology, and the same has been noticed for other categories of atypical readers, e.g. dyslexics (Fiorin, 2009); thus, there is a probability that the simplification, according to the intended target, will alter the distribution of verbal features over paired sentences, as occurred in (11).

(11) O: *Sembrava veramente che il fiume **stesse** per straripare.*

S: *Il fiume **stava** per straripare*.

4 Conclusions and Perspectives

We have illustrated the first annotation scheme for Italian that includes a wide set of simplification rules spanning across different levels of linguistic description. The scheme was used to annotate the only existing Italian parallel corpus. We believe such a resource will give valuable insights into human text simplification and create the prerequisites for automatic text simplification. Current developments are devoted to refine the annotation scheme, on the basis of a qualitative and quantitative analysis of the annotation results; we are also testing the suitability of the annotation scheme with respect to other corpora we are also gathering in a parallel fashion. Based on the statistical findings on the productivity of each rule, we will investigate whether and in which way certain combinations of rules affect the distribution of multi-leveled linguistic features between the original and the simplified texts. In addition, we intend to explore the relation between text simplification and a related task, i.e. readability assessment, with the aim of comparing the effects of such combinations of rules on the readability scores.

Acknowledgments

The research reported in this paper has been partly supported by a grant from the project "intelligent Semantic Liquid eBook - iSLe", POR CReO 2007 – 2013, Regione Toscana, Italy.

References

S. M. Aluísio, L. Specia, T.A. Pardo, E.G. Maziero, and R. P. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceeding of the eighth ACM symposium on Document engineering*, pp. 240-248.

G. Barlacchi and S. Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing*-2013), pp. 476-487.

T.G. Bever. 1970. The cognitive basis for linguistic structures. In *J.R.Hayes (ed.) Cognition and the development of Language*. New York, Wiley.

S. Bott and H. Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation*, Volume 48, Issue 1, pp. 93-120, Springer Netherlands.

L. Brouwers, D. Bernhard, A-L. Ligozat, and T. François. 2014. Syntactic Sentence Simplification for French. In *The 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*. Gothenburg, Sweden, 27 April.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Association for the Advancement of Artificial Intelligence (AAAI).

H. Caseli, T. Pereira, L. Specia, T. Pardo, C. Gasperin, and S. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009)*, March 01–07, Mexico City.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the international conference on computational Linguistics*, pp. 1041–1044.

L. Ciccarelli. 1998. *Comprensione del linguaggio, dei processi di elaborazione e memoria di lavoro: uno studio in età prescolare*, PhD dissertation, University of Padua.

P. Cipriani, A. M. Chilosi, P. Bottari, and L. Pfanner. 1993. *L'acquisizione della morfosintassi in italiano*: *fasi e processi*. Padova: Unipress.

J. De Belder and M-F Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*.

J. De Belder, K. Deschacht, and M-F Moens. 2010. Lexical simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In *SLPAT 2011 - Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* (Edimburgo, UK, July 2011), pp. 73-83. Association for Computational Linguistics Stroudsburg, PA, USA.

B. Drndarevic, S. Stajner, and H. Saggion. 2012. Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of "Easy to read on the web" online symposium*.

G. Fiorin. 2009. The Interpretation of Imperfective Aspect in Developmental Dyslexia. In *Proceedings of the 2nd International Clinical Linguistics Conference*, Universidad Autónoma de Madrid, Universidad Nacional de Educación a Distancia, and Euphonia Ediciones.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing*, ACL 2003.

W. Kintsch and J. Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, pp. 257-274.

S. Klerke and A. Søgaard. 2012. Danish parallel corpus for text simplification. In *Proceedings of Language Resources and Evaluation Conference (LREC 2012)*.

M. Maratsos. 1974. Children who get worse at understanding the passive: A replication to Bever. *Journal of Psycholinguistic Research*, 3, pp. 65-74.

Y. Ozuru, K. Dempsey, and D. McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19, pp. 228-242.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Speech and Language Technology for Education (SLaTE)*.

A. Siddharthan. 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC 2002)*.

D. I. Slobin and R. G. Bever. 1982. Children use canonical sentence schemas. A cross-linguistic study of word order and inflections. In *Cognition*, 12(3), pp. 229-265.

F. Volpato. 2010. *The acquisition of relative clauses and phi-features: evidence from hearing and hearing-impaired populations*. PhD dissertation. Ca' Foscari University of Venice.

10.12871/CLICIT2014119

Senso Comune as a Knowledge Base of Italian language: the Resource and its Development

Tommaso Caselli
VU Amsterdam
t.caselli@vu.nl

Isabella Chiari
Università di Roma 'Sapienza'
isabella.chiari@uniroma1.it

Aldo Gangemi
CNR ISTC
agangemi@loa-cnr.it

Elisabetta Jezek
Università di Pavia
jezek@unipv.it

Alessandro Oltramari
Carnegie Mellon University
aoltrama@andrew.cmu.edu

Guido Vetere
IBM Italia
gvetere@it.ibm.com

Laure Vieu
CNRS IRIT
vieu@irit.fr

Fabio Massimo Zanzotto
Università di Roma 'Tor Vergata'
zanzotto@info.uniroma2.it

Abstract

English. *Senso Comune* is a linguistic knowledge base for the Italian Language, which accommodates the content of a legacy dictionary in a rich formal model. The model is implemented in a platform which allows a community of contributors to enrich the resource. We provide here an overview of the main project features, including the lexical-ontology model, the process of sense classification, and the annotation of meaning definitions (glosses) and lexicographic examples. Also, we will illustrate the latest work of alignment with MultiWordNet, to illustrate the methodologies that have been experimented with, to share some preliminary result, and to highlight some remarkable findings about the semantic coverage of the two resources.

Italiano. *Senso Comune è una base di conoscenza della lingua italiana, che offre il contenuto di un dizionario tradizionale in un ricco modello formale. Il modello è implementato in una piattaforma che consente di arricchire la risorsa ad una comunità di contributori. Qui forniamo una panoramica delle principali caratteristiche del progetto, compreso il modello lessicale-ontologico, il processo di classificazione dei sensi, l'annotazione delle definizioni (glosse) ed degli esempi d'uso lessicografici. Tratteremo inoltre del lavoro di allineamento con MultiWordNet, illustrando le metodologie che sono state sperimentate, e riportando alcune considerazioni circa la copertura semantica delle due risorse.*

1 Introduction

Senso Comune[1] is an open, machine-readable knowledge base of the Italian language. The lexical content has been extracted from a monolingual Italian dictionary[2], and is continuously enriched through a collaborative online platform. The knowledge base is freely distributed. *Senso Comune* linguistic knowledge consists in a structured lexicographic model, where senses can be qualified with respect to a small set of ontological categories. *Senso Comune*'s senses can be further enriched in many ways and mapped to other dictionaries, such as the Italian version of MultiWordnet, thus qualifying as a linguistic Linked Open Data resource.

1.1 General principles

The *Senso Comune* initiative embraces a number of basic principles. First of all, in the era of user generated content, lexicography should be able to build on the direct witness of native speakers. Thus, the project views at linguistic knowledge acquisition in a way that goes beyond the exploitation of textual sources. Another important assumption is about the relationship between language and ontology (sec. 2.1). The correspondence between linguistic meanings, as they are listed in dictionaries, and ontological categories, is not direct (if any), but rather *tangential*. Linguistic senses commit to the existence of various

[1] www.sensocomune.it

[2] T. De Mauro, Grande dizionario italiano dell'uso (GRADIT), UTET 2000

kinds of entities, but should not be in general confused with (and collapsed to) logical predicates directly interpretable on these entities. Finally, we believe that, like the language itself, linguistic knowledge should be owned by the entire community of speakers, thus they are committed to keep the resource open and fully available.

2 Senso Comune Essentials

2.1 Lexicon and ontology

In compliance with recent trends of research in integrating ontologies and lexical resources (see e.g. (Oltramari et al., 2013) and (Prévot et al., 2010)) *Senso Comune* model includes a lexicon and an ontology as independent semantic layers. Instead of providing *synsets* with formal specifications aimed at qualifying them as ontological classes (Gangemi et al., 2003), *Senso Comune* adopts a notion of *ontological commitment*, which can be summarized as follows:

If the sense S commits to ($\mapsto$) the concept C, then there are entities of type C to which occurrences of S may refer to.

$$(S \mapsto C) \Leftrightarrow \exists s, c | S(s) \wedge C(c) \wedge refers\_to(s, c)$$

This way, linguistic senses are not modelled as logical predicates to be directly interpreted with respect to individuals in some domain of quantification, but rather as *semiotic objects* that occur in texts or communication acts, whose relationship with other real world entities is mediated by cognitive structures, emotional polarity and social interactions.

As a consequence of this model, lexical relations such as synonymy, which hold among senses, do not bear any direct ontological import; conversely, ontological axioms, such as disjointness, do not have immediate linguistic side-effects. This approach allows senses of different types to be freely put into lexical relations, without the need of assigning the same (complex) type to every member of the synonymy relation; on the other hand, it prevents the system from directly inferring ontological relations out of linguistic evidences, which might be a limitation in many cases. Anyway, if the equivalence of linguistic senses to logic predicates is desired (e.g. for technical, monosemic portions of the dictionary), this condition can be specifically formalized and managed.

2.2 Sense classification

Meanings from De Mauro's core Italian lexicon have been clustered and classified according to ontological categories belonging to *Senso Comune* model, through a supervised process we called TMEO, a tutoring methodology to support sense classification by means of interactive enrichment of ontologies (Oltramari, 2012). TMEO is based on broad foundational distinctions derived from a simplified version of DOLCE[3] (Masolo et al., 2002) (Chiari et al., 2013). The overarching goal is to support users that, by design, have only access to the lexical level of the resource, in the task of selecting the most adequate category of the *Senso Comune* ontology as the super-class of a given lexicalized concept: different answer paths lead to different mappings between the lexical and the ontological layer of *Senso Comune* knowledge base.

Ongoing work on TMEO focuses on extending the coverage of the methodology and refining both the category distinctions in the ontology and the questions in the decision tree. In a previous experiment reported in (Chiari et al., 2010), we observed that users have a high degree of confidence and precision in classifying the concepts referring to the physical realm, while they face several problems in identifying abstract notions like 'company', 'text', 'beauty', 'duration', 'idea', etc. Accordingly, the new scheme, already tested in our last experiment (Jezek et al., 2014) summarized below, mainly improves the *Senso Comune* ontology in the abstract realm. It substitutes the too vague category `Idea` with the more generic `SocialOrMentalObject`, within which `InformationObject` and `Organization` are distinguished subcategories. In addition, the remaining abstract categories `TemporalQuality`, `Quality` and `Function` are complemented and grouped under a more general category `PropertyOrRelation`. Finally, we added the possibility to distinguish, for each category, a singular and a collective sense, thus allowing to annotate the main senses of the lemmas 'popolo' (people) and 'gregge' (herd) with the categories `Person` and `Animal` (adding a 'collective' tag). The results are a richer taxonomy and better organized decision tree.

[3] `http://www.loa.istc.cnr.it/old/DOLCE.html`

2.3 Annotation of lexicographic examples and definitions

Ongoing work in *Senso Comune* focuses on manual annotation of the usage examples associated with the sense definitions of the most common verbs in the resource, with the goal of providing *Senso Comune* with corpus-derived verbal frames. The annotation task, which is performed through a Web-based tool, is organized in two main subtasks. The first (task 1) consists in identifying the constituents that hold a relation with the target verb in the example and to annotate them with information about the type of phrase and grammatical relation. In semantic annotation (task 2), users are asked to attach a semantic role, an ontological category and the sense definition associated with the argument filler of each frame participant in the instances. For this aim, we provide them with a hierarchical taxonomy of 24 coarse-grained semantic roles based on (Bonial et al., 2011), together with definitions and examples for each role, as well as decision trees for the roles with rather subtler differences. The TMEO methodology is used to help them selecting the ontological category in a new simplified ontology based on *Senso Comune*'s top-level. For noun sense tagging, the annotator exploits the senses already available in the resource. Drawing on the results of the previous experiment on nouns senses, we allow multiple classification in all the three semantic subtasks, that is, we allow the users to annotate more than one semantic role, ontological category and sense definition for each frame participant. Up to now we performed two pilot experiments to release the beta version of the annotation scheme. The results of IA agreement are very good for the syntactic dependency annotation task and fair for the semantic task, the latter especially so since these tasks are notoriously difficult (see (Jezek et al., 2014) for details). Once completed, the annotated data will be used to conduct an extensive study of the interplay between thematic role information and ontological constraints associated with the participants in a frame; to refine the ontologisation of nouns senses in *Senso Comune* by assigning ontological classes to nouns in predicative context instead of nouns in isolation; to investigate systematic polysemy effects in nominal semantics on a quantitative basis. Our long-term goal is to enrich the resource with a rich ontology for verb types, informed by the empirical data provided by the annotated corpus.

3 Word Sense Alignment: Towards Semantic Interoperability

As a strategy to enrich the Senso Comune Lexicon (SCL) and make it interoperable with other Lexico-semantic resources (LSRs), two experiments of Word Sense Alignment (WSA) have been conducted: a manual alignment and an automatic one. WSA aims at creating a list of pairs of senses from two (or more) lexical-semantic resources where each pair of aligned senses denotes the same meaning (Matuschek and Gurevych, 2013). The target resource for the alignment is MultiWordNet (MWN) (Pianta et al., 2002).
SCL and MWN are based on different models[4]. The alignment aims at finding a semantic portion common to the set of senses represented in SCL by the conjunction of glosses and usage examples and in MWN by the synset words and their semantic relationships (hypernyms, hyponyms, etc.). Since semantic representation in the form of lexicographic glosses and in the form of synsets cannot be considered in any respect homomorphic the procedure of alignment is not biunique in any of the two directions. Thus, there are single SCL glosses aligned to more than one MWN synsets and single MWN synsets aligned with more than one SCL gloss. Another goal of the alignment experiments is the integration of high quality Italian glosses in MWN, so as to make available an enhanced version of MWN to NLP community, which could help improving Word Sense Disambiguation (WSD) and other tasks.

3.1 Manual aligment

On going work on the manual alignment of SCL and MWN synsets aims at providing associations between SCL glosses and synsets for all 1,233 nouns labelled as belonging to the basic vocabulary. The alignment is performed through the online platform that allows for each SCL word sense the association with one or more MWN synset.
At the time of this writing, 584 lemmas of SCL have been processed for manual alignment, for a total of 6,730 word senses (glosses), about 3.64 average word senses for each lemma. The alignment involves all SCL word senses, including

[4]Readers are referred to (Vetere et al., 2011) and (Caselli et al., 2014) for details on the two resources and their differences.

word senses not labelled as fundamental (about 29% of all word senses). Preliminary results show that only 2,131 glosses could be aligned with at least one MWN synset (31.7%) and 2,187 synsets could be aligned to at least one gloss. Exclusively biunique relationships among SCL glosses and MWN synsets involve 1,093 glosses. Each SCL gloss is associated to one synset in 1,622 cases (76.1%), to two synsets in 367 cases (17.2%), to three synsets 108 cases (5%), to four 25 (1,1%), to five in four cases, to six in three cases and to seven synsets in one case. While on the other side each MWN synset is associated to one SCL gloss in 1,681 cases (76.8%), to two glosses in 400 cases (18.2%), to three glosses in 85 cases (3.8%), to four in 17 cases, to five in three cases, and to six glosses in one case. The picture portrayed by the asymmetry of relationship between the granularity of SCL and MWN appears very similar, meaning that there is no systematic difference in the level of detail in the two resources aligned, as far as this preliminary analysis reveals. Attention should be drawn to the fact that biunique associations do not directly entail that the semantic representation deriving from the SCL gloss and the MWN synset are semantically equivalent or that they regard the same set of senses. These association only indicate that there is no other gloss or synset that can properly fit another association procedure. Levels of abstraction can be significantly different. Furthermore, as data show, there is a large number of SCL glosses not aligned to any MWN synset, and vice versa. This mismatch probably derives from the fact that MWN synsets are modelled on the English WN. Many WN synsets could be aligned to Italian senses outside the basic vocabulary; however, in general, we think that this mismatch simply reflects the semantic peculiarity of the two languages.

3.2 Automatic alignment

We conducted two automatic alignment experiments by applying state-of-the-art WSA techniques. The first technique, Lexical Match, aims at aligning the senses by counting the number of overlapping tokens between two sense descriptions, normalized by the length of the strings. We used `Text::Similarity` v.0.09 The second technique, Sense Similarity, is based on computing the cosine score between the vector representations of the sense descriptions. Vector representations have been obtained by means of the Personalized Page Rank (PPR) algorithm (Agirre et al., 2014) with WN30 extended with the "Princeton Annotated Gloss Corpus" as knowledge base[5]. The evaluation of the automatic alignments is performed with respect to two manually created Gold Standards, one for verbs and one for nouns, by means of standard Precision (P), Recall (R) and F1 score. The verb Gold Standard contains 350 sense pairs over 44 lemmas, while the noun Gold Standard has 166 sense pairs for 46 lemmas. The two gold standards have been independently created with respect to the manual alignment described in Section 3.1 and took into account only fundamental senses. Concerning the coverage of in terms of aligned entries, as for verbs MWN covers 49.76% of the SCDM senses while for nouns MWN covers 62.03% of the SCDM senses. The best results in terms of F1 score have been obtained by merging the outputs of the two approaches together, namely we obtained an F1 equals to 0.47 for verbs (P=0.61, R=0.38) and of 0.64 for nouns (P=0.67, R=0.61).

4 Conclusion

In this paper, we have introduced *Senso Comune* as an open cooperative knowledge base of Italian language, and discussed the issue of its alignment with other linguistic resources, such as WordNet. Experiments of automatic and manual alignment with the Italian MultiWordNet have shown that the gap between a native Italian dictionary and a WordNet-based linguistic resource may be relevant, both in terms of coverage and granularity. While this finding is in line with classic semiology (e.g. De Saussure's *principle of arbitrariness*), it suggests that more attention should be paid to the semantic peculiarity of each language, i.e. the specific way each language constructs a conceptual view of the World. One of the major features of *Senso Comune* is the way linguistic senses and ontological concepts are put into relation. Instead of equalising senses to concepts, a formal relation of *ontological commitment* is adopted, which weakens the ontological import of the lexicon. Part of our future research will be dedicated to leverage on this as an enabling feature for the integration of different lexical resources, both across and within national languages.

[5]Readers are referred to (Caselli et al., 2014) for details on the two methods used and result filtering.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

C. Bonial, W. Corvey, M. Palmer, V.V. Petukhova, and H. Bunt. 2011. A hierarchical unification of lirics and verbnet semantic roles. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 483–489. IEEE.

Tommaso Caselli, Carlo Strapparava, Laure Vieu, and Guido Vetere. 2014. Aligning an italianwordnet with a lexicographic dictionary: Coping with limited data. In *Proceedings of the Seventh Global WordNet Conference*, pages 290–298.

Isabella Chiari, Alessandro Oltramari, and Guido Vetere. 2010. Di che cosa parliamo quando parliamo fondamentale? In S. Ferreri, editor, *Atti del Convegno della Societ di Linguistica Italiana*, pages 185–202, Roma. Bulzoni.

Isabella Chiari, Aldo Gangemi, Elisabetta Jezek, Alessandro Oltramari, Guido Vetere, and Laure Vieu. 2013. An open knowledge base for italian language in a collaborative perspective. In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, DH-CASE '13, pages 14:1–14:6, New York, NY, USA. ACM.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer.

E. Jezek, L. Vieu, F. M. Zanzotto, G. Vetere, A. Oltramari, A. Gangemi, and R. Varvara. 2014. Enriching senso comune with semantic role sets. In *Proceedings of the Tenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Reykjavik, Iceland (May 26, 2014)*, pages 88–94.

C. Masolo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. 2002. WonderWeb Deliverable D17: The WonderWeb Library of Foundational Ontologies. Technical report.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-wsa: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 2:to appear.

A. Oltramari, P. Vossen, L. Qin, and E. Hovy, editors. 2013. *New Trends of Research in Ontologies and Lexical Resources*, volume XV of *Theory and Applications of Natural Language Processing*. Springer, Heidelberg.

Alessandro Oltramari. 2012. An introduction to hybrid semantics: The role of cognition in semantic resources. In Alexander Mehler, Kai-Uwe Khnberger, Henning Lobin, Harald Lngen, Angelika Storrer, and Andreas Witt, editors, *Modeling, Learning, and Processing of Text Technological Data Structures*, volume 370 of *Studies in Computational Intelligence*, pages 97–109. Springer Berlin Heidelberg.

Emanuele Pianta, Luisa Bentivogli, and Cristian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari, editors. 2010. *Ontology and the Lexicon*. Cambridge University Press.

Guido Vetere, Alessandro Oltramari, Isabella Chiari, Elisabetta Jezek, Laure Vieu, and Fabio Massimo Zanzotto. 2011. Senso Comune, an open knowledge base for italian. *Traitement Automatique des Langues*, 53(3):217–243.

10.12871/CLICIT2014120

CorEA: Italian News Corpus with Emotions and Agreement

Fabio Celli
University of Trento
Via Sommarive 5,
Trento, Italy
fabio.celli@unitn.it

Giuseppe Riccardi
University of Trento
Via Sommarive 5,
Trento, Italy
riccardi@disi.unitn.it

Arindam Ghosh
University of Trento
Via Sommarive 5,
Trento, Italy
arindam.ghosh@unitn.it

Abstract

English. In this paper, we describe an Italian corpus of news blogs, including bloggers' emotion tags, and annotations of agreement relations amongst blogger-comment pairs. The main contributions of this work are: the formalization of the agreement relation, the design of guidelines for its annotation, the quantitative analysis of the annotators' agreement.

Italiano. *In questo articolo descriviamo la raccolta di un corpus di blog giornalistici in Italiano che include le emozioni etichettate dai blogger e l'annotazione manuale con la relazione di approvazione tra commenti. I contributi principali di questo articolo sono: la formalizzazione della relazione di approvazione, le linee guida per la sua annotazione e l'analisi quantitativa dell'accordo tra annotatori.*

1 Introduction

Online news media, such as journals and blogs, allow people to comment news articles, to express their own opinions and to debate about a wide variety of different topics, from politics to gossips. In this scenario, commenters express approval and dislike about topics, other users and articles, either in a linguistic form and/or using like precoded actions (e.g. *like* buttons). Corriere is one of the most visited Italian news websites, attracting over 1.6 million readers everyday[1]. The peculiarity of *corriere.it* with respect to most news websites, is that it contains metadata on emotions expressed by the readers about the articles. The emotions (amused, satisfied, sad, preoccupied and indignated) are annotated directly by the readers on a voluntary basis. They can express one emotion per article. In this paper, we describe the collection of a corpus from *corriere.it*, that combines emotions and agreement/disagreement.

The paper is structured as follows: in section 2 we will provide an overview of related work, in sections 3 and 4 we will define the agreement/disagreement relation, describe the corpus, comparing it to related work, and provide the annotation guidelines. In section 5 we will draw some conclusions.

2 Background and Related Work

The CorEA corpus combines emotions and agreement/disagreement in a social media domain. Emotions and sentiment in corpora are usually annotated manually or automatically at message level. Examples of manually annotated corpora are Affective Text (Strapparava and Mihalcea, 2007), that contains annotation of news titles with emotion labels (anger, disgust, fear, joy, sadness, surprise), and sentiTUT (Bosco et al., 2013), that combines sentiment (positive/negative message polarity) and irony. Automatically and semi-automatically annotated corpora, like TWITA (Basile and Nissim, 2013), usually exploit external resources such as senticNet (Cambria et al., 2012). The peculiarity of CorEA is that emotions are annotated directly by commenters on a voluntary basis. These ground truth emotion labels (amused, satisfied, sad, preoccupied and indignated) are not at message level, but at author level. In other words are part of the bloggers' personal profile and describe all the emotions they declared after reading articles.

There are not many corpora of agreement/disagreement. The ICSI corpus of multi-party conversation (Shriberg et al., 2004), is a collection of 75 meetings between 53 unique speakers, annotated with

[1] source 'http://en.wikipedia.org/wiki/Corriere della Sera' retrieved in Jan 2014.

dialogue acts (including 4 labels for strong and weak agreement/disagreement) by 2 raters. A specific inter-annotator agreement for the agreement/disagreement relation is not reported. More recent corpora with agreement/disagreement labels are the AAWD corpus of Wikipedia talk pages (Bender et al., 2011), the AACD chat corpus (Morgan et al., 2013) and the IAC/ARGUE corpus of political debates (Abbott et al., 2011) (Walker et al., 2012). AAWD is a collection of asynchronous conversations from Wikipedia in English, Russian and Mandarin Chinese (about 500 threads and 325k tokens in total). It is annotated with 2 classes (agreement/disagreement, called positive/negative alignment) and authority claims by 2 annotators. AACD is a small corpus (12 threads, 14k tokens in total) of elicited chat dialogues in the same languages, annotated in the same way. The average inter-annotator agreement for alignment over the three languages of AAWD is Cohen's *k*=0.5 (Cohen, 1977). IAC/ARGUE is a large corpus in English (about 2700 authors, 11k threads) sampled from 4forums.com and annotated with Amazon's Mechanical Turk[2]. It combines agreement/disagreement, emotionality (subjective/objective), sarcasm, attack (objective/offensive language) and attitude (nice/nasty). Agreement/disagreement in IAC/ARGUE has been annotated with a scale +5, -5 and the inter-annotator agreement is α=0.62 (Krippendorff, 2004).

In all these corpora agreement/disagreement is at message level (post or utterance). There is also a corpus that combines LiveJournal and Wikipedia (118 threads) (Andreas et al., 2012), annotated with agreement/disagreement labels at sentence level (segments or chunks of messages). They reported inter-annotator agreement on 3 classes (agree/disagree/neutral) between 2 annotators as Cohen's *k*=0.73. CorEA corpus aggregates self-reported anntotations, such as emotions and likes, and metadata information, (ids, time stamps, etc.) about the conversation and human annotation of the agreement/disagreement relation.

3 Definition of Agreement/Disagreement

During debates in social media, participants attack or support the content of other participants' messages (Herring, 2007). This practice can be modeled in two different actions: 1) refer-to-message and 2) expression of agreement/disagreement. Refer-to-message, depicted as connection lines with round heads in figure 1, are directed links between pairs of messages. This information can be encoded as metadata in the message exchange strucure - as in Corriere - or as surface realizations in text in the form of coreference expressions (i.e. @ Lettore_10108563, see figure 1). Here we define agreement/disagreement as a relation, built on refer-to-message links, between a set of participants $P = \{p_i, ..., p_n\}$ to a conversation C that generate a set of messages $M = \{m_{ij}, ..., m_{nm}\}$, where m_{ij} is the j^{th} message of participant p_i. The conversation contains a set of

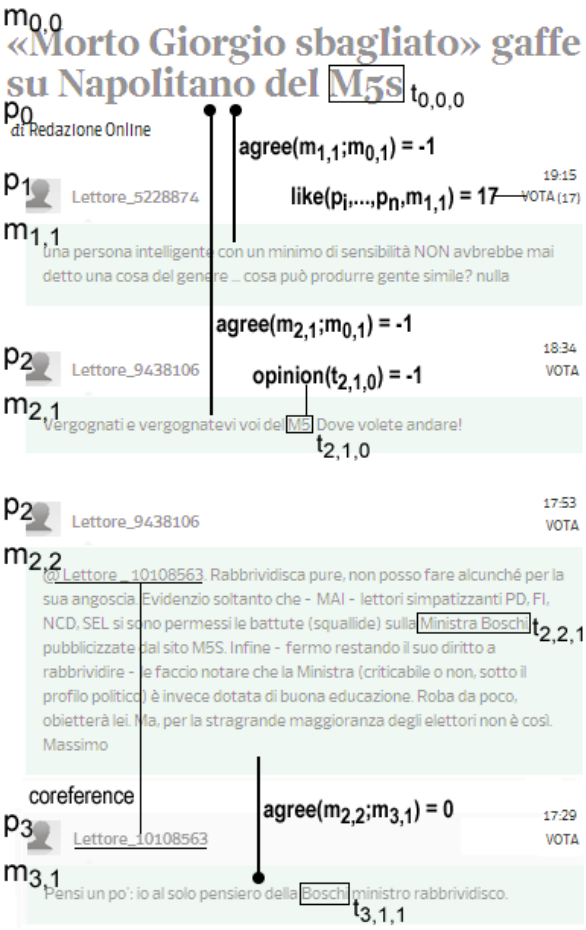


Figure 1: Example of asynchronous conversation in Corriere with participants $P = \{p_i, ..., p_n\}$, the messages they produce $M = \{m_{ij}, ..., m_{nm}\}$, sorted by time from bottom to top, and topics within messages $T = \{t_{ijk}, ..., t_{nmo}\}$. Connection lines with round heads are refer-to-message links, occasionally corresponding to coreferences. The agreement/disagreement relation is defined at message level as the *agree* functions $agree(m_{ij}; m_{i'j'})$ that maps pairs of participants/messages to values (+1,0,-1). *Opinion* is a function that maps a topic to a positive or negative polarity (+1,-1). *Like* is a function that measures the appreciation of participants to a message.

[2]https://www.mturk.com/mturk/welcome

topics $T = \{t_{ijk}, ..., t_{nmo}\}$, where t_{ijk} is the k^{th} topic of the conversation contained into the j^{th} message of participant p_i. We define topics as recurrent chunks or named entities appearing in different messages of C (see figure 1). We formalize agreement/disagreement as the *agree* function, that maps pairs of participants and messages to values between 1 (agree) and -1 (disagree), where 0 is neutral, as reported below:

$agree(m_{ij}; m_{i'j'}) = \{-1, 0, 1\}$

where m_{ij} is the parent participant/message pair, and $m_{i'j'}$ is the child participant/message pair. The parent m_{ij} precedes the child $m_{i'j'}$ in a time sequence. The child $m_{i'j'}$ is the j'^{th} message of $p_{i'}$ referred to the j^{th} message of p_i. The *agree* function is different from opinion expression and from like. The *opinion* function maps a topic to a positive or negative polarity (+1,-1):

$opinion(t_{ijk}) = \{-1, 1\}$

The *like* function measures the appreciation of a subset of participants to a message:

$like(p_i, ..., p_n; m_{ij}) = \{0, inf\}$

It is possible to define a more fine-grained function at topic level $agree(t_{ijk}; t_{i'j'k}) = \{-1, 0, 1\}$, where two (portions of) different messages m_{ij} and $m_{i'j'}$, connected by a refer-to-message link, are generated by two different participants (p_i and $p_{i'}$), and contain the same topic. The annotation of agreement/disagreement at topic level requires much more effort than at message level, we plan to annotate CorEA at topic level in the future. The agreement/disagreement relation concerns participants, messages and topics. Since participants are an important part of the relation, the *agree* function should exploit also information about them. This is why we combined emotions and agreement/disagreement relations in a single corpus. In Corriere, and social media in general, users/commenters/bloggers/authors are participants, comments/posts are messages, threads are conversations and articles are the first message of a conversation. In the next section we describe the procedure for the annotation of agreement/disagreement in CorEA.

4 Data, Annotation Schema and Guidelines

The CorEA corpus is a collection of news articles and comments from Corriere. It contains 27 news articles, about 1660 unique authors and more than 2900 posts (comments and articles) for a total of 135.6k tokens. Details are reported in table 1. We selected articles from all the main categories of news, in order to have a balance between categories that generate many comments, such as politics, and categories that generate few comments, such as culture and science. The corpus contains the data reported in table 2.

topics	articles	tokens	comments
technology	4	11.6k	266
culture	3	9.3k	215
politics	3	39.2k	876
science	2	2.6k	70
economics	3	30.1k	578
news	6	31.6k	560
gossip	3	4.4k	168
sport	3	6.8k	154
total	**27**	**135.6k**	**2887**

Table 1: Details of the CorEA corpus.

We performed a manual annotation of the agreement/disagreement relations at message level on each child participant/message pair, using the following guidelines:

field	description
Mid	message Id
Pid	participant Id
Pname	participant's nickname
Mtype	article/comment
text	text
timestamp	date/time
category	macro-topic
refer-to-P	Id of parent participant
refer-to-M	Id of parent message
avatar	link to participant's picture
replies-count	replies to the message
likes	like count of the message
agree	agree/disagre labels
Pday-activity	participant's activity score
Pinterests	count of interests of participant
Pviews	participant page views
Pcomments	count of messages of participant
Pshares	count of shares
Pcomments-votes	count of participant's votes
emo-indig	indignation score
emo-disapp	disappointment score
emo-worried	preoccupation score
emo-amused	amusement score
emo-satisfied	satisfaction score

Table 2: Corpus data schema.

1) Read and understand the content of the article and its title.

2) Read the messages of each child pair one by one, sorted by time from the oldest to the newest.

3) For each child pair, check the refer-to-message link finding the corresponding parent pair.

4) read the parent pair, understand the semantics of the relation between child and parent.

5) Annotate with a "NA" label (not applicable) if the child falls under one or both the following conditions: a) **broken refer-to-mesage**: cannot find the parent (e.g. the message is not referred to any other); b) **mixed agreement** (e.g. "I partly agree with you but ..").

6) Judge the agreement/disagreement expressed in the child with respect to the parent. Annotate the child pair with the corresponding label: agree (1), disagree (-1) neutral (0). We did not use any annotation tool. An example of annotation follows:

```
  1:  5 Stars Movement party
returns 2.5 milions Euros to
Italian citizens.
  2:  great!!!.  [agree(2,1)=1]
  3:  http://xyz.com see this :)
ha ha [NA]
  4:  what has to do this link
with the topic?  [agree(4,3)=-1]
  5:  if only every party did
it!..  [agree(5,1)=1]
  6:  would not change anything.
[agree(6,5)=-1]
  7:  what do you mean?
[agree(7,6)=0]
```

We computed the inter-annotator agreement between two Italian native speaker raters, for 50 and 100 instances with 2 (+1, -1) and 3 classes (+1, -1, 0). The "NA" labels were reannotated into the other classes to include all cases into the evaluation. We used Fliess' k (Fleiss et al., 1981), comparable to Cohen's k (used in most of previous work) but generalized over individual raters, like Krippendorf's α (Artstein and Poesio, 2008). Results are reported in table 3. In particular, we noticed that the neutral class is the main source of disagreement between annotators. Figure 2 reports the distribution of the agreement/disagreement labels between annotators and in the corpus. We annotated again the examples using only 2 classes: inter-annotator agreement rose from moderate, in line with (Morgan et al., 2013), to substantial.

type	classes	instances	score
inter	3	50	k=0.6
inter	3	100	k=0.58
inter	2	50	k=0.87
inter	2	100	k=0.93
intra	3	100	k=0.87
intra	2	100	k=0.91

Table 3: Inter- and intra- annotator agreement for the agreement/disagreement relation annotation in CorEA.

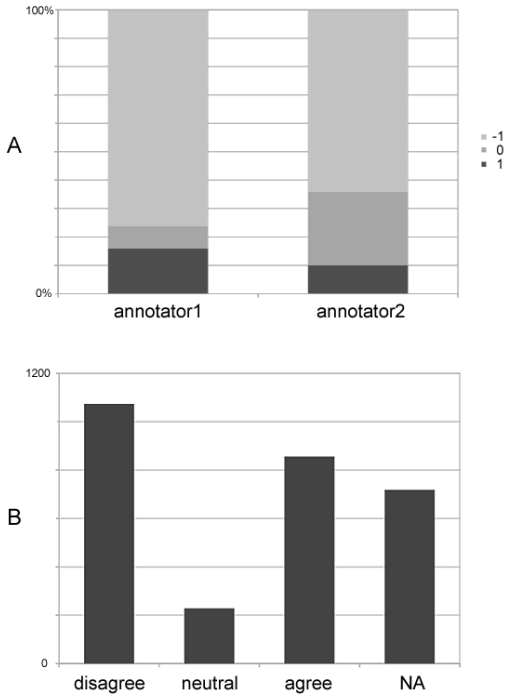


Figure 2: A) Distribution of agreement/disagreement labels between 2 annotators (50 comments, 3 classes) and B) distribution of labels in the corpus.

We labeled twice a set of 100 comments to compute intra-annotator agreement, reported in table 3 as well.

5 Conclusion

We presented the CorEA corpus, a resource that combines agreement/disagreement at message level and emotions at participant level. We are not aware of any other resource of this type for Italian. We found that the best way to annotate agreement/disagreement is with binary classes, filtering out "NA" and neutral cases.

In the future, we would like to annotate CorEA at topic level and develop classifiers for agreement/disagreement. We plan to make available the corpus at the end of the project.

Aknowledgements

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n 610916 SENSEI.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11. Association for Computational Linguistics.

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *LREC*, pages 818–822.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, dec.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. *WASSA 2013*, page 100.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis and opinion mining: the case of irony and senti-tut. *IEEE Intelligent Systems*, page 1.

Erik Cambria, Catherine Havasi, and Amir Hussain. 2012. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS Conference*, pages 202–207.

Jacob Cohen. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press, New York.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Susan C. Herring. 2007. A faceted classification scheme for computer-mediated discourse. *Language@ internet*, 4(1):1–37.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality & quantity*, 38:787–800.

Jonathan T. Morgan, Meghan Oxley, Emily Bender, Liyi Zhu, Varya Gracheva, and Mark Zachry. 2013. Are we there yet?: The development of a corpus annotated for social acts in multilingual online discourse. *Dialogue & Discourse*, 4(2):1–33.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, DTIC Document.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

10.12871/CLICIT2014121

Detecting Attribution Relations in Speech: a Corpus Study

Alessandra Cervone[1], Silvia Pareti[23], Peter Bell[2], Irina Prodanof[1], Tommaso Caselli[4]
Dept. of Humanities, University of Pavia[1], School of Informatics, University of Edinburgh[2], Google Inc.[3], Trento RISE[4]
alessandra.cervone01@universitadipavia.it; s.pareti@sms.ed.ac.uk; peter.bell@ed.ac.uk; irina.prodanof@unipv.it; t.caselli@trentorise.eu

Abstract

English. In this work we present a methodology for the annotation of Attribution Relations (ARs) in speech which we apply to create a pilot corpus of spoken informal dialogues. This represents the first step towards the creation of a resource for the analysis of ARs in speech and the development of automatic extraction systems. Despite its relevance for speech recognition systems and spoken language understanding, the relation holding between quotations and opinions and their source has been studied and extracted only in written corpora, characterized by a formal register (news, literature, scientific articles). The shift to the informal register and to a spoken corpus widens our view of this relation and poses new challenges. Our hypothesis is that the decreased reliability of the linguistic cues found for written corpora in the fragmented structure of speech could be overcome by including prosodic clues in the system. The analysis of SARC confirms the hypothesis showing the crucial role played by the acoustic level in providing the missing lexical clues.

Italiano. *In questo lavoro viene presentata una metodologia di annotazione delle Relazioni di Attribuzione nel parlato utilizzata per creare un corpus pilota di dialoghi parlati informali. Ció rappresenta il primo passo verso la creazione di una risorsa per l'analisi delle ARs nel parlato e lo sviluppo di sistemi di estrazione automatica. Nonostante la sua rilevanza per i sistemi di riconoscimento e comprensione del parlato, la relazione esistente tra le citazioni e le opinioni e la loro fonte è stata studiata ed estratta soltanto in corpora scritti, caratterizzati da un registro formale (articoli di giornale, letteratura, articoli scientifici). Lo studio di un corpus parlato, caratterizzato da un registro informale, amplia la nostra visione di questa relazione e pone nuove sfide. La nostra ipotesi é che la ridotta affidabilitá degli indizi linguistici trovati per lo scritto nella struttura frammentata del parlato potrebbe essere superata includendo indizi prosodici nel sistema. L'analisi di SARC conferma quest'ipotesi mostrando il ruolo cruciale interpretato dal livello acustico nel fornire gli indizi lessicali mancanti.*

1 Introduction

Our everyday conversations are populated by other people's words, thoughts and opinions. Detecting quotations in speech represents the key to "one of the most widespread and fundamental topics of human speech" (Bakhtin, 1981, p. 337).

A system able to automatically extract a quotation and attribute it to its truthful author from speech would be crucial for many applications. Besides Information Extraction systems aimed at processing spoken documents, it could be useful for Speaker Identification systems, (e.g. the strategy of emulating the voice of the reported speaker in quotations could be misunderstood by the system as a change of speaker). Furthermore, attribution extraction could also improve the performance of Dialogue parsing, Named-Entity Recognition and Speech Synthesis tools. On a more basic level, recognizing citations from speech could be useful for sentence boundaries automatic detection systems, where quotations, being sentences embedded in other sentences, could be a source of confusion.

So far, however, attribution extraction systems have been developed only for written corpora.

Extracting the text span corresponding to quotations and opinions and ascribing it to their proper source within a text means to reconstruct the Attribution Relations (ARs, henceforth) holding between three constitutive elements (following Pareti (2012)):

- the Source
- the Cue, i.e. the lexical anchor of the AR (e.g. *say, announce, idea*)
- the Content

(1) This morning [$_{\text{Source}}$ John] [$_{\text{Cue}}$ told] me: [$_{\text{Content}}$ "It's important to support our leader. I trust him."].

In the past few years ARs extraction has attracted growing attention in NLP for its many potential applications (e.g.Information Extraction, Opinion Mining) while remaining an open challenge. Automatically identifying ARs from a text is a complex task, in particular due to the wide range of syntactic structures that the relation can assume and the lack of a dedicated encoding in the language. While the content boundaries of a direct quotation are explicitly marked by quotation markers, opinions and indirect quotations only partially have syntactically, albeit blurred, boundaries as they can span intersententially. The subtask of identifying the presence of an AR could be tackled with more success by exploiting the presence of the cue as a lexical anchor establishing the links to source and content spans. For this reason, cues are the starting point or a fundamental feature of extraction systems (Pareti et al., 2013; Sarmento and Nunes, 2009; Krestel, 2007).

In our previous work (Pareti and Prodanof, 2010; Pareti, 2012), starting from a flexible and comprehensive definition (Pareti and Prodanof, 2010, p. 3566) of AR, we created an annotation scheme which has been used to build the first large annotated resource for attribution, the Penn Attribution Relations Corpus (PARC)[1], a corpus of news articles.

In order to address the issue of detecting ARs in speech, we started from the theoretical and annotation framework from PARC to create a comparable resource. Section 2 explains the issues connected with extracting ARs from speech. Section 3 describes the Speech Attribution Relations Corpus (SARC, henceforth) and its annotation scheme. The analysis of the corpus is presented in Section 4. Section 5 reports an example of how prosodic cues can be crucial to identify ARs in speech. Finally, Section 6 draws on the conclusions and discusses future work.

[1]The corpus adds to and further completes the annotation of attribution in the PDTB (Prasad et al., 2008).

2 The challenge of detecting Attribution Relations in speech

The shift from written to spoken language makes the task of ARs extraction much harder. Current approaches to ARs detection rely heavily on lexical cues and punctuation to identify ARs and in particular the Content span boundaries. In the fragmented structures full of disfluencies typical of speech, however, lexical cues become less reliable, sometimes being completely absent.

On the other hand, punctuation, in most cases crucial in giving the key to the correct interpretation of ARs, is replaced in speech by prosody. While punctuation is a formal symbolic system, prosody is a continuous system which could greatly vary due to language-specific, diatopic, diaphasic and idiosyncratic reasons, thus much harder to process for a tool.

Our working hypothesis focused on the role of prosody in marking the presence and boundaries of quotations in speech. In particular, we considered that it would be possible to find acoustic cues to integrate the linguistic ones in order to improve the task of correctly reconstructing the ARs in a spoken corpus.

Preliminary support for our hypothesis can be found in previous studies which aimed at identifying acoustic correlates of reported speech. However, these approaches, which suggest shift in pitch, intensity and pauses duration as possible prosodic indicators of quotations, offer only fragmented insights on the phenomenon of Attribution. Some of these studies analyze only the variations in pitch (Jansen et al., 2001; Bertrand et al., 2002), others analyze only the ending boundary of quotations (Oliveira and Cunha, 2004) and most of them consider only direct reported speech (Bertrand et al., 2002; Oliveira and Cunha, 2004). Even if the results of these studies are encouraging, the acoustic cues they propose need to be tested and further investigated in a larger project which consider different types of reported speech along with all the prosodic features which could be linked to quotations (pitch, intensity and pauses).

3 Description of the corpus

SARC is composed by four informal telephone conversations between English speakers. The dialogues have a mean duration of 15 minutes where every speaker is recorded on a different track (totally about 2 hours of recordings and 8 speakers). Table 1 shows the main aspects which differentiate SARC from PARC.

	SARC	**PARC**
Register	Informal	Formal
Medium	Oral	Written
Genre	Dialogue	News
Tokens	16k, 2h	1139k
ARs Frequency (ARs per k tokens)	(223/16k) 13.9	(10k/1139k) 9.2

Table 1: Differences between SARC and PARC.

While PARC displays a rather formal English register, typical of the news genre and of the written medium, SARC portrays a radically different one, the coloured, fragmented and highly contextualized register used in informal conversations. The impact of these differences in the type of language presented in our corpus have lead to an adaptation, summarized in Table 2, of the annotation scheme created for PARC (Pareti, 2012).

Attribution Elements	Source	
	Cue	
	Content	*Direct*
		Indirect
		Fading Out
Relation		

Table 2: Annotation scheme for SARC.

All the basic annotation elements (source, cue, content) from PARC have been kept in order for the results to be comparable. The content has been further subdivided into 3 types, of which the last one, the Fading out, never used previously in attribution extraction schemes, is a category introduced by Bolden (2004) to identify those cases typical of dialogues in which the ending boundary of a quotation is left purposely ambiguous by the speaker. We adopted PARC (Pareti, 2012; Pareti, forthcoming) annotation guidelines, with the following modifications: in PARC cases like *"I say"* or *"I think"* are considered quotations of the author himself, while in our annotation, where every sentence is considered a personal opinion of the speaker, they are not (see Klewitz and Couper-Kuhlen(1999, p. 4)). The annotation has been performed with MMAX2 (Müller and Strube, 2006) by one annotator (who was also trained on PARC scheme and guidelines). For further details about the construction and annotation process of SARC we refer you to Cervone (2014).

4 Analysis of SARC

The analysis of SARC (see the chart in Figure 1) shows how in about 10% of the cases in our corpus the cue is completely missing, while in PARC such cases were rare (only in 4% of the cases was the source missing). Therefore, at least 1 out of 10 ARs in SARC is impossible to identify without the aid of prosodic cues. Furthermore, due to

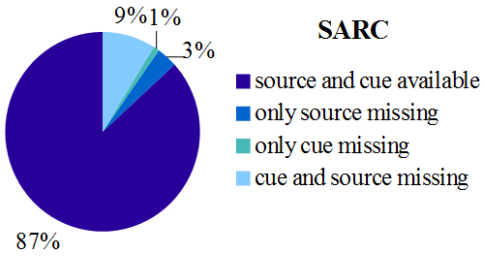


Figure 1: Cases of missing AR elements in SARC.

the absence of punctuation all the boundary clues found for written corpora are missing. We cannot rely any more on quotation marks, without punctuation we have no clue about the sentence structure (crucial for indirect quotes) and due to disfluencies the syntactic structure is less reliable and complete (some ARs are syntactically encoded). This means that even that 87% of cases in which in SARC no element of the AR is missing are more problematic than almost 50% of PARC cases (3,262 direct, 1,549 mixed) where punctuation defines the content. If we rely only on the lexical level for detecting ARs in speech, we have no assurance that the boundaries of the content span we identified out of many possible interpretations are the correct ones.

5 Prosodic cues of Attribution

The analysis of SARC has shown how much the shift to a spoken corpus can make the task of detecting ARs harder, displaying the need to find other cues to improve the performance of an attribution extraction system for speech. In Section 2

we indicated prosody as a possible source for cues of attribution. This section details how prosodic information can be used to identify ARs in speech.

Example 2 presents an utterance transcribed from SARC where ARs could be present. Considering only the lexical level, however, the sentence could be subject to many possible interpretations (e.g. there are at least 3 different possible lexical cues (represented by verbs of saying)).

(2) I *said* to him when you left do you remember I *told* you I *said* to him don't forget Dave if you ever get in trouble give us a call you never know your luck.

To choose the correct interpretation, we employed the judgement of a human annotator who listened to the recording and then we conducted an analysis of the acoustic features suggested in previous studies using Praat (Boersma and Weenink, 2014).

As shown in Figure 2, the waveform of the reported example is divided into two phases by a pause (0.7 seconds) (between the dotted lines) which occurs between the second *I said to him* and *don't forget*. The presence of the pause seems to mark the beginning of a new prosodic unit, which, directly following the lexical cue *said*, could be reported speech. The two graphs in Figure 3

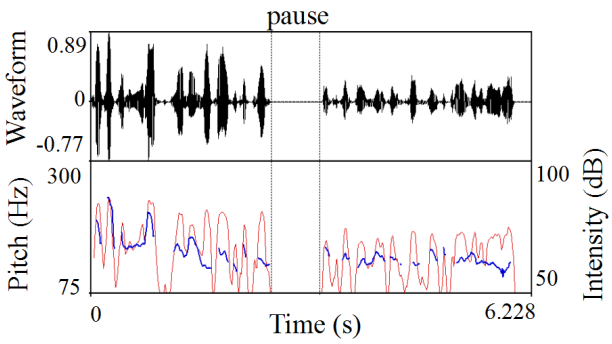


Figure 2: Rawdata of Ex(2).

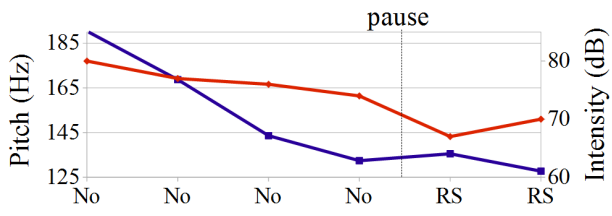


Figure 3: Means of Pitch and Intensity in Ex(2).

shows the variation in the means of respectively pitch (Hz)(blue) and intensity (dB)(red) along the timespan of the excerpt, elaborated from the rawdata in Figure 2. On the x-axis is displayed the presence (RS) or not (No) of reported speech according to our intepretation of the pause (dotted line) marking. The means of pitch and intensity show a similar tendency: a decrease of the mean with a stabilisation to a lower level after the pause. All the acoustic features seem therefore to suggest a difference in the prosodic marking between the first time span (No) and the second one (RS). This interpretation matches the one given by the human annotator. Thanks to the integration of the lexical cues with the acoustic analysis of the three prosodic factors combined it was possible to achieve the correct identification of the quotation (*don't forget Dave if you ever get in trouble give us a call you never know your luck*) out of at least three possible interpretations (considering only the verbs of saying). The full corpus contains many similar examples which demonstrate the importance of accessing to the acoustics for disambiguation of ARs in speech and how the judgements of human annotators can be analyzed by looking at the prosodic features.

6 Conclusions and future work

The analysis of SARC, the first resource developed to study ARs in speech, has helped to highlight a major problem of detecting attribution in a spoken corpus: the decreased reliability of the lexical cues crucial in previous approaches (completely useless in at least 10% of the cases) and the consequential need to find reliable prosodic clues to integrate them. The example provided in Section 5 has showed how the integration of the acoustic cues could be useful to improve the accuracy of attribution detection in speech.

As a future project we are going to perform a large acoustic analysis of the ARs found in SARC, in order to see if some reliable prosodic cues can in fact be found and used in order to develop a software able to extract attribution from speech.

Acknowledgments

We wish to thank the NLP group at the School of Informatics of the University of Edinburgh, where this project has been developed thanks to the Erasmus Placement Scholarship, and especially Bonnie Webber, Bert Remijsen and Catherine Lai.

References

Mikhail M. Bakhtin. 1981. *The dialogic imagination: Four essays*. University of Texas Press.

Claude Barras and Edouard Geoffrois and Zhibiao Wu and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. *First International Conference on Language Resources and Evaluation (LREC)*. 1373–1376.

Sabine Bergler, Monia Doandes, Christine Gerard and René Witte. 2004. Attributions. *Proceedings of the Eight International Conference on Language Resources and Exploring Attitude and Affect in Text: Theories and Applications, Technical Report SS-04-07*. 16–19.

Roxane Bertrand, Robert Espesser and others. 2002. Voice diversity in conversation: a case study. *Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence, France.

Paul Boersma and David Weenink. 2014. Praat: Doing Phonetics by Computer [Computer Program]. Version 5.3.63. Available online at: http://www.praat.org/.

Galina Bolden. 2004. The quote and beyond: defining boundaries of reported speech in conversational Russian. *Journal of pragmatics*. Elsevier. 36(6): 1071–1118.

Alessandra Cervone. 2014. Attribution Relations Extraction in Speech: A Lexical-Prosodic Approach. Thesis of the Master in Theoretical and Applied Linguistics. University of Pavia, Pavia.

David K. Elson and Kathleen McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. *AAAI*.

Edouard Geoffrois and Claude Barras and Steve Bird and Zhibiao Wu. 2000. Transcribing with Annotation Graphs. *Second International Conference on Language Resources and Evaluation (LREC)*. 1517–1521.

Wouter Jansen, Michelle L Gregory, Jason M Brenier. 2001. Prosodic correlates of directly reported speech: Evidence from conversational speech. *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*. Molly Pitcher Inn, Red Bank, NJ, USA.

Gabriele Klewitz and Elizabeth Couper-Kuhlen. 1999. *Quote-unquote? The role of prosody in the contextualization of reported speech sequences*. Universität Konstanz, Philosophische Fakultät, Fachgruppe Sprachwissenschaft.

Ralf Krestel. 2007. Automatic analysis and reasoning on reported speech in newspaper articles. Tesis de Magister Universität Karlsruhe. Karlsruhe. Available at http://www. semanticsoftware. info/system/files/believer. pdf.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Proceedings of the Eight International Conference on Language Resources and Corpus Technology and Language Pedagogy: NewResources, New Tools, New Methods*. 3: 197–214.

Miguel Oliveira, Jr. and Dòris A. C. Cunha. 2004. Prosody as marker of direct reported speech boundary. *Speech Prosody 2004, International Conference*.

Silvia Pareti and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. *Proceedings of LREC10*.

Silvia Pareti. 2012. A Database of Attribution Relations. *Proceedings of the Eight International Conference on Language Resources and Evaluation*. 3213–3217.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irene Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. *Proceedings of the 2013 Conference in Empirical Methods in Natural Language Processing*. 989–999.

Silvia Pareti. Forthcoming. Attribution: A Computational Approach. PhD Thesis, School of Informatics, the University of Edinburgh.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*. Citeseer. 47(2): 43–64.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC08*.

Luís Sarmento and Sérgio Nunes. 2009. Automatic extraction of quotes and topics from news feeds. *DSIE'09-4th Doctoral Symposium on Informatics Engineering*.

10.12871/CLICIT2014122

Adattamento al Progetto dei Modelli di Traduzione Automatica nella Traduzione Assistita

Mauro Cettolo **Nicola Bertoldi** **Marcello Federico**
FBK - Fondazione Bruno Kessler
Trento, Italy
cognome@fbk.eu

Abstract

Italiano. L'integrazione della traduzione automatica nei sistemi di traduzione assistita è una sfida sia per la ricerca accademica sia per quella industriale. Infatti, i traduttori professionisti percepiscono come cruciale l'abilità dei sistemi automatici di adattarsi al loro stile e alle loro correzioni. In questo articolo proponiamo uno schema di adattamento dei sistemi di traduzione automatica ad uno specifico documento sulla base di una limitata quantità di testo, corretto manualmente, pari a quella prodotta giornalmente da un singolo traduttore.

English. *The effective integration of MT technology into computer-assisted translation tools is a challenging topic both for academic research and the translation industry. Particularly, professional translators feel crucial the ability of MT systems to adapt to their feedback. In this paper, we propose an adaptation scheme to tune a statistical MT system to a translation project using small amounts of post-edited texts, like those generated by a single user in even just one day of work.*

1 Introduzione

Nonostante i significativi e continui progressi, la traduzione automatica (TA) non è ancora in grado di generare testi adatti alla pubblicazione senza l'intervento umano. D'altra parte, molti studi hanno confermato che nell'ambito della traduzione assistita la correzione di testi tradotti automaticamente permette un incremento della produttività dei traduttori professionisti (si veda il paragrafo 2). Questa applicazione della TA è tanto più efficace quanto maggiore è l'integrazione del sistema di traduzione automatico nell'intero processo di traduzione, che può essere ottenuta specializzando il sistema sia al particolare testo da tradurre sia alle caratteristiche dello specifico traduttore e alle sue correzioni. Nell'industria della traduzione, lo scenario tipico è quello di uno o più traduttori che lavorano per alcuni giorni su un dato *progetto di traduzione*, ovvero su un insieme di documenti omogenei. Dopo un giorno di lavoro, le informazioni contenute nei testi appena tradotti e le correzioni apportate dai traduttori possono essere immesse nel sistema automatico con l'obiettivo di migliorare la qualità delle traduzioni automatiche proposte il giorno successivo. Chiameremo questo processo *adattamento al progetto*. L'adattamento al progetto può essere ripetuto quotidianamente fino al termine del lavoro, in modo da sfruttare al meglio tutte le informazioni che implicitamente i traduttori mettono a disposizione del sistema.

Questo articolo presenta uno dei risultati del progetto europeo MateCat,[1] nel cui ambito abbiamo sviluppato un sistema per la traduzione assistita basato sul Web integrante un modulo di TA che si auto-adatta allo specifico progetto. Gli esperimenti di validazione che andremo ad illustrare sono stati effettuati su quattro coppie di lingue, dall'inglese all'italiano (IT), al francese (FR), allo spagnolo (ES) e al tedesco (DE), e in due domini, tecnologie dell'informazione e della comunicazione (TIC) e legale (LGL).

Idealmente, i metodi di adattamento proposti dovrebbero essere valutati misurando il guadagno in termini di produttività su progetti di traduzione reali. Pertanto, per quanto possibile, abbiamo eseguito delle *valutazioni sul campo* in cui dei traduttori professionisti hanno corretto le traduzioni ipotizzate da sistemi automatici, adattati e non. L'adattamento è stato eseguito sulla base di una porzione del progetto tradotto durante una fase preliminare, in cui allo stesso traduttore è stato chiesto di correggere le traduzioni fornite da un sistema di partenza non adattato.

[1] http://www.matecat.com

Siccome le valutazioni sul campo sono estremamente costose, esse non possono essere eseguite frequentemente per confrontare tutte le possibili varianti degli algoritmi e dei processi. Abbiamo quindi condotto anche delle *valutazioni di laboratorio*, in cui le correzioni dei traduttori erano simulate dalle traduzioni di riferimento.

Complessivamente, nel dominio legale i miglioramenti osservati in laboratorio hanno anticipato quelli misurati sul campo. Al contrario, i risultati nel dominio TIC sono stati controversi a causa della poca corrispondenza tra i testi usati per l'adattamento e quelli effettivamente tradotti durante la sperimentazione.

2 Lavori correlati

L'idea che la TA possa migliorare la produttività dei traduttori si è consolidata negli anni grazie ai miglioramenti della qualità della TA statistica e ai tanti lavori che hanno sperimentalmente valutato il suo impatto (Guerberof, 2009; Plitt and Masselot, 2010; Federico et al., 2012; Läubli et al., 2013; Green et al., 2013).

Dal punto di vista dei metodi, il nostro lavoro si occupa di adattamento in generale, e di quello incrementale più nello specifico. Senza entrare nel dettaglio per mancanza di spazio, vogliamo qui segnalare il lavoro di Bertoldi et al. (2012), dove i modelli di traduzione vengono adattati incrementalmente su pacchetti di dati nuovi man mano che questi sono disponibili, e i seguenti lavori in qualche modo a quello correlati: (Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Bisazza et al., 2011; Liu et al., 2012; Bach et al., 2009; Niehues and Waibel, 2012; Hasler et al., 2012).

Come vedremo, noi eseguiamo anche una selezione di dati, problema ampiamente investigato dalla nostra comunità scientifica, si veda ad esempio (Yasuda et al., 2008; Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011). Quella che noi applichiamo è una tecnica piuttosto convenzionale (Moore and Lewis, 2010), ma in condizioni inusuali, dove la selezione viene effettuata su un corpus di dati nel dominio di interesse, e quindi non proprio generico, e a partire da una quantità estremamente ridotta di dati specifici.

3 Metodi di adattamento

Selezione di dati - Come accennato sopra, quello della selezione di dati è un problema ampiamente studiato in letteratura. In effetti spesso ci troviamo a dover addestrare dei modelli con dati provenienti da sorgenti eterogenee in termini di dimensione, qualità, dominio, ecc. L'obiettivo di questa tecnica è di selezionare un sottoinsieme dei dati a disposizione che sia pertinente rispetto ad un certo testo, nel nostro caso quello di uno specifico progetto di traduzione. Noi abbiamo implementato la tecnica proposta in (Moore and Lewis, 2010) e l'abbiamo resa disponibile attraverso il pacchetto IRSTLM (Federico et al., 2008). Per applicare l'algoritmo, si parte da un *corpus specifico*, che si suppone rappresentare bene il documento da tradurre, e da un *corpus generico*, molto più grande e in cui si suppone di poter trovare del materiale pertinente al documento da tradurre. Sfruttando due modelli del linguaggio (ML), uno specifico e uno generico, a ogni frase generica viene assegnato un punteggio tanto più alto quanto più essa è specifica e lontana dalla "media" di quelle generiche. Effettuato l'ordinamento su questo punteggio, viene infine selezionata la quantità di frasi generiche che ottimizzano la perplessità di un testo di controllo.
Fill-up dei modelli di traduzione - La selezione di dati è efficace per addestrare modelli di traduzione sui testi più rilevanti per uno specifico progetto. D'altra parte, scartare una porzione dei dati disponibili significa correre il rischio di perdere delle informazioni comunque utili; per evitarlo, si può ricorrere alla tecnica *fill-up*, proposta da Nakov (2008) e raffinata da Bisazza et al. (2011). Essa fonde i modelli di traduzione generico e specifico, unendo gli insiemi delle loro voci e mantenendo le probabilità del modello specifico per le voci in comune.
Mistura di ML - Per l'adattamento dei ML siamo ricorsi alla mistura dei modelli proposta da Kneser e Steinbiss (1993), che consiste nella combinazione convessa di due o più ML; i pesi della mistura sono stimati per mezzo di una validazione incrociata sui dati di addestramento con la quale si simula l'occorrenza di n-grammi nuovi. Il metodo è disponibile nel già citato pacchetto IRSTLM.

4 Dati per gli esperimenti

Coi domini e le coppie di lingue menzionate nell'introduzione abbiamo definito sei configurazioni sperimentali. Qui di seguito forniamo dettagli sui dati di addestramento e di valutazione per ciascuna di esse.
Dati di addestramento - Per l'addestramnento abbiamo usato sia dati paralleli sia memorie di traduzione. Per il dominio TIC, sono stati sfruttati i manuali software del corpus OPUS (Tiedemann, 2012) e una memoria di traduzione proprietaria, fornitaci dal partner industriale di MateCat.

Per il dominio LGL abbiamo acquisito il corpus JRC-Acquis (Steinberger et al., 2006), che include

dominio	coppia	corpus	seg	parole	
				sorgente	obiettivo
TIC	IT	generico	5.4 M	57.2M	59.9M
		selezione	0.36M	3.8M	4.0M
		calibrazione	2,156	26,080	28,137
	FR	generico	2.3 M	35.4M	40.1M
		selezione	0.53M	8.6M	9.5M
		calibrazione	4,755	26,747	30,100
LGL	IT	generico	2.7 M	61.4M	63.2M
		selezione	0.18M	5.4M	5.4M
		calibrazione	181	5,967	6,510
	FR	generico	2.8 M	65.7M	71.1M
		selezione	0.18M	5.5M	5.8M
		calibrazione	600	17,737	19,613
	ES	generico	2.3 M	56.1M	62.0M
		selezione	0.18M	5.6M	6.1M
		calibrazione	700	32,271	36,748
	DE	generico	2.5 M	45.3M	41.8.0M
		selezione	0.18M	5.2M	4.7M
		calibrazione	133	3,082	3,125

Tabella 1: Statistiche sui dati paralleli usati per la preparazione dei sistemi di TA: numero di segmenti e di parole. Il simbolo M sta per 10^6.

la legislazione della UE in 22 lingue.

La tabella 1 riporta alcune statistiche dei testi paralleli impiegati per l'addestramento dei modelli di traduzione e di riordinamento; i ML sono stati stimati sul testo obiettivo. Per ciascuna configurazione sperimentale, la voce *generico* si riferisce alla totalità dei dati a disposizione, mentre *selezione* indica i dati selezionati pertinenti al progetto in esame. I dati per la *calibrazione* sono aggiuntivi e utilizzati per il bilanciamento ottimale dei vari modelli che definiscono il motore di TA.

Dati di valutazione - Per il dominio TIC i dati sono stati forniti dal partner industriale di MateCat che ha selezionato dal suo archivio un progetto di traduzione reale in cui dei documenti in inglese erano già stati tradotti in italiano e francese senza l'ausilio del sistema MateCat. Per il dominio LGL, abbiamo selezionato un documento della legislazione europea[2] per il quale erano disponibili le traduzioni nelle quattro lingue di nostro interesse. Nella tabella 2 sono raccolte le statistiche dei testi da tradurre nella fase preliminare e in quella di validazione vera e propria del sistema adattato.

5 Valutazioni di laboratorio

Sistemi di TA - I sistemi di TA sviluppati sono statistici e costruiti col pacchetto Moses (Koehn et al., 2007). I modelli di traduzione e di riordinamento sono stati addestrati sui dati bilingue della tabella 1 nei modi descritti in seguito. Per modellare il linguaggio, le distribuzioni dei 5-grammi

[2]2013/488/EU: "Council Decision of 23 September 2013 on the security rules for protecting EU classified information".

dominio	coppia	fase	seg	parole	
				sorgente	obiettivo
TIC	IT	preliminare	342	3,435	3,583
		validazione	1,614	14,388	14,837
	FR	preliminare	342	3,435	3,902
		validazione	1,614	14,388	15,860
LGL	IT	preliminare	133	3,082	3,346
		validazione	472	10,822	11,508
	FR	preliminare	134	3,084	3,695
		validazione	472	10,822	12,810
	ES	preliminare	131	3,007	3,574
		validazione	472	10,822	12,699
	DE	preliminare	133	3,082	3,125
		validazione	472	10,822	10,963

Tabella 2: Statistiche sui dati di valutazione.

sono state stimate sul testo obiettivo e applicandovi la tecnica di smoothing Kneser-Ney (Chen and Goodman, 1999). La calibrazione dei sistemi, ovvero la stima dei pesi dell'interpolazione dei vari modelli, è stata effettuata su opportuni testi aggiuntivi (voci *calibrazione* in tabella 1).

Per ciascuna delle sei configurazioni, sono stati valutati due sistemi TA, uno di riferimento (RIF) e uno adattato (ADA). I modelli del RIF sono stati addestrati sui dati corrispondenti alle voci *generico* della tabella 1. Abbiamo quindi selezionato una porzione dei dati generici usando come corpus specifico il testo bilingue della fase preliminare e la parte sorgente del testo di validazione, ottenendo i testi *selezione* della tabella 1. Sulla concatenazione dei testi della fase preliminare e di quelli selezionati abbiamo successivamente addestrato i modelli specifici che sono stati combinati coi modelli generici per mezzo del fill-up (modelli di traduzione/riordinamento) e della mistura (ML) al fine di costruire il sistema ADA.

Risultati - La tabella 3 quantifica la qualità della TA fornita dai sistemi RIF e ADA in termini di Bleu, Ter e Gtm, misurati sui documenti di validazione rispetto alle traduzioni manuali.

coppia	TA	dominio TIC			dominio LGL		
		Bleu	Ter	Gtm	Bleu	Ter	Gtm
IT	RIF	55.3	29.2	77.8	31.0	53.1	61.8
	ADA	57.5	26.3	78.6	35.0	49.1	64.6
FR	RIF	41.3	38.3	69.5	33.9	52.2	63.0
	ADA	41.4	37.9	69.9	36.4	49.1	65.1
ES	RIF	–	–	–	35.5	50.7	65.7
	ADA	–	–	–	36.4	50.2	65.6
DE	RIF	–	–	–	18.3	68.4	50.5
	ADA	–	–	–	19.7	66.6	52.3

Tabella 3: Prestazioni TA sui testi di validazione .

Nel dominio LGL il miglioramento fornito dal processo di adattamento è rilevante. Ad esempio, il Bleu migliora del 12.9% (da 31.0 a 35.0) nella traduzione in italiano, del 7.4% (da 33.9 a 36.4)

verso il francese, del 2.5% (da 35.5 a 36.4) verso lo spagnolo e del 7.7% (da 18.3 a 19.7) verso il tedesco.

Al contrario, nel TIC si osserva un certo miglioramento solo per l'italiano (4%, da 55.3 a 57.5), mentre è nullo per il francese. L'analisi riportata in (Bertoldi et al., 2013) mostra che qui il problema è originato dal fatto che i testi tradotti nella fase preliminare, e quindi usati per la selezione, sono poco rappresentativi del documento da tradurre nella fase di validazione.

6 Valutazioni sul campo

In questo paragrafo relazioniamo sugli esperimenti effettuati per valutare l'impatto dell'adattamento al progetto sulla produttività di traduttori professionisti. La valutazione sul campo ha riguardato la traduzione dall'inglese all'italiano di documenti nei due domini TIC e LGL.

Protocollo - La valutazione sul campo è stata eseguita con il sistema di ausilio alla traduzione sviluppato nell'ambito del progetto MateCat che integra i sistemi di TA auto-adattantisi al progetto, come descritto in questo articolo. L'esperimento è stato organizzato su due giorni ed ha coinvolto quattro traduttori per ciascun dominio. Durante il primo giorno – la fase preliminare – per la traduzione della prima parte del progetto i suggerimenti di TA venivano forniti dal sistema RIF; nel secondo giorno – la fase di validazione –, durante il quale è stata tradotta la seconda parte del progetto, i suggerimenti di TA provenivano dal sistema ADA. L'impatto dello schema di adattamento proposto in questo articolo è stato misurato confrontando la produttività dello stesso traduttore nel primo e nel secondo giorno, misurata in termini di time-to-edit (TTE)[3] e post-editing effort (PEE).[3]

Risultati - I risultati sono raccolti nella tabella 4. Per due traduttori su quattro nel dominio TIC (t1 e t4) e per tre su quattro nel LGL (t2-t4) migliorano significativamente entrambe le misure. La maggior parte delle riduzioni del TTE (cinque su otto) sono statisticamente significative (p-value$<$0.05), mentre lo stesso accade solo per due delle variazioni del PEE. Guardando alle medie, nel dominio TIC si registra un guadagno dell'11.2% del TTE e del 6.5% del PEE, mentre nel LGL i miglioramenti sono rispettivamente del 22.2% e del 10.7%. Infine, la buona correlazione osservata tra PEE e TTE nelle diverse condizioni sperimentate mostra come sia verosimile che i traduttori abbiano tratto beneficio dai suggerimenti provenienti dal sistema di TA adattato, dato che il PEE è migliorato in sette casi su otto.

[3]In breve, il TTE è il tempo medio (in secondi) di traduzione per parola, il PEE la percentuale di parole che sono state corrette.

mtrc	dmn	usr	prlmnr	vldzn	p-value	Δ
TTE	TIC	t1	4.70	3.36	0.001	28.51%
		t2	2.26	2.47	0.220	-9.29%
		t3	3.17	3.11	0.450	1.89%
		t4	4.77	3.64	0.006	23.69%
	LGL	t1	5.20	5.63	0.222	-8.27%
		t2	5.42	3.92	0.002	27.68%
		t3	5.86	4.32	0.000	26.28%
		t4	6.60	3.73	0.000	43.48%
PEE	TIC	t1	34.27	30.99	0.060	9.57%
		t2	38.50	39.52	0.330	-2.65%
		t3	32.53	30.17	0.133	7.25%
		t4	32.22	28.44	0.040	11.73%
	LGL	t1	26.47	24.57	0.212	7.18%
		t2	29.11	26.25	0.140	9.82%
		t3	35.65	34.11	0.247	4.32%
		t4	22.72	18.07	0.011	20.47%

Tabella 4: TTE e PEE di ciascun traduttore nelle due sessioni, la preliminare (*prlmnr*) e di validazione (*vldzn*). Sono riportate anche la differenza dei valori tra le due sessioni e la sua significatività statistica in termini di p-value, calcolato tramite la versione randomizzata del test di permutazione (Noreen, 1989).

7 Conclusioni

Un argomento di ricerca particolarmente attuale per l'industria della traduzione assistita è come dotare i sistemi di traduzione automatica della capacità di auto-adattamento. In questo lavoro abbiamo presentato uno schema di auto-adattamento ed i risultati della sua validazione non solo in esperimenti di laboratorio ma anche sul campo, col coinvogimento di traduttori professionisti, grazie alla collaborazione col partner industriale di MateCat.

I risultati sperimentali hanno confermato l'efficacia della nostra proposta, essendosi ottenuti guadagni di produttività fino al 43%. Tuttavia, il metodo funziona solo se i testi utilizzati come base per la selezione di dati specifici su cui eseguire l'adattamento è rappresentativo del documento che si vuol far tradurre. Infatti, laddove tale condizione non fosse verificata, com'era nei nostri esperimenti inglese-francese/TIC, i modelli adattati possono risultare incapaci di migliorare quelli di partenza; ad ogni modo anche in queste condizioni critiche non abbiamo osservato alcun deterioramento delle prestazioni, a dimostrazione del comportamento conservativo del nostro schema.

Ringraziamenti

Questo lavoro è stato possibile grazie al progetto MateCat, finanziato dalla Commissione Europea nell'ambito del Settimo programma quadro.

References

A. Axelrod, X. He, and J. Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *EMNLP*, pp. 355–362, Edinburgh, UK.

N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A. W. Black. 2009. Incremental Adaptation of Speech-to-Speech Translation. In *NAACL HLT (Short Papers)*, pp. 149–152, Boulder, US-CO.

N. Bertoldi, M. Cettolo, M. Federico, and C. Buck. 2012. Evaluating the Learning Curve of Domain Adaptive Statistical MachineTranslation Systems. In *WMT*, pp 433–441, Montréal, Canada.

N. Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Federico, and H. Schwenk. 2013. D5.4: Second report on lab and field test. Deliverable, MateCat project. http://www.matecat.com/wp-content/uploads/2014/06/D5.4_Second-Report-on-Lab-and-Field-Test_v2.pdf.

A. Bisazza, N. Ruiz, and M. Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *IWSLT*, pp. 136–143, San Francisco, US-CA.

S. F. Chen and J. Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech and Language*, 4(13):359–393.

M. Federico, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Interspeech*, pp. 1618–1621, Melbourne, Australia.

M. Federico, A. Cattelan, and M. Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *AMTA*, San Diego, US-CA.

G. Foster and R. Kuhn. 2007. Mixture-model Adaptation for SMT. In *WMT*, pp. 128–135, Prague, Czech Republic.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *EMNLP*, pp. 451–459, Cambridge, US-MA.

S. Green, J. Heer, and C. D Manning. 2013. The efficacy of human post-editing for language translation. In *SIGCHI Conference on Human Factors in Computing Systems*, pp. 439–448, Paris, France.

A. Guerberof. 2009. Productivity and quality in MT post-editing. In *MT Summit - Beyond Translation Memories: New Tools for Translators Workshop*.

E. Hasler, B. Haddow, and P. Koehn. 2012. Sparse lexicalised features and topic adaptation for SMT. In *IWSLT*, pp. 268–275, Hong-Kong (China).

R. Kneser and V. Steinbiss. 1993. On the dynamic adaptation of stochastic language models. In *ICASSP*, volume II, pp. 586–588, Minneapolis, US-MN.

P. Koehn and J. Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *WMT*, pp. 224–227, Prague, Czech Republic.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL: Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic.

S. Läubli, M. Fishel, G. Massey, M. Ehrensberger-Dow, and M. Volk. 2013. Assessing Post-Editing Efficiency in a Realistic Translation Environment. In *MT Summit Workshop on Post-editing Technology and Practice*, pp 83–91, Nice, France.

L. Liu, H. Cao, T. Watanabe, T. Zhao, M. Yu, and C. Zhu. 2012. Locally Training the Log-Linear Model for SMT. In *EMNLP-CoNLL*, pp. 402–411, Jeju Island, Korea.

S. Matsoukas, A.-V. I. Rosti, and B. Zhang. 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *EMNLP*, pp. 708–717, Singapore.

R. C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pp. 220–224.

P. Nakov. 2008. Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. In *WMT*, pp. 147–150, Columbus, US-OH.

J. Niehues and A. Waibel. 2012. Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT. In *AMTA*, San Diego, US-CA.

E. W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley Interscience.

M. Plitt and F. Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufi, and D. Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *LREC*, pp. 2142–2147, Genoa, Italy.

J. Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *LREC*, Istanbul, Turkey.

K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *IJCNLP*, Hyderabad, India.

10.12871/CLICIT2014123

The New Basic Vocabulary of Italian as a Linguistic Resource

Isabella Chiari
Dip. di Scienze documentarie, linguistico-filologiche e geografiche dell'Università di Roma "La Sapienza"
pl.le Aldo Moro, 5, 00185 Roma
isabella.chiari@uniroma1.it

Tullio De Mauro
Dip. di Scienze documentarie, linguistico-filologiche e geografiche dell'Università di Roma "La Sapienza"
pl.le Aldo Moro, 5, 00185 Roma
tullio.demauro@uniroma1.it

Abstract

English. The New Basic Vocabulary of Italian (NVdB) is a reference linguistic resource for contemporary Italian describing most used and understood words of the language. The paper offers an overview of the objectives of the work, its main features and most relevant linguistic and computational applications.

Italiano. Il Nuovo Vocabolario di Base della lingua italiana (NVdB) costituisce una risorsa linguistica di riferimento dell'italiano contemporaneo che descrive le parole più usate e conosciute dalla maggioranza della popolazione italiana con una istruzione media inferiore. Il contributo descrive le ragioni dell'impianto del NVdB, le caratteristiche della risorsa e le principali applicazioni linguistiche e computazionali.

1 Introduction

Core dictionaries are precious resources that represent the most widely known (in production and reception) lexemes of a language. Among the most significant features characterizing basic vocabulary of a language is the high textual coverage of a small number of lexemes (ranging from 2,000 to 5,000 top ranking words in frequency lists), their large polysemy, their relationship to the oldest lexical heritage of a language, their relevance in fist and second language learning and teaching and as reference tools for lexical analysis.

Many recent corpus based works have been produced to provide up-to-date core dictionaries to many European languages (e.g. the Routledge frequency dictionary series). Italian language has a number reference frequency lists all of which are related to corpora and collections of texts dating 1994 or earlier (among the most relevant Bortolini et al., 1971; Juilland and Traversa, 1973; De Mauro et al., 1993; Bertinetto et al. 2005).

The Basic Vocabulary of Italian (VdB, De Mauro, 1980) first appeared as an annex to *Guida all'uso delle parole* and has been subsequently included in all lexicographic works directed by Tullio De Mauro, with some minor changes.

VdB has benefited from a combination of statistical criteria for the selection of lemmas (both grammatical and content words) mainly based on a frequency list of written Italian, LIF (Bortolini et al., 1972) and later on a frequency list of spoken Italian, LIP (De Mauro et al., 1993) – and independent evaluations further submitted to experimentation on primary school pupils.

The last version of VdB was published in 2007 in an additional tome of GRADIT (De Mauro, 1999) and counts about 6,700 lemmas, organised in three vocabulary ranges.

Fundamental vocabulary (FO) includes the highest frequency words that cover about 90% of all written and spoken text occurrences [*appartamento* 'apartment', *commercio* 'commerce', *cosa* 'thing', *fiore* 'flower', *improvviso* 'sudden', *incontro* 'meeting', *malato* 'ill', *odiare* 'to hate'], while high usage vocabulary (AU) covers about 6% of the subsequent high frequency words [*acciaio* 'steel', *concerto* 'concert', *fase* 'phase', *formica* 'ant', *inaugurazione* 'inauguration', *indovinare* 'to guess', *parroco* 'parish priest', *pettinare* 'to comb']. On the contrary high availability (AD) vocabulary is not based on textual statistical resources but is derived from a

psycholinguistic insight experimentally verified, and is to be intended in the tradition of the *vocabulaire de haute disponibilité*, first introduced in the *Français fondamentale* project (Michéa, 1953; Gougenheim, 1964). VdB thus integrates high frequency vocabulary ranges with the so-called high availability vocabulary (*haute disponibilité*) and thus provides a full picture of not only written and spoken usages, but also purely mental usages of word (commonly regarding words having a specific relationship with the concreteness of ordinary life) [*abbaiare* to 'bark', *ago* 'needle', *forchetta 'fork', mancino* 'left-handed', *pala* 'shovel', *pescatore* 'fisherman'].

From the first edition of VdB many things have changed in Italian society and language: Italian language was then used only by 50% of the population. Today Italian is used by 95% of the population. Many things have changed in the conditions of use of the language for the speakers and the relationship between Italian language and dialects have been deeply transformed.

The renovated version of VdB, NVdB (Chiari and De Mauro, in press), will be presented and previewed in this paper. NVdB is a linguistic resource designed to meet three different purposes: a linguistic one, to be intended in both a theoretical and a descriptive sense, an educational-linguistic one and a regulative one, for the development of guidelines in public communication.

The educational objective is focused on providing a resource to develop tools for language teaching and learning, both for first and second language learners. The descriptive lexicological objective is providing a lexical resource that can be used as a reference in evaluating behaviour of lexemes belonging to different text typologies, taking into account the behaviour of different lexemes both from an empirical-corpus based approach and an experimental (intuition based) approach and enable the description of linguistic changes that affected most commonly known words in Italian from the Fifties up to today. The descriptive objective is tightly connected to the possible computational applications of the resource in tools able to process general language and take into account its peculiar behaviour. The regulative objective regards the use of VdB as a reference for the editing of administrative texts, and in general, for easy reading texts.

2 Overview of the resource

NVdB is characterised by a number of methodological choices that make it a unique tool both for educational, descriptive and computational linguistics. A major feature of NVdB is its stratification in vocabulary ranges. While other lexicographic works contain only a plain list of frequent words, NVdB is organised internally and reveals different statistical and non statistical properties of the elements of the lexicon. The stratification of NVdB, though complex methodologically, allows isolating the different textual behaviour of lexemes in context, their coverage power and dispersion, and also taking into account separately known words that rarely appear in text corpora but that are generally available to native speakers and that necessitate experimental methods to be acquired.

A new experimentation of high availability words completes and redefines the role of frequency and usage introducing a receptive and psycholinguistic perspective in the third layer of the core dictionary.

In order to facilitate applicative uses of NVDB all data will be distributed both in paper and in an open source electronic versions in multiple formats.

2.1 The corpus and linguistic processing

The first two layers of NVdB (FO, AU) are derived by the analysis of a specifically built corpus of contemporary Italian (written and spoken), of 18,000,000 words. The corpus is organized in 6 subcorpora of similar size, further normalized: press (newspapers and periodicals), literature (novels, short stories, poetry), nonfiction (textbooks, essays, encyclopaedia), entertainment (theatre, cinema, songs, and TV shows), computer mediated communication (forum, newsgroup, blog, chat and social networks), spoken language.

Subcorpora	Occurrences
PRESS (newspapers and periodicals)	3,000,000
LITERATURE (novels, short stories and poetry)	3,000,000
NONFICTION (textbooks, essays and encyclopedia)	3,000,000
ENTERTAINMENT (theatre, cinema, songs and TV shows)	3,000,000
COMPUTER MEDIATED COMMUNICATION (forum, newsgroup, blog, chat and social networks)	3,000,000
SPOKEN LANGUAGE	3,000,000

Figure 1: NVdB corpus

The chronological span of the texts included in the corpus range from 2000 to 2012, not diachronically balanced, with a polarization on the last two years. The general criteria for the selec-

tion of texts were maximum variability in authors' and speakers' characteristics. Texts produced during the last years were preferred to older ones. For printed materials we have chosen texts from widely known sources (for example using book charts and prize-winners, most read periodicals and TV shows, statistics of blogs and forum access, etc.). As for length, to have to maximize variability of text features we have preferred shorter works over longer ones, always trying to include texts in their integrity.

The corpus has been POS tagged and extensively corrected manually for all entries belonging to the NVdB (Chiari and De Mauro, 2012). POS tagging has been performed in different sessions. The TreeTagger (Schmid, 1994) has been used with the Italian parameter file provided by Marco Baroni as a first step. Errors were corrected and a new reference dictionary has been built in order to perform further correction sessions. Lemmatization procedures and principles were conducted using GRADIT as the main reference tool and thus follows the guidelines of traditional lexicography (De Mauro, 1999). The main consequence of this choice is that VdB does not appear as a flat frequency list in which each line is a couple lemma-POS, but is a hierarchical list o lemmas (as will be discussed in the next paragraph).

Extensive manual correction has involved correction of proper names tagging, of unknown forms, of incorrect lemma-POS attributions, especially regarding high frequency errors. Manual correction has been performed by checking and disambiguating cases in concordances produced for each item in the list (lemma or word form). A special session of lexical homograph disambiguation has been performed fully manually in order to assure complete alignment of the VdB resource results to GRADIT dictionary.

An evaluation of the amount of manual correction of data is fully provided in the documentation.

2.2 Organization of linguistic data in the resource

One of the most significant improvements in the linguistic resource relies on the fact that all data (relative frequency, usage, dispersion) is given in detail for all subcorpora in order to evaluate different behaviour of lexical units in different subcorpora.

The criteria for the organization of the entries in the lexicon follow lexicographic principles and are perfectly aligned to the entries of GRADIT (De Mauro, 1999). Thus while an ordinary frequency list is a flat list of couples represented by the citation form of a lemma and its grammatical qualification (e.g. *cattivo* noun, *fare* verb appear as different entries – and ranks – from *cattivo* adjective and *fare* noun), the internal organization of NVdB is hierarchical: each entry is conceived as a full lexical entry (presumably as saved in the mental lexicon) where each lemma/entry can be associated to more than one grammatical qualification. In NVdB the entry *cattivo* has a general rank, frequency, usage deriving from the sum of its different grammatical realizations, and will also provide detailed information on the frequency and usage of each of the grammatical qualification for the overall corpus and all subcorpora.

Furthermore for the first time in a frequency list and in a core dictionary extensive account of lexical/absolute homographs has been provided (by disambiguating concordance lines manually for all top ranked lemmas). While textual/relative homography is generally addressed in POS tagging, absolute homography is still a significant challenge and cannot be performed adequately by automatic tools. Thus entries in NVdB and their quantitative data make distinction between *riso* noun ('risata', 'alimento'); *calcio* noun ('gioco/pedata', 'elemento chimico'); *asse* noun ('tavola', 'linea…'); *avanzare* verb ('andare avanti', 'essere in sovrabbondanza'); *buono* noun ('il bene o persona buona', 'documento che dà diritto a ricevere un servizio'). Manual disambiguation of lexical homographs touched about 8,3% of all occurrences in the corpus.

Full processing (cumulative and relative) of formal orthographic variants especially needed in case of loanwords (e.g. *goal*, *gol*; *email*, *e-mail*) is provided.

Moreover one of the major novelties in NVdB is the processing and inclusion of multiword expressions (idioms, fixed expressions, named entities) in the lemma list, both marked independently (lemmatised) and cross referenced under main lemma entries (e.g. *al fine di* is a conjunctional idiom lemmatised autonomously and cross-referenced under the headword *fine*). Multiword expressions included in the NVdB follow the main threshold of the AU layer of the general vocabulary list. Data on multiwords belonging to all grammatical categories have been provided by projecting lemmatized version of the reference list of multiwords included in the largest lexicographic work available for Italian

(GRADIT, De Mauro 1999), 67,678 multiword lemmas, also taking into account possible modifiers occurring between multiwords. Data on multiwords has been fully manually checked in order to exclude multiword sequences that are not used idiomatically in the form of fixed expression. Multiwords belonging to the basic vocabulary are provided in a separate lemmatized list.

The final layers of NVdB describe about 7,400 lexemes: about 2,000 fundamental lexemes, about 3,000 high usage lexemes and about 2,400 high availability lexemes.

3 A short overview of data

Interpreting the comparison between VdB and NVdB can be a very difficult task since there are methodological and linguistic (internal) factors that interact inextricably in results. The main problems derive from the different size and design of the two corpora used and internally comparison of lexical differences in usage is insufficient to provide a full interpretation of the new data presented in the NVdB. It is thus capital to merge quantitative and qualitative analysis and to interconnect lexical stableness, shifts and changes to cultural and social changes that occurred in Italy in the past fifty years.

A rough snapshot of the stableness of Vdb (1980) can be seen by observing how much of the old layers still belongs to the same layer in NVdB. 73.3% of the old FO is stable, 47% of AU is preserved. Most new entries in the new FO layer previously belonged to AU (15% of the overall new FO). Examples of AU lexemes that migrated to FO layer are: *adulto* 'adult', *anziano* 'old' ', *assenza* 'absence', *camion* 'truck', *buco* 'hole', *cassa* 'box', *codice* 'code', *concerto* 'concert', *individuo* 'individual', *insegnante* 'teacher', *lavoratore* 'worker', *letteratura* 'literature', *maggioranza* 'majority', *paziente* 'patient', *procedura* 'procedure', *reagire* 'to react', *ruolo* 'role', *ritmo* 'rhythm', *strumento* 'instrument', *telefonata* 'phone call', *turno* 'turn'.

Other words dropped from FO: *aggiustare* 'to fix', *agricoltura* 'agriculture', *animo* 'soul', *calma* 'calmness', *carità* 'charity', *collina* 'hill', *cretino* 'idiot', *ebbene*, *educare* 'to educate', *fidanzato* 'fiancée', *guaio* 'trouble', *illuminare* 'to illuminate', *ladro* 'thief', *mela* 'apple', *noioso* 'boring', *occupazione* 'occupation', *patria* 'homeland', *pietà* 'pity', *provinciale* 'provincial', *valigia* 'suitcase', *vasto* 'wide', [1]*volgare* 'vulgar'.

4 Conclusion and future developments

The NVdB of Italian is be distributed as a frequency dictionary of lemmatized lexemes and multiword, with data on coverage, frequency, dispersion, usage labels, grammatical qualifications in all subcorpora. A linguistic analysis and comparison with previous data is also provided with full methodological documentation.

The core dictionary and data are also distributed electronically in various formats in order to be used as a reference tool for different applications.

Future work will be to integrate data from the core dictionary with new lexicographic entries (glosses, examples, collocations) in order to provide a tool useful both for first and second language learners and for further computational applications.

References

Umberta Bortolini, Carlo Tagliavini and Antonio Zampolli, 1971. *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia, Milano.

Isabella Chiari and Tullio De Mauro. 2012. The new basic vocabulary of Italian: problems and methods, in «Statistica applicata», 22 (1), 2012, pp. 21-35.

Isabella Chiari and Tullio De Mauro. In press. *Il Nuovo Vocabolario di Base della lingua italiana*. Casa Editrice Sapienza, Roma.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'uso*. UTET, Torino.

Tullio De Mauro, Federico Mancini, Massimo Vedovelli and Miriam Voghera, 1993. *Lessico di frequenza dell'italiano parlato (LIP)*, Etas libri, Milano.

Georges Gougenheim. 1964. *L'élaboration du francais fondamental (1er degré): Étude sur l'établissement d'un vocabulaire et d'une grammaire du base*. Didier, Paris.

Juilland, Alphonse G. and Vincenzo Traversa. 1973. *Frequency Dictionary of Italian Words*, Mouton, The Hague,.

René Michéa. 1953. Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage. *Les langues modernes*. (47): 338-44.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Paper presented to the Proceedings of International Conference on New Methods in Language Processing*: 44-49.

10.12871/CLICIT2014124

Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi

Francesca Chiusaroli
Università di Macerata
f.chiusaroli@unimc.it
francesca.chiusaroli@gmail.com

Abstract

Italiano Il contributo presenta un'analisi linguistica della categoria dell'hashtag in Twitter, con particolare riguardo alle forme italiane, allo scopo di osservarne le caratteristiche morfotattiche nel corpo del testo ed anche le potenzialità semantiche ai fini della possibile interpretazione delle forme in chiave tassonomica. Il percorso di indagine si articolerà all'interno dell'orizzonte teorico definito dal concetto di Scritture Brevi come si trova elaborato in Chiusaroli and Zanzotto 2012a e 2012b ed ora in www.scritturebrevi.it

English *The paper presents a linguistic analysis of Twitter hashtag, with special reference to an Italian case study. The aim is to observe its morphological and syntactical features and also its semantic role to the purpose of a taxonomic categorization. This study moves from the definition of Scritture Brevi (Short Writings) as developed in Chiusaroli and Zanzotto 2012a e 2012b and now at www.scritturebrevi.it*

1. Introduzione

Nella definizione del cosiddetto "gergo di Twitter" si colloca a buon diritto la categoria dell'hashtag, per le tipiche difficoltà alla immediata e pratica leggibilità del tweet poste dalla forma preceduta dal cancelletto. Particolarmente la presenza dell'hashtag, insieme all'occorrenza dell'account (indirizzo dell'utente preceduto dalla chiocciola), connotano la struttura artefatta del testo di Twitter rispetto alla scrittura ordinaria e convenzionale, poiché la stringa frasale risulta concretamente alterata da tali figure tradizionalmente non contemplate nelle regole ortografiche della lingua standard. La cripticità è confermata dal facile esperimento del trasferimento di un tweet contenente hashtag ed account fuori dall'ambiente di Twitter, lì dove è immediatamente percepibile la mancata integrazione delle forme ed anche sostanzialmente viene compromesso il procedimento di lettura e comprensione. Tale difficoltà, incontrata dal neofita del mezzo, appare di fatto superabile con la pratica, mentre alcune caratteristiche speciali dell'hashtag possono suscitare problematiche quanto alla decodifica formale, o automatica, dei testi.

Il presente contributo si propone di fornire una descrizione delle proprietà linguistiche dell'hashtag, che è il più peculiare elemento testuale di Twitter (Chiusaroli, 2014), con particolare riguardo alle espressioni in italiano. La considerazione dei valori grammaticali e delle funzioni semantiche consente di delineare le regole di lettura del testo, come pure di valutare la rilevanza e la necessità, per l'analisi, di una interpretazione in chiave tassonomica utile per la sistematica classificazione di tale recente forma dell'odierna lingua del web, particolarmente importante oggi tra le fenomenologie della scrittura della rete (Pistolesi, 2014; Antonelli, 2007; Maraschio and De Martino, 2010; Tavosanis, 2011). Il percorso di indagine vede rientrare l'hashtag nella definizione di Scritture Brevi come si trova elaborata in Chiusaroli and Zanzotto 2012a e 2012b ed ora in www.scritturebrevi.it:

"L'etichetta Scritture Brevi è proposta come categoria concettuale e metalinguistica per la classificazione di forme grafiche come abbreviazioni, acronimi, segni, icone, indici e simboli, elementi figurativi, espressioni testuali e codici visivi per i quali risulti dirimente il principio della 'brevità' connesso al criterio dell''economia'. In particolare sono comprese nella categoria Scritture Brevi tutte le manifestazioni grafiche che, nella dimensione sintagmatica, si sottraggono al principio della linearità del significante, alterano le regole morfotattiche convenzionali della lingua scritta, e intervengono nella costruzione del messaggio nei termini di

'riduzione, contenimento, sintesi' indotti dai supporti e dai contesti. La categoria ha applicazione nella sincronia e nella diacronia linguistica, nei sistemi standard e non standard, negli ambiti generali e specialistici."

L'analisi si avvarrà inoltre dell'esperienza su Twitter maturata con account @FChiusaroli e hashtag *#scritturebrevi* (dalla data del 26 dicembre 2012) e altri hashtag correlati (ancora elaborati e/o discussi in www.scritturebrevi.it).

2. Morfosintassi

La proprietà fondamentale dell'hashtag è di costituirsi come elemento capace di generare link, ciò che ne determina lo statuto di "aggregatore" del mezzo, indispensabile elemento per la costituzione del dibattito interno e della composizione della relativa comunità social. Proprio tale specifico attributo si pone alla base delle componenti formali dell'hashtag, da cui originano le difficoltà di decifrazione.

Risulta infatti fondamentale l'inscindibilità degli elementi che compongono la forma con hashtag, poiché il simbolo del cancelletto può mantenere insieme le parti a patto che queste non contengano elementi "scissori".

Tale presupposto comporta, conseguentemente, nell'atto di creazione dell'hashtag, la tipica attivazione di processi di accorpamento allo scopo di evitare l'introduzione di elementi grafici che impediscano l'unità, e naturalmente, ed innanzi tutto, al fine di eliminare l'immissione dello spazio bianco separatore.

In tal senso è significativa la perdita di funzione del simbolo "&" oppure del classico trattino medio "-", tradizionalmente segni ortografici di unione i quali tuttavia su Twitter vengono a perdere il proprio statuto poiché inadeguati a preservare l'unità formale.

La decadenza funzionale dei segni di punteggiatura, ereditata dai linguaggi dell'informatica, colpisce alcune regole fondamentali della scrittura tradizionale, corrompendo così la soluzione standard particolarmente in favore della prevalente *scriptio continua* (ammesso il trattino basso, non ammessi il numerale isolato e alcuni segni diacritici, irrilevanti i caratteri accenti). Di qui la scelta di forme morfologicamente agglutinate, come *#scritturebrevi*, rispetto alle versioni "staccate" **#scritture brevi* o **#scritture-brevi*, **#scritture&brevi*. La maiuscola appare ininfluente per la distinzione tra hashtag *(#scritturebrevi = #ScrittureBrevi*), così che l'alternanza tra maiuscola e minuscola è deputata alla questione percettiva, ovvero ad esempio a segnalare il confine o il nucleo di parola *(#TwImago*), anche con ricercati effetti di distintività semantica: si vedano casi come *#narrArte* rispetto a *#narrarTe*.

La funzione "Cerca" di Twitter legge uniformando tali varianti intese come differenze di superficie e, significativamente, può ridurre alla medesima fonte le forme separate: *Cerca #scritturebrevi* fornisce gli stessi risultati di *Cerca scritture brevi*, e altresì *Cerca scritturebrevi* include anche i risultati con hashtag).

Tale operazione di aggregazione attiva più rigidi meccanismi nel caso di forme con refuso (particolarmente diffuse per la nota velocità di digitazione e per la struttura dei dispositivi mobili su touch screen): il servizio *Cerca* risulta per questo dotato di opzione *Forse cercavi*, mentre l'aggregazione diretta per automatica correzione del refuso non è ovviamente consentita.

La perdita di rilievo da parte dei tradizionali strumenti ortografici, e contestualmente la tendenza comune a formare hashtag nelle produzioni (tweet) individuali (livello saussuriano della *parole*), ed anche l'uniformazione operata dal motore di ricerca interno di Twitter, tutto ciò può invece favorire erronee identificazioni, come nel caso di forme omografiche (*#Nello* riferito a Nello Ajello nell'occasione della morte del giornalista il 12 agosto 2013 è stato dai sistemi di ricerca confuso con la preposizione *#nello*, per la mancata distinzione maiuscola vs minuscola). Per questi motivi particolare attenzione deve porsi nella scelta dell'hashtag allorché si sia nelle condizioni di elaborare un hashtag dedicato che non sia confondibile con altri.

L'appartenenza delle forme con hashtag alle più varie categorie grammaticali determina un particolare trattamento delle stesse nella struttura testuale, dal momento che l'hashtag può valere come forma unica ed unitaria, sintagmatica o sintematica (*Grazie da #scritturebrevi*; *Leggerò con #scritturebrevi; Una #scritturebrevi; Un #fatespazio; #adottaunsegno è un hashtag di #scritturebrevi*), oppure nell'economia del tweet le componenti possono essere recuperate e trattate come forme sintatticamente integrate (*Le #scritturebrevi sono molto interessanti; #fatespazio alla virgola; C'è chi #adottaunsegno all'ora*).

Tali applicazioni negli usi mettono in luce la rapida e conseguente evoluzione della funzione originaria dell'hashtag, inizialmente fondato col valore di *topic*, cui si è presto associata (tanto da apparire spesso indistinguibile) la funzione correlata di *comment*: in tali casi l'hashtag appare forma isolata ed isolante, ed individuabile per la

collocazione, di norma, all'estremità (in fondo) al tweet, molto frequentemente privo di punteggiatura di separazione (*Forma interessante #scritturebrevi = Forma interessante. #scritturebrevi = Forma interessante per #scritturebrevi*).

3. Semantica

La peculiarità della immodificabilità della forma con hashtag, per le funzioni di aggregatore assolte dal simbolo, induce una riflessione sulle conseguenze derivanti da tale implicita rigidità, non soltanto, come abbiamo già osservato, alla luce di motivazioni linguistiche, ma anche per la dimensione semantica annessa alla forma in Twitter.

La distintività propria dell'elemento determina certamente alcune difficoltà per quanto attiene alla sua capacità comunicativa. Ad esempio è tecnicamente impossibile "catturare" (comprendere, includere) con un hashtag un suo possibile equivalente in una lingua diversa, a meno di optare per la soluzione di citare nello stesso tweet entrambe le forme: così *#scritturebrevi* conosce un suo eventuale sinonimo nell'inglese *#shortwritings* e tuttavia non si può sostituire l'uno con l'altro se non perdendo il legame con l'aggregazione di partenza.

Ecco perché anche i tweet in inglese facenti capo alla serie di *#scritturebrevi* avranno il tag originale *#scritturebrevi*, con interessanti occasioni oggettive per la diffusione della forma italiana oltre confine. La soluzione di citare con doppio hashtag (in due lingue) fornisce una pratica condizione veicolare, utile per le prime produzioni ed eventualmente per il conteggio generale.

Ma la resistenza della struttura rivela altre difficoltà, interne alla lingua, come nei casi di hashtag a base sinonimica. Alcuni prendono avvio per refuso o per errore di digitazione, o per disattenzione o deriva, ma tanto basta per determinare la immediata scissione delle trafile (ad esempio *#ilibricheamo* da *#libricheamo*, oppure *#sonounalettrice* da *#sonounlettore*) con conseguenze sulla tenuta o fortuna o addirittura semantica della serie. Presumibilmente derivati da iniziative improvvisate o occasionali sono i casi di hashtag circolanti nelle ricorrenze o festività pubbliche, per i quali non risulti programmata o ricostruibile una forma ufficiale (*#buonprimomaggio*; *#unomaggio*; *#1maggio*; *#buon1*; cui si è aggiunta la sottospecie *#concertoprimomaggio*.

Tali condizioni non paritarie tra le espressioni suscitano difficoltà nelle operazioni di valutazione dello statuto semantico degli hashtag, ma come tali non possono non intervenire nella definizione dell'universo del discorso in questione, ad esempio ai fini dell'importante conteggio dei tweet, di cui si occupano applicazioni deputate.

Per la collocazione di queste fenomenologie Scritture Brevi ha definito la categoria metalinguistica "plurihashtag", con lo scopo di radunare e censire le forme sinonimiche. Le analisi fanno verificare il ruolo prevalente del contributo individuale nel processo di ideazione dell'hashtag, un processo che attiene al livello della *parole* con conseguenti difficoltà nell'operazione predittiva e di investigazione.

Significativo è il caso della recente esperienza degli hashtag dedicati ai Mondiali di calcio 2014, per i quali all'interno di Twitter è stata predisposta una speciale *Lista* da seguire (titolo: *Esplora I Mondiali*), così da accorpare gli interventi e le discussioni in un'unica cronologia. All'interno della rubrica, accanto all'ufficiale *#Mondiali2014*, risultano associati di volta in volta gli hashtag dedicati alle specifiche partite giocate. Ad esempio il 9 luglio occorrevano: *#OlandaArgentina; #ArgentinaOlanda; #NEDvsARG; #NEDARG*. Ogni tweet poteva contenerne uno o più di uno; ogni hashtag costituiva nondimeno tecnicamente serie autonoma. Oltre a questi modelli di base, riprodotti per tutta la serie dei Mondiali *(#GermaniaBrasile; #BrasileGermania; #GERvsBRA; #GERBRA*) appare interessante l'abbondanza e varietà interna alla lista, come si evidenzia dalle forme contestualmente usate come *#Brasil2014, #WorldCup, #Mondiali, #Ottavidifinale #Quartidifinale, #Semifinale, #coppadelmondo* e casi di episodi specifici giudicati degni di nota, come ad esempio, il 28 giugno 2014, contemporaneamente in tendenza, *#BrasileCile, #JulioCesar, #Pinilla, #rigori, #Medel*.

4. Conclusioni

La necessità di considerare gli elementi con hashtag per il loro valore sia formale che semantico si conferma indispensabile per una corretta valutazione dei prodotti di lingua (Cann, 1993, e, per le basi, Fillmore, 1976; Lyons, 1977; Chierchia and McConnell Ginet, 1990), in particolare, ma non solo, per poter giudicare l'impatto reale e concreto del fenomeno della scrittura della rete sulle forme e sugli usi, anche nella più ampia prospettiva del mutamento diacronico (Simone, 1993).

Lì dove l'hashtag è importante elemento isolato e come tale capace di radunare contenuti ed

idee, appare incompleta ogni analisi che non tenga conto dell'appartenenza dell'hashtag a più categorie della lingua, dal nome comune semplice o composto, al nome proprio, semplice o composto, al nesso sintematico e frasale, con naturali conseguenze di trattamento morfosintattico (Grossmann, 2004; Doleschal and Thornton, 2000; Recanati, 2011).

Una analisi appropriata non può inoltre prescindere da una classificazione semantica in senso gerarchico e tassonomico delle voci (Cardona, 1980, 1985a, 1985b), ovvero che tenga conto dei gradi delle relazioni tra gli elementi, dei rapporti di sinonimia ovvero di iperonimia e iponimia (Basile, 2005 e Jezek, 2005), ed anche dei rapporti soltanto formali, omografici e omonimici (Pazienza, 1999; Nakagawa and Mori, 2003; Pazienza and Pennacchiotti and Zanzotto, 2005), e infine dei rimandi in termini di corrispondenze in altre lingue (Smadja, McKeown and Hatzivassiloglou, 1996), specialmente l'inglese, per il suo ruolo di idioma veicolare della rete (Crystal, 2003).

Se è vero che la rete e la conoscenza nella rete si formano secondo procedimenti non più lineari o monodimensionali, bensì con andamento in profondità e per strati (Eco, 2007), appare indispensabile inserire nell'orizzonte dell'analisi, oltre all'elemento formale, numerico e quantitativo, anche la valutazione della struttura semantica e prototipica attraverso la ricostruzione degli elementi minimi o "primi" della conoscenza, un metodo ben noto alla storia della linguistica con il termine di *reductio* (Chiusaroli, 1998 e 2001), che per altro si pone all'origine dell'algoritmo del motore di ricerca (Eco, 1993). Proprio la struttura del web e l'organizzazione interna alla CMC consentono di utilizzare l'hashtag di Twitter come uno studio di caso emblematico per testare l'efficacia di un metodo che unisca la considerazione della potenza funzionale della stringa grafica con la rilevanza del piano contenutistico semantico: una intersezione di fattori diversi che devono risultare reciprocamente dipendenti per la corretta verifica dei dati; una teoria integrata della (web-)conoscenza basata sulla scrittura (Ong, 2002).

Reference

Giuseppe Antonelli. 2007. *L'italiano nella società della comunicazione*. Il Mulino, Bologna.

Grazia Basile. 2005. *Le parole nella mente. Relazioni semantiche e struttura del lessico*. Franco Angeli, Milano.

Ronnie Cann. 1993. *Formal Semantics*. *An Introduction*. Cambridge UP, Cambridge-New York.

Giorgio Raimondo Cardona. 1985a. *I sei lati del mondo. Linguaggio ed esperienza*. Laterza, Roma-Bari [nuova ed. 2006].

Giorgio Raimondo Cardona. 1985b. *La foresta di piume. Manuale di etnoscienza*, Laterza. Roma-Bari, Laterza.

Giorgio Raimondo Cardona. 1980. *Introduzione all'etnolinguistica*. Il Mulino. Bologna [nuova ed. Torino, UTET, 2006].

Gennaro Chierchia, G. and Sally McConnell Ginet. 1990. *Meaning and Grammar. An Introduction to Semantics*, MIT, Cambridge (Mass.) [2nd ed. 2000].

Francesca Chiusaroli. 1998. *Categorie di pensiero e categorie di lingua. L'idioma filosofico di John Wilkins*. Il Calamo, Roma.

Francesca Chiusaroli. 2001. *Una trafila secentesca di* reductio. In Vincenzo Orioles (ed.). *Dal 'paradigma' alla parola. Riflessioni sul metalinguaggio della linguistica.* Atti del Convegno, Università degli studi di Udine - Gorizia, 10-11 febbraio 1999. Il Calamo, Roma: 33-51.

Francesca Chiusaroli and Fabio Massimo Zanzotto. 2012a. *Scritture brevi di oggi*. Quaderni di Linguistica Zero, 1. Università degli studi di Napoli "L'Orientale", Napoli.

Francesca Chiusaroli and Fabio Massimo Zanzotto. 2012b. *Informatività e scritture brevi del web*. In Francesca Chiusaroli and Fabio Massimo Zanzotto (eds.). *Scritture brevi nelle lingue moderne*. Quaderni di Linguistica Zero, 2. Università degli studi di Napoli "L'Orientale", Napoli: 3-20.

Francesca Chiusaroli. 2014. *Scritture Brevi di Twitter: note di grammatica e di terminologia*. In Vincenzo Orioles, Raffaella Bombi and Marika Brazzo (eds.). *Metalinguaggio. Storia e statuto dei costrutti della linguistica*. Il Calamo, Roma: 435-448.

David Crystal. 2003[2]. *English as a Global Language*. Oxford UP, Oxford.

Ursula Doleschal, U. and Anna M. Thornton. (eds.). 2000. *Extragrammatical and Marginal Morphology*. Lincom, München.

Umberto Eco. 1993. *La ricerca della lingua perfetta nella cultura europea*. Laterza, Roma-Bari.

Umberto Eco. 2007. *Dall'albero al labirinto. Studi storici sul segno e l'interpretazione*. Bompiani, Milano.

Charles J. Fillmore. 1976. *The Need for a Frame Semantics within Linguistics*. In Hans Karlgreen (ed.). *Statistical Method in Linguistics*. Sprakforlaget Skriptor, Stockholm: 5-29.

Maria Grossmann and Franz Rainer. 2004. *La formazione delle parole in italiano*. Niemeyer, Tubingen.

Elisabetta Jezek. 2005. *Lessico. Classi, strutture, combinazioni*. Il Mulino, Bologna.

John Lyons. 1977. *Semantics*. Cambridge UP, Cambridge-New York.

Nicoletta Maraschio and Domenico De Martino (eds.). 2010. *Se telefonando… ti scrivo / I giovani e la lingua, L'italiano al telefono, dal parlato al digitato*. Atti dei Convegni, Firenze, Accademia della Crusca, 11 maggio 2007 / 26 novembre 2007. Accademia della Crusca, Firenze.

Hiroshi Nakagawa and Tatsunori Mori. 2003. Automatic Term Recognition Based on Statistics of Compound Nouns and their Components. *Terminology*, 9(2): 201-219.

Walter J. Ong. 2002[2]. *Orality and Literacy. The Techonologizing of the Word*. Routledge, New York.

Maria Teresa Pazienza. 1999. A Domain Specific Terminology Extraction System. *International Journal of Terminology*. 5(2): 183-201.

Maria Teresa Pazienza, Marco Pennacchiotti and Fabio Massimo Zanzotto. 2005. *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. In Spiros Sirmakessis (ed.). *Knowledge Mining. Series: Studies in Fuzziness and Soft Computing*. Springer Verlag, Berlin-Heidelberg. 185: 5-20.

Elena Pistolesi. 2014. *Scritture digitali*. In Giuseppe Antonelli, Matteo Motolese and Lorenzo Tomasin (eds.). *Storia dell'italiano scritto*. Vol. III: *Italiano dell'uso*. Roma, Carocci: 349-375

François Recanati. 2011. *Compositionality, Flexibility and Context-Dependence*. In Markus Werning, Wolfram Hinzen and Edouard Machery (eds.). *The Oxford Handbook of Compositionality*. Oxford UP, Oxford: 175-191.

Raffaele Simone. 1993. *Stabilità e instabilità nei caratteri originali dell'italiano*. In Alberto A. Sobrero (ed.). *Introduzione all'italiano contemporaneo, Struttura e variazioni*. Vol. I. Laterza, Roma-Bari: 41-100.

Frank A. Smadja, Kathleen McKeown and Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1): 1-38.

Mirko Tavosanis. 2011. *L'italiano del web*. Carocci, Roma.

10.12871/CLICIT2014125

Annotation of Complex Emotions in Real-Life Dialogues: The Case of Empathy

Morena Danieli
Department of Information Engineering and Computer Science, University of Trento, Italy
danieli@disi.unitn.it

Giuseppe Riccardi
Department of Information Engineering and Computer Science, University of Trento, Italy
riccardi@disi.unitn.it

Firoj Alam
Department of Information Engineering and Computer Science, University of Trento, Italy
alam@disi.unitn.it

Abstract

English. In this paper we discuss the problem of an-notating emotions in real-life spoken conversations by investigating the special case of empathy. We propose an annotation model based on the situated theories of emotions. The annotation scheme is directed to ob-serve the natural unfolding of empathy during the conversations. The key component of the protocol is the identification of the annotation unit based both on linguistic and paralinguistic cues. In the last part of the paper we evaluate the reliability of the annotation model.

Italiano. *In questo articolo illustriamo il problema dell'annotazione delle emozioni nelle conversazioni reali, illustrando il caso particolare dell'empatia. Proponiamo un modello di annotazione basato sulla teoria situazionale delle emozioni. Lo schema di an-notazione è diretto all'osservazione al naturale dipa-namento dell'empatia nel corso della conversazione. La componente principale del protocollo è l'identificazione dell'unità di annotazione basata sul contenuto linguistico e paralinguistico dell'evento emozionale. Nell'ultima parte dell'articolo riportiamo i risultati relativi all'affidabilità del modello di annotazione*.

1 Introduction

The work we present is part of a research project aiming to provide scientific evidence for the situated nature of emotional processes. In particular we investigate the case of complex social emotions, like empathy, by seeing them as relational events that are recognized by observers on the basis of their unfolding in human interactions. The ultimate goals of our research project are a) understanding the multidimensional signals of empathy in human conversations, and b) generating a computational model of basic and complex emotions. A fundamental requirement for building such computational systems is the reliability of the annotation model adopted for coding real life conversations. Therefore, in this paper, we will focus on the annotation scheme that we are using in our project by illustrating the case of empathy annotation.

Empathy is often defined by metaphors that evoke the emotional or intellectual ability to identify another person's emotional states, and/or to understand states of mind of the others. The word "empathy" was introduced in the psychological literature by Titchener in 1909 for translating the German term "Einfühlung". Nowadays it is a common held opinion that empathy encompasses several human interaction abilities. The concept of empathy has been deeply investigated by cognitive scientists and neuroscientists, who proposed the hypothesis according to which empathy underpins the social competence of reconstructing the psychic processes of another person on the basis of the possible identification with his/her internal world and actions (Sperber & Wilson, 2002; Gallese, 2003).

Despite the wide use of the notion of empathy in the psychological research, the concept is still vague and difficult to measure. Among psychologists there is little consensus about which signals subjects rely on for recognizing and echoing empathic responses. Also the uses of the concept by the computational attempts to repro-

duce empathic behavior in virtual agents seem to be suffering due to the lack of operational definitions.

Since the goal of our research is addressing the problem of automatic recognition of emotions in real life situations, we need an operational model of complex emotions, including empathy, focused on the *unfolding* of the emotional events. Our contribution to the design of such a model assumes that processing the discriminative characteristics of acoustic, linguistic, and psycholinguistic levels of the signals can support the automatic recognition of empathy in situated human conversations.

The paper is organized as follows: in the next Section we introduce the situated model of emotions underlying our approach, and its possible impact on emotion annotation tasks. In Section 3 we describe our annotation model, its empirical bases, and reliability evaluation. Finally, we discuss the results of lexical features analysis and ranking

2 Situated theories of emotions and emotion annotation

The theoretical model of situated cognition is an interesting framework for investigating complex emotions. Recently, both neuropsychologists and neuroscientists used the situated model for experimenting on the emotional experiences. Some results provided evidences supporting the thesis that complex emotions are mental events which are construed within situated conceptualizations (Wilson-Mendenhall et al. 2011). According with this view, a subject experiences a complex emotion when s/he conceptualizes an instance of affective feeling. In other terms, experiencing and recognizing an emotion is an act of categorization based on embodied knowledge about how feelings unfold in situated interactions (Barrett 2006). In this view experiencing an emotion is an event emerging at the level of psychological description, but causally constituted by neurobiological processes (Barrett & Lindquist 2008; Wambach and Jerder, 2004).

The situated approach is compatible with the modal model of emotions by Gross (Gross 1998; Gross & Thompson 2007), which emphasizes the attentional and appraisal acts underlying the emotional process. According to Gross, emotions arise in situations where interpersonal transactions can occur. The relevant variables are the behavior of the participating subjects, including their linguistic behavior, and the physical context, including the physiological responses of the participating speakers. The situation compels the attention to the subject, implies a particular meaning for the person, and gives rise to coordinated and malleable responses.

The framework mentioned above has important implications for our goal because it focuses on the process underlying the emotional experience. Actually one of the problems of annotating emotions is related with the difficulty of capturing how the emotional events feel like and how they arise in verbal and non-verbal interactions.

In the field of spoken language processing we have several collections of annotated emotional databases. Rao and Koolagudi (2013), and El Ayadi (2011) provide well informed survey of emotional speech corpora. From their analysis it results that there is a significant disparity among such data collections, in terms of explicitness of the adopted definitions of emotions, of complexity of the annotated emotions, and of definition of the annotation units. Most of the available emotional speech databases have been designed to perform specific tasks, e.g. emotion recognition or emotional speech synthesis (Tesser et al. 2004; Zovato et al. 2004), and the associated annotation schemes mostly depend from the specific tasks as well. A common feature shared by many emotional corpora is their focus on discrete emotion categorizations. To the best of our knowledge no one provides specific insights for annotating the process where emotions unfold. Also more comprehensive models either base their annotation schemes on sets of basic emotions, like the one developed within the HUMAINE project (Douglas-Cowie *et al.* 2003), or they present data collected in artificial human-virtual agent interactions, like the SEMAINE corpus (McKeown et al. 2007).

In the field of human computer interaction, the present models of empathy aim to identify different "sentiment features" such as affect, personality and mood (Ochs et al. 2007). Few, if any, of those works investigate the differential contribution of speech content and emotional prosody to the recognition of empathy, in spite of the evidences that the interplay between verbal and non-verbal features of behavior are probably the best candidate *loci* where human emotions reveal themselves in social interactions (a view supported by many studies, including Magno-Caldognetto 2002; Zovato *et al.* 2008; Danieli, 2007; Kotz & Paulmann 2007; Brück *et al.* 2012; Gili-Fivela & Bazzanella, 2014 among others).

3 Annotation scheme for complex emotions

We argue that the difficult problem of providing guidelines for complex emotion annotation can benefit from focusing the annotators' attention on the emotional process. This requires the identification of the annotation units that are more promising from the point of view of supporting the observer's evaluation on when and how a given emotion arises.

3.1 In search of the annotation units

For pursuing the research described in this paper we investigated if any of the available psychometric scales or questionnaires were usable in our data analysis, both for empathy and for other complex social emotions like satisfaction, and frustration.

As for empathy, we found that among psychologists there are some fundamental concerns about the adequacy of the various scales. For example, no significant correlation was found between the scores on empathy scales and the measurement of empathic accuracy (Lietz *et al*. 2011). The *de-facto* standardized available tests, such as the one referenced in Bahron-Cohen et al. 2013, seem to be effective mostly for clinical applications within well-established experimental settings. However, they can hardly be adapted to judge the empathic abilities of virtual agents and to evaluate human empathic behavior in everyday situations by an external observer.

Given the problematic applicability of psychological scales and computational coding schemes, for capturing in real-life conversations the unfolding of the emotional process, we chose to focus on the interplay between speech content and voice expression. It is well known that the paralinguistic features of vocal expression convey a great deal of information in spoken interactions. In different kinds of interpersonal communication, the accessibility to the facial expressions (in terms of visual frames) is not available. In such cases we usually rely on spoken content and on the paralinguistic events of the spoken utterances. Therefore, in our research we focused on acoustic, lexical and psycholinguistic features for the automatic classification of empathy in conversations, but we chose to rely only on the perception of affective prosody for the annotation task.

3.2 *The empirical base*s

For designing the annotation scheme we made an extensive analyses on a large corpus of real human-human, dyadic conversations collected in a call center in Italy. Each conversation length was around 7 minutes. An expert psycholinguist, Italian native speaker, listened to one hundred of such conversations. She focused on dialog segments where she could perceive emotional attitudes in one of the speakers. The expert annotator's goal was to pay attention to the onset of prosodic variations and judge their relevance with respect to empathy. In doing that she evaluated the communicative situation in terms of appraisal of the transition from a neutral emotional state to an emotional connoted state. Let us clarify this with a dialogue excerpt from the annotated corpus. The fragment is reported in Figure 1, where "C" is the Customer, and "A" is the Agent. The situation is the following: C is calling because a payment to the company is overdue, he is ashamed for not being able to pay immediately, and his speech is plenty of hesitations. This causes an empathic echoing by A: that emerges from the intonation profile of A's reply, and from her lexical choices. For example in the second question of A's turn, she uses the hortatory first plural person instead of the first singular person. Also the rhetorical structure of A's turn, i.e., the use of questions instead of assertions, conveys her empathic attitude.

C:	Senta … ho una bolletta scaduta di 833 euro eh… vorrei sapere se … come posso rateizzarla?
A:	Ma perché non ha chiesto prima di rateizzarla? Proviamo a farlo adesso, ok? […]

Figure 1: An excerpt of a conversation

The expert annotator thus perceived the intonation variation, and marked the speech segment corresponding to the intonation unit outlined in the example, where the word "proviamo" (*let us try*) is tagged as onset of the emotional process.

The results of this listening supported the hypothesis that the relevant speech segments were often characterized by significant transitions in the prosody of speech. As expected, such variations sometimes co-occurred with emotionally connoted words, but also with functional parts of speech like Adverbs and Interjections. Also

phrases and Verbs, as in the example, could play the role of lexical supports for the manifestation of emotions.

On the basis of those results, we designed the annotation scheme for empathy by taking into account *only* the acoustic perception of the variations in the intonation profiles of the utterances. Two expert psychologists, Italian native speakers, performed the actual annotation task. They were instructed to mark the relevant speech segments with empathy tags where they perceived a transition in the emotional state of the speaker, by paying attention to the speech melody, the speaker's tone of voice and only limited attention to the semantic content of the utterance. In the analyzed corpus 785 calls were tagged with respect to the occurrence of empathy. The annotators used the EXMARaLDA Partitur Editor (Schmidt 2004) for the annotation task.

3.3 *Evaluation*

To measure the reliability of this coding scheme we calculated inter-annotator agreement by using the Cohen's kappa statistics, as discussed in Carletta, 1996. For the evaluation, two psychologists worked independently over a set of 64 spoken conversations. We found reliable results with kappa = 0.74. In particular, the comparison showed that 31.25% of the annotated speech segments were exactly tagged by the two annotators at the same positions of the time axis of the waveforms. 53.12% was the percentage of cases where the two annotators perceived the empathic attitude of the speaker occurring in different time frames of the same dialog turns. No other disagreement was reported.

4. Lexical feature analysis and ranking

For the feature analysis, we extracted lexical features from manual transcription consisting of a lexicon of size 13K. Trigram features were extracted to understand whether there are any linguistically relevant contextual manifestations while expressing empathy. For the analysis of the lexical features we used Relief feature selection algorithm (Kononenko, 1994), which has been effective in personality recognition from speech (Alam & Riccardi 2013). Prior to the feature selection we have transformed the raw lexical features into bag-of-words (vector space model), which is a numeric representation of text that has been introduced in text categorization (Joachims, 1998) and is widely used in behavioral signal processing (Shrikanth *et al.* 2013). Each word in the text can be represented as an element in a vector in the form of either Boolean zero/one or frequency. In case of using frequency, it can be transformed into various forms such as logarithmic term frequency (tf), inverse document frequency (idf) or combination of both (tf-idf). For this study, the frequency in the feature vector was transformed into tf-idf, the product of tf and idf. After that, feature values were discretized into 10 equal frequency bins using un-supervised discretization approach to get the benefits in feature selection and classification. Then, we used Relief feature selection algorithm and ranked the features, based on the score computed by the algorithm. In Table 1, we present a selection of the top ranked lexical features selected using the Relief feature selection, which are highly discriminative for the automatic recognition of empathy.

Lexical Features	Score
posso aiutarla	0.17
se lei vuole	0.10
assolutamente sì	0.10
vediamo	0.07
sicuramente	0.06

Table 1: Excerpt from top-ranked lexical features using Relief feature selection algorithm.

As we can see from Table 1, the selected lexical features highlight the type of sentences that are commonly used in customer care services by the Agents, like "posso aiutarla" (*can I help you*), but also less common phrases like "se lei vuole" (*if you want*, including the courtesy Italian pronoun *lei*), and the use of the first plural form of Verbs, like "vediamo" (*let us see*).

5. Conclusions

In this paper we propose a protocol for annotating complex social emotions in real-life conversations by illustrating the special case of empathy. The definition of our annotation scheme is empirically-driven and compatible with the situated models of emotions. The difficult goal of annotating the unfolding of the emotional processes in conversations has been approached by capturing the transitions between neutral and emotionally connoted speech events as those transitions manifest themselves in the melodic variations of the speech signals.

Acknowledgements

The research leading to these results has received funding from the European Union - Seventh Framework Program (FP7/2007-2013) under grant agreement n 610916 SENSEI.

References

Alam, F., Riccardi, G. 2013. Comparative Study of Speaker Personality Traits Recognition in Conversational and Broadcast News Speech, *Proceedings of Interspeech-2013*

Alba-Ferrara, L., Hausmann, M., Mitchell, R.L., and Weis, S. 2011. The Neural Correlates of Emotional Prosody Comprehension: Disentangling Single from Complex Emotions. *PLoS ONE* 6(12): e28701. doi: 10.1371/journal.pone.0028701

Baron-Cohen, S., Tager-Flusberg, H., and Lomberdo, M.(Eds). 2013. *Understanding Other Minds*. London: Oxford Univ. Press

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1), 20-46.

Barrett, L. F., & Lindquist, K. A. (2008). The embodiment of emotion. In Semin, G..R. & Smith, E.R. (Eds) *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*, Cambridge Univ. Press, New York, 237-262.

Brück, C., Kreifelts, B., & Wildgruber, D. 2012. From evolutionary roots to a broad spectrum of complex human emotions: Future research perspectives in the field of emotional vocal communication. Reply to comments on. *Physics of Life Reviews*, 9, 9-12.

Carletta, J., 1996. "Assessing agreement on classification tasks: the kappa statistics", *Computational linguistics* 22.2: 249-254.

Danieli, M. 2007. "Emotional speech and emotional experience". In Turnbull, O., & Zellner, M. 2010. International Neuropsychoanalysis Society: Open Research Days, 2002-2009. *Neuropsychoanalysis: An Interdisciplinary Journal for Psychoanalysis and the Neurosciences*, 12(1), 113-124.

Douglas-Cowie, E., Cowie, R., Schröder, M., 2003. The description of naturally occurring emotional speech. In: *Proceedings of the 15th International Conference on Phonetic Sciences*, Barcelona, Spain, pp. 2877-2880

El Ayadi, M., Kamel, M. S., & Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572-587.

Gallese, V. (2003). The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, 36(4), 171-180.

Gili Fivela, B., & Bazzanella, C. 2014. The relevance of prosody and context to the interplay between intensity and politeness. An exploratory study on Italian. *Journal of Politeness Research*, 10(1), 97-126.

Gross, J. J. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3), 271.

Gross, J. J., & Thompson, R. A. 2007. Emotion regulation: Conceptual foundations. In J.J. Gross (Ed) *Handbook of emotion regulation*. The Guildford Press, New York.

Joachims, T., 1998. *Text categorization with support vector machines: Learning with many relevant features*, Springer.

Kononenko, I. 1994. "Estimating Attributes: Analysis and Extensions of RELIEF", *European Conference on Machine Learning*, 171-182.

Kotz, S.A., and Paulmann, S. 2007. When Emotional Prosody and Semantics Dance Cheek-to-Cheek: ERP Evidence. *Brain Research*, vol. 1151, pages 107-118.

Lietz. C., Gerdes, K.E, Sun, F., Geiger Mullins J., Wagaman, A., and Segal, E.A. 2011. The Empathy Assessment Index (EAI): A Confirmatory Factor Analysis of a Multidimensional Model of Empathy, *Journal of the Society for Social Work and Research*. Vol. 2, No. 2, pp. 104-124, 2011.

Magno Caldognetto, E. 2002. I correlati fonetici delle emozioni. In C. Bazzanella & P.Kobau, 2002. *Passioni, emozioni, affetti*, Mc Graw Hill, Milano, 197-213.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1), 5-17.

Ochs, M., Pelachaud, C., & Sadek, D. 2007. Emotion elicitation in an empathic virtual dialog agent. In *Prooceedings of the Second European Cognitive Science Conference (EuroCogSci).*

Rao, K. S. and Koolagudi, S.G. 2013. *Emotion Recognition Using Speech Features*. Springer. New York.

Schmidt, T. 2004. Transcribing and annotating spoken language with EXMARaLDA, *Proc. of LREC 2004 Workshop on XML-based Richly Annotated Corpora*.

Shrikanth S. Narayanan and Panayiotis G. Georgiou. 2013. Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language, *Proceedings of IEEE*, 101 (5) : 1203-1233.

Sperber, D., & Wilson, D. 2002. Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1 - 2), 3-23.

Tesser, F., Cosi, P., Drioli, C., Tisato, G., & ISTC-CNR, P. 2004. Modelli Prosodici Emotivi per la

Sintesi dell'italiano. *Proc. of AISV 2004*: http://www2.pd.istc.cnr.it/Papers/PieroCosi/tf-AISV2004.pdf

Titchener, E. B. 1909. *Experimental Psychology of the Thought Processes*. Macmillan, London.

Wambach, I.J.A., and Jerger, J.F.. 2004. Processing of Affective Prosody and Lexical Semantics in Spoken Utterances as Differentiated by Event-Related Potentials. *Cognitive Brain Research*, 20 (3): 427-437.

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, 49(5), 1105-1127.

Zovato, E., Sandri, S., Quazza, S., & Badino, L. 2004. Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis. *Proc. ICSLP 2004*, Vol. 2: 1453-1457.

Zovato, E., Tini-Brunozzi, F., and Danieli, M. 2008. "Interplay between pragmatic and acoustic level to embody expressive cues in a text-to-speech system", *Proc. Symposium on Affective Language in Human and Machine,* AISB 2008.

10.12871/CLICIT2014126

Evaluating ImagAct-WordNet mapping for English and Italian through videos

Irene De Felice, Roberto Bartolini, Irene Russo, Valeria Quochi, Monica Monachini
Istituto di Linguistica Computazione A. Zampolli, ILC CNR Pisa
firstname.lastname@ilc.cnr.it

Abstract

English. In this paper we present the results of the evaluation of an automatic mapping between two lexical resources, WordNet/ItalWordNet and ImagAct, a conceptual ontology of action types instantiated by video scenes. Results are compared with those obtained from a previous experiment performed only on Italian data. Differences between the two evaluation strategies, as well as between the quality of the mappings for the two languages considered in this paper, are discussed.

Italiano. *L'articolo presenta i risultati della valutazione di un mapping automatico realizzato tra due risorse lessicali, WordNet/ItalWordNet e ImagAct, un'ontologia concettuale di tipi azionali rappresentati per mezzo di video. Tali risultati vengono confrontati con quelli ottenuti da un precedente esperimento, condotto esclusivamente sull'italiano. Vengono inoltre discusse le differenze tra le due strategie di valutazione, così come nella qualità del mapping proposto per le due lingue qui considerate.*

1 Introduction

In lexicography, the meaning of words is represented through words: definitions in dictionaries try to make clear the denotation of lemmas, reporting examples of linguistic usages that are fundamental especially for function words like prepositions. Corpus linguistics derives definitions from a huge amount of data. This operation improves words meaning induction and refinements, but still supports the view that words can be defined by words.

In the last 20 years dictionaries and lexicographic resources such as WordNet have been enriched with multimodal content (e.g. illustrations, pictures, animations, videos, audio files). Visual representations of denotative words like concrete nouns are effective: see for example the ImageNet project, that enriches WordNets glosses with pictures taken from the web.

Conveying the meaning of action verbs with static representations is not possible; for such cases the use of animations and videos has been proposed (Lew 2010). Short videos depicting basic actions can support the users need (especially in second language acquisition) to understand the range of applicability of verbs. In this paper we describe the multimodal enrichment of ItalWordNet and WordNet 3.0 action verbs entries by means of an automatic mapping with ImagAct (www.imagact.it), a conceptual ontology of action types instantiated by video scenes (Moneglia et al. 2012). Through the connection between synsets and videos we want to illustrate the meaning described by glosses, specifying when the video represents a more specific or a more generic action with respect to the one described by the gloss. We evaluate the mapping watching videos and then finding out which, among the synsets related to the video, is the best to describe the action performed.

2 ImagAct and ItalWordNet/WordNet: general principles

In ImagAct, concrete verbs meanings are represented as 3D videos and, from a theoretical point of view, different meanings of the same verb are intended as different conceptual basic action types. ImagAct action types have been derived bottom-up, by annotating occurrences of 600 high frequency Italian and English action verbs, previously extracted from spoken corpora. All occurrences have been manually clustered into action types, on the basis of body movements and objects

involved. Each lemma usually has more than one action type: for example, for the verb to open we have 7 basic action types, each of them denoting a different physical action and applicable to different sets of objects. This process was carried out in parallel on English and Italian data; finally, Italian and English action types were mapped onto one another and refinements or adjustments were made in order to stabilise the ontology. In this way, 1100 basic action types have been identified.

The ontologys nodes (action types) consist of videos created as 3D animations, each one provided with the sentence that best exemplifies it, according to annotators; each short video represents a particular type of action (e.g. a man taking a glass from a table) and it is related to a list of Italian and English verbs that can be used to describe that action (all the lemmas associated to a scene can thus be seen as something quite similar to WordNet synsets). The 3D animations represent the gist of an action in terms of movements and interactions with the object in a pragmatically neutral context. Sometimes, high level actional concepts could not be represented with a video: in this case, an ontological node is created and associated to a scene ID as well as to a list of Italian and English verbs, but no video is uploaded in the resource. This said, it is evident that ImagAct is a lexical resource structured in a multimodal way: videos represent the core of the resource.

If in WordNet (Fellbaum 1998) and in ItalWordNet (Roventini et al. 2000) lexicographic principles guide the individuation of meanings, ImagAct aims to list the different concepts (one or more) which we refer to when using action verbs. Furthermore, WordNet aims to describe all different uses of a verb, including idiomatic or metaphorical expressions, whereas ImagAct is specifically focused on linguistic uses related to concrete actions. Being aware of the differences between the two resources, we want to map ImagAct on WordNet not only to make clear how the focus on the perceptual aspect of actions can cause the induction of different verbs' senses, but also to enrich WordNet with videos depicting the actions denoted by glosses.

3 Methods

We describe an approach inspired by ontology matching methods for the automatic mapping of ImagAct video scenes onto WordNet/ItalWordNet. The aim of the mapping is to automatically establish correspondences between WN verbal synsets and ImagAct basic action types. This can be done by measuring the semantic proximity between video scenes and synsets in terms of overlap between the class of verbs (lemmas) associated to a scene in ImagAct and the set of synonyms in WordNet synsets (together with their hypernyms and hyponyms).

The ImagAct dataset used for the mapping consists of 1120 video scenes, with a total of 1100 associated Italian verb types (500 lemmas, with an average of 2.4 verb lemmas per scene). For English, we have 1163 video scenes, with a total of 1181 associated English verb types (543 lemmas, with an average of 2.2 verb lemmas per scene). The difference between Italian and English number of scenes is due to the fact that some action types have only been identified for English and cannot be mapped on any Italian action types.

Concerning WordNet, we consider as relevant information: verbal synsets, verb senses, hyponymic and hypernymic relations. Altogether, the ItalWordnet database (hosted at CNR-ILC) contains 8903 verbal synsets and 14086 verb senses (8121 lemmas, with an average of 1.1 verb lemmas per synset) that are potential candidates for the mapping.

As described in Bartolini et al. (2014), we implemented an algorithm inspired by Rodriguez and Egenhofer (2003), based on set-theory and feature-based similarity assessment (Jaccard, 1912; Tversky, 1977), which proved particularly interesting for the mapping of different and independent ontologies and especially fit for lexical resources, as it is primarily based on word matching (for details about the mapping algorithm, see Bartolini et al., 2014). In that paper we presented the mapping between ImagAct and ItalWordNet. The evaluation was performed on a gold standard of 260 Italian verb lemmas corresponding to 358 action types, which mapped onto a total of 343 ItalWordNet synsets. This gold standard was created by mapping verb action types (not scenes) to ItalWordNet synsets. The performance of the algorithm was assessed on the same task of mapping verb types onto synsets: a similarity score was calculated between the verbs contained in a synset and those related to an action type; the best candidate synset is thus the synset with the bigger overlap with the action type, as this overlapping

is measured by the algorithm. In terms of performance, our evaluation results (recall 0.61, precision 0.69) proved that, at least for WordNet-like lexical resources, differences in the synonym sets are relevant for assessing the proximity or distance of concepts.

Since results from this first experiment were encouraging, we adopted the same algorithm also for the English mapping (ImagAct-Princeton WordNet). The database of WordNet 3.0 contains 13767 verbal synsets and 25047 verb senses (11529 lemmas, with an average of 1,19 verb lemmas per synset), as potential candidates for the mapping. For this task we had no gold standard previously created, thus a new evaluation strategy was assessed and then conducted on both English and Italian data. We think that this will not only improve the judgment of the quality of the mapping proposed, but also allow us to compare results from two different kinds of evaluating methods.

4 Evaluation

In this paper we propose a new evaluation of both the English and the Italian mapping. The evaluation was conducted by two authors, respectively on Italian and English data. To test the quality of the mapping proposed by the algorithm, we decided to select a group of ImagAct scenes related to the actions of putting and then to manually assign a judgement to the definitions of the candidate synsets proposed for the mapping for both languages. The two steps of the evaluation were carried out in parallel, one that considers the mapping proposed between ImagAct scenes and ItalWordNet synsets, and the other that considers the mapping proposed between the same ImagAct scenes and English WordNet synsets.

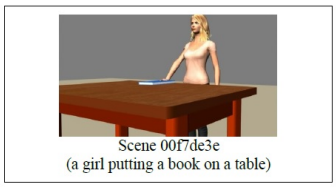


Figure 1: Examples for Imagact-(Ital)WordNet mapping evaluation: equivalence relation.

We expected four possible cases of acceptable mapping (for each one we report, when possible, examples from the two languages):

1. The synset's gloss perfectly describes the scene (equivalence relation (see Figure 1).
2. The synsets gloss describes an event that is more general than that represented by the scene (WordNet more generic).

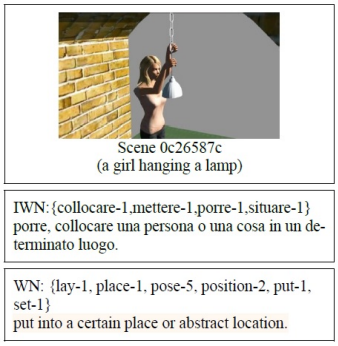


Figure 2: Examples for ImagAct-(Ital)WordNet mapping evaluation: WordNet more generic.

3. The synsets gloss describes an event that is more specific than that represented by the scene (WordNet more specific).

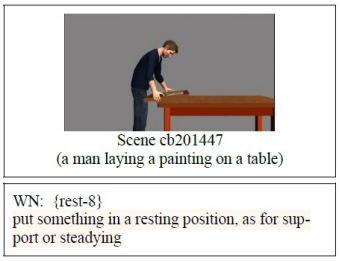


Figure 3: Examples for Imagact-(Ital)WordNet mapping evaluation: WordNet more specific.

4. The synset's gloss is unrelated to the scene (no relation).

Details about the evaluation are reported in Table 1.

	Scene (tot.)	**Scene without videos**
IT	108	27
EN	111	29

Table 1: Scenes evaluated.

The difference in terms of scenes between Italian and English depends on the fact that it is possible that one scene is pointed to only by English verbs, thus this scene cannot be mapped on ItalWordNet.

The results of the evaluation are summarised in the Table 2: in each column is reported the number

of scenes that can be described exactly (=) with the gloss of the first, second or third synset in the mapping, Considering that for each scene the average number of mapped synsets is 60 for Italian and 65 for English, and that we chose to evaluate a group of scenes representing actions that include very generic verbs such as to put and to bring, results for Italian are very good: in the vast majority of cases the right synset is among the first three synsets evaluated as appropriate by the mapping algorithm. Only in the 14.8% of cases no possible match was found. The main factor that impacts on the results for English depends on the way WordNet is structured: in WordNet we find more synonyms with respect to ItalWordNet and as a consequence the mapping algorithm has a different performance. An example of the mapping resulted from the evaluation is available at http://tinyurl.com/q32cps6.

	Italian		English	
	=	all	=	all
First result	41	64	15	33
Second result	2	4	2	5
Third result	1	1	2	5
All	69 (85.2%)		43 (52.4%)	

Table 2: Evaluation results.

5 Conclusions

Mutual enrichments of lexical resources is convenient, especially when different kinds of information are available. In this paper we describe the mapping between ImagAct videos representing action verbs' meanings and WordNet/ItalWordNet, in order to enrich the glosses multimodally. Two types of evaluation have been performed, one based on a gold standard that establishes correspondences between ImagActs basic action types and ItalWordNets synsets (Bartolini at al. 2014) and the other one based on the suitability of a synsets gloss to describe the action watched in the videos. The second type of evaluation suggests that for Italian the automatic mapping is effective in projecting the videos on ItalWordNet's glosses. For what regards the mapping for English, as future work we plan to change the settings, in order to test if the number of synonyms available in WordNet has a negative impact on the quality of the mapping.

Acknowledgments

This research was supported by the MODELACT project, funded by the Futuro in Ricerca 2012 programme (Project Code RBFR12C608); website: http://modelact.lablita.it.

References

Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, and Monica Monachini. 2014. From Synsets to Videos: Enriching ItalWordNet Multimodally. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3110-3117.

Christiane Fellbaum (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Jaccard Paul, 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37-50.

Massimo Moneglia, Gloria Gagliardi, Alessandro Panunzi, Francesca Frontini, Irene Russo, and Monica Monachini. 2012. IMAGACT: Deriving an Action Ontology from Spoken Corpora. In *Proceedings of the Eighth Joint ACL - ISO Work- shop on Interoperable Semantic Annotation*, 42-47.

Andrea M. Rodrguez and Max J. Egenhofer. 2003. Determining Semantic Similarity among Entity Classes from Different Ontologies. In *IEEE Transactions on Knowledge and Data Engineering*, 15(2): 442-456..

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. ItalWordNet: a Large Semantic Database for Italian. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 783-790.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327-352.

10.12871/CLICIT2014127

CLaSSES: a new digital resource for Latin epigraphy

Irene De Felice
Department of Philology, Literature and Linguistics
University of Pisa
irene_def@yahoo.it
ILC (CNR, Pisa)
irene.defelice@ilc.cnr.it

Margherita Donati
Department of Philology, Literature and Linguistics
University of Pisa
margherita.donati@for.unipi.it

Giovanna Marotta
Department of Philology, Literature and Linguistics
University of Pisa
gmarotta@ling.unipi.it

Abstract

English. CLaSSES (Corpus for Latin Sociolinguistic Studies on Epigraphic textS) is an annotated corpus for quantitative and qualitative sociolinguistic analyses on Latin inscriptions. It allows specific researches on phonological and morphophonological phenomena of non-standard Latin forms with crucial reference to the typology of the text, its origin and chronological collocation. This paper presents the first macrosection of CLaSSES, focused on the inscriptions from the archaic-early period.

Italiano. *CLaSSES (Corpus for Latin Sociolinguistic Studies on Epigraphic textS) è un corpus annotato finalizzato all'analisi sociolinguistica quantitativa e qualitativa delle epigrafi latine. Permette di analizzare i fenomeni fonologici e morfofonologici che caratterizzano le forme latine non standard, in relazione alla tipologia testuale, all'area geografica di provenienza e alla datazione delle iscrizioni. L'articolo presenta la prima macrosezione di CLaSSES, incentrata sulle iscrizioni risalenti al periodo preletterario e arcaico.*

1 Digital resources for Latin inscriptions

Available digital resources for Latin epigraphy include some important databases. The Clauss-Slaby database (*http://www.manfredclauss.de/gb/index.html*) records almost all Latin inscriptions (by now 696.313 sets of data for 463.566 inscriptions from over 2.480 publications), including also some pictures. It can be searched by records, province, place and specific terms, thus providing users with quantitative information. The Epigraphic Database Roma EDR (*http://www.edr-edr.it/English/index_en.php*) is part of the international federation of Epigraphic Databases called Electronic Archive of Greek and Latin Epigraphy (EAGLE). It is possible to look through EDR both as a single database or together with its partner databases accessing EAGLE's portal (*www.eagle-eagle.it*).[1]

Although they collect a large amount of data, these resources cannot provide linguists with rich qualitative and quantitative linguistic information focused on specific phenomena. The need for a different kind of information automatically extracted from epigraphic texts is particularly pressing when dealing with sociolinguistic issues.

There is a current debate on whether inscriptions can provide direct evidence on actual linguistic variations occurring in Latin society or they cannot. As Herman (1985) points out, the debate on the linguistic representativity of inscriptions alternates between totally skeptical and too optimistic approaches. Following Herman (1970, 1978a, 1978b, 1982, 1985, 1987, 1990, 2000), we believe that epigraphic texts can be regarded as a fundamental source for studying variation phenomena, provided that one adopts a critical approach. Therefore, we cannot entirely agree with the skeptical view adopted by Adams (2013: 33-34), who denies the role of inscriptions as a source for sociolinguistic variation in the absence of evidence also from metalinguistic comments by grammarians and literary authors.

That said, the current state-of-the-art digital resources for Latin epigraphic texts does not allow researchers to evaluate the relevance of inscriptions for a sociolinguistic study that would

[1] As regards the representation of epigraphic texts in digital form, the international project *EpiDoc* provides guidelines for encoding scholarly and educational editions in XML (*http://sourceforge.net/p/epidoc/wiki/Home/).*

like to rely on direct evidence. Furthermore, it is worth noting that within the huge amount of epigraphic texts available for the Latin language not every inscription is equally significant for linguistic studies: e.g., many inscriptions are very short or fragmentary, others are manipulated or intentionally archaising. Obviously, a (socio)linguistic approach to epigraphic texts should take into account only linguistically significant texts.

2 Aims of the corpus

The resource we present is part of a research project devoted to the sociolinguistic variation in the Latin language (see Donati et al., in press, for further details on this project). Sociolinguistic variation of Latin in Rome and the Empire is a promising research area (Rochette, 1997; Adams et al., 2002; Adams, 2003; Adams, 2007; Biville et al., 2008; Dickey and Chahoud, 2010; Clackson, 2011; Adams, 2013). Since the seminal work by Campanile (1971), many scholars have underlined that sociolinguistic categories and methods can be usefully applied to ancient languages (Lazzeroni, 1984; Vineis, 1984, 1993; Giacalone Ramat, 2000; Molinelli, 2006), even if cautiously.

Assuming this methodological perspective, our empirical analysis of Latin texts is focused on identifying and classifying specific sociolinguistic variants, mostly at the phonological and the morphophonological level. Being aware of the debate on the reliability of inscriptions currently ongoing (§ 1), we intend to investigate whether it is possible to find out relevant evidence for sociolinguistic variation in Latin *via* integration of the modern quantitative and correlative sociolinguistics with a corpus-based approach. Since digital resources devoted to this particular kind of research are actually lacking, our first step was the creation of an original resource for sociolinguistic research on Latin epigraphic texts.

First of all, we collected a corpus including a quite large amount of linguistic and metalinguistic data, to allow grounded quantitative analyses. Our hypothesis is that sociolinguistic aspects eventually emerging from the inscriptions can be detected first identifying the occurrence of non-standard forms in terms of frequency, with crucial reference to the typology of text, its origin and chronological collocation (§ 3), and then also comparing them with their standard variants.[2]

In our analysis of the inscriptions from the archaic and the early period, we considered as non-standard those forms which deviate from the standard as it will be established between the 3rd and the 1st century BCE. For this reason we prefer here the more neutral term "non-standard" (instead of "substandard", used e.g. in Cuzzolin and Haverling, 2009), in the sense of "non-classical", i.e. not present in standard/classical Latin (for a more detailed discussion of this terms see Donati et al., in press).[3] So, e.g. in CIL I^2, inscription 8 (*L Cornelio L f Scipio aidiles cosol cesor*, ca. 250-200 BCE), *Cornelio* can be identified as a non-standard nominative form for the standard *Cornelius*.

3 Methods

3.1 The Corpus CLaSSES

As a first step, we collected the texts of the inscriptions we were interested in and built a corpus. Inscriptions are from the *Corpus Inscriptionum Latinarum* (CIL), the main and most comprehensive source for Latin epigraphy research. Here we present the work carried out during the first phase of our project, corresponding to one macrosection of CLaSSES.

As for the chronology, inscriptions selected are dated from 350 to 150 BCE with most of them falling into the 3rd century BCE (i.e. Archaic-Early Latin). The volumes of CIL covering this chronological segments that were systematically examined are the following: CIL I^2 *Pars II, fasc. I*, section *Inscriptiones vetustissimae*; CIL

[2] It is worth noting that assuming non-standard forms is not a trivial epistemic operation for every phase of Latin, in particular for the archaic (7th century BCE-ca. 240 BCE) and the early period (ca. 240 BCE-ca. 90 BCE). A Latin linguistic and literary standard gradually emerges between the second half of the 3rd century BCE and the 1st century BCE, culminating in the Classical period (Mancini, 2005, 2006; Clackson and Horrocks, 2007; Cuzzolin and Haverling, 2009).

[3] Even if assuming non-standard forms in archaic and early Latin may seem anachronistic in some way, this choice is based on two fundamental aspects: a) many phenomena occurring in these "deviant" forms seem to represent the basis for diachronic developments occurring from Late Latin to the Romance Languages, thus revealing some continuity at least at some (sociolinguistic?) level from the Early to the Late Latin (this point is not uncontroversial, see e.g. Adams, 2013: 8); b) in any case, they provide evidence for phonological and morphophonological variation within archaic epigraphs, thus presumably indicating different levels in the diasystem.

I² *Pars II, fasc. II, Addenda Nummi Indices*, section *Addenda ad inscriptiones vetustissimas;* CIL I² *Pars II, fasc. III, Addenda altera Indices*, section *Addenda ad inscriptiones vetustissimas;* CIL I² *Pars II, fasc. IV, Addenda tertia*, section *Addenda ad inscriptiones vetustissimas.*

It is worth noting that the texts offered by CIL were also revised and checked by means of the available philological resources for Latin epigraphy of this period (Warmington, 1940; Degrassi, 1957, 1963; Wachter, 1987), in order to guarantee the most reliable and updated philological accuracy.

Since inscriptions are not all equally relevant for (socio)linguistic studies, the following texts have been excluded: 1) legal texts, since generally prone to be intentionally archaising; 2) too short (single letters, initials) or fragmentary inscriptions; 3) inscriptions from the necropolis of Praeneste, since containing only an-throponyms in nominative form.

To sum up, the final number of inscriptions in the archaic-early section of CLaSSES is 379 (1804 words). These 379 inscriptions are classified into four textual typologies:

1. *tituli sepulcrales* (n. 27), i.e. epitaphs;
2. *tituli honorarii* (n. 18), i.e. inscriptions celebrating public people;
3. *tituli sacri* (n. 96), i.e. votive inscriptions;
4. *instrumenta domestica* (n. 238), i.e. inscriptions on domestic tools.

The entire collected corpus was then manually tokenized and an index was created, so that each token of the corpus is univocally associated to a token-ID containing the CIL volume, the number of the inscription and the position in which the token occurs within the inscription. Each epigraphic text of CLaSSES was also enriched with metalinguistic information, regarding its geographic origin, its textual typology and its dating. For example, in CIL I^2, inscription 45 (*Diana mereto noutrix Paperia*), *mereto* is identified by the string CIL-I^2-45/2, while CIL-I^2-45 is associated to the following data: loc.: *Gabii*, text. typ.: *tit. sacr.*, dat.: 250-200 BCE.

3.2 Annotation of non-standard forms

In a second step, CLaSSES has been linguistically analysed (for textual interpretation of inscriptions, we mainly referred to the rich information included within CIL, as well as to Warmington, 1940; Degrassi, 1957, 1963; Wachter, 1987). This is the core part of the annotation phase, that provides the corpus with a rich set of qualitative data.

Each non-standard form (already identified by its token-ID) was manually retrieved by two annotators, then also associated to both its corresponding standard form and its lemma, e.g. *cosulibus* (non-standard dat. pl.) - *consulibus* (standard dat. pl.) - *consul* (lemma). Uncertain cases were discussed by the annotators to achieve consensus.

5	Publ_ID	Token_ID	Text. Typ.	Origin	Dating	Non-standard Form	Standard Form	Lemma
6	CIL-I2-7	31	*Tit. Sep.*	Roma	200-150	*abdoucit*	*abduxit*	*abduco*
7	CIL-I2-8	2	*Tit. Sep.*	Roma	250-200	*Cornelio*	*Cornelius*	*Cornelius*
8	CIL-I2-8	6	*Tit. Sep.*	Roma	250-200	*aidiles*	*aedilis*	*aedilis*
9	CIL-I2-8	7	*Tit. Sep.*	Roma	250-200	*cosol*	*consul*	*consul*
10	CIL-I2-9	1	*Tit. Sep.*	Roma	200	*honc*	*hunc*	*hic*
11	CIL-I2-9	2	*Tit. Sep.*	Roma	200	*oino*	*unum*	*unus*

Table 1. Sample excerpt from the Excel sheet containing the annotation of CLaSSES non-standard forms.

Then, all non-standard forms were classified into three classes: vocalism, consonantism and morphophonology, according to the level in which they deviate from the standard form. For example, the nominative *consol* shows a vocalic phenomenon, because it deviates from the standard *consul* for the vowel lowering *u>o*.

A finer-grained analysis of non-standard forms led to a sub-classification of the phenomena investigated. Relevant categories adopted for this classification are the following:

1. for vowels, timbric alterations (lowering, raising), length (*apex*, I *longa*, gemination), syncope, deletion, insertion, monophthongization and archaic spellings of diphthongs;
2. for consonants, final consonant deletion (*-s, -m, -t, -r*), nasal deletion (*-ns->-s-, -nf->-f-*), insertion, assimilation, dissimilation, length (gemination, degemination), voice (voiceless *pro* voiced and voiced *pro* voiceless stops), deaspiration.

Some of these phenomena are especially relevant in the current discussion about social stratification of Latin, namely vowel lowering (*i>e, u>o*), monophthongization (*ae>e, au>o*), synchope, final *-s* and *-m* deletion (cf. among others Campanile, 1971; Pulgram, 1975; Leumann, 1977; Vineis, 1984; Herman, 1987; Weiss, 2009; Loporcaro, 2011a, 2011b; Adams, 2013; Benedetti and Marotta, 2014). Data related to vocalism and consonantism were also classified according to morphophonology: for example, the non-standard nominative *Cornelio* for *Cornelius* is annotated for the lowering *u>o*, for

the final *-s* deletion and for the non-standard *-o* ending nominative of the second declension.

This fine-grained annotation allows researchers to evaluate the statistical incidence of these discussed non-standard phenomena with respect to the corresponding standard forms, also with reference to textual typology, period, geographical origin. Thus, the linguistic annotation of CLaSSES is original and innovative, because it provides not only a list of non-standard occurrences, but especially a collection of data well-suited for a systematically grounded quantitative and qualitative analysis.

4 Possible applications

The data collected so far, together with those deriving from our future work (§ 5), will be the input for the creation of a database that will allow users to make different queries through a web interface.

There are many possible operations that can be done on what we already have. For example, as a conclusion of the annotation work conducted on texts from CIL I^2, we automatically created a *Lexicon* (that will be shortly published) of non-standard forms that contains 340 lemmas. For each lemma, all inflected non-standard forms are reported, with their corresponding inflected standard form, the indication of the inscription they belong to, the indication of their position within the inscription, e.g. *curo*: *coiraveront* (*curaverunt*), CIL I^2, 364(20); *coraveron* (*curaverunt*), CIL I^2, 59(5).

Comparing the total number of non-standard tokens with the *Index* resulting from tokenization, we are also allowed to highlight the proportion of standard and non-standard forms for a given lemma. We registered a 38,4% presence of non-standard forms in the overall corpus:

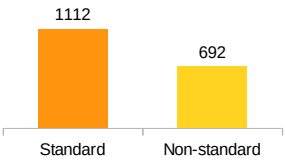


Figure 1. Non-standard vs. standard forms in the corpus (tot. 1804 words).

Similarly, for those interested in particular linguistic issues (such as vowel raising or lowering, monophthongization, etc.), a frequency count of the occurrences of a given phenomenon can be easily done, with or without considering the position of the word within the inscription.

Finally, cross-researches that take into account not only linguistic information (lemma, morphological form, phenomena) but also metalinguistic information (origin, dating, textual typology) are supported. This is one of the strongest points of our resource, because it allows to find correlations among categories. For instance, the following graph shows the percentages of non-standard forms over the total number of forms with respect to the different typologies of text:

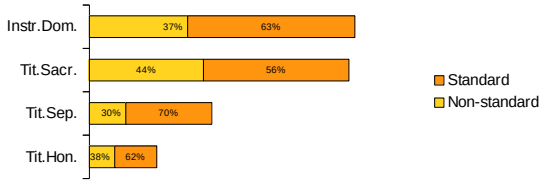


Figure 2. Percentages of non-standard and standard forms with respect to the different typologies of inscriptions.

Moreover, it is also possible to analyze the correlation between a particular phenomenon and the dating of an inscription, or its typology (whether it is classified among *instrumenta domestica*, *tituli sacri*, etc.).

This is exactly the kind of evidence we need to foster a sociolinguistic approach to epigraphic texts. These examples of possible queries follow the belief that quantitative evidence is a necessary requirement for a grounded, systematic linguistic study, even in the case of a corpus language.

5 Conclusion

CLaSSES is an epigraphic Latin corpus for quantitative and qualitative sociolinguistic analyses on Latin inscriptions, that can be useful for both historical linguists and philologists. It is annotated with linguistic and metalinguistic features which allow specific queries on different levels of non-standard Latin forms.

We have here presented the first macrosection of CLaSSES, containing inscriptions from the archaic-early period. In the next future we will collect comparable sub-corpora for the Classical and the Imperial period. Moreover, data will be organized in a database available on the web.

Aknowledgments

This research is part of a project of the University of Pisa developed within the PRIN *Linguistic representations of identity. Sociolinguistic mod-*

els and historical linguistics, coordinator Piera Molinelli (PRIN2010, prot. 2010HXPFF2_001). The research and the results related to the project are presented at *http://www.mediling.eu/*.

References

Adams, James N. 2003. *Bilingualism and the Latin Language*, Cambridge University Press, Cambridge.

Adams, James N. 2007. *The regional diversification of Latin, 200 BC-AD 600*. Cambridge University Press, Cambridge.

Adams, James N. 2013. *Social Variation and the Latin Language*. Cambridge University Press, Cambridge.

Adams, James N., Mark Janse, and Simon Swain (eds.). 2002. *Bilingualism in ancient society. Language contact and the written word*. Oxford University Press, Oxford.

Benedetti, Marina and Giovanna Marotta. 2014. Monottongazione e geminazione in latino: nuovi elementi a favore dell'isocronismo sillabico. In Molinelli, Piera, Pierluigi Cuzzolin, and Chiara Fedriani (eds.). *Latin Vulgaire–Latin Tardif, Actes du Xe Colloque International sur le Latin Vulgaire et Tardif.* Sestante Edizioni, Bergamo: 25-43.

Biville, Frédérique, Jean-Claude Decourt, and Georges Rougemont (eds.). 2008. *Bilinguisme gréco-latin et épigraphie. Actes du colloque international*. Maison de l'Orient et de la Mediterranée-J. Pouilloux, Lyon.

Campanile, Enrico. 1971. Due studi sul latino volgare. *L'Italia Dialettale*, 34:1-64.

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. I Inscriptiones Latinae antiquissimae* (Lommatzsch, 1918 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. II, Addenda Nummi Indices* (Lommatzsch, 1931 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. III, Addenda altera Indices* (Lommatzsch, 1943 ed.).

CIL I² *Inscriptiones Latinae antiquissimae ad C. Caesaris mortem, Pars II, fasc. IV, Addenda tertia* (Degrassi and Krummrey, 1986 eds.).

Clackson, James (ed.). 2011. *A Companion to the Latin Language*. Blackwell Publishing, Malden, Mass.

Clackson, James and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Oxford and Carlton, Malden, Mass.

Cuzzolin, Pierluigi and Gerd Haverling. 2009. Syntax, sociolinguistics, and literary genres. In Baldi, Philip and Pierluigi Cuzzolin (eds.). *New Perspectives on Historical Latin Syntax*: *Syntax of the Sentence*. De Gruyter, Berlin-New York:19-64.

Degrassi, Attilio. 1957, 1963. *Inscriptiones latinae liberae rei publicae*. La Nuova Italia, Firenze.

Dickey, Eleonor and Anna Chahoud (eds.). 2010. *Colloquial and literary Latin*. Cambridge University Press, Cambridge.

Donati, Margherita, Giovanna Marotta, and Francesco Rovai. in press. Prospettive sociolinguistiche sul latino: un corpus per l'analisi dei testi epigrafici. In *Proceedings of the Eleventh International Conference on Latin vulgaire - Latin tardif (LVLT 11)*, Oviedo 2014.

Giacalone Ramat, Anna. 2000. Mutamento linguistico e fattori sociali: riflessioni tra presente e passato. In Cipriano, Palmira, Rita D'Avino, and Paolo Di Giovine (eds.). *Linguistica Storica e Sociolinguistica*. Il Calamo, Roma:45-78.

Herman, József. 1970. *Le latin vulgaire*. Press Universitaires de France, Paris.

Herman, József. 1978a. Évolution a>e en latin tardif? Essai sur les liens entre la phonétique historique et la phonologie diachronique. *Acta Antiquae Academiae Scientiarum Hungariae*, 26:37-48 [also in Herman 1990:204-216].

Herman, József. 1978b. Du Latin épigraphique au Latin provincial: essai de sociologie linguistique sur la langue des inscriptions. In *Étrennes de septantaine: Travaux de linguistique et de grammaire comparée offerts à Michel Lejeune*. Éditions Klincksieck, Paris:99-114 [also in Herman 1990: 35-49].

Herman, József. 1982. Un vieux dossier réouvert: les transformations du système latin des quantités vocaliques. *Bulletin de la Société de Linguistique de Paris*, 77:285-302 [also in Herman 1990:217-231].

Herman, József. 1985. Témoignage des inscriptions latines et préhistoire des langues romanes: le cas de la Sardaigne. *Mèlanges Skok*. Jugoslavenska Akademija Znanosti i Umjetnosti, Zagreb:207-216 [also in Herman 1990:183–194].

Herman, József. 1987. La disparation de *-s* et la morphologie dialectale du latin parlé. In Herman, József (ed.). *Latin vulgaire – Latin tardif: Actes du Ier Colloque international sur le latin vulgaire et tardif, Pécs, 2–5 septembre 1985*. Tübingen:97-108.

Herman, József. 1990. *Du latin aux langues romanes: études de linguistique historique*. Niemeyer, Tübingen.

Herman, József. 2000. Differenze territoriali nel latino parlato dell'Italia: un contributo preliminare. In Herman, József and Anna Marinetti (eds.). *La preistoria dell'italiano. Atti della Tavola Rotonda di Linguistica Storica. Università Ca' Foscari di Venezia 11-13 giugno 1998*. Niemeyer, Tübingen:123-135.

Lazzeroni, Romano. 1984. Lingua e società in Atene antica. *Studi classici e orientali*, 34:16-26.

Leumann, Manu. 1977. *Lateinische Laut- und Formenlehre*. Beck, München.

Loporcaro, Michele. 2011a. Syllable, segment and prosody. In Maiden, Martin, John Charles Smith, and Adam Ledgeway (eds.). *The Cambridge History of the Romance Languages. I: Structures*. Cambridge University Press, Cambridge:50–108.

Loporcaro, Michele. 2011b. Phonological Processes. In Maiden, Martin, John Charles Smith, and Adam Ledgeway (eds.). *The Cambridge History of the Romance Languages. I: Structures*. Cambridge University Press, Cambridge:109–154.

Mancini, Marco. 2005. La formazione del neostandard latino: il caso delle *differentiae uerborum*. In Kiss, Sándor, Luca Mondin, and Giampaolo Salvi (eds.). *Latin et langues romanes, Etudes linguistiques offertes à J. Herman à l'occasion de son 80ème anniversaire*. Niemeyer, Tübingen:137-155.

Mancini, Marco. 2006. *Dilatandis Litteris*: uno studio su Cicerone e la pronunzia 'rustica'. In Bombi, Raffaella, Guido Cifoletti, Fabiana Fusco, Lucia Innocente, and Vincenzo Orioles (eds.). *Studi linguistici in onore di Roberto Gusmani*. Ed. dell'Orso, Alessandria:1023-1046.

Molinelli, Piera. 2006. Per una sociolinguistica del latino. In Arias Abellán, Carmen (ed.). *Latin vulgaire - Latin tardif VII*. Secretariado de Publicaciones Univ. de Sevilla, Sevilla:463-474.

Pulgram, Ernst. 1975. *Latin-Romance Phonology: Prosodics and Metrics*. Wilhelm Fink Verlag, Munich.

Rochette, Bruno. 1997. *Le latin dans le mond grec*. Latomus, Bruxelles.

Vineis, Edoardo. 1984. Problemi di ricostruzione della fonologia del latino volgare. In Vineis, Edoardo (ed.). *Latino volgare, latino medioevale, lingue romanze*. Giardini, Pisa:45-62.

Vineis, Edoardo. 1993. Preliminari per una storia (e una grammatica) del latino parlato. In Stolz, Friedrich, Albert Debrunner, and Wolfgang P. Schmidt (eds.). *Storia della lingua latina*. Pàtron, Bologna:xxxvii-lviii.

Wachter, Rudolf. 1987. *Altlateinische Inschriften. Sprachliche und epigraphische Untersuchungen zu den Dokumenten bis etwa 150 v. Chr*. Peter Lang, Bern-Frankfurt am Main-New York-Paris.

Warmington, Eric Herbert. 1940. *Remains of Old Latin. Vol. 4, Archaic inscriptions*. Harvard University Press-Heinemann, Cambridge MA-London.

Weiss, Michael. 2009. *Outline of the historical and comparative grammar of Latin*. Beech Stave press, New York.

10.12871/CLICIT2014128

Online and Multitask Learning for Machine Translation Quality Estimation in Real-world Scenarios

José G. C. de Souza[(1,2)] **Marco Turchi**[(1)]
Antonios Anastasopoulos[(3)] **Matteo Negri**[(1)]
[(1)] FBK - Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[(2)] University of Trento, Italy
[(3)] University of Notre Dame, Indiana, USA
{desouza,turchi,negri}@fbk.eu
aanastas@nd.edu

Abstract

English. We investigate the application of different supervised learning approaches to machine translation quality estimation in realistic conditions where training data are not available or are heterogeneous with respect to the test data. Our experiments are carried out with two techniques: *online* and *multitask* learning. The former is capable to learn and self-adapt to user feedback, and is suitable for situations in which training data is not available. The latter is capable to learn from data coming from multiple domains, which might considerably differ from the actual testing domain. Two focused experiments in such challenging conditions indicate the good potential of the two approaches.

Italiano. *Questo articolo descrive l'utilizzo di tecniche di apprendimento supervisionato per stimare la qualità della traduzione automatica in condizioni in cui i dati per l'addestramento non sono disponibili o sono disomogenei rispetto a quelli usati per la valutazione. A tal fine si confrontano due approcci: online e multitask learning. Il primo consente di apprendere da feedback degli utenti, dimostrandosi adatto a situazioni di assenza di dati. Il secondo consente l'apprendimento da dati provenienti da più dominii, anche molto diversi da quello in cui il sistema verrà valutato. I risultati di due esperimenti in tali scenari suggeriscono l'efficacia di entrambi gli approcci.*

1 Introduction

Quality Estimation (QE) for Machine Translation (MT) is the task of estimating the quality of a translated sentence at run-time and without access to reference translations (Specia et al., 2009; Soricut and Echihabi, 2010; Bach et al., 2011; Specia, 2011; Mehdad et al., 2012; C. de Souza et al., 2013; C. de Souza et al., 2014a). As a quality indicator, in a typical QE setting, automatic systems have to predict either the time or the number of editing operations (*e.g.* in terms of HTER[1]) required to a human to transform the translation into a syntactically/semantically correct sentence. In recent years, QE gained increasing interest in the MT community as a possible way to: *i)* decide whether a given translation is good enough for publishing as is, *ii)* inform readers of the target language only whether or not they can rely on a translation, *iii)* filter out sentences that are not good enough for post-editing by professional translators, or *iv)* select the best translation among options from multiple MT and/or translation memory systems.

So far, despite its many possible applications, QE research has been mainly conducted in controlled lab testing scenarios that disregard some of the possible challenges posed by real working conditions. Indeed, the large body of research resulting from three editions of the shared QE task organized within the yearly Workshop on Machine Translation (WMT – (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014)) has relied on simplistic assumptions that do not always hold in real life. These assumptions include the idea that the data available to train QE models is: *i) large* (WMT systems are usually trained over datasets of 800/1000 instances) and *ii) representative* (WMT training and test sets are always drawn from the same domain and are uniformly distributed).

[1]The HTER (Snover et al., 2006) measures the minimum edit distance between the MT output and its manually post-edited version in the [0,1] interval. Edit distance is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower HTER values indicate better translations.

In order to investigate the difficulties of training a QE model in realistic scenarios where such conditions might not hold, in this paper we approach the task in situations where: *i)* training data is not available at all (§2), and *ii)* training instances come from different domains (§3). In these two situations, particularly challenging from the machine learning perspective, we investigate the potential of online and multitask learning methods (the former for dealing with the lack of data, and the latter to cope with data heterogeneity), comparing them with the batch methods currently used.

2 How to obtain a QE model *without* training data?

Our first experiment addresses the problem of building a QE model from scratch, when training data is not available (*i.e.* by only learning from the test set). In this scenario, we apply the online learning protocol as a way to build our model and stepwise refine its predictions by exploiting user feedback on the processed test instances.

In the online framework, differently from the batch mode where the model is built from an available training set, the learning algorithm sequentially processes an unknown sequence of instances $X = x_1, x_2, ..., x_n$, returning a prediction $p(x_i)$ as output at each step. Differences between $p(x_i)$ and the true label $\hat{p}(x_i)$ obtained as feedback are used by the learner to refine the next prediction $p(x_{i+1})$. In our experiment we aim to predict the quality of the suggested translations in terms of HTER. In this scenario:

- The set of instances X is represented by (*source, target*) pairs;
- The prediction $p(x_i)$ is the automatically estimated HTER score;
- The true label $\hat{p}(x_i)$ is the actual HTER score calculated over the target and its post-edition.

At each step of the process, the goal of the learner is to exploit user post-editions to reduce the difference between the predicted HTER values and the true labels for the following (*source, target*) pairs. Similar to (Turchi et al., 2014), we do it as follows:

1. At step *i*, an unlabelled (*source, target*) pair x_i is sent to a feature extraction component. To this aim, we used an adapted version (Shah et al., 2014) of the open-source QuEst tool (Specia et al., 2013). The tool, which implements a large number of features proposed by participants in the WMT QE shared tasks, has been modified to process one sentence at a time;
2. The extracted features are sent to an online regressor, which returns a QE prediction score $p(x_i)$ in the [0,1] interval (set to 0 at the first round of the iteration);
3. Based on the post-edition done by the user, the true HTER label $\hat{p}(x_i)$ is calculated by means of the TERCpp[2] open source tool;
4. The true label is sent back to the online algorithm for a stepwise model improvement. The updated model is then ready to process the following instance x_{i+1}.

Online vs batch algorithms. We compare the results achieved by OnlineSVR (Parrella, 2007)[3] with those obtained by a batch strategy based on the Scikit-learn implementation of Support Vector Regression (SVR).[4] Our goal is to check to what extent the online approach (which learns from scratch from the test set) can approximate the batch results obtained, in more favourable conditions, with different amounts of training data.

Feature set. Our feature set consists of the seventeen features proposed in (Specia et al., 2009). These features, fully described in (Callison-Burch et al., 2012), take into account the complexity of the source sentence (*e.g.* number of tokens, number of translations per source word) and the fluency of the translation (*e.g.* language model probabilities). The results of previous WMT QE shared tasks have shown that these baseline features are particularly competitive in the regression task.

Performance indicator. Performance is measured by computing the Mean Absolute Error (MAE), a metric for regression problems also used in the WMT QE shared tasks. The MAE is the average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction of the model and y_i is the true value for the i^{th} instance.

Dataset. Our dataset is drawn from the WMT12 English-Spanish corpus and consists of: *i)* three training sets of different size (200, 600, and 1500

[2] goo.gl/nkh2rE

[3] http://www2.imperial.ac.uk/~gmontana/onlinesvr.htm

[4] http://scikit-learn.org/

instances) used to build the batch models, and *ii)* one "test" set of 754 instances used to build the online models and compare the results obtained with the two strategies. The HTER labels used as feedback by the online approach are calculated using the post-edited version of the target sentences, which is also provided in the WMT12 dataset.

Results. Table 1 reports the MAE results achieved by SVR models (batch strategy) obtained from the three training sets, and the result achieved by the OnlineSVR model (online strategy) obtained by learning only from the test set (since the model is always trained on the same test set, this result of 13.5% MAE is always the same).

As can be seen from the table, similar MAE values show a similar behaviour for the two strategies. This holds even when the batch method can take advantage of the largest dataset to learn from (1500 instances, twice the size of the data used by OnlineSVR).[5] For the batch method, this is an ideal condition not only due to the large amount of data to learn from, but also due to the high homogeneity of the training and test sets (indeed, WMT data come from the same domain and are uniformly distributed). In spite of this, when moving from 600 to 1500 training instances, SVR performance gets stable to a value (12.5% MAE) that is not significantly better than the performance achieved by OnlineSVR. Finally, it's worth noting that, since they are calculated over the entire test set, OnlineSVR results can be highly affected by completely wrong predictions returned for the first instances (recall that at the first step the model returns 0 as a default value). These results, particularly interesting from an application-oriented perspective, indicate the potential of online learning to deal with situations in which training data is not available.

Train	Test	SVR	OnlineSVR
200	754	13.2	13.5*
600	754	12.7	13.5*
1500	754	12.7	13.5*

Table 1: QE performance (MAE) of three batch models (SVR) built from different amounts of training data, and one online model (OnlineSVR) that only learns from the test set.

[5]The online results marked with the "*" symbol are NOT statistically significant compared to the corresponding batch model. Statistical significance at $p \leq 0.005$ has been calculated with approximate randomization (Yeh, 2000).

3 How to obtain a QE model from *heterogeneous* training/test data?

The dominant QE framework presents some characteristics that can limit models' applicability in real-world scenarios. First, the scores used as training labels (HTER, time) are costly to obtain because they are derived from manual post-editions of MT output. Such requirement makes it difficult to develop models for domains in which there is a limited amount of labelled data. Second, the learning methods currently used assume that training and test data are sampled from the same distribution. Though reasonable as a first evaluation setting to promote research in the field, this controlled scenario is not realistic because different data in real-world applications might be post-edited by different translators whose different attitudes have to be modelled (Cohn and Specia, 2013; Turchi et al., 2013; Turchi et al., 2014), the translations might be generated by different MT systems and the documents being translated might belong to different domains or genres.

To overcome these limitations, which represent a major problem for current batch approaches, a reasonable research objective is to exploit techniques that: *i)* allow domains and distributions of features to be different between training and test data, and *ii)* cope with the scarce amount of training labels by sharing information across domains.

In our second experiment we investigate the use of techniques that can exploit training instances from different domains to learn a QE model for a specific target domain for which there is a small amount of labelled data. As suggested in (C. de Souza et al., 2014b) this problem can be approached as a *transfer learning* problem in which the knowledge extracted from one or more source tasks is applied to a target task (Pan and Yang, 2010). *Multitask learning*, a special case of transfer learning, uses domain-specific training signals of related tasks to improve model generalization (Caruana, 1997). Although it was not originally thought for transferring knowledge to a new task, MTL can be used to achieve this objective due to its capability to capture task relatedness, which is important knowledge that can be applied to a new task (Jiang, 2009). When applied to domain adaptation, the approach is transformed in a standard learning problem by augmenting the source and target feature set. The feature space is transformed to be a cross-product of the features of

the source and target domains augmented with the original target domain features. In *supervised* domain adaptation, out-of-domain labels and a small amount of available in-domain labelled data are exploited to train a model (Daumé III, 2007). This is different from the *semi-supervised* case, in which in-domain labels are not available.

Multitask vs single task algorithms. Our approach falls in the supervised domain adaptation framework, for which we apply the Robust MTL approach (RMTL – (Chen et al., 2011)). Our goal is to check to what extent this approach can improve over single task learning strategies. To this aim, RMTL is compared with: *i)* a regressor built only on the available in-domain data (SVR In-domain), and *ii)* a regressor trained by pooling together the training data from all domains, without any kind of task relationship notion (SVR Pooling). These two regressors are built using the implementation of Scikit-learn (Pedregosa et al., 2011).

Feature set and performance indicator. In this experiment we use the same feature set (Specia et al., 2009) and the same performance indicator (MAE) used in §2.

Dataset. Our experiments focus on the English-French language pair and encompass three very different domains: newswire text (henceforth News), transcriptions of Technology Entertainment Design talks (TED) and Information Technology manuals (IT). Such domains represent a challenging combination for adaptive systems since they come from very different sources spanning speech and written discourse (TED and News/IT, respectively) as well as a very well defined and controlled vocabulary in the case of IT. Each domain is composed of 363 tuples formed by the source sentence in English, the French translation produced by an MT system and a human post-edition of the translated sentence. For each pair (translation, post-edition) we compute the HTER to be used as label. For the three domains we use half of the data for training (181 instances) and the other half for testing (182 instances). The reduced amount of instances for training contrasts with the 800 or more instances of the WMT evaluation campaigns and is closer to real-world applications where the availability of large training sets is far from being guaranteed. The sentence tuples for the first two domains were randomly sampled from the Trace corpus[6]. The translations were generated by two different MT systems and post-edited by up to four different translators as described in (Wisniewski et al., 2013). The IT texts come from a software user manual translated by a statistical MT system based on the state-of-the-art phrase-based Moses toolkit (Koehn et al., 2007) trained on about 2M parallel sentences. The post-editions were collected from one professional translator operating in real working conditions with the MateCat tool (Federico et al., 2014).

Results. Table 2 reports the MAE results achieved by the three models (RMTL, SVR In-domain, SVR Pooling). As can be seen from the table, RMTL always outperforms the other methods with statistically significant improvements. These results provide a strong evidence about the higher suitability of multitask learning to deal with real-world contexts that require robust methods to cope with scarce and heterogeneous training data.

Method	TED	News	IT
30 % of training data (54 instances)			
SVR In-Domain	0.2013	0.1753	0.2235
SVR Pooling	0.1962	0.1899	0.2201
RMTL	**0.1946**	**0.1685**	**0.2162**
50% of training data (90 instances)			
SVR In-Domain	0.1976	0.1711	0.2183
SVR Pooling	0.1951	0.1865	0.2191
RMTL	**0.1878**	**0.1653**	**0.2119**
100% of training data (181 instances)			
SVR In-Domain	0.1928	0.1690	0.2081
SVR Pooling	0.1927	0.1849	0.2203
RMTL	**0.1846**	**0.1653**	**0.2075**

Table 2: Average performance (MAE) of fifty runs of the models (multitask RMTL and the single-task SVR In-domain and SVR Pooling) on 30, 50 and 100 percent of training data.

4 Conclusion

We investigated the problem of training reliable QE models in particularly challenging conditions from the learning perspective. Two focused experiments have been carried out by applying: *i)* online learning to cope with the lack of training data, and *ii)* multitask learning to cope with heterogeneous training data. The positive results of our experiments suggest that the two paradigms should be further explored (and possibly combined) to overcome the limitations of current methods and make QE applicable in real-world scenarios.

[6] http://anrtrace.limsi.fr/trace_postedit.tar.bz2

Acknowledgments

This work has been partially supported by the EC-funded project MateCat (ICT-2011.4.2-287688).

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a Method for Measuring Machine Translation Confidence. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 211–219. The Association for Computer Linguistics.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, WMT-2013, pages 1–44, Sofia, Bulgaria.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria, August. Association for Computational Linguistics.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014a. FBK-UPV-UEdin Participation in the WMT14 Quality Estimation Shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014b. Machine Translation Quality Estimation Across Domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 409–420, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 10–51, Montréal, Canada.

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Jianhui Chen, Jiayu Zhou, and Jieping Ye. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 42, New York, New York, USA. ACM Press.

Trevor Cohn and Lucia Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL-2013, pages 32–42, Sofia, Bulgaria.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Conference of the Association for Computational Linguistics (ACL)*.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. THE MATECAT TOOL. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Jing Jiang. 2009. Multi-Task Transfer Learning for Weakly-Supervised Relation Extraction. In *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, number August, pages 1012–1020.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*, ACL '07, pages 177–180.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 171–180, Montréal, Canada.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Francesco Parrella. 2007. Online support vector regression. *Master's Thesis, Department of Information Science, University of Genoa, Italy*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kashif Shah, Marco Turchi, and Lucia Specia. 2014. An Efficient and User-friendly Tool for Machine Translation Quality Estimation. In *Proceedings of the 9^th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the 48^{th} Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 612–621.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13^{th} Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.

Lucia Specia, Kashif Shah, José G.C. de Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51^{st} Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. pages 73–80.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the 8^{th} Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland, June. Association for Computational Linguistics.

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Edition. In *Machine Translation Summit XIV*, pages 117–124.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Proceedings of the 18th conference on Computational linguistics (COLING 2000) - Volume 2*, pages 947–953, Saarbrucken, Germany.

10.12871/CLICIT2014129

A Computational Approach to Poetic Structure, Rhythm and Rhyme

Rodolfo Delmonte
Department of Linguistic Studies and Comparative Cultures
Ca' Bembo 1075 – Università Ca' Foscari – 30123 Venezia
Email: delmont@unive.it

Abstract

English. In this paper we present *SPARSAR*, a system for the automatic analysis of English and Italian poetry. The system can work on any type of poem and produces a set of parameters that are then used to compare poems with one another, of the same author or of different authors. In this paper, we will concentrate on the second module, which is a rule-based system to represent and analyze poetic devices. Evaluation of the system on the basis of a manually created dataset - including poets from Shakespeare's time down to T.S.Eliot and Sylvia Plath - has shown its high precision and accuracy approximating 90%.

Italiano. *In questo lavoro presentiamo SPARSAR, un sistema per l'analisi automatica di poesia inglese e italiana. Il sistema è in grado di lavorare su qualunque poesia e produce un insieme di parametri che vengono poi usati per confrontare poesie e autori tra di loro. In questo lavoro ci concentreremo sul secondo modulo che consiste in un sistema a regole per rappresentare e analizzare i dispositivi e le tecniche poetiche.*

Introduction

In this paper we present *SPARSAR*[1], a system for the automatic analysis of English and Italian poetry. The system can work on any type of poem and produces a set of parameters that are then used to compare poems with one another, of the same author or of different authors. The output can be visualized as a set of coloured boxes of different length and width and allows a direct comparison between poems and poets. In addition, parameters produced can be used to evaluate best similar candidate poems by different authors by means of Pearson's correlation coefficient. The system uses a modified version of *VENSES*, a semantically oriented NLP pipeline (Delmonte et al., 2005). It is accompanied by a module that works at sentence level and produces a whole set of analysis both at quantitative, syntactic and semantic level. The second module is a rule-based system that converts each poem into phonetic characters, it divides words into stressed/unstressed syllables and computes rhyming schemes at line and stanza level. To this end it uses grapheme to phoneme translations made available by different sources, amounting to some 500K entries, and include CMU dictionary, MRC Psycholinguistic Database, Celex Database, plus our own database made of some 20,000 entries. Out of vocabulary words are computed by means of a prosodic parser we implemented in a previous project (Bacalu & Delmonte, 1999a,b).

The system has no limitation on type of poetic and rhetoric devices, however it is dependent on language: Italian line verse requires a certain number of beats and metric accents which are different from the ones contained in an English iambic pentameter. Rules implemented can demote or promote word-stress on a certain syllable depending on selected language, line-level syllable length and contextual information. This includes knowledge about a word being part of a dependency structure either as dependent or as head. A peculiar feature of the system is the use of prosodic measures of syllable durations in msec, taken from a database created in a previous project(Bacalu & Delmonte, 1999a,b). We produce a theoretic prosodic measure for each line and stanza using mean durational values associated to stressed/ unstressed syllables. We call this index, "prosodic-phonetic density index", because it contains count of phones plus count of theoretic durations: the index is intended to characterize the real speakable and audible consistency of each line of the poem. A statistics is issued at different levels to evaluate distributional properties in terms of standard deviations, skewness and kurtosis. The final output of the system is a parameterized version of the poem which is then read aloud by a TTS

[1] The system is available at sparsar.wordpress.com and will soon be made interactive via a webservice.

system: parameters are generated taking into account all previous analysis including sentiment or affective analysis and discourse structure, with the aim to produce an expressive reading.

This paper extends previous conference and demo work (SLATE, Essem, EACL), and concentrates on the second module which focuses on poetic rhythm. The paper is organized as follows: the following section 2 is devoted to present the main features of the prosodic-phonetic system with some example; we then present a conclusion and future work.

2 The prosodic-phonetic module of the system

As R.Tsur(2012) comments in his introduction to his book, iambic pentameter has to be treated as an abstract pattern and no strict boundary can be established. The majority of famous English poets of the past, while using iambic pentameter have introduced violations, which in some cases – as for Milton's Paradise Lost – constitute the majority of verse patterns. Instead, the prosodic nature of the English language needs to be addressed, at first. English is a stress-timed language as opposed to Spanish or Italian which are syllable-timed languages. As a consequence, what really matters in the evaluation of iambic pentameters is the existence of a certain number of beats – 5 in normal cases, but also 4 in deviant ones. Unstressed syllables can number higher, as for instance in the case of exceptional feminine rhyme or double rhyme, which consists of a foot made of a stressed and an unstressed syllable (very common in Italian), ending the line - this is also used by Greene et al. 2010 to loosen the strict iambic model. These variations are made to derive from elementary two-syllable feet, the iamb, the trochee, the spondee, the pyrrich. According to the author, these variations are not casual, they are all motivated by the higher syntactic-semantic structure of the phrase. So there can be variations as long as they are constrained by a meaningful phrase structure.

In our system, in order to allow for variations in the metrical structure of any line, we operate on the basis of syntactic dependency and have a stress demotion rule to decide whether to demote stress on the basis of contextual information. The rule states that word stress can be demoted in dependents in adjacency with their head, in case they are monosyllabic words. In addition, we also have a promotion rule that promotes function words which require word stress. This applies typically to ambiguously tagged words, like "there", which can be used as expletive pronoun in preverbal position, and be unstressed; but it can also be used as locative adverb, in that case in postverbal position, and be stressed. For all these ambiguous cases, but also for homographs not homophones, tagging and syntactic information is paramount.

Our rule system tries to avoid stress clashes and prohibits sequences of three stressed/three unstressed syllables, unless the line syntactic-semantic structure allow it to be interpreted otherwise. Generally speaking, prepositions and auxiliary verbs may be promoted; articles and pronouns never. An important feature of English vs. Italian is length of words in terms of syllables. As may be easily gathered, English words have a high percentage of one-syllable words when compared to Italian which on the contrary has a high percentage of 3/4-syllable words.
In the two tables below we show percentages of

	1-syll. words	2-syll. words	Total 1+2	Total words	Percent
English CELEX	34269	102204	136,473	213,266	63%
English CMU	15945	55426	71371	115,000	62%
Italian PHONit	1496	15258	16,754	120,000	13.96%
Italian SIWL	30	2432	2462	31291	7.9%
Italian ITDict	3012	3989	7001	56000	12%
Totals	53256	164051	217307	535,557	40.58%

Table 1. English/Italian Quantitative 1- 2-Syllable Word Statistics

	Tot 3-5 syll. words	Total words	Perc.
Italian PHONit	97,485	120,000	81.23%
Italian SIWL	22861	31291	73.06%
Italian ITDict	44098	56000	78.75%
Totals	217307	535,557	40.58%

Table 2. Italian Quantitative 3- 5-Syllable Word Statistics

syllable length as contained in phonetic dictionaries of Italian vs English[2].

2.1 Computing Metrical Structure and Rhyming Scheme

Any poem can be characterized by its rhythm which is also revealing of the poet's peculiar style. In turn, the poem's rhythm is based mainly on two elements: meter, that is distribution of stressed and unstressed syllables in the verse, presence of rhyming and other poetic devices like alliteration, assonance, consonance, enjambments, etc. which contribute to poetic form at stanza level. This level is combined then with syntax and semantics to produce the adequate breath-groups and consequent subdivision: these will usually coincide with line stop words, but they may continue to the following line by means of enjambments.

What is paramount in our description of rhythm, is the use of the acoustic parameter of duration. The use of acoustic duration allows our system to produce a model of a poetry reader that we implement by speech synthesis. The use of objective prosodic rhythmic and stylistic features, allows us to compare similar poems of the same poet and of different poets both prosodically and metrically. To this aim we assume that syllable acoustic identity changes as a function of three parameters: internal structure in terms of onset and rhyme which is characterized by number of consonants, consonant clusters, vowel or diphthong; position in the word, whether beginning, end or middle; primary stress, secondary stress or unstressed.

The analysis starts by translating every poem into its phonetic form - see Figure 1. in the Appendix. After reading out the whole poem on a line by line basis and having produced all phonemic transcription, we look for rhetoric devices. Here assonances, consonances, alliterations and rhymes are analysed and then evaluated. Then we compute metrical structure, that is the alternation of beats: this is computed by considering all function or grammatical words which are monosyllabic as unstressed. We associate a "0" to all unstressed syllables, and a value of "1" to all stressed syllables, thus including both primary and secondary stressed syllables. We try to build syllables starting from longest possible phone sequences to shortest one. This is done heuristically trying to match pseudo syllables with our syllable list. Matching may fail and will then result in a new syllable which has not been previously met. We assume that any syllable inventory will be deficient, and will never be sufficient to cover the whole spectrum of syllables available in the English language. For this reason, we introduced a number of phonological rules to account for any new syllable that may appear. To produce our prosodic model we take mean durational values. We also select, whenever possible, positional and stress values. We also take advantage of syntactic information computed separately to highlight chunks' heads as produced by our bottomup parser. In that case, stressed syllables take maximum duration values. Dependent words on the contrary are "demoted" and take minimum duration values.

Durations are then collected at stanza level and a statistics is produced. Metrical structure is used to evaluate statistical measures of its distribution in the poem. As a final result, we found out that it is difficult to find lines with identical number of syllables, identical number of metrical feet and identical metrical verse structure. If we consider the sequence "01" as representing the typical iambic foot, and the iambic pentameter as the typical verse metre of English poetry, there is no poem strictly respecting it in our transcription. On the contrary we find trochees, "10", dactyls, "100", anapests, "001"and spondees, "11". At the end of the computation, the system is able to measure two important indices: "mean verse length" and "mean verse length in no. of feet" that is mean metrical structure.

Additional measures that we are now able to produce are related to rhyming devices. Since we intended to take into account structural internal rhyming schemes and their persistence in the poem we enriched our algorithm with additional data. These measures are then accompanied by information derived from two additional component: word repetition and rhyme repetition at stanza level. Sometimes also refrain may apply, that is the repetition of an entire line of verse. Rhyming schemes together with metrical length, are the strongest parameters to consider when assessing similarity between two poems.

Eventually we reconstruct the internal structure of metrical devices used by the poet: in

[2] For English we use the CMU syllable dictionary, the MRC Psycholinguistic Database, the database contained in the CELEX LDC distribution CD. For Italian, we used our own material amounting to some 100K phonetically transcribed lemmata and wordforms taken from a frequency list computed on 500K tokens of text. We also use PhoneItalia data (see Goslin et al., 2013)

some cases, also stanza repetition at poem level may apply. We then use this information as a multiplier. The final score is tripled in case of structural persistence of more than one rhyming scheme; it is doubled for one repeated rhyme scheme. With no rhyming scheme there will be no increase in the linear count of rethorical and rhyming devices. To create the rhyming scheme we assign labels to each couple of rhyming line and then match recursively each final phonetic word with the following ones, starting from the closest to the one that is further apart. Each time we register the rhyming words and their distance. In the following pass we reconstruct the actual final line numbers and then produce an indexed list of couples, Line Number-Rhyming Line for all the lines, stanza boundaries included. Eventually, we associate alphabetic labels to the each rhyming verse starting from A to Z. A simple alphabetic incremental mechanism updates the rhyme label. This may go beyond the limits of the alphabet itself and in that case, double letters are used.

What is important for final evaluation, is persistence of a given rhyme scheme, how many stanzas contain the same rhyme scheme and the length of the scheme. A poem with no rhyme scheme is much poorer than a poem that has at least one, so this needs to be evaluated positively and this is what we do. Rhetorical and rhyming devices are then used, besides semantic and conceptual indices, to match and compare poems and poets.

SPARSAR visualizes differences by increasing the length and the width of each coloured bar associated to the indices (see Figure 2. in the Appendix). Parameters evaluated and shown by coloured bars include: Poetic Rhetoric Devices (in red); Metrical Length (in green); Semantic Density (in blue); Prosodic Structure Dispersion (in black); Deep Conceptual Index (in brown); Rhyming Scheme Comparison (in purple). Their extension indicates the dimension and size of the index: longer bars are for higher values. In this way it is easily shown which component of the poem has major weight in the evaluation.

Parameters related to the Rhyming Scheme (RS) multiply metrical structure which includes: a count of metrical feet and its distribution in the poem; a count of rhyming devices and their distribution in the poem; a count of prosodic evaluation based on durational values and their distribution. RS is based on the regularity in the repetition of a rhyming scheme across the stanzas or simply the sequence of lines in case the poem is not divided up into stanzas. We don't assess different RSs even though we could: the only additional value is given by the presence of a Chain Rhyme scheme, that is a rhyme present in one stanza which is inherited by the following stanza. Values to be computed are related to the Repetition Rate (RR), that is how many rhymes are repeated in the scheme or in the stanza: this is a ratio between number of verses and their rhyming types. For instance, a scheme like AABBCC, has a higher repetition rate (corresponding to 2) than say AABCDD (1.5), or ABCCDD (1.5). The RR is a parameter linked to the length of the scheme, but also to the number of repeated schemes in the poem: RS may change during the poem and there may be more than one scheme. A higher evaluation is given to full rhymes, which add up the number of identical phones, with respect to half-rhymes which on the contrary count only half that number. We normalize final evaluation to balance the difference between longer vs. shorter poems, where longer poems are rewarded for the intrinsic difficulty of maintaining identical rhyming schemes with different stanzas and different vocabulary.

In Figure 3. in the Appendix, general graded evaluation is shown for the first 53 Shakespeare's sonnets. Position in the space is determined by values of each of the six macro-indices as well as the overall skewness and kurtosis. Most valued sonnets are placed at the top and in the middle of the space, thus indicating the even distribution of their parameters. It is interesting to see that best ranked sonnet is no.29, which has always been regarded as one of the best of the collection.

3 Evaluation and Conclusion

We have done a manual evaluation by analysing a randomly chosen sample of 50 poems out of the 500 analysed by the system. The evaluation has been made by a secondary school teacher of English literature, expert in poetry. We asked the teacher to verify the following four levels of analysis: 1. phonetic translation; 2. syllable division; 3. feet grouping; 4. metrical rhyming structure. Results show a percentage of error which is around 5% as a whole, in the four different levels of analysis, thus subdivided: 1.8 for parameter 1; 2.1 for parameter 2; 0.3 for parameter 3; 0.7 for parameter 4.

References

Bacalu, C. & Delmonte R. (1999a), Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database, in Atti 9° Convegno GFS-AIA, Venezia.

Bacalu, C. & Delmonte R. (1999b), Prosodic Modeling for Speech Recognition, in Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.45- 55.

Baayen, R. H., R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium.

Byrd, Roy J. and Martin S. Chodorow. 1985. Using an online dictionary to find rhyming words and pronunciations for unknown words. In Proceedings of the 23rd Annual Meeting of ACL, 277–283, Chicago, Illinois, USA, July.

Delmonte R. (1999), A Prosodic Module for Self-Learning Activities, Proc.MATISSE, London, 129-132.

Delmonte, R. 2013. SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, SLATE 2013, Demonstration Track.

Delmonte R., Computing Poetry Style, in C.Battaglino, C.Bosco, E.Cambria, R.Damiano, V.Patti, P.Rosso(eds.), Proceedings of 1st International Workshop ESSEM 2013, CEUR Workshop Proc., n.1096, 148-155, Aachen.

Delmonte, R. & Anton Maria Prati, SPARSAR: An Expressive Poetry Reader. 2014. Proceedings EACL, Demonstration Track.

Delmonte R., Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot, Emanuele Pianta, 2005. VENSES – a Linguistically-Based System for Semantic Evaluation, in J. Quiñonero-Candela et al.(eds.), 2005. Machine Learning Challenges. LNCS, Springer, Berlin, 344-371.

Genzel, Dmitriy, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" statistical machine translation: Rhyme and meter. In Proceedings of EMNLP.

Greene, Erica, Tugba Bodrumlu, Kevin Knight. 2010. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation, in Proceedings of the 2010 Conference on EMNLP, 524–533, MIT, Massachusetts, USA, 9-11 October 2010.

Goslin, Jeremy, Claudia Galluzzi, Cristina Romani, (2013), PhonItalia: a phonological lexicon for Italian, in Behavior Research Methods, vol. 45, no. 3, pp.17.

Hayward, M. (1991). A connectionist model of poetic meter. Poetics, 20, 303-317.

Hayward, M. (1996). Application of a connectionist model of poetic meter to problems in generative metrics. Research in Humanities Computing 4. (pp. 185-192). Oxford: Clarendon P.

Kaplan, D., & Blei, D. (2007). A computational approach to style in American poetry. In *IEEE Conference on Data Mining.*

Keppel-Jones, David, 2001. The Strict Metrical Tradition: Variations in the Literary Iambic Pentameter from Sidney and Spenser to Matthew Arnold, Mcgill Queens Univ. Pr., 280.

Sonderegger, Morgan, 2011. Applications of graph theory to an English rhyming corpus. Computer Speech and Language, 25:655–678.

Sravana Reddy & Kevin Knight, 2011. Unsupervised Discovery of Rhyme Schemes, in Proceedings of the 49th Annual Meeting of ACL: shortpapers, 77-82, Portland, Oregon, June 19-24, 2011.

Reuven Tsur, 2012. Poetic Rhythm: Structure and Performance: An Empirical Study in Cognitive Poetics, Sussex Academic Press, 472.

APPENDIX 1.

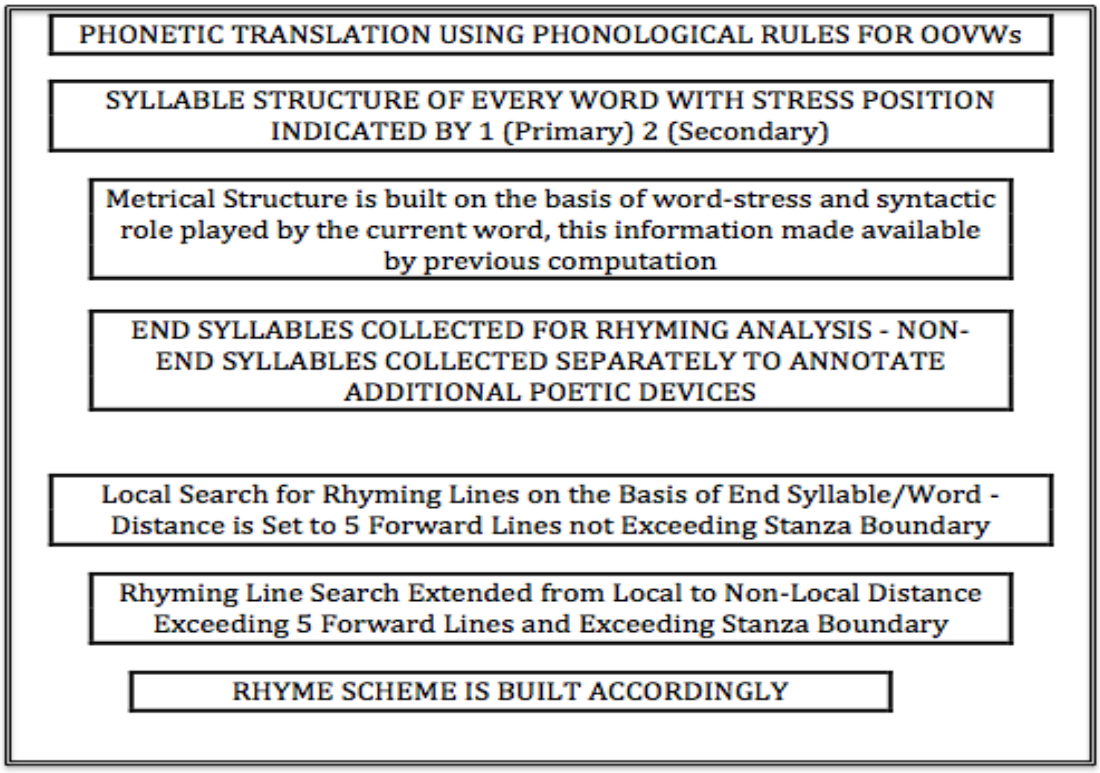


Figure 1. The Rhythm and Rhyme Module of SPARSAR Poetic Analyzer

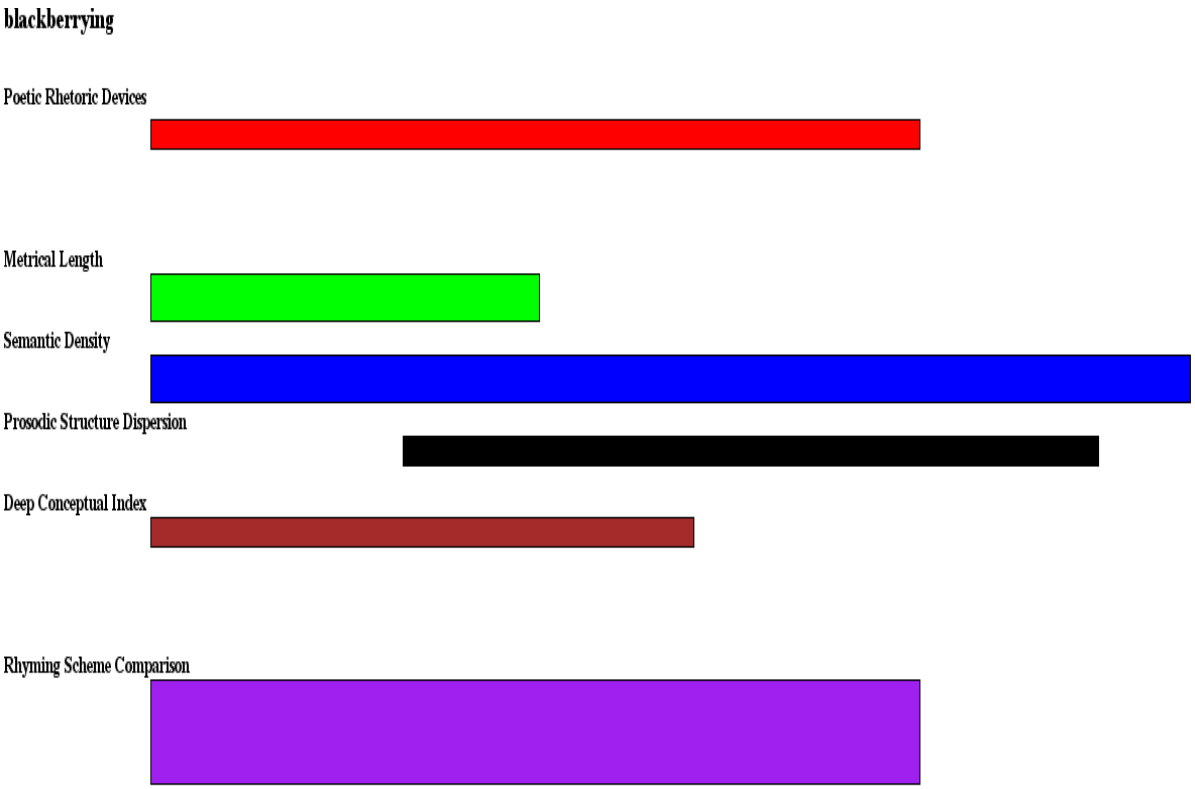


Figure 2. SPARSAR's six macroindices for Sylvia Plath's Blackberrying

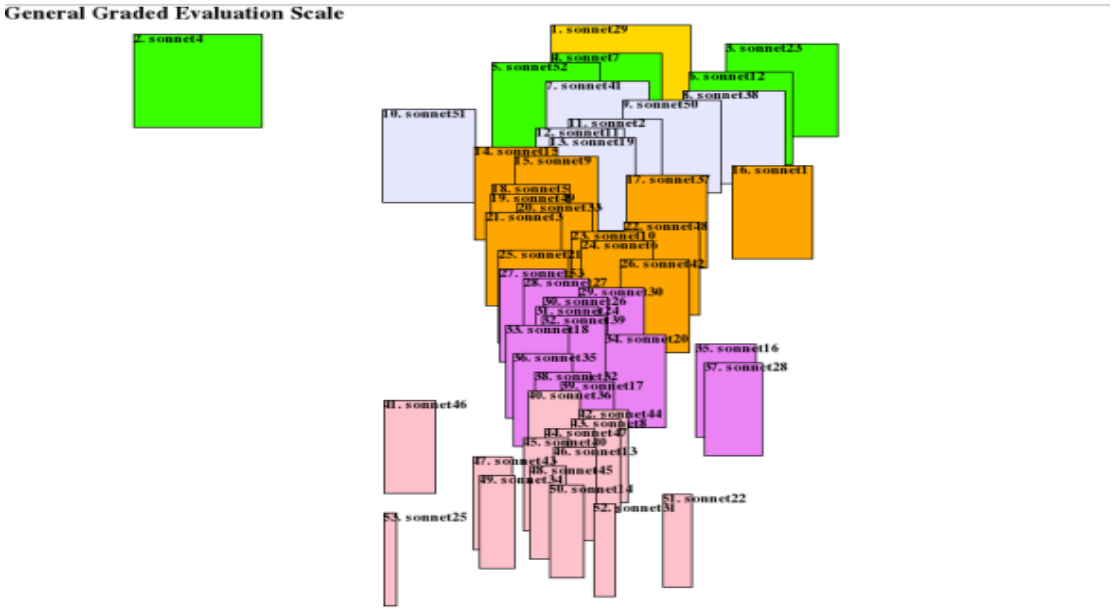


Figure 3. Graded Evaluation of 53 sonnets by William Shakespeare

10.12871/CLICIT2014130

A Reevaluation of Dependency Structure Evaluation

Rodolfo Delmonte
Dipartimento di Studi Linguistici e Culturali Comparati
Ca' Bembo – Dorsoduro 1075
Università Ca' Foscari – 30123 VENEZIA
Email: delmont@unive.it

Abstract

English. In this paper we will develop the argument indirectly raised by the organizer of 2014 Dependency Parsing for Information Extraction task when they classify 19 relations out of 45 as those semantically relevant for the evaluation, and exclude the others which confirms our stance which considers the new paradigm of Dependency parsing evaluation favoured in comparison to the previous parsing scheme based mainly on constituent or phrase structure evaluation. We will also speak in favour of rule-based dependency parsing and against statistically based dependency parsers for reasons related to the role played by the SUBJect relation in Italian.

Italiano. *In questo lavoro svilupperemo un argomento indirettamente sollevato dagli organizzatori del task "2014 Dependency Parsing for Information Extraction", quando classificano 19 relazione come semanticamente rilevanti delle 45 presenti ed escludono le altre. Questo conferma la nostra posizione che considera il paradigma della valutazione dei parser a dipendenze favorito se confrontato con il precedente schema di parsing basato principalmente sulla valutazione della costituenza o strutture sintagmatiche. Parleremo anche a favore del parsing a dipendenze basato su regole e contro i parser a dipendenze solo statistici per ragioni relative al ruolo giocato dal ruolo di SOGGetto in italiano.*

1 Introduction

In this paper I will question the currently widely spread assumption that Dependency Structures (hence DS) are the most convenient syntactic representation, when compared to phrase or constituent structure. I will also claim that evaluation metrics applied to DS are somehow "boasting" its performance with respect to phrase structure (hence PS) representation, without a real advantage, or at least it has not yet been proven there is one. In fact, one first verification has been achieved by this year Evalita Campaign which has introduced a new way of evaluating Dependency Structures, called DS for Information Extraction - and we will comment on that below[1].

In the paper I will also argue that some features of current statistical dependency parsers speak against the use of such an approach to the parsing of languages like Italian which have a high percentage of non-canonical structures (hence NC). In particular I will focus on problems raised by the way in which SUBJect arguments are encoded. State of the art systems are using more and more dependency representations which have lately shown great resiliency, robustness, scalability and great adaptability for semantic enrichment and processing. However, by far the majority of systems available off the shelf don't support a fully semantically consistent representation and lack Empty or Null Elements (see Cai et al. 2001)[2].

O.Rambow (2010) in his opinion paper on the relations between dependency and phrase structure representation has omitted to mention the most important feature that differentiates them. PS evaluation is done on the basis of Brackets, where each bracket contains at least one HEAD, but it may contain other Heads nested inside. Of course, it may also contain a certain number of minor categories which however don't count for evaluation purposes. On

[1] As we read in the details of the call published on the Evalita website:

> "The output of participant systems will be evaluated on the basis of two scoring mechanisms focusing respectively on the parsing performance and suitability for IE… In particular, evaluation will focus on a selection of relations (19 out of a total of 45) chosen according to the following general criteria:
> a. semantic relevance of the relation (i.e. nsubj, dobj ...)
> b. exclusion of syntactic easy to identify relations (i.e. det, aux ...);
> c. exclusion of sparse and difficult to identify relations (i.e. csubj)"

[2] Additional problems are raised by the existence of Non-projective relations which amount to a consistent number of displaced constituents, both as Arguments and as Adjuncts, as discussed below.

the contrary, DS evaluation is done on the basis of head-dependent relations intervening between a pair of TOKENs. So on the one side, F-measure evaluates number of brackets which coincide with number of Heads; on the other side it evaluates number of TOKENS. Now, the difference in performance is clearly shown by percent accuracy obtained with PS evaluation which for Italian was contained in a range between 70% to 75% in Evalita 2007, and between 75% and 80% in Evalita 2009 – I don't take into account 2011 results which are referred to only one participant. DS evaluation reached peaks of 95% for UAS and in between 84% and 91% for LAS evaluation. Since data were the same for the two campaigns, one wonders what makes one representation more successful than the other.

Typically, constituent parsing is evaluated on the basis of constituents, which are made up of a head and an internal sequence of minor constituents dependent on the head. What is really important in the evaluation is the head of each constituent and the way in which PS are organized, and this corresponds to bracketing. On the contrary, DS are organized on the basis of a "word level grammar", so that each TOKEN constributes to the overall evaluation, including punctuation (not always). Since minor categories are by far the great majority of the tokens making up a sentence – in Western languages, but no so in Chinese, for instance (see Yang & Xue, 2010)– the evaluation is basically made on the ability of the parser to connect minor categories to their heads.

What speaks in favour of adopting DS is the clear advantage gained in the much richer number of labeled relations which intervene at word level, when compared to the number of constituent labels used to annotate PS relations[3]. It is worth while noting that DS is not only a much richer representation than PS, but it encompasses different levels of linguistic knowledge. For instance, punctuation may be used to indicate appositions, parentheticals, coordinated sets, elliptical material, subdivision of complex sentences into main and subordinate clause. The same applies to discourse markers which may be the ROOT of a sentence. These have all to be taken into account when computing DS but not with PS parsing.

2 Hypotheses about Dependency Evaluation Success Story

In every Western language, the number of SEMANTIC heads – Nouns, Verbs, Adjectives and Adverbs – is very low when compared to the number of tokens. Rank lists for Italian and English in their upper part are cluttered with articles, prepositions, conjunctions, quantifiers and other determiners. Semantically relevant words only come below a certain frequency threshold. To ascertain these proportions, we decided to look into the dependency treebank made freely available for the current Evalita campaign: here below is a statistics of heads which play the role of Arguments and then those that play both the role of Argument and that of Adjuncts. Percentages are obtained by dividing each relation total occurrency with the total number of tokens, which is 158485.

As can be easily noticed, Core Arguments only make 10% of all tokens and even in a 90% accuracy test result all of them might be wrong. Notice that NSUBJs include 1049 NCUSBJPASS.

	Occur.	**Percent**
nsubj	7549	4.5518%
dobj	5519	3.3278%
iobj	852	0.5137%
xcomp	1036	0.6242%
acomp	1020	0.6150%
TotCore	***15976***	***10.008%***
pobj	23313	14.058%
TotalC+P	***39289***	***24.79%***
csubj	187	
vmod	7920	
rcmod	1945	
TotalAdj	***10052***	***6.343%***
ROOT	7399	
TotalC+P+A+R	***56740***	***35.801%***

Table 1. Grammatical Relations in SIDT

POBJ include both Oblique arguments – a small part - and circumstantial Adjuncts – the great majority. We know for sure that Oblique arguments usually occur with intransitive verbs, which are a small percentage of all verbs. They may also occur as arguments of Ditransitive verbs, but also these are a small percentage of Italian verbs. So we may well say that core Argument grammatical relations only cover some

[3] In particular, then, there is at least one relation lacking in PS representation and is coordination, which on contrary is always represented in DS. As for grammatical relations like SUBJect and OBJect, they are usually not available in PS but they actually appear in PennTreebank II and so they can be learned.

12% of all tokens. Considering that the 64% of all tokens have minor or secondary dependency relations – those based on minor categories or punctuation -, we come up with the conclusion that the remaining 27% needed to cover the best accuracy result obtained so far (91%) is scattered amongst Arguments and Adjuncts. But Arguments and Adjuncts head-dependent relations only constitute 36% of all dependency relations and 27% will only cover 75% of them, no more. So eventually, an evaluation based on semantically relevant heads of Arguments and Adjuncts will achieve worse results than one based on phrase structures.

Now consider ROOT heads which include also root heads of fragments, typically nominal heads. The total number of Inflected verbs in the treebank amounts to 10800 heads. This means that the percentage of null subject elements is 30.102% of all inflected clauses - we subtract expressed subjects from total inflected verbs 10800-7549 = 3251. This 30% of missing SUBJect arguments deteriorates any evaluation. Then we need to consider that there will be another 30% of subjects which are difficult to get because they are placed in noncanonical position – this is derived from a statistics based on VIT(see Delmonte et al. 2007)[4]. It is a fact that in this way, the semantics of the representation used and produced at runtime becomes inconsistent and will reduce dramatically its usefulness in real life applications like Information Extraction, Q/A and other semantically driven fields by hampering the mapping of a complete logical form. Statistical models for DS only encode lexically expressed subjects, null elements being strictly forbidden.

Coming now to general results of the Relations Task in Evalita - specific results are not yet available -, we see that Precision best percentage almost reaches 82%, while the worst is around 78%. However seen that Recall is in the range of 90-85% the accuracy would average 80%. F1 is subsequently contained in the range 86-83%. Data are then equal to if not worse than those of PS evaluation. Even though we don't have available a detailed distribution of the data in the different categories, we may definitely say that they confirm our stance. Thus we expect minor categories like DET to be correct at 98%; not so for those relevant relations corresponding to semantic heads.

2.1 Problematic issues for statistical parsers of Italian

There are two types of Dependency parsers: rule-based symbolic parsers which can also make use of statistics at some step of computation; and statistically only parsers which make use of a classifier and a model to decide how to process the input word (see Delmonte, 1999; 2000; 2002; 2005). The second one could also be – as is the case of Stanford parser (see De Marneff et al., 2011) - a phrase structure probabilistic parser with a mapping or conversion step of syntactic constituents into DS.

Statistical dependency parsers are trained on annotated treebank data and make predictions on the basis of the model. They tap their knowledge from a training corpus which leads to the creation of a model using a classifier. The fundamental idea is the ability of the parser to use the model in a predictive way in order to generalize the encoded information to new and unseen linguistic material[5]. Even if it is obvious that a statistical model can represent linguistic knowledge at any depth and level of representation by increasing the number of features, this is not always convenient both on grounds of efficiency and overall performance. However, linguistic knowledge is split into two main components: the grammar and the lexicon. It is reasonable to assume that learning can only be achieved for the grammatical component and only for regular linguistic phenomena. The other important component, the lexicon, is on the contrary not predictable by definition. Lexical knowledge is idiosyncratic and unpredictable: for instance,

[4] We verified the proportion of null subject in VIT is even higher. We derived data of non-canonical structures in Italian from the treebank called VIT - Venice Italian Treebank (see Delmonte et al., 2007; Delmonte 2009) and compared them to PennTreebank data. In particular, VIT has 36.46% of NC structures vs. 13.01% for PT. As to lexically unexpressed or null subject, VIT has 51.31% vs 0.27% for PT. NC structures are measured against the number of total utterances that is 10,200 for VIT, and 55,600 for PT. On the contrary, Null Subjects are counted against simple sentences, that is 19,099 for VIT and 93,532 for PT. As for Subjects, there were 6230 canonical - i.e. strictly in preverbal position with no linguistic material intervening between Subject and inflected Verb - lexically expressed SUBJects out of the total 10,100 lexically expressed SUBJects. This means that non-canonical subjects constitutes 1/3 of all expressed SUBJects.

[5] In fact, when used in a different domain, the same parser is usually susceptible to serious performance degradation which can get as high as 14% (see Lease & Charniak, 2004). This problem has been partially solved by introducing several parser adaptation techniques to new domains.

knowing that certain verbs belong to the class of atmospheric or impersonal verbs and are associated to special constructions simply requires knowing which they are. Grammatical knowledge is on the contrary predictable being associated either to grammatical or functional words - which are very frequent, or to the presence of specific morphemes[6].

Given the great variety of possible structures in Italian sentences, it is quite reasonable to assume that they may suffer from problems related to the SUBJect relation. Parsers of Italian are in general unable to detect duplicate subjects, and can erroneously licence a proposition or clause without a subject even if one is available but not in canonical position. Since Italian has null subjects, this may happen quite frequently. Just for the matter of documenting the phenomenon, we will show one such example below. The example is useful for two reasons: it shows two different approaches to parsing (one without and one with null elements); secondly it helps documenting the phenomenon.

We take Null subject in Italian to be a feature that speaks in favour of rule-based parsers. Rule-based parser have more resiliency and don't need any training. They can base their knowledge on the lexicon where selectional preferences are encoded, and can produce empty categories. We will use one such parser as example, and we are here referring to TULE TUT parser documented in Lombardo, Lesmo's (1998) paper. In order to show how this may affect the output representation, we report in the appendix one sentence parsed by TALN/DeSR parser (see Attardi, 2006; Attardi et al. 2009), available as webservice at http://tanl.di.unipi.it/ it/. This parser is regarded one of the best statistical dependency parsers of Italian, achieving best results in Evalita campaigns. The output is reported in Appendix 1.

(1) E dovranno riportare per ogni unità urbana anche i dati di superficie espressi in metri quadri in conformità alle istruzioni che saranno fornite in seguito, poiché questo sarà in futuro il parametro in base al quale sarà decretato l'esborso del contribuente al posto dei <vani utili> che andranno in soffitta. [7]

I used half of a very long sentence taken from an Italian administrative bill expressed in a style which is considered "burocratic" style. In DeSR output I marked with double stars all cases of wrong argument selection, and by indenting all cases of relative clauses which have no indication of argument nor grammatical relation – COMP is a generic label and should have been substituted by NSUBJ, DOBJ or POBJ according to the grammatical relation held by the relative pronoun. Wrong argument selection in one case of double subject assignment, as well as two cases of no subject assignment. Errors may be partly due to wrong tagging disambiguation.

The same sentence has been parsed by TULETUT parser which is able to process the two relative structures with almost no error. One of the reasons for this difference, maybe because it uses subcategorization information. In addition, TULETUT parser also correctly produces empty subject categories and traces for long distance dependencies. However, also this representation has one error, and it is the missing link between the relative pronoun and its governing verb: as it is usually the case, the relative pronoun is linked to the verb of the internal clause and the verb is linked to antecedent. This does not happen with relative prepositional object IN_BASE_AL QUALE.

3 Conclusion

In this paper I tried to highlight critical issues on the current way of evaluating DS which indirectly "boasts" the performance of the parsers when compared to phrase structure evaluation. I assume this is due to the inherent shortcoming of DS evaluation not considering semantically relevant grammatical relations as being more important than minor categories. Statistical dependency parsers may have more problems in encoding features of Italian Subject because of its multiple free representations. For this reasons, I argued in favour of rule-based dependency parsers and I presented in particular, one example from TULETUT, a deep parser of Italian.

[6] In fact, for some linguistic theories - the constructional theory of grammar - also syntactic constructions are part of lexical knowledge (see Goldberg, 2006).

[7] This is a literal translation: "And they should include for each urban unit also surface data expressed in square meters in compliance with the instructions that will be made available later, because this will be in future the parameter on the basis of which will be decided the payment by the tax-payers in place of <useful rooms> which will be abandoned."

References

Attardi, G. 2006. Experiments with a Multilanguage Non-Projective Dependency Parser. Proc. of the Tenth Conference on Natural Language Learning, New York, (NY), 2006.

Attardi, G., F. Dell'Orletta, M. Simi, J. Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. Proc. of Workshop Evalita.

Cai, Shu, David Chiang, Yoav Goldberg, 2011. Language-Independent Parsing with Empty Elements, in Proceedings of the 49th Annual Meeting of the ACL, 212–216.

De Marneffe, M.C., B. MacCartney, C. D. Manning, Generating typed dependency parses from phrase structure parses, Proceedings of LREC, 2006/5, 449-454.

Delmonte R. 1999. From Shallow Parsing to Functional Structure, in Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.8-19.

Delmonte R. 2000. Parsing Preferences and Linguistic Strategies, in LDV-Forum - Zeitschrift fuer Computerlinguistik und Sprachtechnologie - "Communicating Agents", Band 17, 1,2, pp. 56-73.

Delmonte R. 2002. GETARUN PARSER - A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG, Proc. LFG2002 Conference, Athens, pp.130-153, at http://cslipublications.stanford.edu/hand/miscpubs online.html.

Delmonte R., 2005, Deep & Shallow Linguistically Based Parsing, in A.M.Di Sciullo(ed), UG and External Systems, John Benjamins, Amsterdam/Philadelphia, pp.335-374.

Delmonte, R., A. Bristot, S. Tonelli 2007. VIT - Venice Italian Treebank: Syntactic and Quantitative Features, in K. De Smedt, Jan Hajic, Sandra Kübler(Eds.), Proc. Sixth International Workshop on TLT, Nealt Proc. Series Vol.1, 43-54.

Delmonte R., 2009. Treebanking in VIT: from Phrase Structure to Dependency Representation, in Sergei Nirenburg (ed.), Language Engineering for Lesser-Studied Languages, IOS Press, Amsterdam, The Netherlands, 51-80.

Goldberg, A. E. 2006. Constructions at work: The nature of generalization in language. Oxford: Oxford University Press.

Lease M. & Eugene Charniak, 2005. Parsing biomedical literature. in Proceedings IJCNLP ’05, Second International Joint Conference, Jeju Island, Korea. pp 58-69.

Lombardo, V. & L. Lesmo, 1998. Formal aspects and parsing issues of dependency theory. In Coling-ACL ’98, 787-793, Montreal, Canada.

Rambow, Owen (2010). The simple truth about dependency and phrase structure representations: an opinion piece. In Proceedings of NAACL-HLT, 1-4.

Yang, Yaqin and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the Chinese Treebank.In Proc. COLING.

APPENDIX 1. TALN/DeSR parser output of sentence (1)

ID	FORM	LEMMA	CPOS	POS	MORPH	HEAD	DEP	PH	PD
1	e	e	C	CC	_	0	ROOT	_	_
2	dovranno	dovere	V	VM	num=p|per=3|mod=i|ten=f	3	aux	_	_
3	riportare	riportare	V	V	mod=f	1	conj	_	_
4	per	per	E	E	_	3	comp	_	_
5	ogni	ogni	D	DI	num=s|gen=n	6	mod	_	_
6	unità	unità	S	S	num=n|gen=f	4	prep	_	_
7	urbana	urbano	A	A	num=s|gen=f	6	mod	_	_
8	anche	anche	B	B	_	10	mod	_	_
9	i	il	R	RD	num=p|gen=m	10	det	_	_
10	dati	dato	S	S	num=p|gen=m	3	obj	_	_
11	di	di	E	E	_	10	comp	_	_
12	superficie	superficie	S	S	num=s|gen=f	11	prep	_	_
13	espressi	esprimere	V	V	num=p|mod=p|gen=m	10	mod	_	_
14	in	in	E	E	_	13	comp	_	_
15	metri	metro	S	S	num=p|gen=m	14	prep	_	_
**16	quadri	quadro	S	S	num=p|gen=m	13	subj_pass	_	_
17	in	in	E	E	_	13	comp	_	_
18	conformità	conformità	S	S	num=n|gen=f	17	prep	_	_
19	alle	al	E	EA	num=p|gen=f	18	comp	_	_
	20 istruzioni	istruzione	S	S	num=p|gen=f	19	prep	_	_

**21	che	che	P	PR	num=n|gen=n	23	comp	_ _
	22 saranno	essere	V	VA	num=p|per=3|mod=i|ten=f	23	aux	_ _
	23 fornite	fornire	V	V	num=p|mod=p|gen=f	20	relcl	_ _
24	in	in	E	E	_	23	comp	_ _
25	seguito	seguito	S	S	num=s|gen=m	24	prep	_ _
26	,	,	F	FF	_	3	punc	_ _
27	poiché	poiché	C	CS	_	3	mod	_ _
**28	questo	questo	P	PD	num=s|gen=m	29	subj	_ _
29	sarà	essere	V	V	num=s|per=3|mod=i|ten=f	27	sub	_ _
30	in	in	E	E	_	29	pred	_ _
31	futuro	futuro	S	S	num=s|gen=m	30	prep	_ _
32	il	il	R	RD	num=s|gen=m	33	det	_ _
**33	parametro	parametro	S	S	num=s|gen=m	39	subj	_ _
	34 in	in	E	E	_	33	comp	_ _
	35 base	base	S	S	num=s|gen=f	34	concat	_ _
	36 al	al	E	EA	num=s|gen=m	35	concat	_ _
**37	quale	quale	P	PR	_	36	prep	_ _
	38 sarà	essere	V	V	num=s|per=3|mod=i|ten=f	39	aux	_ _
	39 decretato	decretare	V	V	num=s|mod=p|gen=m	29	conj	_ _
40	l'	il	R	RD	num=s|gen=n	41	det	_ _
**41	esborso	esborso	S	S	num=s|gen=m	39	subj_pass	_ _
42	del	di	E	EA	num=s|gen=m	41	comp	_ _
43	contribuente	contribuente	S	S	num=s|gen=n	42	prep	_ _
44	al	al	E	EA	num=s|gen=m	41	comp	_ _
45	posto	posto	S	S	num=s|gen=m	44	prep	_ _
46	dei	di	E	EA	num=p|gen=m	45	comp	_ _
47	<	<	N	N	_	48	mod	_ _
	48 vani	vano	S	S	num=p|gen=m	46	prep	_ _
	49 utili	utile	A	A	num=p|gen=n	50	mod	_ _
	50 >	>	S	S	num=p|gen=m	48	mod	_ _
**51	che	che	P	PR	num=n|gen=n	52	comp	_ _
	52 andranno	andare	V	V	num=p|per=3|mod=i|ten=f	48	relcl	_ _
53	in	in	E	E	_	52	comp	_ _
54	soffitta	soffitta	S	S	num=s|gen=f	53	prep	_ _
55	.	.	F	FS	_	1	punc	_ _

APPENDIX 2. TULETUT parser output of sentence (1)

1 E (E CONJ COORD COORD) [0;TOP-CONJ]
2 dovranno (DOVERE VERB MOD IND FUT INTRANS 3 PL) [1;COORD2ND]
2.10 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [2;VERB-SUBJ]
3 riportare (RIPORTARE VERB MAIN INFINITE PRES TRANS) [2;VERB+MODAL-INDCOMPL]
3.10 t [2.10f] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [3;VERB-SUBJ]
4 per (PER PREP MONO) [3;RMOD]
5 ogni (OGNI ADJ INDEF ALLVAL SING) [4;PREP-ARG]
6 unità (UNITà NOUN COMMON F ALLVAL) [5;DET+QUANTIF-ARG]
7 urbana (URBANO ADJ QUALIF F SING) [6;ADJC+QUALIF-RMOD]
8 anche (ANCHE ADV CONCESS) [9;ADVB+CONCESS-RMOD]
9 i (IL ART DEF M PL) [3;VERB-OBJ]
10 dati (DATO NOUN COMMON M PL) [9;DET+DEF-ARG]
11 di (DI PREP MONO) [10;PREP-RMOD]
12 superficie (SUPERFICIE NOUN COMMON F SING) [11;PREP-ARG]
13 espressi (ESPRIMERE VERB MAIN PARTICIPLE PAST TRANS PL M) [10;VERB-RMOD+RELCL+REDUC]
13.10 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [13;VERB-OBJ/VERB-SUBJ]
13.11 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [13;VERB-SUBJ/VERB-INDCOMPL-AGENT]
14 in (IN PREP MONO) [13;RMOD]
15 metri (METRO NOUN COMMON M PL) [14;PREP-ARG]
16 quadri (QUADRO ADJ QUALIF M PL) [15;ADJC+QUALIF-RMOD]
17 in (IN PREP MONO) [13;RMOD]
18 conformità (CONFORMITà NOUN COMMON M SING) [17;PREP-ARG]
19 alle (A PREP MONO) [13;RMOD]
19.1 alle (IL ART DEF F PL) [19;PREP-ARG]
20 istruzioni (ISTRUZIONE NOUN COMMON F PL) [19.1;DET+DEF-ARG]
21 che (CHE PRON RELAT ALLVAL ALLVAL LSUBJ+LOBJ) [23;VERB-OBJ/VERB-SUBJ]
22 saranno (ESSERE VERB AUX IND FUT INTRANS 3 PL) [23;AUX]
23 fornite (FORNIRE VERB MAIN PARTICIPLE PAST TRANS PL F) [20;VERB-RMOD+RELCL]
23.10 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [23;VERB-SUBJ/VERB-INDCOMPL-AGENT]
23.11 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [23;VERB-INDOBJ]
24 in (IN PREP MONO) [23;RMOD]
25 seguito (SEGUITO NOUN COMMON M SING) [24;PREP-ARG]
26 , (#\, PUNCT) [23;SEPARATOR]
27 poiché (POICHé‰ CONJ SUBORD CAUS) [23;RMOD]

28 questo (QUESTO PRON DEMONS M SING LSUBJ+LOBJ+OBL) [29;VERB-SUBJ]
29 sarà (ESSERE VERB MAIN IND FUT INTRANS 3 SING) [27;CONJ-ARG]
30 in (IN PREP MONO) [29;RMOD]
31 futuro (FUTURO NOUN COMMON M SING) [30;PREP-ARG]
32 il (IL ART DEF M SING) [29;VERB-PREDCOMPL+SUBJ]
33 parametro (PARAMETRO NOUN COMMON M SING) [32;DET+DEF-ARG]
34 in (IN_BASE_A PREP POLI LOCUTION) [29;RMOD]
35 base (IN_BASE_A PREP POLI LOCUTION) [34;CONTIN+LOCUT]
36 al (IN_BASE_A PREP POLI LOCUTION) [35;CONTIN+LOCUT]
36.1 al (IL ART DEF M SING) [34;PREP-ARG]
** 37 quale (QUALE PRON RELAT ALLVAL SING 3 LSUBJ+LOBJ+OBL) [36.1;DET+DEF-ARG]
38 sarà (ESSERE VERB AUX IND FUT INTRANS 3 SING) [39;AUX]
39 decretato (DECRETARE VERB MAIN PARTICIPLE PAST TRANS SING M) [33;VERB-RMOD+RELCL]
39.10 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [39;VERB-SUBJ/VERB-INDCOMPL-AGENT]
40 l' (IL ART DEF M SING) [39;VERB-OBJ/VERB-SUBJ]
41 esborso (ESBORSO NOUN COMMON M SING) [40;DET+DEF-ARG]
42 del (DI PREP MONO) [41;PREP-RMOD]
42.1 del (IL ART DEF M SING) [42;PREP-ARG]
43 contribuente (CONTRIBUENTE NOUN COMMON ALLVAL SING) [42.1;DET+DEF-ARG]
44 al (AL_POSTO_DI PREP POLI LOCUTION) [39;RMOD]
45 posto (AL_POSTO_DI PREP POLI LOCUTION) [44;CONTIN+LOCUT]
46 dei (AL_POSTO_DI PREP POLI LOCUTION) [45;CONTIN+LOCUT]
46.1 dei (IL ART DEF M PL) [44;PREP-ARG]
47 < (#\< PUNCT) [46.1;SEPARATOR]
48 vani (VANO NOUN COMMON M PL) [46.1;DET+DEF-ARG]
49 utili (UTILE ADJ QUALIF ALLVAL PL) [48;ADJC+QUALIF-RMOD]
50 > (#\> PUNCT) [48;SEPARATOR]
51 che (CHE PRON RELAT ALLVAL ALLVAL LSUBJ+LOBJ) [52;VERB-SUBJ]
52 andranno (ANDARE VERB MAIN IND FUT INTRANS 3 PL) [48;VERB-RMOD+RELCL]
53 in (IN PREP MONO) [52;VERB-INDCOMPL-LOC+TO]
54 soffitta (SOFFITTA NOUN COMMON F SING) [53;PREP-ARG]
55 . (#\. PUNCT) [1;END]

10.12871/CLICIT2014131

Analisi Linguistica e Stilostatistica – Uno Studio Predittivo sul Campo

Rodolfo Delmonte
Dipartimento di Studi Linguistici e Culturali Comparati
Ca' Bembo – Dorsoduro 1715
Università Ca' Foscari – 30123 Venezia
Email: delmont@unive.it

Abstract

Italiano. In questo lavoro presentiamo uno studio sul campo per definire uno schema di valutazione preciso per la stilistica del testo che è stato usato per stabilire una graduatoria di diversi documenti sulla base della loro abilità di persuasione e facilità di lettura. Lo studio concerne i documenti dei programmi politici pubblicati su un forum pubblico dai candidati a Rettore dell'Università Ca' Foscari – Venezia. I documenti sono stati analizzati dal nostro sistema ed è stata creata una graduatoria sulla base di punteggi associati a undici parametri. Dopo la votazione, abbiamo creato la graduatoria e abbiamo scoperto che il sistema aveva previsto il nome del reale vincitore in anticipo. I risultati sono apparsi su un giornale locale[1].

English. *This paper presents a case study defining a precise evaluation scheme for text stylistics to be used to rank different documents in terms of persuasiveness and easyness of reading. The study concerns political program documents published on a public forum by candidates to rector of the University Ca' Foscari – Venice. The documents have been analysed by our system and a rank list has been created on the basis of scores associated to eleven parameters. After voting has taken place, we graded the different analyses and discovered that the system had predicted the name of the actual winner in advance. The result has been published on a local newspaper.*

1. Introduzione

L'analisi parte dall'idea che lo stile di un documento programmatico sia composto da elementi quantitativi a livello di parola, da elementi derivati dall'uso frequente di certe strutture sintattiche nonché da caratteristiche squisitamente semantiche e pragmatiche come l'utilizzo di parole e concetti che ispirano positività. Ho eseguito l'analisi partendo dai testi disponibili su web o ricevuti dai candidati, utilizzando una serie di parametri che ho creato per l'analisi del discorso politico nei quotidiani italiani durante la penultima e ultima crisi di governo. I risultati sono pubblicati in alcuni lavori a livello nazionale e internazionale che ho elencato in una breve bibliografia. L'analisi utilizza dati quantitativi classici come il rapporto types/tokens e poi introduce informazioni derivate dal sistema GETARUNS che compie un parsing completo dei testi dal punto di vista sintattico, semantico e pragmatico. I dati riportati nelle tabelle sono derivati dai file di output del sistema. Il sistema produce un file per ogni frase, un file complessivo per l'analisi semantica del testo e un file con la versione verticalizzata del testo analizzato dove ogni parola è accompagnata da una classificazione sintattico-semantica-pragmatica. Il sistema è composto da un parser a reti di transizione aumentate da informazioni di sottocategorizzazione, che costruisce prima i chunks e poi a cascata le strutture a costituenti complesse più alte fino a quella di frase. Questa rappresentazione viene passata a un altro parser che lavora a isole, partendo da ciascun complesso verbale, corrispondente al costituente verbale. Il parser a isole individua la struttura predicato-argomentale, includendo anche gli aggiunti sulla base delle informazioni contenuto in un lessico di sottocategorizzazione per l'italiano costruito in precedenti progetti, contenente circa 50mila entrate verbali e aggettivali a diversi livelli di profondità. Viene utilizzata anche una lista di preferenze di selezione per verbi, nomi e aggettivi ricavata dai treebanks di italiano disponibili e contenente circa 30mila entrate. Riportiamo in Tabella 1. i dati in numeri assoluti.

[1] "Il mio sondaggio aveva già dato la vittoria a Bugliesi – I risultati di una ricerca di un docente di Linguistica" - Il Gazzettino – VeneziaMestre, mercoledì 11.06.2014,p.7.

Candidato	Tokens	Types	Rare words	Frasi	Little Pro	Strutture Proposiz.
Bertinetti	4992	1561	1341	162	200	585
Brugiavini	2841	987	852	91	119	308
Bugliesi	13210	2483	1899	463	541	1232
Cardinaletti	5346	1479	1243	167	159	469
LiCalzi	14376	3120	2516	769	720	1624

Tabella 1. Dati assoluti dei testi analizzati

2. I risultati dell'analisi

Mostriamo in questa sezione i risultati comparati dell'analisi su tutti i livelli linguistici. Commentiamo ogni grafico descrivendo il contenuto in forma verbale senza fornire alcuna valutazione oggettiva, cosa questa che faremo in una sezione finale del lavoro. Questi risultati sono quelli resi pubblici sul forum dei candidati rettore prima dell'elezione. In Fig.1 nell'Appendice, il grafico associato ai dati sintattico-semantici. I dati sono riportati in frequenze relative in modo da poter essere confrontabili, visto che i testi dei programmi sono di lunghezza diversa. Partendo dall'alto il parametro NullSubject misura le frasi semplici a verbo flesso (quindi non quelle a tempo indefinito come le infinitive) che non hanno soggetto espresso. Il secondo parametro misura le frasi che esprimono un punto di vista soggettivo (sono quindi rette da un verbo del tipo di "pensare, credere" ecc.). Il terzo parametro individua le frasi semplici o clausole o proposizioni a polarità negativa, che quindi contengono una negazione a livello verbale (un NON, ma anche MAI ecc.). In questo caso, sono considerate non a polarità negativa le frasi rette da un verbo con significato negativo lessicalizzato accompagnate dalla negazione. Mi riferisco a verbi del tipo di "distruggere, rifiutare, odiare", ecc.

Il quarto parametro misura le frasi non fattive o non fattuali, che cioè non descrivono un fatto avvenuto o che esiste nel mondo. Queste frasi sono rette da verbi di modo irreale (come il condizionale, il congiuntivo, ma anche dal tempo futuro) o sono in forma non dichiarativa, come le domande o le imperative. Come si può evincere dal grafico, Cardinaletti e Bugliesi hanno il maggior numero di frasi non fattive. Invece LiCalzi fa un uso più elevato di frasi a soggetto nullo assieme a Bugliesi, e di frasi soggettive.

Nel secondo grafico (vedi Fig.2 in Appendice) sono analizzati gli aspetti affettivi – questa analisi è chiamata Sentiment Analysis, e contiene anche dati semantici sulla complessità testuale. Sulla base di un lessico specializzato per l'italiano si contano le parole che hanno "prevalentemente" un valore negativo vs. positivo. Il lessico è composto da SentiWordNet (Esuli, Sebastiani, 2006), opportunamente corretto per tutte le parole con valore ambiguo; e contiene un lessico specializzato di circa 70mila entrate prodotto manualmente da me sulla base dell'analisi di testi di giornale per 1 milione di tokens. Le parole con valore neutro non vengono prese in considerazione. Un terzo parametro è quello dell'uso della diatesi passiva, che ha come funzione testuale di permettere la cancellazione dell'agente per far risaltare l'oggetto del verbo e trasformarlo in Argomento principale (o Topic) del discorso. Come si evince dal grafico, il numero maggiore di parole positive è di Brugiavini e Bugliesi, mentre il maggior numero di parole negative è di LiCalzi e Brugiavini. Bugliesi ne utilizza meno di tutti. Per quanto riguarda la forma passiva di nuovo Bugliesi è quello che ne usa di meno, invece Cardinaletti ne usa più di tutti. Per quanto riguarda la complessità, viene riportata la proporzione di proposizioni semantiche per frase, includendo in questo frasi semplici, clausole o complessi predicativi composti da un verbo a tempo indefinito e suoi argomenti. La maggior complessità spetta a Brugiavini e Bertinetti. LiCalzi ha quella più bassa.

Nel terzo grafico (Fig.3 in Appendice) si mostrano dati quantitativi della Vocabulary Richness (VR) in basso, derivati dal conteggio del numero di occorrenze di forme di parola singole chiamate Types, rispetto al totale delle occorrenze chiamate Tokens (queste includono anche la punteggiatura), e indicate dall'abbreviazione TT. La formula in alto invece rappresenta il rapporto che interviene tra i Types e le Rare Words (RW), che sono tutte le forme di parola che ricorrono una volta, due volte e tre volte nel testo, e sono anche chiamate Hapax, Dis e Tris Legomena. I dati rappresentati vedono Brugiavini e Bertinetti come quelli con la più alta ricchezza di vocabolario, e Bugliesi con i valori più bassi. Il rapporto Tokens/Sentence ci dice che LiCalzi ha quello più basso seguito da Bugliesi, mentre gli altri tre testi sono più o meno allo stesso livello.

Per studiare meglio nel dettaglio i concetti che hanno caratterizzato i vari programmi abbiamo

quindi fatto ricorso a una comparazione delle Rank List ricavate dalle liste di frequenza – che non possiamo qui includere per mancanza di spazio. La Rank List è la lista delle parole Types fatta sulla base della loro frequenza. La posizione nella lista indica la rilevanza che la parola assume all'interno del testo. Benché le frequenze assolute siano diverse da testo a testo, la posizione nella rank list permette di valutare le differenze/somiglianze tra testi diversi nella utilizzazione di certe parole chiave.
Tutti i candidati hanno la stessa parola all'inizio della rank list, "ateneo". Anche le posizioni reciproche di "ricerca" e "didattica" e "studenti" sono molto vicine e sono rispettivamente in terza posizione "ricerca" (seconda per Cardinaletti), e in quarta posizione "didattica" (seconda Bertinetti e quinta Bugliesi). Poi le liste si differenziano: la parola "dipartimenti" viene trattata in maniera diversa da LiCalzi che la posiziona molto in alto, mentre Bertinetti la posiziona in basso e in Brugiavini non la si ritrova nelle prime 30. La parola "personale" appare nei testi dei primi tre candidati ma non in quelli di Bertinetti e Brugiavini. Lo stesso dicasi per "valutazione". Invece per quanto riguarda la parola "lavoro" vediamo che essa risulta in alto nella lista dei due candidati Brugiavini e Bertinetti, a differenza di quanto avviene nelle liste degli altri tre candidati dove si trova spostata in basso. Tornerò al contenuto della Rank List più avanti.

3. Calcolo della Correlazione

Infine ho eseguito il calcolo della correlazione tra i vari candidati. Ho utilizzato i vettori delle 11 feature presentate prima, in valori assoluti come parametri di confronto. Nella valutazione, ho considerato solo i casi in cui l'indice R supera 0.998. Il risultato più alto è stato ottenuto dal confronto Brugiavini e Bertinetti, seguono LiCalzi e Bugliesi.

1. ***Brugiavini/Bertinetti***
R = 0.9988378753376379.
2. ***Bugliesi/LiCalzi***
R = 0.9988321771943326.

I valori della ***Cardinaletti*** sono risultati vicini solo a quelli di ***Brugiavini***.

R = 0.9961624024306578.

4. Analisi Semantica Dettagliata di un concetto: PERSONALE

Ho verificato nel dettaglio i dati relativi al concetto PERSONALE che indico in basso. I dati sono limitati ai quattro documenti dei candidati che parlano di PERSONALE in maniera consistente. Abbiamo escluso dal conteggio i due candidati Bertinetti e Brugiavini perché i numeri assoluti nel loro caso sono così esigui rispetto al numero complessivo di TYPES che ricadono nelle cosiddette Rare Words, cioè le parole utilizzate come Hapax, Dis o Trislegomena. Queste parole fanno parte della coda della distribuzione e non contribuiscono a caratterizzare il testo.
Ho contato le volte che la parola viene utilizzata come Nome, come Aggettivo, e come parte della Forma Polirematica Personale_Docente.

	Nome	Aggettivo	Multiword	Totale
LiCalzi	22	4	5	17
Cardin.	11	2	4	7
Bugliesi	37	2	5	32

Tabella 1. Utilizzo del concetto *Personale*

Se si considera quindi che il significato voluto della parola PERSONALE si ottiene solo quando è utilizzata come Nome escludendo le occorrenze dello stesso nome nella forma polirematica, abbiamo Bugliesi primo, seguito da LiCalzi e Cardinaletti.
Nei testi però, si utilizzano descrizioni linguistiche diverse per riferirsi alla stessa entità - persona, organizzazione, località o istituzione. Per quanto riguarda l'uso di coreferenti al concetto PERSONALE abbiamo considerato i due iponimi, PTA e CEL, di cui elenchiamo le seguenti quantità assolute e relative:

- Cardinaletti 8 0.72
- LiCalzi 7 0.32
- Bugliesi 6 0.16

Per ricavare valori relativi, abbiamo fatto la proporzione tra l'uso del riferimento generico PERSONALE e i suoi iponimi. Si conferma l'ordine sulla base dei dati assoluti.

5. Valutazione e Conclusione

Volendo fare una graduatoria complessiva, si può considerare che ciascun parametro possa avere valore positivo o negativo. Nel caso fosse

positivo la persona con la quantità maggiore si vedrà assegnare come ricompensa il valore 5 e gli altri a scalare un valore inferiore di un punto, fino al valore 1. Nel caso invece che il parametro avesse valenza negativa, il candidato con la quantità maggiore riceverà al contrario il punteggio inferiore di 1 e gli altri a scalare un valore superiore di un punto fino a 5. La graduatoria complessiva verrà quindi stilata facendo la somma di tutti i punteggi singoli ottenuti. L'assegnazione della polarità a ciascun parametro segue criteri linguistici e stilistici, ed è la seguente:

1. NullSubject - positive: La maggior quantità di soggetti nulli indica la volontà di creare un testo molto coeso e di non sovraccaricare il riferimento alla stessa entità con forme ripetute o coreferenti.
2. Subjective Props - negative: La maggior quantità di proposizioni che esprimono un contenuto soggettivo indica la tendenza da parte del soggetto di esporre le proprie idee in maniera non oggettiva.
3. Negative Props - negative: Il maggior uso di proposizioni negative, cioè con l'utilizzo della negazione o di avverbi negativi, è un tratto stilistico che non è propositivo ma tende a contrastare quanto affermato o fatto da altri.
4. Nonfactive Props - negative: L'utilizzo di proposizioni non fattive indica la tendenza stilistica ad esporre le proprie idee utilizzando tempi e modi verbali irreali - congiuntivo, condizionale, futuro e tempi indefiniti.
5. Props / Sents - negative: Il rapporto che indica il numero di proposizioni per frase viene considerato in maniera negativa a significare che più è elevato maggiore è la complessità dello stile.
6. Negative Ws - negative: Il numero di parole negative utilizzate in proporzione al numero totale di parole ha un valore negativo.
7. Positive Ws - positive: Il numero di parole positive utilizzate in proporzione al numero totale di parole ha un valore positivo.
8. Passive Diath - negative: Il numero di forme passive utilizzate viene considerato in maniera negativa in quanto oscura l'agente dell'azione descritta.
9. Token / Sents - negative: Il numero di token in rapporto alle frasi espresse viene trattato come fattore negativo di nuovo in riferimento al problema della complessità indotta.
10. Vr - Rw - negative: Questa misura considera la ricchezza di vocabolario sulla base delle cosiddette RareWords, o numero complessivo di Hapax/Dis/Tris Legomena nella Rank List. Maggiori sono le parole uniche o poco frequenti più lo stile è complesso.
11. Vr - Tt - negative: Come sopra, questa volta considerando il numero totale dei Tipi.
L'assegnazione del punteggio sulla base dei criteri indicati definisce la seguente graduatoria finale:

Bugliesi	47
LiCalzi	36
Brugiavini	28
Cardinaletti	27
Bertinetti	27

Tabella 2. Graduatoria finale sulla base degli 11 parametri (vedi Tab. 2.1 in Appendice 2)

Volendo includere anche i punteggi relativi all'uso di PERSONALE e dei suoi iponimi avremo questo risultato complessivo:

Bugliesi	53
LiCalzi	44
Brugiavini	37
Cardinaletti	31
Bertinetti	30

Tabella 3. Graduatoria finale sulla base dei 13 parametri (vedi Tab. 3.1 in Appendice 2)

Utilizzando i parametri come elementi di giudizio per classificare lo stile dei candidati e assegnando una valutazione a parole, si ottengono i due giudizi sottostanti.
1. Bugliesi ha vinto perché ha utilizzato uno stile più coeso, con un vocabolario più semplice, delle strutture sintattiche semplici e dirette, esprimendo i contenuti in maniera concreta e fattuale, parlando a tutti i livelli di parti interessate, docenti e non docenti. Inoltre ha utilizzato meno espressioni e frasi negative e più espressioni positive.
I dati ci dicono anche che il programma di Bugliesi è in forte correlazione con quello di LiCalzi ma non con quello degli altri candidati.
2.Cardinaletti ha scritto un programma che utilizza uno stile poco coeso, con un vocabolario alquanto elaborato, con strutture sintattiche abbastanza più complesse, esprimendo i contenuti in maniera molto meno concreta e molto meno fattuale, parlando a tutti i livelli di parti interessate, docenti e non docenti. Inoltre ha utilizzato poche espressioni e frasi negative e relativamente poche espressioni positive. Infine il programma della Cardinaletti è in buona correlazione con il programma della Brugiavini.

Bibliografia

Delmonte, R., 2013, Extracting Opinion and Factivity from Italian political discourse, in B. Sharp, M. Zock, (eds), Proceedings 10th International Workshop NLPCS, Natural Language Processing and Cognitive Science, 162-176, Marseille.

Delmonte, R., 2012. Predicate Argument Structures for Information Extraction from Dependency Representations: Null Elements are Missing, 2013. in C. Lai, A. Giuliani and G. Semeraro (eds.), DART 2012: Revised and Invited Papers, "Studies in Computational Intelligence", Springer Verlag, 1-25.

Delmonte, R., Daniela Gifu, Rocco Tripodi, 2013. Opinion and Factivity Analysis of Italian political discourse, in R. Basili, F. Sebastiani, G. Semeraro (eds.), Proc. 4th Italian Information Retrieval Workshop, IIR2013, Pisa. CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org, vol. 964, 88-99.

Delmonte, R., D. Gifu, R. Tripodi, 2012. A Linguistically-Based Analyzer of Print Press Discourse, International Conference on Corpus Linguistics, Saint Petersburg, at http://corpora.phil.spbu.ru/talks2013.

Delmonte R. & Daniela Gîfu, 2013. Opinion and Sentiment Analysis of Italian Print Press, in International Journal Of Advanced Computer And Communication Technology (IJACCT), Vol. 1, Issue. 1, 24-38, http://www.scribd.com/doc/.

Delmonte R. & Vincenzo Pallotta, 2011. Opinion Mining and Sentiment Analysis Need Text Understanding, in Pallotta, V., Soro, A., Vargiu, E. (Eds.), "Advances in Distributed Agent-based Retrieval Tools: Studies in Computational Intelligence, Vol. 361, Springer, 81-96.

Delmonte, R., 2010. Keynote Speaker, OPINION MINING, SUBJECTIVITY and FACTUALITY, ALTA (Australasian Language Technology Association) Workshop, ICT University of Melbourne,http://www.alta.asn.au/events/alta2010 / alta-2010-program.html.

Esuli A, Sebastiani F, 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of LREC 2006 – 5th Conference on Language Resources and Evaluation.

APPENDICE 1.

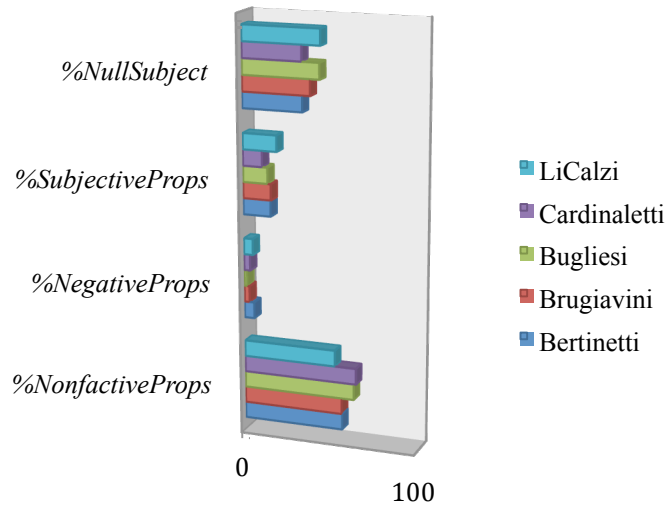


Figura 1. Dati Sintattico-Semantici.

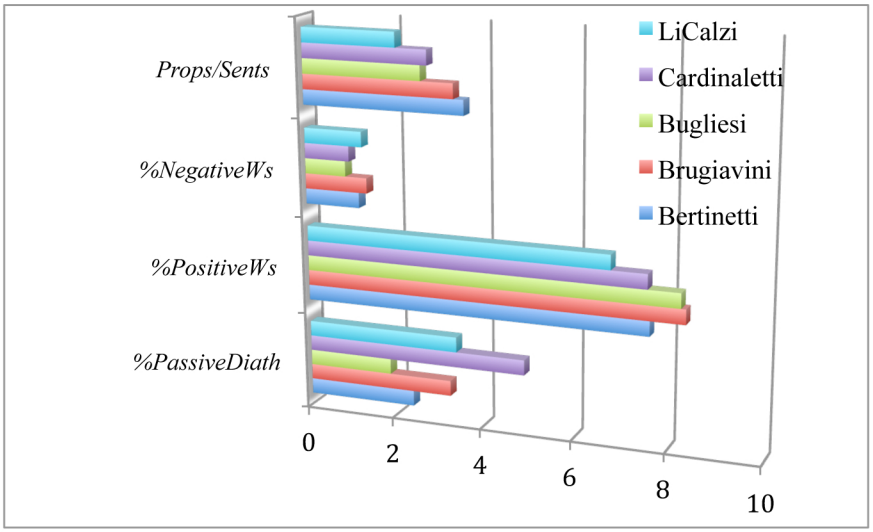


Figura 2. Dati Affettivi e Semantici.

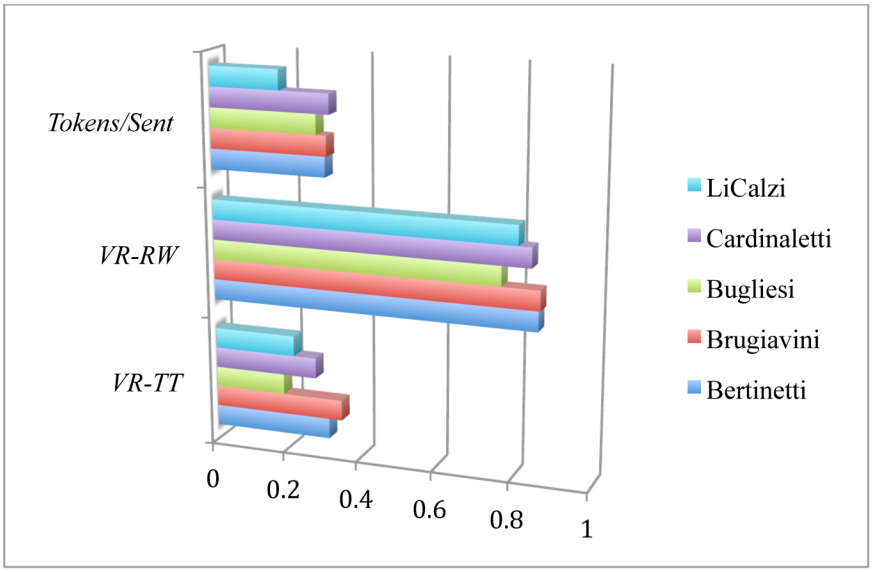


Figura 3. Dati Quantitativi.

APPENDICE 2

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	Totale
Bugliesi	4	4	5	2	4	5	4	5	4	5	5	47
LiCalzi	5	1	2	5	5	2	1	2	5	4	4	36
Brugiavini	3	2	4	4	2	1	5	3	2	1	1	28
Cardinaletti	2	5	3	1	3	4	2	1	1	3	3	27
Bertinetti	1	3	1	3	1	3	3	4	3	2	2	27

Tabella 2.1 Graduatoria finale sulla base degli 11 parametri.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	Totale
Bugliesi	4	4	5	2	4	5	4	5	4	5	5	3	3	53
LiCalzi	5	1	2	5	5	2	1	2	5	4	4	4	4	44
Cardinaletti	2	5	3	1	3	4	2	1	1	3	3	5	5	37
Brugiavini	3	2	4	4	2	1	5	3	2	1	1	2	1	31
Bertinetti	1	3	1	3	1	3	3	4	3	2	2	1	2	30

Tabella 3.1 Graduatoria finale sulla base dei 13 parametri.

10.12871/CLICIT2014132

An adaptable morphological parser for agglutinative languages

Marina Ermolaeva
Lomonosov Moscow State University
marinkaermolaeva@mail.ru

Abstract

English. The paper reports the state of the ongoing work on creating an adaptable morphological parser for various agglutinative languages. A hybrid approach involving methods typically used for non-agglutinative languages is proposed. We explain the design of a working prototype for inflectional nominal morphology and demonstrate its work with an implementation for Turkish language. An additional experiment of adapting the parser to Buryat (Mongolic family) is discussed.

Italiano. *Il presente articolo riporta lo stato dei lavori nel corso della creazione di un parser morfologico adattabile per diverse lingue agglutinanti. Proponiamo un approccio ibrido che coinvolge i metodi tipicamente utilizzati per le lingue non-agglutinanti. Spieghiamo lo schema di un prototipo funzionante per la flessione morfologica nominale e dimostriamo il suo funzionamento con un'implementazione per la lingua turca. Infine viene discusso un ulteriore esperimento che consiste nell'adattare il parser alla lingua buriata (la famiglia mongolica).*

1 Introduction

The most obvious way to perform morphological parsing is to make a list of all possible morphological variants of each word. This method has been successfully used for non-agglutinative languages, e.g. (Segalovich 2003) for Russian, Polish and English.

Agglutinative languages pose a much more complex task, since the number of possible forms of a single word is theoretically infinite (Jurafsky and Martin 2000). Parsing languages like Turkish often involves designing complicated finite-state machines where each transition corresponds to a single affix (Hankamer 1986; Eryiğit and Adalı 2004; Çöltekin 2010; Sak et al. 2009; Sahin et al. 2013). While these systems can perform extremely well, a considerable redesigning of the whole system is required in order to implement a new language or to take care of a few more affixes.

The proposed approach combines both methods mentioned above. A simple finite-state machine allows to split up the set of possible affixes, producing a finite and relatively small set of sequences that can be easily stored in a dictionary.

Most systems created for parsing agglutinative languages, starting with (Hankamer 1986) and (Oflazer 1994), process words from left to right: first stem candidates are found in a lexicon, then the remaining part is analyzed. The system presented in this paper applies the right-to-left method (cf. (Eryiğit and Adalı 2004)): affixes are found in the first place. It can ultimately work without a lexicon, in which case the remaining part of the word is assumed to be the stem; to improve precision of parsing, it is possible to compare it to stems contained in a lexicon. A major advantage of right-to-left parsing is the ability to process words with unknown stems without additional computations.

Multi-language systems (Akın and Akın 2007; Arkhangelskiy 2012) are a relatively new tendency. With the hybrid approach mentioned above, the proposed system fits within this trend. As the research is still in progress, the working prototype of the parser (written in Python language) is currently restricted to nominal inflectional morphology. Within this scope, it has been implemented for Turkish; an additional experiment with Buryat language is discussed in the section 5.

2 Turkish challenges

The complexity of Turkish morphology is easily perceptible in nouns. The word stem itself can be complex. Compounding of "adjective + noun" or "noun + noun" structure is a productive way of

word formation, which means that this problem cannot be solved by listing all known compounds in a dictionary.

Due to the vowel harmony and assimilation rules, most affixes have multiple allomorphs distributed complementarily according to the phonological context; e.g. the locative case marker has 4 forms (two harmonic variants of the vowel and a voiced/voiceless alternation).

A nominal stem can receive number, possession and case affixes. Moreover, certain other affixes (e.g. copular and person markers) can attach to these forms to form predicates:

(1) ev-ler-imiz-de-ymiş-ler[1]
home-PL-P1PL-LOC-COP.EV-3PL
Apparently they are/were at our homes.

An interesting option is the affix *-ki*, which can be recursively attached to a nominal form containing a genitive or locative marker:

(2) ev-de-ki-ler-in-ki
home-LOC-KI1-PL-GEN-KI2[2]
the one belonging to those at home

3 System design

3.1 Data representation

The language-specific data necessary to implement a new language includes:

- Phonology description (phoneme inventory, harmony, etc.)
- Morphology description: a list of all allomorphs. For each allomorph its category, gloss and possible (morpho)phonological context is stored.
- Lexicon: a list of stems with part-of-speech tags. If a stem has multiple phonological variants, they are stored as separate entries along with data about contexts they can be used in. The lexicon is optional, yet it significantly improves precision of parsing.

The parser itself is language-independent and does not require any custom coding to implement new languages.

For Turkish, the system uses a relatively small lexicon of 16000 nominal and adjectival stems. The modest size of the lexicon is mostly compensated by the ability to analyze morphology even is the stem is absent in the lexicon. In this case, parses for all possible stems are output.

The exceedingly long morpheme sequences that can attach to a stem are split up into shorter chains. The whole set of grammatical categories is represented as a set of slots, each of them containing categories that have strictly fixed order(s):

- two stem slots (for nominal compounds)
- noun inflection
- noun loop (the recursive suffix *-ki*)
- nominal verb suffixes (e.g. copulas and adverbial markers)

The number and order of categories within slots can be changed without modifying the system itself, which simplifies implementing new languages.

For each slot, a list of possible affix sequences is obtained. At this step all the checks of morphotactic and phonological compatibility of the affixes within a slot are performed, so they do not have to be applied at runtime. The lists are converted into tries in order to speed up the search. All the sequences are stored inverted, so that the trie could be searched during the parsing process. A fragment of the nominal morphology trie and the sequences compatible with it are shown in Figure 1 and Table 1 respectively.

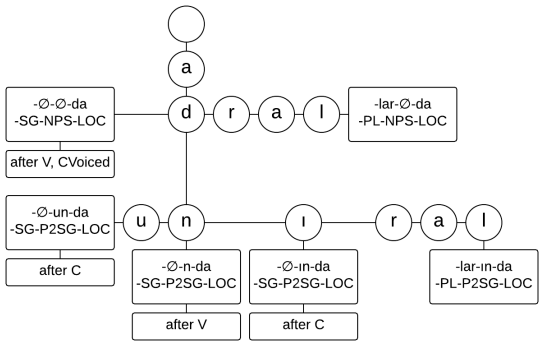


Figure 1. A fragment of the nominal affix trie

Sequence	Gloss	Context
-∅-∅-da	-SG-NPS-LOC	after vowels and voiced consonants
-∅-un-da	-SG-P2SG-LOC	after consonants
-∅-n-da	-SG-P2SG-LOC	after vowels
-lar-ın-da	-PL-P2SG-LOC	(no restrictions)
-∅-ın-da	-SG-P2SG-LOC	after consonants
-lar-∅-da	-PL-NPS-LOC	(no restrictions)

Table 1. Sequence list for Figure 1

Similarly, the lexicon is stored as a set of tries. Stems are also inverted, in order to effectively find stem boundaries within

[1] Examples (1)-(4) are from (Göksel and Kerslake 2005)

[2] According to Hankamer (2004), *-ki* has different properties when attached to a locative form and to a genitive form; therefore, two separate *-ki*'s are postulated. In this paper, they are referred to as KI1 and KI2 respectively.

compounds. Stems with multiple phonological variants are included in the lexicon as a set of separate entries; each entry receives special labels determining possible phonological context. For instance, *his* "sensation" appears in the form *hiss* before vowels and in the vocabulary form in other cases. A fragment of the lexicon trie is represented in Figure 2; it corresponds to the list of stems in Table 2.

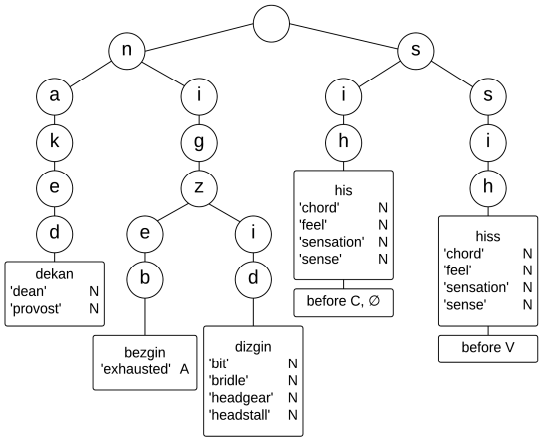


Figure 2. A fragment of the lexicon trie

Sequence	Translation(s)	Context
dekan	dean, provost	(no restrictions)
bezgin	exhausted	(no restrictions)
dizgin	bit, bridle, …	(no restrictions)
his	chord, feel, …	before consonants; at the word's end
hiss	chord, feel, …	before vowels

Table 2. Sequence list for Figure 2

3.2 Parsing algorithm

The transitions between slots are performed via a (very simple) finite-state machine shown in Figure 3:

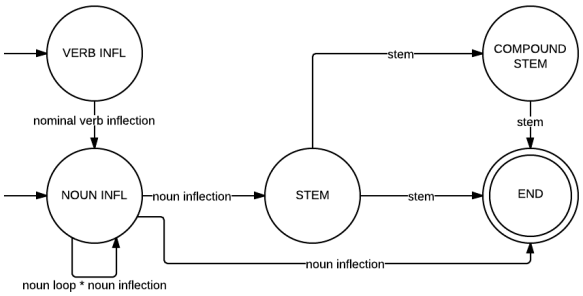


Figure 3. The finite-state machine

Each transition corresponds to a sequence of affixes rather than of a single affix; Each transition involves finding all possible candidate sequences using an appropriate stem or affix trie. Checks of compatibility are only done between slot sequences; at other points, no linguistic information is used. The simplified algorithm of analysis includes following steps:

1. Find all affix sequences that match the input word form.
2. For each hypothetical parse, try to find a stem in the lexicon using the unglossed part at the word's left end. If a stem is found and there are no "leftover" characters at the left end of the word, output all such parses. If a stem is found, yet some part of the word remains unglossed, go to step 3. If no stem is found at all, assume that the stem is unknown and output all hypothetical parses.
3. Assume that the stem is compound; for the remaining unglossed part, try to find another stem. If a stem is found and no unprocessed characters are left, output all such parses. Else discard the hypothetical compound parses and output all parses with no stem found.

Some examples of different decisions made by the algorithm are demonstrated below. In (3), the input is ambiguous. For two of the possible stem-affix boundaries (*adam-dı* and *ada-mdı*), a known stem has been found in the lexicon:

(3) **input**: *adamdı*
decision: single stem
output:
1. adam-Ø-Ø-Ø-dı-Ø
man-SG-NPS-NOM-COP.PST-3
2. ada-Ø-m-Ø-dı-Ø
island-SG-P1SG-NOM-COP.PST-3

Even if there is no single stem matching the input in the lexicon, like in (4), a suitable parse might be found under the assumption that there is an additional boundary within the stem:

(4) **input**: *kızarkadaş*
decision: compound
output:
1. kız-arkadaş-Ø-Ø-Ø
girl-friend-SG-NPS-NOM
2. kız-arkadaş-Ø-Ø-Ø-Ø-Ø
girl-friend-SG-NPS-NOM-COP.PRS-3

Finally, the pseudo-word in (5) has two feasible stem-affix boundaries (with hypothetical stems *fefe* and *fef*), but no single or compound match in the lexicon for any of them. The stem is considered unknown, and all parses are output:

(5) **input**: *fefe*
decision: unknown stem
output:
1. fef-Ø-Ø-e
FEF-SG-NPS-DAT

2. fef-∅-∅-e-∅-∅
   FEF-SG-NPS-DAT-COP.PRS-3
3. fefe-∅-∅-∅
   FEF-SG-NPS-NOM
4. fefe-∅-∅-∅-∅-∅
   FEF-SG-NPS-NOM-COP.PRS-3

4 Evaluation

Turkish is known for a significant level of morphological ambiguity. For example, it is impossible to disambiguate (6) and (7) without appealing to the context:

(6) ev-in
    house-GEN
    'of the house'

(7) ev-in
    house-P2SG
    'your house'

Since the system does not perform disambiguation, it must output all possible parses for each word. To take this into account, the evaluation method described in (Paroubek 2007) has been used. First, precision (P) and recall (R) values for each word w_i in the test sample are obtained:

$P(w_i) = \frac{t_i \cap r_i}{t_i}$; $R(w_i) = \frac{t_i \cap r_i}{r_i}$,

where t_i is the number of parses for w_i output by the parser and r_i is the number of correct parses.

After that, mean values for the whole sample are calculated. As most derivational affixes are currently not regarded, the internal structure of the stem was not considered. A parse was accepted if all inflectional affixes had been correctly found and properly labelled.

The Turkish implementation was evaluated with a testing sample of 300 nouns and noun-based predicates and yielded precision and recall values of 94,8% and 96,2% respectively.

5 Implementing new languages

Since Turkic languages are quite similar among themselves, applying the parser to a non-Turkic agglutinative language can help test its universality.

As an experiment, a small part of Buryat morphology has been modelled. Like Turkish, Buryat language poses more challenges than Turkish in some respects. The processing is complicated by a vast number of (mor)phonological variants of both stems and affixes, more complex phonological rules and a harmony system with subtler distinctions (e.g. a distinction between vowels in different syllables).

Crucially, the Buryat implementation did not require any custom coding or language-specific modifications of the parser itself; the only custom elements were phonology description, morpheme list and dictionary. The morphology model was evaluated on a small sample of Buryat nouns, resulting in precision value of approximately 91% and recall value of 96%.

6 Future work

At the moment, the top-importance task is lifting the temporary limitations of the parser by implementing other parts of speech (finite and non-finite verb forms, pronouns, postpositions etc.) and derivational suffixes.

Although the slot system described in 3.1 has been sufficient for both Turkish and Buryat, other agglutinative languages may require more flexibility. This can be achieved either by adding more slots (thus making the slot system nearly universal) or by providing a way to derive the slot system automatically, from plain text or a corpus of tagged texts; the latter solution would also considerably reduce the amount of work that has to be done manually.

Another direction of future work involves integrating the parser into a more complex system. DIRETRA, an engine for Turkish-to-English direct translation, is being developed on the base of the parser. The primary goal is to provide a word-for-word translation of a given text, reflecting the morphological phenomena of the source language as precisely as possible. The gloss lines output by the parser are processed by the other modules of the system and ultimately transformed into text representations in the target language:

input	adamlarınkiler
parser output	man-PL-GEN-KI2-PL
DIRETRA output	ones.owned.by.men

Table 3. An example of DIRETRA output

Though the system is being designed for Turkish, the next step planned is to implement other Turkic languages as well.

Abbreviations

1 – first person, 2 – second person, 3 – third person, COP.EV – evidential copula, COP.PRS – present tense copula, COP.PST – past tense copula, DAT – dative, GEN – genitive, KI1 – -ki suffix after locative,

KI2 – -ki suffix after genitive, LOC – locative, NOM – nominative, NPS – non-possession, P – possession, PL – plural, SG – singular.

References

Ahmet Afşın Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source NLP framework for Turkic Languages.

Timofey Arkhangelskiy. 2012. Printsipy postrojenija morfologicheskogo parsera dlja raznostrukturnyx jazykov [Principles of building a morphological parser for different-structure languages] Abstract of thesis cand. phil. sci. Moscow.

Çağrı Çöltekin. 2010. A Freely Available Morphological Analyzer for Turkish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.

Gülşen Eryiğit and Eşref Adalı. 2004. An Affix Stripping Morphological Analyzer for Turkish. In: IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, 299-304.

Aslı Göksel and Celia Kerslake. 2005. Turkish: A Comprehensive Grammar.

Jorge Hankamer. 1986. Finite state morphology and left-to-right phonology. In: Proceedings of the Fifth West Coast Conference on Formal Linguistics, Stanford, CA, 29-34.

Jorge Hankamer. 2004. Why there are two ki's in Turkish. In: Imer and Dogan, eds., Current Research in Turkish Linguistics, Eastern Mediterranean University Press, 13-25.

Daniel Jurafsky and James H. Martin. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Prentice Hall.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. In Literary and Linguistic Computing, vol. 9, no. 2, 137-148.

Patrick Paroubek. 2007. Chapter 4 - Evaluating Part Of Speech Tagging and Parsing. In: Evaluation of Text and Speech Systems, eds. Laila Dybkjær, Holmer Hemsen, Wolfgang Minker, series: Text, Speech and Language Technology, vol. 36, Kluwer Academic Publisher, 97-116.

Muhammet Şahin, Umut Sulubacak and Gülsen Eryigit. 2013. Redefinition Of Turkish Morphology Using Flag Diacritics. In: Proceedings of the Tenth Symposium on Natural Language Processing (SNLP-2013).

Haşim Sak, Tunga Güngör and Murat Saraçlar. 2009. A stochastic finite-state morphological parser for Turkish. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, August 04-04, 2009, Suntec, Singapore.

Ilya Segalovich. 2003. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. MLMTA, 273-280. CSREA Press.

10.12871/CLICIT2014133

Distributed Smoothed Tree Kernel

Lorenzo Ferrone
University of Rome "Tor Vergata"
Via del Politecnico 1
00133 Roma, Italy
lorenzo.ferrone@gmail.com

Fabio Massimo Zanzotto
University of Rome "Tor Vergata"
Via del Politecnico 1
00133 Roma, Italy
fabio.massimo.zanzotto@uniroma2.it

Abstract

English. In this paper we explore the possibility to merge the world of Compositional Distributional Semantic Models (CDSM) with Tree Kernels (TK). In particular, we will introduce a specific tree kernel (*smoothed tree kernel*, or STK) and then show that is possibile to approximate such kernel with the dot product of two vectors obtained compositionally from the sentences, creating in such a way a new CDSM.

Italiano. *In questo paper vogliamo esplorare la possibilità di unire il mondo dei metodi di semantica distribuzione composizionale (CDSM) con quello dei tree Kernel (TK). In particolare introdurremo un particolare tree kernel e poi mostreremo che possibile approssimare questo kernel tramite il prodotto scalare tra due vettori ottenuti composizionalmente a partire dalle frasi di partenza, creando così di fatto un nuovo modello di semantica distribuzionale composizionale.*

1 Introduction

Compositional distributional semantics is a flourishing research area that leverages distributional semantics (see Baroni and Lenci (2010)) to produce meaning of simple phrases and full sentences (hereafter called *text fragments*). The aim is to scale up the success of word-level relatedness detection to longer fragments of text. Determining similarity or relatedness among sentences is useful for many applications, such as multi-document summarization, recognizing textual entailment (Dagan et al., 2013), and semantic textual similarity detection (Agirre et al., 2013; Jurgens et al., 2014). Compositional distributional semantics models (CDSMs) are functions mapping text fragments to vectors (or higher-order tensors). Functions for simple phrases directly map distributional vectors of words to distributional vectors for the phrases (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Zanzotto et al., 2010). Functions for full sentences are generally defined as recursive functions over the ones for phrases (Socher et al., 2011). Distributional vectors for text fragments are then used as inner layers in neural networks, or to compute similarity among text fragments via dot product.

CDSMs generally exploit structured representations t^x of text fragments x to derive their meaning $f(t^x)$, but the structural information, although extremely important, is obfuscated in the final vectors. Structure and meaning can interact in unexpected ways when computing cosine similarity (or dot product) between vectors of two text fragments, as shown for full additive models in (Ferrone and Zanzotto, 2013).

Smoothed tree kernels (STK) (Croce et al., 2011) instead realize a clearer interaction between structural information and distributional meaning. STKs are specific realizations of convolution kernels (Haussler, 1999) where the similarity function is recursively (and, thus, compositionally) computed. Distributional vectors are used to represent word meaning in computing the similarity among nodes. STKs, however, are not considered part of the CDSMs family. As usual in kernel machines (Cristianini and Shawe-Taylor, 2000), STKs directly compute the similarity between two text fragments x and y over their tree representations t^x and t^y, that is, $STK(t^x, t^y)$. The function f that maps trees into vectors is

only implicitly used, and, thus, $STK(t^x, t^y)$ is not explicitly expressed as the dot product or the cosine between $f(t^x)$ and $f(t^y)$.

Such a function f, which is the underlying reproducing function of the kernel (Aronszajn, 1950), is a CDSM since it maps trees to vectors by using distributional meaning. However, the huge nality of $\mathbb{R}^n$ (since it has to represent the set of all possible subtrees) prevents to actually compute the function $f(t)$, which thus can only remain *implicit*.

Distributed tree kernels (DTK) (Zanzotto and Dell'Arciprete, 2012) partially solve the last problem. DTKs approximate standard tree kernels (such as (Collins and Duffy, 2002)) by defining an *explicit* function DT that maps trees to vectors in $\mathbb{R}^m$ where $m \ll n$ and $\mathbb{R}^n$ is the explicit space for tree kernels. DTKs approximate standard tree kernels (TK), that is, $\langle DT(t^x), DT(t^y)\rangle \approx TK(t^x, t^y)$, by approximating the corresponding reproducing function. Thus, these distributed trees are small vectors that encode structural information. In DTKs tree nodes u and v are represented by nearly orthonormal vectors, that is, vectors $\overrightarrow{u}$ and $\overrightarrow{v}$ such that $\langle \overrightarrow{u}, \overrightarrow{v}\rangle \approx \delta(\overrightarrow{u}, \overrightarrow{v})$ where δ is the Kroneker's delta. This is in contrast with distributional semantics vectors where $\langle \overrightarrow{u}, \overrightarrow{v}\rangle$ is allowed to be any value in $[0, 1]$ according to the similarity between the words v and u. In this paper, leveraging on distributed trees, we present a novel class of CDSMs that encode both structure and distributional meaning: the distributed smoothed trees (DST). DSTs carry structure and distributional meaning on a rank-2 tensor (a matrix): one dimension encodes the structure and one dimension encodes the meaning. By using DSTs to compute the similarity among sentences with a generalized dot product (or cosine), we implicitly define the distributed smoothed tree kernels (DSTK) which approximate the corresponding STKs. We present two DSTs along with the two smoothed tree kernels (STKs) that they approximate. We experiment with our DSTs to show that their generalized dot products approximate STKs by directly comparing the produced similarities and by comparing their performances on two tasks: recognizing textual entailment (RTE) and semantic similarity detection (STS). Both experiments show that the dot product on DSTs approximates STKs and, thus, DSTs encode both structural and distributional semantics of text fragments in tractable rank-2 tensors. Experiments on STS and RTE show that distributional semantics encoded in DSTs increases performance over structure-only kernels. DSTs are the first positive way of taking into account both structure and distributional meaning in CDSMs. The rest of the paper is organized as follows. Section 2.1 introduces the basic notation used in the paper. Section 2 describe our distributed smoothed trees as compositional distributional semantic models that can represent both structural and semantic information. Section 4 reports on the experiments. Finally, Section 5 draws some conclusions.

2 Distributed Smoothed Tree Kernel

We here propose a model that can be considered a compositional distributional semantic model as it transforms sentences into matrices that can then used by the learner as feature vectors. Our model is called *Distributed Smoothed Tree Kernel* (Ferrone and Zanzotto, 2014) as it mixes the distributed trees (Zanzotto and Dell'Arciprete, 2012) representing syntactic information with distributional semantic vectors representing semantic information.

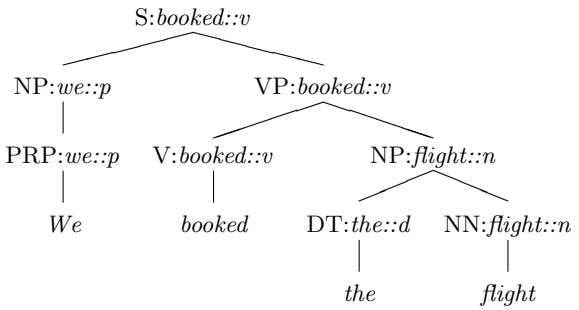


Figure 1: A lexicalized tree

2.1 Notation

Before describing the *distributed smoothed trees* (DST) we introduce a formal way to denote constituency-based *lexicalized parse trees*, as DSTs exploit this kind of data structures. *Lexicalized trees* are denoted with the letter t and $N(t)$ denotes the set of non terminal nodes of tree t. Each non-terminal node $n \in N(t)$

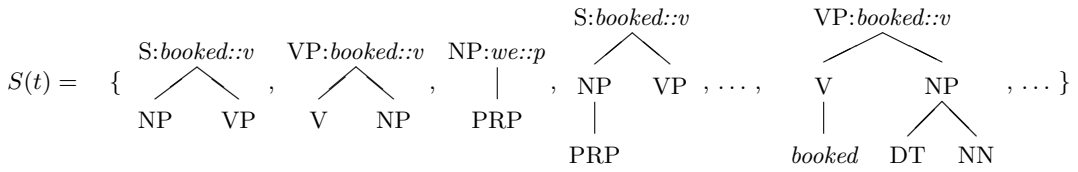


Figure 2: Subtrees of the tree t in Figure 1 (a non-exhaustive list)

has a label l_n composed of two parts $l_n = (s_n, w_n)$: s_n is the syntactic label, while w_n is the semantic headword of the tree headed by n, along with its part-of-speech tag. Terminal nodes of trees are treated differently, these nodes represent only words w_n without any additional information, and their labels thus only consist of the word itself (see Fig. 1). The structure of a DST is represented as follows: Given a tree t, $\mathsf{h}(t)$ is its root node and $\mathsf{s}(t)$ is the tree formed from t but considering only the syntactic structure (that is, only the s_n part of the labels), $c_i(n)$ denotes i-th child of a node n. As usual for constituency-based parse trees, pre-terminal nodes are nodes that have a single terminal node as child.

Finally, we use $\vec{w_n} \in \mathbb{R}^k$ to denote the *distributional* vector for word w_n.

2.2 The method at a glance

We describe here the approach in a few sentences. In line with tree kernels over structures (Collins and Duffy, 2002), we introduce the set $S(t)$ of the subtrees t_i of a given lexicalized tree t. A subtree t_i is in the set $S(t)$ if $\mathsf{s}(t_i)$ is a subtree of $\mathsf{s}(t)$ and, if n is a node in t_i, all the siblings of n in t are in t_i. For each node of t_i we only consider its syntactic label s_n, except for the head $\mathsf{h}(t_i)$ for which we also consider its semantic component w_n (see Fig. 2). The functions DSTs we define compute the following:

$$DST(t) = \mathbf{T} = \sum_{t_i \in S(t)} \mathbf{T}_i$$

where $\mathbf{T}_i$ is the matrix associated to each subtree t_i. The similarity between two text fragments a and b represented as lexicalized trees t^a and t^b can be computed using the Frobenius product between the two matrices $\mathbf{T}^a$ and $\mathbf{T}^b$, that is:

$$\langle \mathbf{T}^a, \mathbf{T}^b \rangle_F = \sum_{\substack{t_i^a \in S(t^a) \\ t_j^b \in S(t^b)}} \langle \mathbf{T}_i^a, \mathbf{T}_j^b \rangle_F \qquad (1)$$

We want to obtain that the product $\langle \mathbf{T}_i^a, \mathbf{T}_j^b \rangle_F$ approximates the dot product between the distributional vectors of the head words ($\langle \mathbf{T}_i^a, \mathbf{T}_j^b \rangle_F \approx \langle \vec{\mathsf{h}}(t_i^a), \vec{\mathsf{h}}(t_j^b) \rangle$) whenever the syntactic structure of the subtrees is the same (that is $\mathsf{s}(t_i^a) = \mathsf{s}(t_j^b)$), and $\langle \mathbf{T}_i^a, \mathbf{T}_j^b \rangle_F \approx 0$ otherwise. This property is expressed as:

$$\langle \mathbf{T}_i^a, \mathbf{T}_j^b \rangle_F \approx \delta(\mathsf{s}(t_i^a), \mathsf{s}(t_j^b)) \cdot \langle \vec{\mathsf{h}}(t_i^a), \vec{\mathsf{h}}(t_j^b) \rangle \quad (2)$$

To obtain the above property, we define

$$\mathbf{T}_i = \vec{\mathsf{s}}(t_i) \vec{w}_{\mathsf{h}(t_i)}^{\top}$$

where $\vec{\mathsf{s}}(t_i)$ are distributed tree fragment (Zanzotto and Dell'Arciprete, 2012) for the subtree t and $\vec{w}_{\mathsf{h}(t_i)}$ is the distributional vector of the head of the subtree t. Distributed tree fragments have the property that $\vec{\mathsf{s}}(t_i)\vec{\mathsf{s}}(t_j) \approx \delta(t_i, t_j)$. Thus, exploiting the fact that: $\langle \vec{a}\vec{w}^{\top}, \vec{b}\vec{v}^{\top} \rangle_F = \langle \vec{a}, \vec{b} \rangle \cdot \langle \vec{w}, \vec{v} \rangle$, we have that Equation 2 is satisfied as:

$$\begin{aligned} \langle \mathbf{T}_i, \mathbf{T}_j \rangle_F &= \langle \vec{\mathsf{s}}(t_i), \vec{\mathsf{s}}(t_j) \rangle \cdot \langle \vec{w}_{\mathsf{h}(t_i)}, \vec{w}_{\mathsf{h}(t_j)} \rangle \\ &\approx \delta(\mathsf{s}(t_i), \mathsf{s}(t_j)) \cdot \langle \vec{w}_{\mathsf{h}(t_i)}, \vec{w}_{\mathsf{h}(t_j)} \rangle \end{aligned}$$

It is possible to show that the overall compositional distributional model $DST(t)$ can be obtained with a recursive algorithm that exploits vectors of the nodes of the tree.

3 The Approximated Smoothed Tree Kernels

The CDSM we proposed approximates a specific tree kernel belonging to the smoothed tree kernels class. This recursively computes (but, the recursive formulation is not given here) the following general equation:

$$STK(t^a, t^b) = \sum_{\substack{t_i \in S(t^a) \\ t_j \in S(t^b)}} \omega(t_i, t_j)$$

		RTE1	RTE2	RTE3	RTE5	headl	FNWN	OnWN	SMT
STK vs DSTK	1024	0.86	0.84	0.90	0.84	0.87	0.65	0.95	0.77
	2048	0.87	0.84	0.91	0.84	0.90	0.65	0.96	0.77

Table 1: Spearman's correlation between Distributed Smoothed Tree Kernels and Smoothed Tree Kernels

where $\omega(t_i, t_j)$ is the similarity weight between two subtrees t_i and t_j. $DTSK$ approximates STK, where the weights are defined as follows:

$$\omega(t_i, t_j) = \alpha \cdot \langle \overrightarrow{w_{\mathsf{h}(t_i)}}, \overrightarrow{w_{\mathsf{h}(t_j)}} \rangle \cdot \delta(\mathsf{s}(t_i), \mathsf{s}(t_j))$$

Where $\alpha = \sqrt{\lambda^{|N(t_i)|+|N(t_j)|}}$ and λ is a parameter.

4 Experimental investigation

Generic settings We experimented with two datasets: the Recognizing Textual Entailment datasets (RTE) (Dagan et al., 2006) and the the Semantic Textual Similarity 2013 datasets (STS) (Agirre et al., 2013). The STS task consists of determining the degree of similarity (ranging from 0 to 5) between two sentences. The STS datasets contains 5 datasets: headlines, OnWN, FNWN and SMT which contains respectively 750, 561, 189 and 750 RTE is instead the task of deciding whether a long text T entails a shorter text, typically a single sentence, called hypothesis H. It has been often seen as a classification task. We used four datasets: RTE1, RTE2, RTE3, and RTE5. We parsed the sentence with the Stanford Parser (Klein and Manning, 2003) and extracted the heads for use in the lexicalized trees with Collins' rules (Collins, 2003). Distributional vectors are derived with DISSECT (Dinu et al., 2013) from a corpus obtained by the concatenation of ukWaC, a mid-2009 dump of the English Wikipedia and the British National Corpus for a total of about 2.8 billion words. The raw count vectors were transformed into positive Pointwise Mutual Information scores and reduced to 300 dimensions by Singular Value Decomposition. This setup was picked without tuning, as we found it effective in previous, unrelated experiments. To build our DTSKs we used the implementation of the distributed tree kernels[1]. We used 1024 and 2048 as the dimension of the distributed vectors, the weight λ is set to 0.4 as it is a value generally considered optimal for many applications (see also (Zanzotto and Dell'Arciprete, 2012)). To test the quality of the approximation we computed the Spearman's correlation between values produced by our $DSTK$ and by the standard versions of the smoothed tree kernel. We obtained text fragment pairs by randomly sampling two text fragments in the selected set. For each set, we produced exactly the number of examples in the set, e.g., we produced 567 pairs for RTE1, etc.

Results Table 1 reports the results for the correlation experiments. We report the Spearman's correlations over the different sets (and different dimensions of distributed vectors) between our $DSTK$ and the STK. The correlation is above 0.80 in average for both RTE and STS datasets. The approximation also depends on the size of the distributed vectors. Higher dimensions yield to better approximation: if we increase the distributed vectors dimension from 1024 to 2048 the correlation between $DSTK$ and STK increases. This direct analysis of the correlation shows that our CDSM are approximating the corresponding kernel function and there is room of improvement by increasing the size of distributed vectors.

5 Conclusions and future work

Distributed Smoothed Trees (DST) are a novel class of Compositional Distributional Semantics Models (CDSM) that effectively encode structural information and distributional semantics in tractable rank-2 tensors, as experiments show. The paper shows that DSTs contribute to close the gap between two apparently different approaches: CDSMs and convolution kernels. This contribute to start a discussion on a deeper understanding of the representation power of structural information of existing CDSMs.

[1] http://code.google.com/p/distributed-tree-kernels/

References

[Agirre et al.2013] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

[Aronszajn1950] N. Aronszajn. 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

[Baroni and Lenci2010] Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, 36(4):673–721, December.

[Baroni and Zamparelli2010] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, October. Association for Computational Linguistics.

[Collins and Duffy2002] Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of ACL02*.

[Collins2003] Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Comput. Linguist.*, 29(4):589–637.

[Cristianini and Shawe-Taylor2000] Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March.

[Croce et al.2011] Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1034–1046, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Dagan et al.2006] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Quiñonero-Candela et al., editor, *LNAI 3944: MLCW 2005*, pages 177–190. Springer-Verlag, Milan, Italy.

[Dagan et al.2013] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[Dinu et al.2013] Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL (System Demonstrations)*, pages 31–36, Sofia, Bulgaria.

[Ferrone and Zanzotto2013] Lorenzo Ferrone and Fabio Massimo Zanzotto. 2013. Linear compositional distributional semantics and structural kernels. In *Proceedings of the Joint Symposium of Semantic Processing (JSSP)*.

[Ferrone and Zanzotto2014] Lorenzo Ferrone and Fabio Massimo Zanzotto. 2014. Towards syntax-aware compositional distributional semantic models. In *Proceedings of Coling 2014*. COLING, Dublin, Ireland, Aug 23–Aug 29.

[Haussler1999] David Haussler. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

[Jurgens et al.2014] David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

[Klein and Manning2003] Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Mitchell and Lapata2008] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

[Socher et al.2011] Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems 24*.

[Zanzotto and Dell'Arciprete2012] F.M. Zanzotto and L. Dell'Arciprete. 2012. Distributed tree kernels. In *Proceedings of International Conference on Machine Learning*, pages 193–200.

[Zanzotto et al.2010] Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional

semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, August,.

10.12871/CLICIT2014134

Polysemy alternations extraction using the PAROLE SIMPLE CLIPS Italian lexicon.

Francesca Frontini **Valeria Quochi** **Monica Monachini**
Istituto di Linguistica Computazionale "A.Zampolli" CNR Pisa
name.surname@ilc.cnr.it

Abstract

English. This paper presents the results of an experiment of polysemy alternations induction from a lexicon (Utt and Padó, 2011; Frontini et al., 2014), discussing the results and proposing an amendment in the original algorithm.

Italiano. *Questo articolo presenta i risultati di un esperimento di induzione di alternamze polisemiche regolari (Utt and Padó, 2011; Frontini et al., 2014), discutendone i risultati e proponendo una modifica all'originale procedura.*

1 Introduction

The various different senses of polysemic words do not always stand to each other in the same way. Some senses group together along certain dimensions of meaning while others stand clearly apart. Machine readable dictionaries have in the past used coarse grained sense distinctions but often without any explicit indication as to whether these senses were related or not. Most significantly, few machine readable dictionaries explicitly encode systematic alternations.

In Utt and Padó (2011) a methodology is described for deriving systematic alternations of senses from WordNet. In Frontini et al. (2014) the work was carried out for Italian using the PAROLE SIMPLE CLIPS lexicon (PSC) (Lenci et al., 2000), a lexical resource that contains a rich set of explicit lexical and semantic relations. The purpose of the latter work was to test the methodology of the former work against the inventory of regular polysemy relations already encoded in the PSC semantic layer. It is important to notice that this was not possible in the original experiment, as WordNet does not contain such information.

The result of the work done on PSC shows how the original methodology can be useful in testing the consistency of encoded polysemies and in finding gaps in individual lexical entries. At the same time the methodology is not infallible especially in distinguishing type alternations that frequently occur in the lexicon due to systematic polysemy from other alternations that are produced by metaphoric extensions, derivation or other non systematic sense shifting phenomena.

In this paper we shall briefly outline the problem of lexical ambiguity; then describe the procedure of type induction carried out in the previous experiments, discussing the most problematic results; finally we will propose a change in the original methodology that seems more promising in capturing the essence of systematic polysemy.

2 Theoretical background on lexical ambiguity

Two main types of lexical ambiguity are usually distinguished, homonymy and polysemy.

The most common definition of homonymy in theoretical linguistics is that two words are homonymous if they share the same form (orthography and/or phonology), but have different, unrelated and mutually underived meanings (Leech, 1974; Lyons, 1977; Saeed, 1997). According to this view, two homonymous words must have different etymologies. Pure homonyms, moreover should manifest both homophony and homography.

The notion of polysemy in contrast foresees a commonality of meaning that is shared between the different senses of the same word. Polysemy has received ample treatment in the literature (Apresjan, 1974; Nunberg and Zaenen, 1992; Copestake and Briscoe, 1995; Nunberg, 1995; Palmer, 1981). Three main types can be identified. **Regular (or logical polysemy)**: Words with two, or more, systematically related meanings. The

meaning of a word is described here in terms of the semantic (or ontological) classes to which the senses of a lexical item refer. Regular polysemy can thus be defined in terms of regularity of type alternations, where the "alternating types" in question are the semantic or ontological categories to which the senses of a lemma belong (Palmer, 1981; Pustejovsky, 1995). Well known cases of regular alternations are ANIMAL−FOOD, BUILDING−INSTITUTION. These systematic meaning alternations are generally salient on conceptual grounds, common to (several) other words[1], and usually derivable by metonymic sense shifts.

Occasional (or irregular) polysemy: a word shows a "derivable" meaning alternation, i.e. there is an evident relation between the meanings, usually again metonymic, but this is not pervasive in the language (e.g. *coccodrillo*, 'crocodile', can be used both to indicate the animal and the (leather) material; this alternation is common to other animal words but is not so pervasive, and is clearly dependent on other world-knowledge factors)

Metaphorical polysemy: a word with meanings that are related by some kind of metaphorical extension. Again, this will not be systematic in the language, although other words may show similar extensions. For example, *fulmine*, 'lightning' NATURAL PHENOMENON, can be used metaphorically to described something or someone as 'very fast' as in *Giovanni è un fulmine*, 'John is as quick as a flash'; *Boa*, 'boa', ANIMAL, can also refer to a feather scarf. The relationship between the two senses of these words is probably one of lexicalized metaphorical extension which it will be hard to generalize to other words.

The distinction between regular polysemy, occasional polysemy and homonymy is somewhat more blurred than it seems at first (Zgusta, 1971; Palmer, 1981; Lyons, 1977; Landau, 1984; Ndlovu and Sayi, 2010), and a continuum can be recognized.

3 Previous experiments

We refer to Utt and Padó (2011) and Frontini et al. (2014) for a precise description of the experiment on English and Italian respectively and of the induction algorithm. Here an intuitive outline is given. If we consider a lemma and all of its senses, each possible sense can be labeled with an ontological class or type and thus each pair of senses of that lemma can be seen as an alternation between two ontological types. Such alternations are called basic alterations (BAs). An instance of BA (i.e. a sense pair within a lemma) may represent a case of regular (systematic) polysemy or a case of simple homonymy. However, when the same BA occurs across many lemmas, this can be taken as evidence of a regular polysemy.

For example, in languages such as English or Italian the presence of a large number of lemmas with two senses, one of which is labeled with the type ANIMAL and the other with the type FOOD provides evidence of the fact that the FOOD#ANIMAL BA is not merely sporadic in such languages but is the product of ANIMAL >FOOD regular polysemy.

The induction algorithm proposed essentially derives the complete list of BAs from a given lexicon by extracting all type alternations occurring within polysemous lemmas (nouns in our case), and ranks them per descending frequency. The assumption is that the most frequent BAs will be polysemous, whereas the less frequent ones will be occasional. The optimal frequency threshold N for a BA to be classified as a regular polysemy is induced by testing it against a set of known homonymous and polysemous seed lemmas. The correct threshold is the one that correctly separates typically homonyms from polysemous words.

In Frontini et al. (2014) we run two experiments with two different sets of seeds and derived two frequency thresholds (≥ 28 and ≥ 21 respectively), identifying a set of overall 36 and 54 Basic Alternations that can be considered polysemous (see the cited paper for the difference between the two thresholds). In the present paper we shall refer mostly to the frequency threshold ≥ 21, which was derived by strictly following the methodology proposed by Utt and Padó (2011), namely using a set of prototypically polysemous/homonymous lemmas drawn form the literature.

In Frontini et al. (2014) we report on the results above the first and second threshold. Each induced BA is compared with all possible relations that are encoded in PSC between senses of words exhibiting that BA. Relations encoded among senses in PSC are of two types: Lexical relations (such as Polysemy itself, Metaphor, Derivation) or Qualia relations (Constitutive, Formal, Telic, Agentive), following the generative lexicon theory (Puste-

[1] Of course some exceptions are possible, e.g. *cow/beef*. (Nunberg, 1995; Copestake and Briscoe, 1995)

jovsky, 1995).

When comparing the induced results with PSC, four cases can be recognized [2]:

A) a BA is matched by one or more polysemy relations

C) no polysemy relation is present but at least another lexical relation (metaphor or derivation) is present

D) only qualia relations exist between the alternating usems of a lemma that expresses a BA

E) no relation at all is encoded in PSC for a BA.

In all cases but (A) it is obviously possible that a regular polysemy is involved that had not been foreseen in the design of PSC. In the first line of Table 1 the results for the original experiment are given.

(A) represents the perfect validation. Classic polysemy cases are to be found here, such as *PolysemySemioticartifact-Information* (e.g., 'letter', 'newspaper'); *PolysemyPlant-Flower*; etc. The presence of qualia relations, often Constitutive, does not impact on the goodness of this result, but shows how some polysemies may be due to meronymic sense shifts.

(C) cases are the more interesting ones, since they illustrate phenomena that may cast a doubt on the frequency based definition of polysemy followed in the present work. Here some very frequent BAs are classified by the lexicographer in terms of zero derivation (such as instrument *violino*, 'violin' INSTRUMENT, used for the PROFESSION, violinist) or of metaphorical extension (such as *coniglio*, 'rabbit', for a cowardly person). Such cases are frequent, probably even semi-productive, but lack the regularity that characterizes systematic polysemy.

(D) cases occur rarely, and the qualia relations listed occur very rarely among the corresponding lemmas. Such lemmas, though not strictly polysemous, represent instances of semanticized metaphoric extension of the sort that may qualify for formal encoding with the *metaphor* relation; so for instance *spada*, 'sword', has a sense typed under AGENT_OF_TEMPORARY_ACTIVITY to indicate uses such as *He is a good sword* meaning 'He is a good swordsman'.

[2]Case A in the present paper merges cases A and B of the previous one; the original labelling for the other cases is maintained for comparison.

(E) cases require careful analysis, since they are the most problematic outcome. Some of them seem to be the result of semi-productive phenomena, despite the lack of lexicographic encoding. So for instance, BODY_PART#PART, with frequency 101, captures the fact that parts of artifacts (e.g. machines, ships, ...) are often denoted in Italian by using words for body parts (such as in *braccio*, used for: 'person's arm', 'gramophone's arm', 'edifice's wing'); PSC lexicographers did not define an explicit relation for such alternations, as they seem more cases of metaphorical extension than of regular polysemy.

Other (E) alternations instead show clearly related senses and a higher level of systematicity. Such is the case with AGENT_OF_PERSISTENT_ACTIVITY#PROFESSION, typical of lemmas such as *pianista*, 'pianist', denoting both someone who plays piano professionally and someone who plays piano regularly, but as an amateur. Another such case is ACT#PSYCH_PROPERTY, with lemmas such as *idiozia*, 'silliness', once listed as the property of associated with being an idiot and then with the act of being idiotic. Such alternations are rarely listed among the known polysemy alternations, and are the product of the semantic richness of PSC and of the SIMPLE ontology, that distinguishes shades of meaning that are normally not taken into account in other resources. At the same time, within the context of PSC, they are quite systematic and may be considered for an explicit encoding.

Finally, some (E) cases are somewhat epiphenomenal: so for instance HUMAN#SUBSTANCE_FOOD is the result of the fact that some animals, typically those familiar animals that are used for food, are also used to metaphorically define properties of humans, such as *pig*, *chicken* and *goat*. In this case, there is a pivotal usem (the ANIMAL one) that is linked to the other two by separate alternations (ANIMAL#HUMAN and ANIMAL#SUBSTANCE_FOOD), producing an indirect alternation (HUMAN#SUBSTANCE_FOOD).

The conclusion drawn from this experiment was that frequency alone is not a sufficient enough a criterion to define *systematic* polysemy. The proposed methodology seems to be more reliable in distinguishing any kind of polysemy alternation between related senses.

4 New experiment and preliminary conclusion

While distinguishing when two senses are totally unrelated may indeed be very useful, the original goal of this research was to be able to automatically detect regular polysemy alternations. In this new experiment we then try to see if the original methodology can be improved in order to make it more capable to single out systematic polysemy, which is characterised by productivity and ontological grounding.

The ontological grounding of polysemy can be assessed in resources such as PSC by checking the qualia relations; indeed many of the officially encoded polysemies in PSC co-occur with qualia relations. Nevertheless this methodology can hardly be automatized or applied to other resources such as WordNet that lack qualia information. As for the productivity, it is clearly related to the directionality of the polysemy rule. If the directionality is from type A to type B we can presume that all words that have a sense of type A can be also used in a sense of type B, but not vice versa. So if the rule is "Animal to Food", then all words for Animal should also have the Food sense, but not vice versa. So *crockodile* can denote food in some contexts, but *spaghetti* cannot be used to refer to an animal.

In a methodology such as the one proposed it is hard to retrieve directionality from polysemy rules, since lexicons are rarely exhaustive. Nevertheless it may be possible to indirectly asses the systematicity of the type alternation by comparing the frequency of the BA with the one of each type separately. An efficient way to treat this problem is to consider measuring the association strength of the two types by using Pointwise Mutual Information, following what has been previously proposed in Tomuro (1998). PMI assigns the maximum value to pairs that occur only together, and in general gets higher values if at least one of the two elements occurs with the other more frequently than alone. It is calculated as:

$$PMI(t1, t2) = \log \frac{\frac{f(t1,t2)}{N}}{\frac{f(t1)}{N} X \frac{f(t2)}{N}} \qquad (1)$$

where t1 and t2 are the number of lemmas in which of each of the two ontological types of a BA occur overall, and (t1,t2) is the number of lemmas in which they occur together. Taking into account the tendency of PMI to promote hapaxes, a raw frequency filter ≥ 5 for co-occurrencies values was implemented. We thus rank the BAs in PSC using descending PMI instead of raw frequency, then we induce the optimal PMI threshold following the standard procedure, using the same set of 12 + 12 typically polysemous/homonymoys lemmas drawn from the literature, and comparing the results. The second line of Table 1 shows the cases obtained from this new experiment, while table 4 presents the complete list. The number of BA induced with the two ranking systems is comparable (49 for PMI vs 54 for raw frequency).

	A	C	D	E	TOT
F $>$ 21	20	11	5	18	54
PMI $>$ 1.8	24	1	8	16	49

Table 1: Comparison between induced BAs and lexical semantic relations in PSC, for both induced thresholds.

First results seem promising. Most significantly PMI ranking promotes only one C case above the threshold, vs 11. LOCATION#OPENING, is indeed a Metaphor occurring only 8 times in PSC, in cases such as "topaia" (rathole) that can be used to metaphorically refer to human abodes in very unflattering terms.

This seems to signify that PMI ranking is more effective in demoting cases unsystematic polysemy. Remarkably PMI ranking demotes one of the most problematic and frequent of the previously discussed BA, BODY_PART#PART, under the threshold while promoting a larger number of the encoded polysemies to the top. In the first 18 positions we find only one gap at position 8 and it turns out that this BA - CONVENTION#MONEY - is actually a good candidate for systematic polysemy, as MONEY is both an artifact and a human convention.

To conclude, such preliminary results actually seem to confirm the hypothesis that measuring the association strength between types, rather than the frequency of their cooccurrence, is useful to capture the systematicity of an alternation.

In future work it may be interesting to test ranking by other association measures (such as Log Likelihood) and with different filternigs. Finally, the original experiment may be repeated on both Italian and English WordNets in order to evaluate the new method on the original lexical resource.

References

Jurij D. Apresjan. 1974. Regular Polysemy. *Linguistics*, 142:5–32.

Ann Copestake and Ted Briscoe. 1995. Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12:15–67.

Francesca Frontini, Valeria Quochi, Sebastian Padó, Monica Monachini, and Jason Utt. 2014. Polysemy Index for Nouns: an Experiment on Italian using the PAROLE SIMPLE CLIPS Lexical Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2955–2963, Reykjavik, Iceland.

S. I. Landau. 1984. *Dictionaries: The Art and Craft of Lexicography*. Charles Scribner's Sons, New York.

G. N. Leech. 1974. *Semantics*. Penguin, Harmondsworth.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4):249–263.

J. Lyons. 1977. *Semantics. Vol 2*. Cambridge University Press, Cambridge.

E. Ndlovu and S. Sayi. 2010. The Treatment of Polysemy and Homonymy in Monolingual General-purpose Dictionaries with Special Reference to "Isichazamazwi SesiNdebele". *Lexikos*, 20:351–370.

Geoff Nunberg and Annie Zaenen. 1992. Systematic polysemy in lexicology and lexicography. In *Proceedings of Euralex II*, pages 387–395, Tampere, Finland.

Geoffrey Nunberg. 1995. Transfers of Meaning. *Journal of Semantics*, 12(2):109–132.

F. R. Palmer. 1981. *Semantics*. Cambridge University Press, Cambridge.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge MA.

J. I. Saeed. 1997. *Semantics*. Blackwell Publishers, Oxford.

Noriko Tomuro. 1998. Semi-automatic induction of systematic polysemy from WordNet. In *Proceedings ACL-98 Workshop on the Use of WordNet in NLP*.

Jason Utt and Sebastian Padó. 2011. Ontology-based Distinction between Polysemy and Homonymy. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, Oxford, UK.

L. Zgusta. 1971. *Manual of Lexicography*. Mouton, The Hague/Paris.

BA	Val.	freq.	PMI
Substance_food#Water_animal	A	58	4.66
Flower#Plant	A	45	4.40
Information#Semiotic_artifact	A	218	4.26
Plant#Vegetable	A	49	4.16
Flavouring#Plant	A	23	4.13
Color#Flower	A	7	3.94
Color#Fruit	A	9	3.85
Convention#Money	D	16	3.82
Fruit#Plant	A	29	3.62
Building#Institution	A	63	3.39
Amount#Container	A	79	3.39
Language#People	A	174	3.35
Earth_animal#Substance_food	A	34	3.33
Color#Vegetable	A	5	3.27
Convention#Semiotic_artifact	A	44	3.09
Artifactual_drink#Plant	A	28	3.04
Human_Group#Institution	A	53	3.01
Color#Natural_substance	A	30	2.98
Area#VegetalEntity	D	6	2.92
Concrete_Entity#Transaction	D	5	2.83
Cause_Change_of_State#Material	E	14	2.82
Artifactual_material#Earth_animal	A	38	2.80
Color#Plant	A	21	2.75
Air_animal#Substance_food	A	12	2.71
Artwork#Color	E	6	2.71
Location#Opening	C	8	2.71
Copy_Creation#Semiotic_artifact	D	5	2.62
Cause_Constitutive_Change# Constitutive_Change	E	5	2.60
Area#Artifactual_area	E	6	2.42
Artwork#Symbolic_Creation	D	9	2.40
Act#Psych_property	E	54	2.40
Artifactual_material#Substance_food	E	10	2.39
Food#Time	D	5	2.37
Amount#D_3_Location	E	8	2.30
Constitutive#Shape	D	7	2.22
Artifactual_material#Artwork	A	8	2.17
Convention#Institution	E	7	2.16
Plant#VegetalEntity	D	10	2.13
Convention#Time	E	15	2.11
Amount#Transaction	E	7	2.05
Time#Unit_of_measurement	E	7	2.04
Cause_Change#Change	E	5	1.98
Artifact#Artifact_Food	E	6	1.96
Abstract_Entity#Metalanguage	E	8	1.95
Artifactual_material#Color	E	5	1.89
Building#Human_Group	A	74	1.86
Natural_substance#Plant	A	39	1.85
Number#Time	E	6	1.84
Convention#Information	A	13	1.83

Table 2: BA induced using PMI as ranking method; letters represent the validation against PSC encoded relations. The order between the two types for each BA is purely alphabetical.

10.12871/CLICIT2014135

Rappresentazione dei concetti azionali attraverso prototipi e accordo nella categorizzazione dei verbi generali. Una validazione statistica.

Gloria Gagliardi
Università degli Studi di Firenze
gloria.gagliardi@unifi.it

Abstract

Italiano. L'articolo presenta i risultati di uno studio volto a valutare la consistenza della categorizzazione dello spazio azionale operata da annotatori madrelingua per un set di verbi semanticamente coesi del database IMAGACT (area semantica di 'girare'). La validazione statistica, articolata in tre test, è basata sul calcolo dell'*inter-tagger agreement* in *task* di disambiguazione di concetti rappresentati mediante prototipi per immagini.

English. *This paper presents the results of a research aimed at evaluating the consistency of the categorization of actions. The study focuses on a set of semantically related verbs of the IMAGACT database ("girare" semantic area), annotated by mother tongue informants. Statistic validation, consisting of three tests, is based on inter-tagger agreement. The task entails the disambiguation of concepts depicted by prototypic scenes.*

1 Introduzione

IMAGACT è un'ontologia interlinguistica che rende esplicito lo spettro di variazione pragmatica associata ai predicati azionali a media ed alta frequenza in italiano ed inglese (Moneglia *et al*., 2014). Le classi di azioni che individuano le entità di riferimento dei concetti linguistici, rappresentate in tale risorsa lessicale nella forma di scene prototipiche (Rosch, 1978), sono state indotte da *corpora* di parlato da linguisti madrelingua, mediante una procedura *bottom-up*: i materiali linguistici sono stati sottoposti ad una articolata procedura di annotazione descritta estesamente in lavori precedenti (Moneglia *et al.*, 2012; Frontini *et al.*, 2012).

L'articolo illustra i risultati di tre test volti a valutare la consistenza della categorizzazione dello spazio azionale proposta dagli annotatori per un *set* ristretto ma semanticamente coerente di verbi della risorsa: tale scelta è stata dettata dalla volontà di studiare ad un alto livello di dettaglio i problemi connessi alla tipizzazione della variazione dei predicati sugli eventi. La predisposizione di questo *case-study* è inoltre propedeutica alla creazione di una procedura standard, estendibile in un secondo tempo a porzioni statisticamente significative dell'ontologia per la sua completa validazione.

Il paragrafo 2 presenterà i coefficienti statistici adottati, nel paragrafo 3 verranno descritti metodologia e risultati dei test realizzati.

2 Coefficienti statistici

La consistenza della categorizzazione è stata valuta mediante il calcolo dell'*inter-tagger agreement* (I.T.A.). Per l'analisi sono stati utilizzati i seguenti coefficienti[1], presentati in maniera congiunta secondo le indicazioni in Di Eugenio and Glass (2004):

- A_o, "*observed agreement*" o "*Index of crude agreement*" (Goodman and Kruskal, 1954);
- π (Scott, 1955);
- k (Cohen, 1960);
- $2A_o$-1 (Byrt *et al.*, 1993);
- α (Krippendorff, 1980);
- multi-k (Davies and Fleiss, 1982);
- multi-π (Fleiss, 1971).

Tali indici, mutuati dalla psicometria, rappresentano ad oggi uno standard *de facto* in linguistica computazionale (Carletta, 1996).

Per l'analisi dei dati è stato utilizzato il modulo "metrics.agreement" di NLTK - Natural Language Toolkit (Bird *et al*., 2009).

Il *dataset* è disponibile all'URL http://www.gloriagagliardi.com/miscellanea/.

[1] Nell'articolo viene adottata la terminologia di Artstein & Poesio (2008), lavoro di riferimento sull'argomento. I coefficienti sono illustrati e discussi in Gagliardi (2014); nel medesimo lavoro vengono inoltre esaminati i parametri che influenzano i livelli di accordo raggiungibili (Bayerl and Paul, 2011; Brown *et al.*, 2010) e i valori di significatività dei coefficienti (Landis and Koch, 1977; Krippendorf, 1980; Carletta, 1996; Reidsma and Carletta, 2008; Bayerl and Paul, 2011), in relazione ai principali studi condotti su I.T.A. per l'italiano e l'inglese in domini di tipo semantico.

3 Validazione

3.1 Test 1 (3 categorie)

Con il test 1 si intende valutare il livello di *agreement* raggiungibile nella categorizzazione di occorrenze verbali nelle categorie 'primario' – 'marcato'.

A due annotatori è stato sottoposto un set di 974 concordanze, riconducibili ad un'area semantica coesa ('girare' e lemmi verbali di significato prossimo). Il *task* consiste in un esercizio di disambiguazione "*coarse-grained*": il protocollo di annotazione prevede che ciascun *coder*, dopo aver letto ed interpretato l'occorrenza verbale in contesto, attribuisca il *tag* PRI (primario) o MAR (marcato), ovvero discrimini tra gli usi fisici ed azionali e quelli metaforici o fraseologici. Nel caso in cui non sia possibile per l'annotatore interpretare l'occorrenza o vi sia un errore di *tagging*, l'istanza deve essere annotata con l'etichetta DEL (delete), analogamente a quanto previsto nel *workflow* di IMAGACT. È inoltre richiesto all'annotatore, per le sole occorrenze PRI, di creare una frase standardizzata che espliciti e sintetizzi l'eventualità predicata. Gli annotatori (tabella 1) hanno un alto livello di esperienza nel *task*.

rater	sesso	età	istruzione	professione
A	F	29	dottorando	assegnista
B	M	29	dottorando	assegnista

Tabella 1: Annotatori test 1.

L'intera procedura è svolta dagli annotatori autonomamente ed indipendentemente. In tabella 2 sono sintetizzati i principali parametri descrittivi del test, ed in tabella 3 i risultati.

TEST 1 – 3 categorie	
numero di *rater*	2
tipologia dei dati	occorrenze verbali e relative concordanze
dimensione del *dataset*	974 occorrenze
categorie	3 (PRI - MAR- DEL)
criteri di selezione dei *rater*	Gli annotatori hanno annotato circa il 90% delle occorrenze verbali di IMAGACT-IT
livello di esperienza dei *rater*	esperti
tipo e intensità del training	intenso
coefficienti statistici	A_o, k, π, $2A_o$-1, α

Tabella 2: test 1, parametri descrittivi.

Lemma	A_o	k	π	$2A_o$-1	α
capovolgere	1.0	1.0	1.0	1.0	1.0
curvare	1.0	1.0	1.0	1.0	1.0
girare	0.90	0.84	0.84	0.84	0.84
mescolare	1.0	1.0	1.0	1.0	1.0
rigirare	0.9	0.85	0.85	0.8	0.85
rivolgere	0.79	0.53	0.52	0.57	0.52
ruotare	0.95	0.91	0.91	0.89	0.91
svoltare	1.0	1.0	1.0	1.0	1.0
volgere	1.0	1.0	1.0	1.0	1.0
voltare	0.91	0.66	0.66	0.82	0.66
TOTALE	0.89	0.83	0.83	0.79	0.83

Tabella 3: test 1 (3 categorie), risultati.

I risultati appaiono molto buoni: i coefficienti calcolati sull'insieme delle occorrenze hanno infatti un valore superiore a 0.8. Anche i valori di *agreement* calcolati per i singoli verbi sono alti: l'accordo è addirittura totale per 5 lemmi su 10. Solo i verbi 'rivolgere' e 'voltare' hanno valori di I.T.A. bassi: per il secondo lemma è però osservabile nei dati una forte prevalenza della categoria PRI (corretta dalla misura $2A_o$-1).

3.2 Test 1 (2 categorie)

In seconda battuta si è deciso di rianalizzare i dati scartando gli *item* a cui almeno un annotatore ha assegnato il *tag* DEL, considerando quindi solo le occorrenze che entrambi i *rater* hanno ritenuto interpretabili.[2] I risultati sono sintetizzati in tabella 4.

Lemma	A_o	k	π	$2A_o$-1	α
capovolgere	1.0	1.0	1.0	1.0	1.0
curvare	1.0	/	/	1.0	/
girare	0.98	0.95	0.95	0.96	0.95
mescolare	1.0	1.0	1.0	1.0	1.0
rigirare	1.0	1.0	1.0	1.0	1.0
rivolgere	0.99	0.93	0.93	0.98	0.93
ruotare	1.0	1.0	1.0	1.0	1.0
svoltare	1.0	1.0	1.0	1.0	1.0
volgere	1.0	/	/	1.0	/
voltare	0.93	0.63	0.63	0.87	0.63
TOTALE	0.98	0.96	0.96	0.91	0.96

Tabella 4: test 1 (2 categorie), risultati.

Il livello di I.T.A., già alto, supera grazie alla riformulazione del task la soglia di 0.9. L'unico lemma problematico resta 'voltare', per il problema di prevalenza già evidenziato.

[2] L'annotatore A ha usato il *tag* DEL 232 volte, l'annotatore B 244. 193 *item* hanno ricevuto il *tag* DEL da entrambi gli annotatori (circa l'80% dei casi).

3.3 Test 2

Con il test 2 si intende verificare il livello di *agreement* raggiungibile da annotatori esperti nell'assegnazione delle frasi standardizzate ai tipi azionali IMAGACT, ovvero la solidità e la coerenza della tipizzazione operata sui lemmi verbali oggetto di studio. A tale scopo è stato creato un set di frasi standardizzate a partire dai materiali annotati nel corso del test 1, secondo la seguente procedura:

- selezione dei lemmi per cui è stata identificata in IMAGACT una variazione primaria;[3]
- selezione dei verbi generali per cui, nel corso del test 1, sono state prodotte standardizzazioni primarie;
- raccolta di tutte le standardizzazioni create nel corso del test 1 per i lemmi rimanenti;
- esclusione delle frasi standardizzate uguali.[4]

Mediante questa serie di selezioni successive è stato estratto un set di 169 frasi standardizzate.

A due mesi di distanza dal primo test, agli stessi *coder* (tabella 1) è stato chiesto di assegnare le frasi standardizzate di ciascun lemma ad un inventario dato di tipi, la variazione primaria identificata nel DB IMAGACT.

In tabella 5 sono sintetizzati i principali parametri descrittivi del test, ed in tabella 6 i risultati.

TEST 2	
numero di *rater*	2
tipologia dei dati	frasi standardizzate
dimensione del *dataset*	169 frasi
categorie	da 3 a 11, in base al lemma
criteri di selezione dei *rater*	gli annotatori hanno annotato circa il 90% delle occorrenze verbali di IMAGACT-IT
livello di esperienza dei *rater*	esperti
tipo e intensità del *training*	intenso
coefficienti statistici	A_o, k, π, $2A_o$-1, α

Tabella 5: test 2, parametri descrittivi.

Il livello di I.T.A. è, in generale, buono: il valore dei coefficienti è infatti complessivamente superiore a 0.8. I due annotatori attribuiscono le standardizzazioni alle classi di azioni in modo sostanzialmente condiviso, pertanto la tipizzazione è considerabile fondata e riproducibile.

Lemma	A_o	k	π	$2A_o$-1	α
girare	0.79	0.77	0.76	0.58	0.76
mescolare	0.87	0.8	0.8	0.75	0.80
rigirare	1.0	1.0	1.0	1.0	1.0
ruotare	0.87	0.83	0.83	0.75	0.84
voltare	1.0	1.0	1.0	1.0	1.0
TOTALE	0.83	0.82	0.82	0.66	0.82

Tabella 6: test 2, risultati.

All'interno di un quadro essenzialmente positivo, com'era facilmente immaginabile il verbo più generale 'girare' appare il più difficile da disambiguare. Analizzando qualitativamente il *disagreement*, i dati evidenziano una forte concentrazione del disaccordo (11 casi su 26) in alcune specifiche categorie, la numero 9 e la numero 10, di cui si riportano i video prototipali in figura 1.

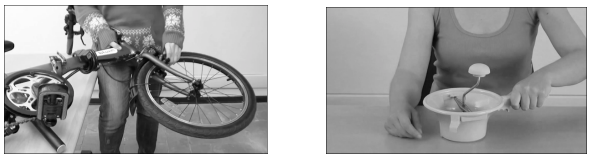


Figura 1: Tipo azionale 9 (a sinistra) e 10 (a destra) del verbo girare.

Vi è un'evidente contiguità tra le due classi di azioni: in entrambi i casi l'agente applica una forza sul tema, imprimendogli movimento rotatorio. Nel tipo 9 il tema è però messo in rotazione mediante un impulso, mentre nel tipo 10 l'agente esercita la forza sull'oggetto in maniera continua. Le tipologie di eventualità, chiaramente distinte sul piano empirico, risultano probabilmente troppo granulari dal punto di vista linguistico, al punto da risultare indistinguibili. Ricalcolando i coefficienti aggregando le due categorie, A_o= 0.879, k = 0.8606, π = 0.8605 ed α= 0.8611.

3.4 Test 3

Si è infine deciso di valutare il livello di *agreement* nell'assegnazione delle frasi standardizzate ai tipi azionali nel caso di annotatori non esperti, per verificare la riconoscibilità e l'effettiva riproducibilità della tassonomia azionale anche per semplici parlanti madrelingua. I *coder* coinvolti non hanno nessuna formazione specifica: gli unici requisiti per la selezione sono stati il livello di istruzione, medio-alto, e la di-

[3] Tali criteri comportano l'esclusione dal *test-set* dei verbi 'capovolgere', 'rivolgere', 'svoltare', 'volgere' e 'curvare'.

[4] La scelta è stata dettata dalla volontà di eliminare, almeno in parte, effetti distorsivi nel campione: alcune frasi, create dagli annotatori da occorrenze del *sub-corpus* di acquisizione LABLITA, si ripetono moltissime volte (es. "Il registratore gira"). Ciò è riconducibile alle modalità espressive del *baby-talk*, non certo ad una maggior frequenza della frase (o del tipo azionale) in italiano standard.

sponibilità a sottoporsi al test senza ricevere alcun compenso. I quattro annotatori reclutati (tabella 7) non sono stati sottoposti ad uno specifico *training*.

rater	sesso	età	istruzione	professione
C	M	32	laurea (LM)	web editor
D	M	28	laurea	web designer
E	M	30	dottorato	insegnante
F	F	26	laurea (LM)	inoccupato

Tabella 7: Annotatori test 3.

Il test segue lo stesso protocollo sperimentale dell'esercizio precedente (tabella 8). In tabella 9 sono sintetizzati i risultati.

TEST 3	
numero di *rater*	4
tipologia dei dati	frasi standardizzate
dimensione del *dataset*	169 frasi
categorie	da 3 a 11, in base al lemma
criteri di selezione dei *rater*	nessuna formazione specifica in linguistica; livello di istruzione medio-alto
livello di esperienza dei *rater*	principianti
tipo e intensità del *training*	Nessun *training*
coefficienti statistici	A_o, multi-*k*, multi-π, α

Tabella 8: test 3, parametri descrittivi.

Lemma	A_o	Multi-*k*	Multi-π	α
girare	0.72	0.69	0.69	0.69
mescolare	0.8	0.7	0.7	0.7
rigirare	0.92	0.88	0.88	0.88
ruotare	0.77	0.69	0.69	0.7
voltare	0.82	0.67	0.66	0.67
TOTALE	0.75	0.73	0.73	0.73

Tabella 9: test 3, risultati.

Valori di *agreement* situati intorno alla soglia di 0.7, pur essendo inferiori ai risultati ottenuti dagli annotatori esperti del test 2, sono comunque da ritenersi accettabili, tanto più se si tiene in considerazione la completa assenza di *training*.

Tutti e quattro i *rater* hanno lamentato una maggior difficoltà nell'annotazione del verbo 'girare' rispetto agli altri lemmi, difficoltà che tuttavia non risulta dai dati. A differenza del test 2, l'unificazione dei tipi 9 e 10 in una unica categoria non porta particolari benefici: A_o= 0.7392, multi-*k* = 0.7064, multi-π = 0.7059, α = 0.7065. Il valore degli indici risulta abbassato, piuttosto, dal comportamento difforme di uno dei *rater*: se, sulla base dei risultati in tabella 9, si selezionassero i migliori tre annotatori (C, E, F) e si ricalcolassero i coefficienti, A_o= 0.8224, multi-*k* = 0.8078, multi-π = 0.8077, α = 0.8081.

	Pairwise agreement					
Lemma	C-D	C-E	C-F	D-E	D-F	C-D
girare	0.61	0.83	0.76	0.60	0.61	0.75
mescolare	0.65	0.70	0.79	0.65	0.56	0.90
rigirare	1.0	0.77	1.0	0.77	1.0	0.77
ruotare	0.67	0.66	0.65	0.83	0.67	0.66
voltare	0.48	0.85	1.0	0.58	0.48	0.85
TOTALE	0.61	0.83	0.76	0.61	0.61	0.75

Tabella 10: test 3, *pairwise agreement*.

4 Conclusioni

Notoriamente i task di annotazione semantica, ed in particolare quelli dedicati al lessico verbale (Fellbaum, 1998; Fellbaum *et al*., 2001), fanno registrare bassi livelli di I.TA.[5] Nel caso in oggetto la possibilità di ottenere valori alti, anche con annotatori non esperti, è con buona probabilità dovuta alla natura esclusivamente azionale e fisica delle classi usate per la categorizzazione.

In seguito alla validazione è stato possibile utilizzare i dati in applicazioni di tipo psicolinguistico (Gagliardi, 2014): il campionario di verbi dell'ontologia, ampio e al tempo stesso formalmente controllato, se integralmente validato potrebbe rappresentare una fonte inedita di dati semantici per le scienze cognitive. A tale scopo, oltre che per un pieno sfruttamento didattico e computazionale della risorsa,[6] in un prossimo futuro la metodologia illustrata verrà estesa ad una porzione quantitativamente e statisticamente significativa del database.

Acknowledgments

Il progetto IMAGACT è stato finanziato dalla regione Toscana nell'ambito del programma PAR.FAS. (linea di azione 1.1.a.3). Ulteriori ricerche, incluso questo articolo, sono state realizzate grazie al contributo del progetto MODELACT (2013-2016, Futuro in Ricerca).

[5] Per una rassegna dei risultati delle maggiori campagne di valutazione si veda Gagliardi (2014).

[6] Ad esempio per l'arricchimento di risorse semantiche esistenti e Word Sense Disambiguation. Si vedano a questo proposito Bartolini *et al.* (2014) e Russo *et al.* (2013).

Reference

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555–596.

Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, Monica Monachini. 2014. From Synsets to Videos: Enriching ItalWordNet Multimodally. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation – LREC'14*, ELRA – European Language Resources Association, pp.3110-3117.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines Inter-Coder Agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4): 699–725.

Steven Bird, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Beijing.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5): 423–9.

Susan Windisch Brown, Travis Rood and Martha Palmer. 2010. Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation? In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Seventh InternationalConference on Language Resources and Evaluation - LREC 2010*. ELRA – European Language Resources Association, pp. 3237-3243.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2): 249–254.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37-46.

Mark Davies and Joseph L. Fleiss. 1982. Measuring Agreement for Multinomial Data. *Biometrics*, 38(4): 1047–1051.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1): 95–101.

Christiane Fellbaum (ed.). 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs and Susan Wolf. 2001. Manual and Automatic Semantic Annotation with WordNet. In: *Proceedings of SIGLEX Workshop on WordNet and other Lexical Resources*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Francesca Frontini, Irene De Felice, Fahad Khan, Irene Russo, Monica Monachini, Gloria Gagliardi and Alessandro Panunzi. 2012. Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In: Michael Zock and Reinhard Rapp (eds.), *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, CogALex III*, pp. 69–80. The COLING 2012 Organizing Committee.

Gloria Gagliardi. 2014. *Validazione dell'Ontologia dell'Azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI)*. PhD thesis, Università degli Studi di Firenze, Italia.

Leo A. Goodman and William H. Kruskal. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268): 732– 764.

Klaus Krippendorff. 1980. *Content Analysis: an introduction to its Methodology*. Sage Publications, Newbury Park, CA, prima edizione.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Massimo Moneglia, Monica Monachini, Omar Calbrese, Alessandro Panunzi, Francesca Frontini, Gloria Gagliardi and Irene Russo. 2012. The IMAGACT cross-linguistic ontology of action. A new infrastructure for natural language disambiguation. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Eigth International Conference on Language Resources and Evaluation – LREC'12*, ELRA – European Language Resources Association, pp. 948-955.

Massimo Moneglia, Susan Windisch Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini and Alessandro Panunzi. 2014. The IMAGACT visual ontology. An extendable multilingual infrastructure for the representation of lexical encoding of action. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation – LREC'14*, ELRA – European Language Resources Association, pp.3425-3432.

Dennis Reidsma and Jean Carletta, 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3): 319–326.

Eleanor Rosch. 1978. Principles of categorization. In: E. Rosch and B. L. Lloyd (eds.), *Cognition and Categorization*, pp. 27-48. Lawrence Erlbaum Associates, Hillsdale, NW.

Irene Russo, Francesca Frontini, Irene De Felice, Fahad Khan, Monica Monachini. 2013. Disambig-

uation of basic action types through Nouns' Telic Qualia. In: Roser Saurì *et al.* (eds.), *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, Association for Computational Linguistics, pp. 70-75.

William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3): 321–325.

10.12871/CLICIT2014136

Correcting OCR errors for German in Fraktur font

Michel Généreux, Egon W. Stemle, Verena Lyding and Lionel Nicolas
EURAC Research
Viale Druso, 1 / Drususallee 1
39100 Bolzano / Bozen - Italy
{michel.genereux, egon.stemle}@eurac.edu
{verena.lyding, lionel.nicolas}@eurac.edu

Abstract

English. In this paper, we present ongoing experiments for correcting OCR errors on German newspapers in Fraktur font. Our approach borrows from techniques for spelling correction in context using a probabilistic edit-operation error model and lexical resources. We highlight conditions in which high error reduction rates can be obtained and where the approach currently stands with real data.

Italiano. *Il contributo presenta esperimenti attualmente in corso che mirano a correggere gli errori di riconoscimento ottico dei caratteri (OCR) in articoli di giornale scritti in lingua tedesca e nel carattere gotico* Fraktur. *L'approccio è basato su tecniche di controllo ortografico contestuale e utilizza un modello probabilistico di correzione degli errori assieme a delle risorse lessicali. Si descrivono le condizioni in cui è possibile ottenere un alto tasso di riduzione degli errori e si illustra infine lo stato di avanzamento attuale mediante dati reali.*

1 Introduction

The OPATCH project (Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage) aims at creating an advanced online search infrastructure for research in an historical newspapers archive. The search experience is enhanced by allowing for dedicated searches on person and place names as well as in defined subsections of the newspapers. For implementing this, OPATCH builds on computational linguistic (CL) methods for structural parsing, word class tagging and named entity recognition (Poesio et al., 2011). The newspaper archive contains ten newspapers in German language from the South Tyrolean region for the time period around the First World War. Dating between 1910 and 1920, the newspapers are typed in the blackletter Fraktur font and paper quality is derogated due to age. Unfortunately, such material is challenging for optical character recognition (OCR), the process of transcribing printed text into computer readable text, which is the first necessary pre-processing step for any further CL processing. Hence, in OPATCH we are starting from majorly error-prone OCR-ed text, in quantities that cannot realistically be corrected manually. In this paper we present attempts to automate the procedure for correcting faulty OCR-ed text.

2 Previous work

Projects, scientific meetings[1] and studies like OPATCH dealing with historical texts (Piotrowski, 2012) are numerous and one recurring theme is the struggle for clean OCR-ed data.

Approaches to post-OCR correction include machine learning (with or without supervision) (Abdulkader and Casey, 2009; Tong and Evans, 1996), merging of more than one system outputs (Volk et al., 2011) or high frequency words (Reynaert, 2008). The approach in Niklas (2010) combines several methods for retrieving the best correction proposal for a misspelled word: A general spelling correction (Anagram Hash), a new OCR adapted method based on the shape of characters (OCR-Key) and context information (bigrams). A manual evaluation of the approach has been performed on The Times Archive of London, a collection of English newspaper articles spanning from 1785 to 1985. Error reduction rates up to 75% and F-Scores up to 88% could be achieved.

For German, an approach akin to ours is Hauser (2007). This approach shares a number of features

[1] For example, DATeCH 2014: Digital Access to Textual Cultural Heritage, May 19-20 2014, Madrid, Spain.

with ours, such as a reference lexicon with similar coverage (90%) and fuzzy lookup matching of potential candidates in the lexicon for correction, based on the Levenshtein distance. However, while our weighting scheme for edit operations is based on an annotated corpus, Hauser (2007) uses a weighting model based on Brill (2000). Our approach also includes contextual information based on bigrams.

Hauser (2007) provides an evaluation of their approach on similar OCR-ed documents, that is from the same period (19th century) and font (Blackletter). Their evaluation on four collections shows error reduction rates from 1.9% to 4.8%, rates quite similar to those we report in tables 2 and 3. However, our results show error reduction rate can go up to 93%, depending on a number of idealized conditions which we will spell out further.

3 Corpora

We used two types of data sources: ten OCR-ed newspaper pages along with their manually corrected version, our Gold Standard (GS), and an independent reference corpus of German.

3.1 OCR-ed pages

Each of the ten pages has been OCR-ed[2] and revised manually[3] , so that for each page we have a scanned image in format TIF, an OCR-ed version in the format METS[4]-ALTO[5] and a manually revised version of the text. The Fraktur font and the decreased paper quality make the translation into text particularly challenging, so the OCR-ed documents are extremely noisy. On average, more than one out of two tokens is misrecognized (see table 3), let alone a substantial number of fragmented and missing tokens. Almost half (48%) of tokens need a minimum of three edit operations for correction (see section 4.1). In total, the OCR-ed documents are made up of 10,468 tokens and 3,621 types. Eight pages (8,324/2,487) are used as training data (section 4) and two pages (2,144/1,134) for testing. One such page is shown in figure 1.


Der Kaiserlich Königlich
privil. legirte
Bothe von Tyrol.
Donnerstag Nro. 69. 28. August 1817.

Figure 1: A typical OCR-ed page

3.2 Reference corpus

The reference corpus is used as a basis for constructing a frequency list of unigrams (a dictionary) and a frequency list of bigrams. We used the SdeWaC corpus (Faaß and Eckart, 2013), a German corpus harvested from the web. We enriched the corpus with texts closer in time to the OCR-ed documents, that is texts in the categories of novels and stories (*Romane und Erzählungen*) from the period 1910-20, a total of 1.3 M tokens.[6] Our final reference corpus is made of 745M tokens from which we derived a dictionary of size 5M and a list of bigrams of 5M.[7] The 5M entries in the dictionary cover 91% of all tokens from the manually corrected OCR-ed files.

4 Approach

The approach consists of three steps: first we build a probabilistic model of edit-operations needed for correction, then we define a procedure to generate candidates for correction based on the model and finally we apply a scoring system to evaluate the most suitable candidate.

[2]Using ABBYY: `http://www.abbyy.com/`

[3]The GSs have not been aligned with the originals, so that there is no trace of where words added, subtracted or simply corrected.

[4]Metadata Encoding & Transmission Standard: `http://www.loc.gov/standards/mets/`

[5]Analyzed Layout and Text Object: `http://www.loc.gov/standards/alto/`

[6]Project Gutenberg: `http://www.gutenberg.org/`

[7]The size of the dictionaries was optimized for running times and computer memory.

4.1 Constitution of the edit-operations probability model

A correction is deemed necessary when a token has no entry in the dictionary. The OCR error correction system uses a probabilistic model built on typical edit errors to generate candidates when a correction is required. To build such a model, our first task is to collate and tally all edit-operations (*delete*, *insert* and *replace*) needed to transform all unrecognized tokens from the training OCR-ed texts to its corrected form in the GS. For example, to transform the token *Veranstaltnngstage* to *Veranstaltungstage* 'days of the event', we must replace the second 'n' with a 'u'. This edit-operation, replacing an 'n' with a 'u', is therefore recorded and tallied. From these counts we build a probability distribution. This part of the system finds its inspiration from Segaran and Hammerbacher (2009). This model defines our alphabet, which includes all the graphemes for German and all spurious symbols generated by the OCR process.[8] All edit-operations recorded constitute the constrained model. The unconstrained model also includes all edit-operations unseen during training, which are assigned a low residual probability.

4.2 Candidate generation

Candidate generation is achieved by finding the closest entries in the dictionary by applying the minimum number of edit-operations to an unrecognized OCR-ed token. The number of candidates is a function of the maximum number of edit-operations allowed and the model used (constrained or not) and may often be by the hundreds, and the sheer number of possibilities makes finding the closest entry more difficults. For example, if presented with the token *wundestc* and asked to generate all candidates within two edit-operations and using the constrained model, the system generates eight candidates, among which: *wundesten* 'sorest', by replacing a 'c' with an 'e' and then inserting an 'n' after the 'e'. When asked to use the unconstrained model, the number of candidates raises to fifteen.

4.3 Selection of the most suitable candidate

We consider the following four features to select the best possible candidate:

- The probability of all edit-operations multiplied together. The more number of operations involved, the lower the probability. In the example we presented in section 4.2, the edit cost to go from *wundestc* to *wundesten* would be the probability of replacing 'c' for 'e' multiplied by the probability of inserting an 'n' after an 'e'.
- The probability of the candidate drawn from the reference corpus (the relative frequency). This would be the frequency of *wundesten* (24) divided by the size of the reference corpus (745M).
- The two probabilities of co-occurrence of the candidate with the immediate left (and also right) neighbour of the token to correct. Probabilities are drawn from the frequency list of bigrams in the reference corpus and processing is carried out left-to-right.

Each candidate is then given a score by simply adding together the values for each of the four features above. In order for each of the features to contribute fairly to the overall score, we normalized the three distributions (*edit*, *unigram* and *bigram*) so that their mean is the same. We also stretched each distribution so that least probable values for a feature tend to zero and most probable to one. The scoring formula for candidate 'c' is:

$$\begin{aligned}&\prod_i prob(edit\_op_i) + prob(c)\\ &+ prob(left\_word + c)\\ &+ prob(c + right\_word)\end{aligned} \quad (1)$$

5 Experiments

In the following experiments we first used a small list of OCR errors to test our model and then we applied the system on the two remaining pages from the ten OCR-ed pages.

5.1 Artificially generated errors

In order to have a comparable set of conditions to evaluate how the system performs, we generated a list of 2,363 errors somewhat artificially. To achieve this we extracted random trigrams from the GS (left context, target, right context) and applied, in reverse, the edit error model. Errors were introduced up to a limit of two per target and contexts. At the end of this process, we have two context words and five candidates, including the target. Table 1 shows the results. When given

[8] Alphabet: üÜöÖäÄß»«èà„ì^()-/016 and [a-z][A-Z]

Five cand.	On the fly, open list of candidates					
	Constr. model			Unconstr. model		
	E1	E2	E3	E1	E2	E3
93%	86%	61%	40%	83%	28%	9%

Table 1: Error reduction rate on 2,363 artificially created errors

five candidates, the system picked the target 93% of the time. The En labels indicate the maximum edit-operations performed[9] to generate candidates. When candidates are generated 'on-the-fly' in variable quantity, we can see a drop in error reduction which was best when we limited the number of candidates (small value for n) and used the constrained model of edit errors. This is hardly surprising, given how the errors were generated in the first place.

5.2 Real errors

We now turn our attention to real errors, those coming from the two remaining pages of our OCR corpus. Table 2 shows the result. The set of 233

Constrained model			Unconstrained model		
E1	E2	E3	E1	E2	E3
16%	18%	15%	20%	16%	9%

Table 2: Error reduction rate on 233 real errors

errors is the result from aligning the two OCR-ed texts with their corresponding GS. We kept only tokens for which we had a clear alignment (see footnote 3 on page 2) and a target which was part of the dictionary. Accuracies dropped drastically for all types of configuration, due to a high proportion of tokens heavily modified by the OCR process (edit distance above 2). Finally, we applied our system to the whole of the two test pages. Evaluating the performance was made difficult because the OCR process may have deleted tokens and fragmented some. Table 3 shows counts of how many tokens have been preserved from the GS to the OCR-ed files as well as to the files corrected by the system (AC). To obtain counts, we compared files line by line as bag of words. Therefore, word order was not taken into account, but the line based comparison mitigated this effect for the text as a whole. Not surprisingly, accuracies were on average 10% lower than those from table

[9]One caveat: n is increased by 1 when the candidate's length is above four. The longer the word we are trying to correct, the more edit-operations necessary.

	Constr. model		Unc. model	
	E1	E2	E1	E2
|GS|	2153	2153	2153	2153
|OCR|	2322	2322	2322	2322
|GS ∩ OCR|	1185	1185	1185	1185
|GS ∩ AC|	1268	1263	1279	1234
Improvement	7%	7%	8%	4%

Table 3: Error reduction rate. | | = size of

2, which can be explained by the fact that not all targets from the test set can be found in the dictionary.

Two final remarks about the evaluation presented. That a token is part of the dictionary does not mean that it is correct. In fact, wrong substitutions constitute a very hard problem with OCR-ed texts and a source of contamination difficult to trace and fix. There is also the problem of misleading tokenization by missing or falsely inserting space characters, producing disrupted and continued tokens which cannot be corrected by comparing words one by one. Work such as Furrer (2013) is an attempt to improve post-correction of OCR-ed texts by using the internal structure of tokens to produce a tokenization scheme less sensitive to segmentation errors produced by the recognition system.

6 Conclusion

The approach we presented to correct OCR errors considered four features of two types: edit-distance and n-grams frequencies. Results show that a simple scoring system can correct OCR-ed texts with very high accuracy under idealized conditions: no more than two edit operations and a perfect dictionary. Obviously, these conditions do not always hold in practice, thus an observed error reduction rate drops to 10%. Nevertheless, we can expect to improve our dictionary coverage so that very noisy OCR-ed texts (i.e. 48% error with distance of at least three to target) can be corrected with accuracies up to 20%. OCR-ed texts with less challenging error patterns can be corrected with accuracies up to 61% (distance 2) and 86% (distance 1).

Acknowledgments

We are grateful to Katalin Szabò from Teßmann for producing the Gold Standard (GS).

References

Ahmad Abdulkader and Mathew R. Casey. 2009. *Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment*. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, pages 576–580, Washington, DC, USA.

Andreas W. Hauser. 2007. *OCR Postcorrection of Historical Texts*. Master Thesis, Ludwig-Maximilians-Universität München.

Gertrud Faaß and Kerstin Eckart. 2013. *Sdewac - a corpus of parsable sentences from the web*. In Gurevych, Biemann and Zesch, editors, GSCL, volume 8105 of Lecture Notes in Computer Science, pages 61–68. Springer.

Eric Brill and Robert C. Moore. 2000. *An improved error model for noisy channel spelling correction*. In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 286–293, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

Kai Niklas. 2010. *Unsupervised Post-Correction of OCR Errors*. PhD Thesis, Leibniz Universität Hannover.

Lenz Furrer. 2013. *Unsupervised Text Segmentation for Correcting OCR Errors*. Master Thesis, Universität Zürich, July 2013.

Martin Reynaert. 2008. *Non-interactive ocr post-correction for giga-scale digitization projects*. In Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, CICLing'08, pages 617–630.

Martin Volk, Lenz Furrer and Rico Sennrich. 2011. *Strategies for reducing and correcting OCR error*. In Language Technology for Cultural Heritage, pages 3–22, Sporleder, Caroline and Bosch, Antal van den and Zervanou, Kalliopi, ISBN 978-3-642-20226-1.

Massimo Poesio, Eduard Barbu, Egon W. Stemle and Christian Girardi. 2011. *Natural Language Processing for Historical Texts*. In Proc. 5th ACL-HLT Work. Lang. Technol. Cult. Heritage, Soc. Sci. Humanit. (LaTeCH 2011), pages 54–62, Portland, OR, USA. Association for Computational Linguistics.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Leibniz Institute of European History, Synthesis Lectures on Human Language Technologies 17, Morgan and Claypool Publishers.

Toby Segaran and Jeff Hammerbacher. 2009. *Beautiful Data: The Stories Behind Elegant Data Solutions*. Theory in practice. O'Reilly Media.

Xion Tong and David A. Evans. 1996. *A Statistical Approach to Automatic OCR Error Correction In Context*. In Proceedings of the Fourth Workshop on Very Large Corpora (WVLC-4), pages 88–100.

10.12871/CLICIT2014137

Some issues on Italian to LIS automatic translation. The case of train announcements

Carlo Geraci
CNRS, Institut Jean-Nicod,
Rue d'Ulm 29,
75005 Paris
carlo.geraci76@gmail.com

Alessandro Mazzei
Università di Torino,
Corso Svizzera 185,
10149 Torino
mazzei@di.unito.it

Marco Angster
Libera Università di Bolzano,
Piazza Università 1,
39100 Bolzano
marco.angster@gmail.com

Abstract

English. In this paper we present some linguistic issues of an automatic translator from Italian to Italian Sign Language (LIS) and how we addressed them.

Italiano. *In questo lavoro presentiamo alcune questioni linguistiche inerenti la traduzione automatica da Italiano a lingua dei segni italiana (LIS).*

1 Introduction

Computational linguistic community showed a growing interest toward sign languages. Several projects of automatic translation into signed languages (SLs) recently started and avatar technology is becoming more and more popular as a tool for implementing automatic translation into SLs (Bangham et al. 2000, Zhao et al. 2000, Huenerfauth 2006, Morrissey et al. 2007, Su and Wu 2009). Current projects investigate relatively small domains in which avatars may perform decently, like post office announcements (Cox et al., 2002), weather forecasting (Verlinden et al., 2002), the jurisprudence of prayer (Almasoud and Al-Khalifa, 2011), driver's license renewal (San-Segundo et al., 2012), and train announcements (e.g. Braffort et al. 2010, Ebling/Volk 2013).

LIS4ALL is a project of automatic translation into LIS where we faced the domain of public transportation announcements. Specifically, we are developing a system of automatic translations of train station announcements from spoken Italian into LIS. The project is the prosecution of ATLAS, a project of automatic translation into LIS of weather forecasting (http://www.atlas.polito.it/index.php/en). In ATLAS two distinct approaches to automatic translation have been adopted, interlingua rule-based translation and statistical translation (Mazzei et al. 2013, Tiotto et al., 2010, Hutchins and Somer 1992). Both approaches have advantages and drawbacks in the specific context of automatic translation into SL. The statistical approach provides greater robustness while the symbolic approaches is more precise in the final results. A preliminary evaluation of the systems developed for ATLAS showed that both approaches have similar results. However, the symbolic approach we implemented produces the structure of the sentence in the target language. This information is used for the automatic allocation of the signs in the signing space for LIS (Mazzei et al. 2013), an aspect not yet implemented in current statistical approaches.

LIS4ALL only uses the symbolic (rule-based) translation architecture to process the Italian input and generate the final LIS string. With respect to ATLAS, two main innovations characterize this project: new linguistic issues are addressed; the translation architecture is partially modified.

As for the linguistic issues: we are enlarging the types of syntactic constructions covered by the avatar and we are increasing the electronic lexicon built for ATLAS (around 2350 signs) by adding new signs (around 120) specific to the railway domain. Indeed, this latter was one of the most challenging aspects of the project especially when the domain of train stations is addressed. Prima facie this issue would look like a special case of proper names, something that should be easily addressed by generating specific signs (basically one for every station). However, the solution is not as simple as it seems. Indeed, several problematic aspects are hidden when looking at the linguistic situation of names in LIS (and more generally in SL). As for the translation architecture, while in ATLAS a real interlingua translation with a deep parser and a FoL meaning representation were used, in LIS4ALL, we decided to employ a regular-expression-based analyzer that produces a simple (non recursive) filler/slot based semantic to parse the Italian input. This is so, because in the train announcement domain, input sentences have a large number of complex noun phrases with several prepo-

sitional phrases, resulting in a degraded parser performance (due to multiple attachment options). Moreover, the domain of application is extremely regular since the announcements are generated by predefined paths (RFI, 2011).
The rest of the paper is organized as follows: Section 2 discusses the linguistic issues, Section 3 discusses the technical issues while Section 4 concludes the paper.

2 Linguistic Issues

The domain of application consists of the messages broadcasted in Italian train stations. Rete Ferroviaria Italiana (RFI) produced a manual, called MAS (Manuale degli Annunci Sonori), that describes the details of each specific message (RFI, 2011). MAS specifies 39 templates that RFI uses to automatically produce the messages: 15 templates deal with leaving trains (A1,..., A15), 13 templates with arriving trains (P1, ..., P13), while 11 messages with special situations (e.g. strikes, R1, ..., R13). The templates have been designed to produce concise and direct messages in Italian. Full relative clauses, coordination and complex structures (e.g. ellipses) are avoided. As a consequence, the domain is that of a controlled language. In Fig. 1 there is a fragment of the template A1, that concerns the leaving of a train without additional information on (in time or place) changes in the schedule.



Figure 1. A fragment of the A1 template (RFI, 2011).

The template includes fixed parts (e.g. “IL TRENO”), variables (e.g. “CATEGORIA” “NUMERO”) and optional parts (e.g. “IN RITARDO”). By analyzing a corpus of 24 hours messages produced at the Torino Porta Nuova Station (5014 messages total) we found that a small number of templates covers the majority of announcements while others are virtually absent (Table 1).

#messages	Template Name	%
1818	A1	36.26
1310	P1	26.13
685	A2	13.66
431	A3	8.60
52	P9	1.04
48	P5	0.96
19	A5	0.38
2	P13	0.04
649	*other templates*	12.94
TOT. 5014		

Table 1. The templates occurrences in 24 hours of Torino Porta Nuova station messages.

2.1 An Italian-LIS parallel corpus

In order to have a minimal but significant bilingual corpus Italian-LIS, we chose a subset of 7 sentences, which have been translated in LIS by a Deaf[1] native signer, supervised by the help of a professional LIS interpreter and a Sign Language linguistics researcher.
Focusing on the nominal domain a number of differences between Italian and LIS emerged. To mention one, consider the quite simplified subject in (1) and its LIS counterpart in (2):
(1) Il treno per Susa …
'the train to Susa … '
(2) TRAIN SUSA GO …
'The train going to Susa'
While the Italian NP is modified by a prepositional phrase, the LIS NP is modified by what we analyzed as a reduced relative clause.
At the clausal level, the syntactic complexity of the subjects in the input language forced the introduction of a pronominal pointing that we analyzed as a resumptive subject clitic, a phenomenon completely absent from Italian.

2.2 The issue of station names

Another crucial linguistic issue concerns the best way to translate the names of the stations in LIS. Indeed, the linguistic situation of names is quite heterogeneous and can be summarized as follows: (1) Sign names fully acknowledged by the Italian Deaf communities; (2) Sign names only acknowledged by (part of) the local Deaf community; (3) There is no sign name even within the local community.

[1]Capital “D” is used to refer to deaf people who are signers and part of the signing community as opposed to people who simply suffer of an acoustic deficit.

The first option illustrates the case of most main stations in big cities. Normally, the name of the station is semantically transparent, as in "Milano centrale", or it involves the name of some prominent character of the Italian history, as in the case of "Milano Porta Garibaldi". However, most of the trains go to and stop at obscure locations. In some cases, local dialects have a specific sign for those stations (normally, the name of the town where the train stops) as in the station of "Castelvetrano". Finally, there are Italian names for which not even the local Deaf community has already developed a local sign name. In those cases, human signers adopt the last resorts at their disposal, namely either they fingerspell the name, or they use mouthing, as in the case of "Rebaudengo Fossata", a very small station in Turin.

Fingerspelling is the typical way in which borrowings from spoken languages are realized (Brentari 2000). However, this practice is not fully adopted by the Italian Deaf communities yet. Indeed, old signers may not know the manual alphabet and in some cases they even refuse to use it, rather preferring the mouthing of word in spoken Italian (Volterra 1987 and Caselli et al. 1996).

Once we leave the domain of human signers and enter the world of signing avatar, additional issues are raised which are specifically connected to fingerspelling and mouthing. Clearly, mouthing is a solution that cannot be usefully pursued for practical reasons: The avatar technology is designed to be portable on different devices including smartphones. Within this framework, lipreading would be almost impossible for most users of the service. Furthermore, working in the domain of public transportation announcements, the timing issue is not trivial. Announcements are normally broadcasted and fingerspelling would introduce additional delay to the sign production, which normally is more time consuming than speech.

After having preliminarily consulted some members of the local Deaf Association of the city where the automatic translation system will be first released (ENS Torino), a twofold solution is going to be adopted: 1. Sign names fully acknowledged by the Italian Deaf communities will be maintained by the signing avatar; 2. Blended written Italian-LIS sign forms will be used (Geraci and Mazzei, 2014).

While names of main stations in big cities are preserved in their original LIS forms, as in Fig. 2., a new strategy is developed for less-familiar stations and gaps in the vocabulary. The avatar will play a classifier sign indicating a wide board while the name of the station will appear in written Italian "centered on the board", as shown in Fig. 3.



Figure 2. Animation for "Milano Centrale"

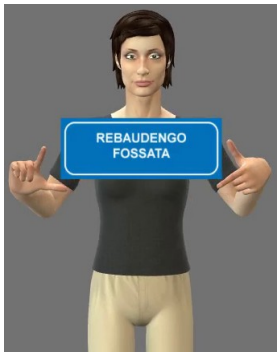


Figure 3. Animation for "Rebaudengo Fossata"

This technical solution blends a manual sign (a generic classifier) with a non-manual component. However, rather than using the standard non-manual channels (facial expressions or body postures), this solution adopts a tool which is not internal to sign language, namely the written form of the dominant language. From the communicative perspective, this solution is much more performative than standard fingerspelling for at least three reasons: 1. It allows a faster assessment of the lexical item since the written input is produced simultaneously and not letter by letter; 2. It does not overload the processing of the entire sentence; 3. It is accessible to all signers, even those with lower levels of literacy. From the timing perspective, blended forms are much quicker to perform than fingerspelling making the entire announcement more alignable with its spoken counterpart. The decision of implementing two separate strategies for train station names rather than extending the blending strategy to all station names has been made after having preliminary consulted our linguistic informants. However, we are planning to assess a broader part of the Deaf community on this specific issue.

3 Technical Issues

Figure 4 illustrates the pipeline of the current architecture and includes five modules: (1) a regular expression parser for Italian; (2) a filler/slot based semantic interpreter; (3) a generator; (4) a spatial planner; (5) an avatar that performs the synthesis of the sequence of signs, i.e. the final LIS sentence. Note that we had access to the MAS manual but we did not have access to the technology used to generate announcements in the station. So, we could not use any additional information, apart from the message, for the translation.

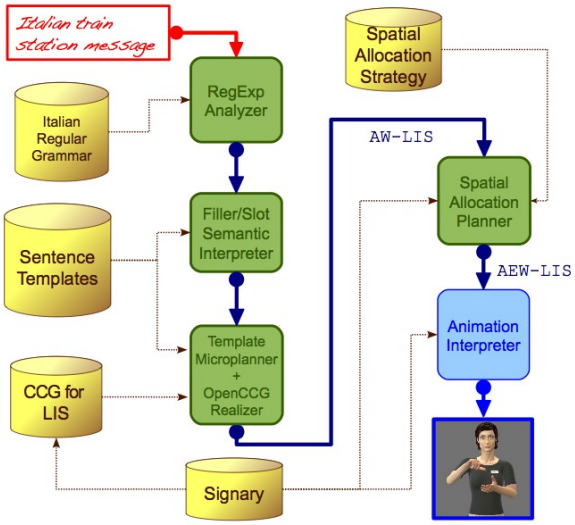


Figure 4: LIS4ALL translation architecture

By using the MAS, we built 8 regular expressions corresponding to the 8 most frequent templates found in our corpus (see Table 1). For each template, we designed a sequence of semantic slots that are filled by lexical items (time, rail, station names, etc.) contained in a specific message. Each singular slot corresponds to a singular variable of the message template (see Fig. 1), that is filled by a domain lexical element (e.g. “milano centrale” or “straordinario”). We plan to cover the remaining templates by the end of the project.

The LIS4ALL generator is composed by two submodules: a microplanner and a realizer (Reiter and Dale, 2000). The microplanner decides about the syntactic organization of the sentences and about which signs to use in the generation. Following Foster and White (2004), we implemented a template based microplanner that is able to exploit the filler/slot structure produced by the semantic analyzer. The output of the microplanner is a hybrid logic formula in a tree-structure (XML), that encodes an abstract syntactic tree. Extending the CCG grammar (Steedman, 2000) designed in the ATLAS (Mazzei 2012), and using the parallel corpus Italian-LIS produced in LIS4ALL, we implemented a new CCG grammar for LIS that can be used by the OpenCCG realizer to produce LIS sentences in the railway domain (White 2006). Finally, the spatial planner accounts for the signs positions in the space by using the same strategy used for ATLAS (this module of the architecture is still in progress.).

In order to implement our solution for stations names we implemented a double access procedure to the signing lexicon in the generator. In a first attempt, the microplanner will search in the lexicon for a direct translation of an Italian station name into LIS (see above "Milano centrale"). If at least one translation is found, then the avatar follows the standard ATLAS communication pipeline and performs the (sequence of) sign(s). If this procedure does not produce results, for instance, when there is a lexical gap in the LIS dictionary for the station name, the microplanner and the realizer command the avatar to produce the Italian-LIS blending for that specific station name in real time. So, we augmented the avatar to allow for the production of a real time Italian-LIS blending from a string (up to 40 characters). Finally, we augmented the communication protocol between SentenceDesigner and the avatar, by adding a new tag <SIGNBOARD> to the AEWLIS (ATLAS Extended Written LIS), i.e. to the XML language in use for the communication between the generator and the avatar.

4 Conclusions

In this paper we considered two issues related to the development of an automatic translator from Italian to LIS in the railway domain. These are: 1) some syntactic mismatches between input and target languages; and 2) how to deal with lexical gaps due to unknown train station names. The first issue emerged in the creation of a parallel Italian-LIS corpus: the specificity of the domain allowed us to use a naive parser based on regular expressions, a semantic interpreter based on filler/slot semantics, a small CCG in generation. The second issue has been addressed by blending written text into a special “sign”. In the next future we plan to quantitatively evaluate our translator.

Acknowledgments

This work has been partially supported by the project LIS4ALL (http://www.poloinnovazioneict.org/index.php?IDpage=5989&IDcontenuto=100) partially funded by Regione Piemonte, Innovation Hub for ICT, 2011-2014, POR-FESR 07-13. Part of the research leading to these results also received funding from the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013) / ERC Grant Agreement N°324115–FRONTSEM (PI: Schlenker). Research was partially conducted at Institut d'Etudes Cognitives (ENS), which is supported by grants ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

This work is dedicated to Leonardo Lesmo who substantially contributed to its realization

Reference

Almasoud, A. M. and Al-Khalifa, H. S. (2011). A proposed semantic machine translation system for translating arabic text to arabic sign language. In Proceedings of the Second Kuwait Conference on e-Services and e-Systems, KCESS '11, pages 23:1–23:6, New York, NY, USA. ACM.

Bangham, J., Cox, S., Elliott, R., Glauert, J., and Marshall, I. (2000). Virtual signing: Capture, animation, storage and transmission – an overview of the VisiCAST project. In. In IEE Seminar on Speech and Language.Braffort, A. et al.: Sign language corpora for analysis, processing and evaluation. In Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.

Brentari, Diane. 2000. Foreign Vocabulary in Sign Languages. (Ed.) Diane Brentari. Mahwah, NJ: Lawrence Erlbaum Associates.

Caselli et al. 1996. Linguaggio e sordità. Il Mulino, Bologna.

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., and Abbott, S. (2002). Tessa, a system to aid communication with deaf people. In Proceedings of the fifth international ACM conference on Assistive technologies, pages 205–212. ACM.

Ebling, S; Glauert, J (2013). Exploiting the full potential of JASigning to build an avatar signing train announcements. In: Third International Symposium on Sign Language Translation and Avatar Technology, Chicago, IL, USA, 18 October 2013 – 19 October 2013.

Foster, M. E. and White, M. (2004). Techniques for text planning with xslt. In Proc, of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology, NLPXML '04, pages 1–8, Stroudsburg, PA, USA. ACL

Geraci, C. and Mazzei, A. (2014). Last train to "Rebaudengo Fossano": The case of some names in avatar translation. In Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages., pages 63–66.

Huenerfauth, M. (2006). Generating American Sign Language classifier predicates for english-to-asl machine translation. PhD thesis, University of Pennsylvania.

Hutchins, W. and Somer, H. L. (1992). An Introduction to Machine Translation. London: Academic Press.

Mazzei, A. (2012). Sign Language Generation with Expert Systems and CCG. In Proceedings of the 7th International Natural Language Generation Conference, pages 105–109, Starved Rock State Park Utica, IL USA. ACL

Mazzei, A., Lesmo, L., Battaglino, C., Vendrame, M., and Bucciarelli, M. (2013). Deep natural language processing for italian sign language translation. In Proc. of XIII Conference of the Italian Association for Artificial Intelligence, volume 8249 of (LNCS), pages 193–204, Turin. Springer.

Morrissey, S., Way, A., Stein, D., Bungeroth, J., and Ney, H. (2007). Combining data-driven mt systems for improved sign language translation. In Proc. XI Machine Translation Summit.

Ong, S. C. W. and Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell., 27(6):873–891.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., and D'Haro, L. F. (2012). Design, development and field evaluation of a spanish into sign language translation system. Pattern Anal. Appl., 15(2):203–224.

Steedman, M. (2000). The syntactic process. MIT Press, Cambridge, MA, USA.

Su, H. . and Wu, C. (2009). Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory. In IEEE Transactions on Audio, Speech and Language Processing, 17 (7), 1305-1315.

Tiotto, G., Prinetto, P., Piccolo, E., Bertoldi, N., Nunnari, F., Lombardo, V., Mazzei, A., Lesmo, L., and Principe, A. D. (2010). On the creation and the annotation of a large-scale Italian-LIS parallel corpus. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Lan-

guage Technologies, Valletta, Malta. ISBN 10: 2-9517408-6-7.

Reiter, E. and Dale, R. (2000). Building natural language generation systems. Cam- bridge University Press, New York, NY, USA.

Rete Ferroviaria Italiana (RFI) (2011). Manuale degli Annunci Sonori - MAS. http://www.rfi.it/cms-file/allegati/rfi/MAS.pdf.

White, M. (2006). Efficient realization of coordinate structures in combinatory categorial grammar. Research on Language and Computation, 2006(4(1)):39—75.

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N. I., and Palmer, M. (2000). A machine translation system from english to american sign language. In Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future, AMTA '00, pages 54–67, London, UK, UK. Springer-Verlag.

10.12871/CLICIT2014138

ConParoleTue: crowdsourcing al servizio di un *Dizionario delle Collocazioni Italiane per Apprendenti* (*Dici-A*)

Andrea Gobbi
Dipartimento di Scienze Politiche, Sociali e della Comunicazione, Università di Salerno
andgobbi@gmail.com

Stefania Spina
Dipartimento di Scienze Umane e Sociali, Università per Stranieri di Perugia
stefania.spina@unistrapg.it

Abstract

English. *ConParoleTue* è un esperimento di uso del crowdsourcing nell'ambito della lessicografia L2. A partire dalla costituzione di un dizionario di collocazioni per apprendenti di italiano L2, *ConParoleTue* rappresenta un tentativo di re-inquadramento di problematiche tipiche dell'elaborazione lessicografica (la qualità e il registro delle definizioni) verso una maggiore centralità delle necessità comunicative di chi apprende. A questo fine una metodologia basata sul crowdsourcing viene sfruttata per la redazione delle definizioni. Questo articolo descrive tale metodologia e presenta una prima valutazione dei suoi risultati: le definizioni ottenute attraverso il crowdsourcing sono quantitativamente rilevanti e qualitativamente adatte a parlanti non nativi dell'italiano.

Italiano. *ConParoleTue* is an experiment of adoption of crowdsourcing techniques applied to L2 lexicography. It started while compiling a dictionary of collocations for learners of Italian as a second language, and it uses crowdsourcing to find new solutions, both quantitatively and qualitatively, to traditional issues connected with lexicography, such as the quality and the register of definitions, towards a more learner-centred approach. This paper describes our methodology and a first evaluation of results: the definitions acquired through crowdsourcing are quantitatively relevant and qualitatively appropriate to non-native speakers of Italian.

1 Introduzione

ConParoleTue (2012) è un esperimento di applicazione del crowdsourcing all'ambito della lessicografia L2, elaborato all'interno del Progetto APRIL (Spina, 2010b) dell'Università per Stranieri di Perugia nel corso della costituzione di un dizionario di collocazioni per apprendenti di italiano L2.

Le collocazioni occupano da alcuni decenni un posto di primo piano negli studi sull'apprendimento di una lingua seconda (Meunier e Granger, 2008). Quella collocazionale è riconosciuta come una competenza chiave per un apprendente, perché svolge un ruolo fondamentale nei due aspetti della produzione (fornisce infatti blocchi lessicali precostituiti e pronti per essere utilizzati, migliorando la fluenza; Schmitt, 2004) e della comprensione (Lewis, 2000). Anche nell'ambito della lessicografia italiana la ricerca sulle collocazioni è stata particolarmente produttiva, ed ha portato, negli ultimi cinque anni, alla pubblicazione di almeno tre dizionari cartacei delle collocazioni italiane: Urzì (2009), nato in ambito traduttivo; Tiberii (2012) e Lo Cascio (2013).

Il *DICI-A* (*Dizionario delle Collocazioni Italiane per Apprendenti*; Spina, 2010a; 2010b) è costituito dalle 11.400 collocazioni italiane estratte dal *Perugia Corpus*, un corpus di riferimento dell'italiano scritto e parlato contemporaneo[1].Tra le tante proposte, la definizione alla base della costituzione del *DICI-A* è quella di Evert (2005), secondo cui una collocazione è "a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon". Le collocazioni del *DICI-A* appartengono a 9 categorie diverse, selezionate sulla base delle sequenze più produttive di categorie grammaticali che le compongono: aggettivo-nome (*tragico errore*), nome-aggettivo (*anno prossimo*), nome-nome (*peso forma*), verbo-(art.)-nome (*fare una domanda/fare pena*), nome-preposizione-nome (*carta di credito*), aggettivo-*come*-nome (*fresco come una rosa*), aggettivo-congiunzione-aggettivo (*sano e salvo*), nome-congiunzione-nome (*carta e penna*), verbo-aggettivo (*costare caro*).

Per ogni collocazione sono stati calcolati gli indici di Juilland di dispersione e di uso (Bortoli-

[1] http://perugiacorpus.unistrapg.it

ni et al., 1971), sulla base dei quali sono state selezionate le collocazioni definitive. Si è presentato dunque il problema di come procedere alla loro definizione. In questo contesto è nata l'idea dell'impiego del crowdsourcing, e all'elaborazione di *ConParoleTue*.

2 La scelta del crowdsourcing

L'adozione del crowdsourcing in linguistica è principalmente legata ad obiettivi di ottimizzazione delle risorse (Snow et al., 2008; Hsueh et al., 2009), in particolare nell'ambito della traduzione (Callison-Burch, 2009), della creazione di corpora (Wang et al., 2012; Post et al.,2012) e della loro annotazione (Munro et al., 2010); tra le metodologie e gli strumenti più utilizzati figurano *Mechanical Turk* di Amazon (Schnoebelen e Kuperman, 2010) e i serious games (Kneissl e Bry, 2012).

Oltre all'aspetto dell'ottimizzazione delle risorse, tuttavia, la scelta del crowdsourcing per il *DICI-A* è stata dettata anche da un preciso approccio alla lingua,che presta particolare attenzione alla natura sociale e condivisa dello strumento linguistico, da cui derivano i suoi specifici processi acquisizionali (Gobbi, 2012; Gobbi, 2013; Gobbi e Spina, 2013).

Il coinvolgimento di una platea molto ampia di collaboratori per acquisire le definizioni delle collocazioni da includere nel dizionario, e il modo stesso con il quale il progetto è stato presentato (ogni richiesta di definizione recitava: "Come lo spiegheresti ad un tuo amico straniero?") era volutamente teso ad elicitare il maggior grado possibile di naturalezza e spontaneità nelle risposte. Da un punto di vista meta lessicografico, ciò ha comportato la decisione di non richiedere ai contributori di conformarsi ad uno stile predeterminato di definizione, allo scopo di perseguire le condizioni di informalità dell'interazione quotidiana. I vantaggi di un tale approccio collaborativo, sviluppato dal basso e mirato alla naturalezza delle definizioni, sono diversi, e di diversa natura: in primo luogo, quello di offrire agli apprendenti e futuri utenti del *DICI-A* uno strumento che fornisca risposte meno accademiche e più formalmente simili a quelle ottenibili nella vita quotidiana, e dunque adeguate ad un contesto interazionale. Un tale approccio, inoltre, si presta alla sensibilizzazione di parlanti nativi su questioni linguistiche, quali il dover riflettere su come definire un'espressione con altre parole, operazione di fatto non semplice (Schafroth, 2011). Infine, lo sviluppo di uno strumento di riferimento per apprendenti di una L2 come un'opera collettiva, sebbene monitorato e revisionato nella sua forma finale, rappresenta una sfida interessante ed ambiziosa, oltre che un esperimento applicativo di metodologie che sempre più spesso si rivelano preziose nella ricerca linguistica.

2.1 Metodologia

Per la realizzazione dell'esperimento, è stata innanzitutto predisposta una piattaforma web dedicata[2]. Dopo una breve schermata di presentazione, attraverso la piattaforma vengono raccolti pochi dati essenziali sui partecipanti (età, sesso, titolo di studio, madrelingua, eventuale livello QCER di italiano), al fine di acquisire alcune informazioni sociolinguistiche di base su ciascuno degli autori delle definizioni.

Il sistema propone quindi, una dopo l'altra, cinque collocazioni da definire, estratte a caso dal database (fig. 1).

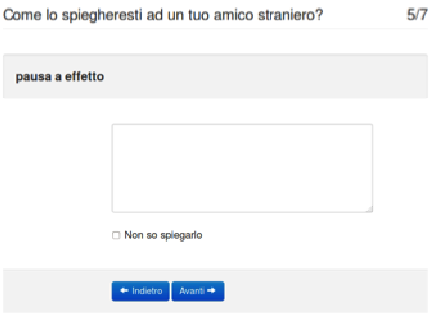


Figura 1 - Esempio di collocazione da definire

Infine, viene chiesto di valutare due definizioni precedentemente elaborate da altri partecipanti, allo scopo di effettuare un primo filtraggio di eventuali definizioni inappropriate (fig. 2).

Il progetto è stato ampiamente diffuso tramite vari social media (una pagina Facebook e un profilo Twitter appositamente creati), una newsletter dedicata, e numerosi contatti istituzionali. Sebbene sia tuttora online, il periodo di maggior attività di *ConParoleTue* è stato quello compreso tra dicembre 2012 ed aprile 2013, data di conclusione del progetto APRIL.

[2]http://elearning.unistrapg.it/conparoletue/

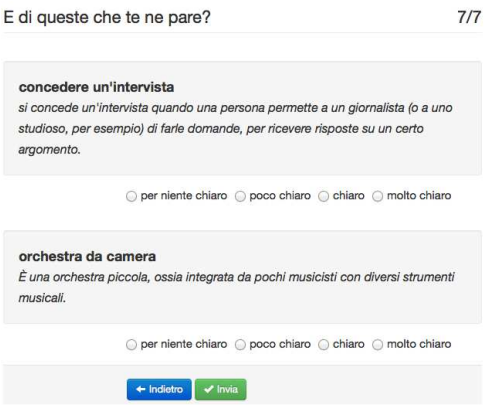


Figura 2 - Esempio di valutazione delle definizioni

3 Risultati

Le definizioni ottenute attraverso l'esperimento di crowdsourcing erano, a marzo 2014, 3.267 (al netto di una ventina redatte in lingue diverse dall'italiano, e di poche altre illeggibili). Per verificare le caratteristiche di tali definizioni, elaborate non da specialisti, ma da semplici parlanti dell'italiano, esse sono state confrontate con un numero identico di definizioni tratte da un dizionario monolingue, il De Mauro Paravia (2000); le 3.267 definizioni del De Mauro sono state estratte in modo casuale tra quelle riferite a una sola delle possibili diverse accezioni di lemmi di marca comune. Il confronto con le definizioni elaborate da lessicografi mira a verificare l'ipotesi di una maggiore naturalezza delle definizioni create da parlanti non specialisti e, di conseguenza, della loro appropriatezza per un dizionario delle collocazioni destinato a parlanti non nativi dell'italiano. Tra le caratteristiche principali di un *learner dictionary*, che ne fanno uno strumento anche concettualmente diverso rispetto ad un dizionario per parlanti nativi (Tarp, 2009), c'è infatti proprio la specificità delle sue definizioni: in quanto rivolte ad un pubblico di parlanti non nativi, esse dovrebbero:

- avere carattere più linguistico che enciclopedico, quindi "evocare un tipo di sapere pre-scientifico, intuitivo, [...] che abbia un valore prototipico, facilmente riconoscibile" (Schafroth, 2011:26);
- essere formate da un lessico semplice, per quanto possibile di base, e da una sintassi poco complessa, adatta alle limitate competenze linguistiche dei destinatari.

Un *learner dictionary* dovrebbe far comprendere ai lettori il significato di un'espressione facendo riferimento quanto più possibile a conoscenza generica e condivisa e non caratteristica della lingua target, fornendo loro il maggior numero di informazioni possibile sui suoi contesti sintagmatici (Schafroth, 2011).

La presenza di queste caratteristiche può essere verificata attraverso alcune misure quantitative calcolate nel corpus di definizioni; nel confronto tra quelle ottenute attraverso l'esperimento di *ConParoleTue* (d'ora in avanti CPT) e quelle del dizionario De Mauro (DM) abbiamo dunque considerato in primo luogo aspetti superficiali dei due testi, come il numero di tokens per definizione e la lunghezza media delle parole, aspetti tradizionalmente associati alla maggiore o minore semplicità di un testo (Franchina e Vacca, 1986). I risultati, riassunti nella tab. 1, mostrano come le definizioni di CPT siano più brevi di quelle di DM,mediamente composte da parole più brevi e da un numero maggiore di frasi più brevi.

	tokens	tokens per definizione	frasi	tokens per frase	lunghezza parole
CPT	38.697	11,8	3.506	11,2	5
DM	42.310	13,2	3.318	13	5,7

Tabella 1 - Misure quantitative di CPT e DM

I tratti superficiali considerati fin qui sono quelli che tradizionalmente concorrono al calcolo dell'indice di leggibilità (Amizzoni e Mastidoro, 1993), che ha appunto l'obiettivo di misurare il grado di facilità con cui un testo viene letto e compreso; uno degli indici di leggibilità più utilizzati per l'italiano, *Gulpease* (Lucisano e Piemontese, 1988), differisce in modo significativo in CPT (68,7) e DM (60,59).

Se tutti questi elementi suggeriscono una maggiore comprensibilità delle definizioni ottenute attraverso il crowdsourcing, vanno comunque considerati i limiti degli indici, che, come quello di *Gulpease*, sono basati esclusivamente su caratteristiche superficiali dei testi, come la lunghezza in caratteri delle parole e quella delle frasi; tali caratteristiche hanno dimostrato di essere indicatori spesso non del tutto attendibili della leggibilità dei testi (vedi ad esempio Feng et al., 2009).

Per valutare in modo più accurato il grado di comprensibilità dei due gruppi di definizioni, in particolare per parlanti non nativi dell'italiano, abbiamo considerato una serie di altri tratti, di tipo lessicale e morfosintattico (Heilman et al., 2007), sulla base di alcune delle indicazioni contenute in Dell'Orletta et al., (2011).

I tratti lessicali comprendono il rapporto tra types e tokens (TTR), che misura la varietà del lessico utilizzato, e la distribuzione dei tokens di CPT e DM nelle tre fasce di frequenza del vocabolario di base. La TTR[3], considerato uno degli indicatori della leggibilità di un testo (Dell'Orletta et al.,2011), è risultata significativamente più elevata in DM (49,4) rispetto a CPT (36.3).

Per misurare la distribuzione dei lemmi delle definizioni nelle tre fasce del vocabolario di base è stata utilizzata la lista di frequenza dei lemmi estratti dal *Perugia Corpus*; in particolare, la fascia dei 2000 lemmi più frequenti (rango 1-2000), che copre il 79% dei lemmi totali del corpus, la fascia dei successivi 2000 lemmi (rango 2001-4000), che aggiunge alla precedente una copertura del 5,9%, e la fascia dei successivi 3000 lemmi (rango 4001-7000), che aggiunge una copertura del 3,4% dei lemmi totali. Le tre fasce, dunque, comprendono i 7000 lemmi più frequenti del *Perugia Corpus*, che totalizzano una copertura dell'88,3% e che sono assunti come vocabolario di base[4]. La fig. 3 rappresenta la diversa distribuzione dei lemmi delle definizioni nelle tre fasce di frequenza; il grafico evidenzia come in CPT siano predominanti i lemmi della fascia più frequente, quindi quelli più verosimilmente già noti a parlanti non nativi di italiano, mentre in DM oltre il 20% dei lemmi è composto da parole non incluse tra le 7000 più frequenti, e in particolare da nomi astratti o poco comuni (*intasamento*, *lamina*, *perno* o *merlatura)*.

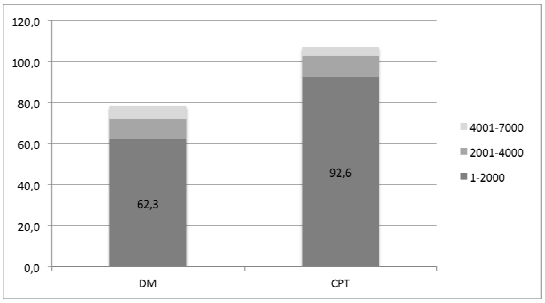


Fig. 3 - La distribuzione dei lemmi di DM e CPT nelle tre fasce del vocabolario di base

Passando infine agli aspetti morfosintattici, nei due corpora di definizioni sono stati misurati i tratti riportati nella tab. 2 (i verbi, i nomi, e tre tipi di frasi subordinate: quelle implicite introdotte da preposizioni, quelle esplicite introdotte da congiunzioni, e le relative). Per ognuno dei tratti è stato calcolata la log-likelihood (Rayson e Garside, 2000), per misurare la significatività delle differenze. Come si evince dalla tab. 2, le definizioni di CPT sono composte da un numero sensibilmente maggiore di verbi (specie di modo finito e per il 90% inclusi nei 2000 lemmi più frequenti) e da un numero minore di nomi; CPT si serve inoltre in misura significativamente maggiore di subordinate, sia implicite che esplicite. Come mostra la coppia di esempi (1) e (2), le definizioni non specialistiche di CPT procedono per brevi subordinate che precisano con parole semplici l'enunciazione della principale, mentre quelle di DM, spesso prive di verbo,sono caratterizzate da un accumulo di sintagmi nominali e preposizionali, per lo più astratti.

(1) Pietra dello scandalo (CPT): qualcuno che è al centro dell'attenzione perché ha fatto qualcosa di grave.

(2) Scandalo (DM): turbamento della coscienza o sconvolgimento della sensibilità.

Tratto	*CPT*	*DM*	*L-L*	*p-value*
Verbi	6185	5746	79,10	0,000
Nomi	8525	9803	11,61	0,001
pre. + sub.	849	388	219,63	0,000
cong. sub.	1516	699	385,92	0,000
rela. ≠CHE	257	183	19,99	0,000

Tabella 2 - Tratti morfosintattici in CPT e DM

4 Conclusioni

L'esperimento descritto, che riguarda l'uso del crowdsourcing per l'acquisizione di definizioni di collocazioni italiane redatte da parlanti generici, si è rivelato efficace sia dal punto di vista quantitativo (oltre 3200 definizioni raccolte in cinque mesi) che da quello della loro appropriatezza ad un pubblico di apprendenti. Un confronto con definizioni redatte da un team di lessicografi ha evidenziato il carattere più intuitivo e naturale delle definizioni dei non specialisti, rispetto alla maggiore astrattezza e complessità delle definizioni dei professionisti. I risultati descritti inducono a proseguire la redazione del dizionario attraverso tale metodologia basata sul crowdsourcing.

[3] La TTR è stata calcolata usando l'indice di Guiraud (Guiraud, 1954), per ovviare alla non omogeneità nel numero dei tokens dei due insiemi di dati.

[4] Il *Vocabolario di Base della lingua italiana* (De Mauro 1980) è in corso di revisione. Per questo si è deciso di utilizzare al posto della sua vecchia versione la lista di frequenza dei lemmi del *Perugia Corpus*, anche se non rappresenta nativamente un vocabolario di base dell'italiano.

References

Maurizio Amizzoni e Nicola Mastidoro. 1993. Linguistica applicata alla leggibilità: considerazioni teoriche e applicazioni. *Bollettino della Società Filosofica Italiana*, n. 149 (maggio - agosto 1993), pp. 49-63.

Umberta Bortolini, Carlo Tagliavini e Antonio Zampolli. 1971. *Lessico di frequenza della lingua italiana contemporanea*. Garzanti, Milano.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286–295.

ConParoleTue. 2012. Home Page del progetto: http://elearning.unistrapg.it/conparoletue.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Tullio De Mauro. 2000. *Dizionario della lingua italiana*. Paravia, Torino.

Stefen Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Dissertation, IMS, University of Stuttgart.

Lijun Feng, Noemie Elhadad and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '09), pp. 229–237

Valerio Franchina e Roberto Vacca. 1986. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi* (3), pp. 47-49

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35.

Fabian Kneissl and François Bry. 2012. MetropolItalia: a crowdsourcing platform for linguistic field research. *Proceedings of the IADIS international conference WWW/internet*.

Andrea Gobbi. 2012. Ipotesi Glottodidattica 2.0. *Journal of e-Learning and Knowledge Society*, 8(3): 47-56.

Andrea Gobbi. 2013. Tweetaliano: a native 2.0 approach to language learning. *ICT for Language Learning 2013, Conference Proceedings*, 282-285.

Andrea Gobbi e Stefania Spina. 2013. Smart Cities and Languages: The Language Network. *Interaction Design and Architecture(s) Journal – IxD&A*. 16: 37-46.

Paul Guiraud. 1954. *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.

Michael J. Heilman, Kevyn Collins and Jamie Callan. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of the Human Language Technology Conference*, pp. 460–467

Michael Lewis. 2000. *Teaching collocation. Further developments in the lexical approach*. Language Teaching Publications, Hove.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano*. John Benjamins, Amsterdam.

Pietro Lucisano e Maria Emanuela Piemontese. 1988. GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana, *Scuola e città*, 3, 31, marzo 1988, pp. 110-124.

Fanny Meunier e Sylviane Granger. 2008. *Phraseology in foreign language learning and teaching*. John Benjamins, Amsterdam.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122-130.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for computational linguistics, 401-409.

Progetto April. 2010. Home Page del progetto: http://april.unistrapg.it/april/.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora*, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.

Elmar Schafroth. 2011. Caratteristiche fondamentali di un learner's dictionary italiano. *Italiano Lingua Due*, 1, pp. 23-52.

Norbert Schmitt (Ed.). 2004. *Formulaic Sequences*. John Benjamins, Amsterdam.

Tyler Schnoebelen and Victor Kuperman. 2010.Using Amazon Mechanical Turk for linguistic research. *Psihologija*, Vol. 43 (4), 441–464.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew T. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natu-

ral language tasks. *EMNLP '08: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Stefania Spina. 2010a. The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform, in Granger S., Paquot M., *eLexicography in the 21st century: New Challenges, New Applications*, Proceeding of eLex 2009 (Louvain-La-Neuve, 22-24 ottobre 2009), Presses Universitaires de Louvain, pp. 273-282.

Stefania Spina. 2010b. The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment, in Calzolari N., Choukri K., Maegaard B., Mariani J., Odjik J., Piperidis S., Rosner M.and Tapias D., 2010, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta, May 2010, European Language Resources Association, pp. 3202-3208 .

Sven Tarp. 2009. The foundations of a theory of learners' dictionaries. In *Lexicographica*, 25, pp. 155-168.

Paola Tiberii. 2012. *Dizionario delle collocazioni*. Zanichelli, Bologna.

Francesco Urzì. 2009. *Dizionario delle Combinazioni Lessicali*. Convivium, Lussemburgo.

William Yang Wang, Dan Bohus, Ece Kamar and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. *Proceedings of the IEEE SLT 2012*, 73-78.

10.12871/CLICIT2014139

Making Latent *SVM^{struct}* Practical for Coreference Resolution

Iryna Haponchyk[1] and **Alessandro Moschitti**[2,1]
[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Qatar Computing Research Institute
iryna.haponchyk@unitn.it, amoschitti@gmail.com

Abstract

English. The recent work on coreference resolution has shown a renewed interest in the structured perceptron model, which seems to achieve the state of the art in this field. Interestingly, while SVMs are known to generally provide higher accuracy than a perceptron, according to previous work and theoretical findings, no recent paper currently describes the use of *SVM^{struct}* for coreference resolution. In this paper, we address this question by solving some technical problems at both theoretical and algorithmic level enabling the use of SVMs for coreference resolution and other similar structured output tasks (e.g., based on clustering).

Italiano. *Ricerca recente sulla risoluzione delle coreferenze linguistiche ha mostrato un rinnovato interesse per l'algoritmo del percettrone strutturato, il quale sembra essere lo stato dell'arte per questa disciplina. È interessante notare che, mentre l'esperienza passata e i risultati teorici mostrano che le SVMs sono piú accurate del percettrone, nessun articolo recente descrive l'uso di SVM^{struct} per la risoluzione di coreferenze. In questo articolo, si prova a dare una risposta a tale domanda, risolvendo alcuni problemi tecnici, sia a livello teorico che algoritmico, cosí consentendo l'utilizzo delle SVMs per la risoluzione delle coreferenze e altri problemi che richiedono l'uso di funzioni di output strutturato (e.g., basati su clustering).*

1 Introduction

Coreference resolution (CR) is a complex task, in which document phrases (mentions) are partitioned into equivalence sets. It has recently been approached by applying learning algorithms operating in structured output spaces (Tsochantaridis et al., 2004). Considering the nature of the problem, i.e., the NP-hardness of finding optimal mention clusters, the task has been reformulated as a spanning graph problem.

First, Yu and Joachims (2009) proposed to (i) represent all possible mention clusters with fully connected undirected graphs and (ii) infer document mention cluster sets by applying Kruskal's spanning algorithm (Kruskal, 1956). Since the same clustering can be obtained from multiple spanning forests (there is no one-to-one correspondence), these latter are treated as hidden or latent variables. Therefore, an extension of the structural SVM – *Latent SVM^{struct}* (LSVM) – was designed to include these structures in the learning procedure.

Later, Fernandes et al. (2012) presented their CR system having a resembling architecture. They do inference on a directed candidate graph using the algorithm of Edmonds (1967). This modeling coupled with the latent structured perceptron delivered state-of-the-art results in the CoNLL-2012 Shared Task (Pradhan et al., 2012).

To the best of our knowledge, there is no previous work on a comparison of the two methods, and the LSVM approach of Yu and Joachims has not been applied to the CoNLL data. In our work, we aim, firstly, at evaluating LSVM with respect to the recent benchmark standards (corpus and evaluation metrics defined by the CoNLL-shared task) and, secondly, at understanding the differences and advantages of the two structured learning models. In a closer look at the LSVM implementation[1], we found out that it is restricted to inference on a fully-connected graph. Thus, we provide an extension of the algorithm enabling to op-

[1] http://www.cs.cornell.edu/~cnyu/latentssvm/

erate on an arbitrary graph: this is very important as all the best CR models exploit heuristics to prefilter edges of the CR graph. Therefore our modification of LSVM allows us to use it with powerful heuristics, which greatly contribute to the achievement of the state of the art. Regarding the comparison with the latent perceptron of Fernandes et al. (2012), the results of our experiments provide evidence that the latent trees derived by Edmonds' spanning tree algorithm better capture the nature of CR. Therefore, we speculate that the use of this spanning tree algorithm within LSVM may produce higher results than those of the current perceptron algorithm.

2 Structured Perceptron vs. *SVM^{struct}*

In this section, we briefly describe the basics of the widely known structured prediction framework. Structured learning algorithms aim at discovering patterns that relate input to complex (thus generally structured) output. Formally, they seek for a mapping $f : X \times Y \rightarrow \mathbb{R}$ over a combined feature space of input variables X and output variables Y, where predictions are derived by finding the $\underset{\mathbf{y}\in Y}{\operatorname{argmax}} f(\mathbf{x}, \mathbf{y})$. The function $f(\mathbf{x}, \mathbf{y})$ is often assumed to be linear with respect to $\Phi(\mathbf{x}, \mathbf{y})$, which is a *joint feature vector* representing an input example together with its associated output. In other words, we have a linear function of the type: $f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y})$. The structured perceptron learning consists in iterating over the entire training set $\{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1,..,l}$ of the following operations: (i) find the optimal output:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in Y}{\operatorname{argmax}} f(\mathbf{x_i}, \mathbf{y})$$

(given the current weight $\mathbf{w}$) and (ii) update $\mathbf{w}$ as follows: $\mathbf{w} \leftarrow \mathbf{w} + \Phi(\mathbf{x_i}, \mathbf{y_i}) - \Phi(\mathbf{x_i}, \hat{\mathbf{y}})$ when prediction errors occur, i.e., $\hat{\mathbf{y}} \neq \tilde{\mathbf{y}}$, where $\tilde{\mathbf{y}}$ is the gold standard output. The structured perceptron algorithm dates back to the early work of Collins (2002), who provided its theoretical guarantees and proof of convergence.

SVMs outperform perceptron in terms of generalization accuracy. They were extended by Tsochantaridis et al. (2004) to deal with structured output spaces. In a standard form, the optimization problem is formulated as

$$\min_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{w}\|^2$$

s.t.$\forall \mathbf{i}, \forall \mathbf{y} \in Y \setminus \mathbf{y_i}$:$\langle \mathbf{w}, \Phi(\mathbf{x_i}, \mathbf{y_i}) - \Phi(\mathbf{x_i}, \mathbf{y})\rangle \geq 1$. The set of margin constraints in the above formulation may be exponentially large or even infinite when Y, for example, is the space of subtrees in syntactic parsing or the space of strings in the sequence labeling task. However, it was shown that using the sparseness of Y and structure and dependencies in Φ, one can drastically reduce the number of constraints to be examined, which makes the optimization feasible. A general SVM algorithm for predicting structured outputs, as well as its instantiations for several complex prediction tasks, was implemented in *SVM^{struct}* and made publicly available[2].

CR is essentially modelled as a clustering problem. Considering a clustering of a document mention set a desired output of a predictor, one can approach the task with a learning algorithm operating in the output space Y of all possible clusterings. Further, we describe two structured learning methods, applied to CR, that were able to overcome the intractability of search for an optimal clustering in Y.

3 Corereference resolution with SVMs

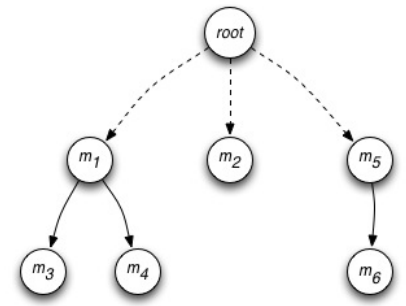
Latent SVM^{struct} was introduced by Yu and Joachims (2009), who construct an undirected graph for each document (Figure 1b). The authors reformulate the *structural SVM* of Tsochantaridis et al. (2004) introducing *latent* variables into a learning procedure. In the LSVM formulation, an input-output example is, thus, described by a tuple $(\mathbf{x}, \mathbf{y}, \mathbf{h})$, where $\mathbf{x}$ is a document mention set, $\mathbf{y}$ is a corresponding clustering and $\mathbf{h}$ is a latent variable. $\mathbf{h}$ is consistent with $\mathbf{y}$ in a way that for training examples, $\mathbf{h}$ contains only links between mention nodes that are coreferent according to $\mathbf{y}$. For test examples a clustering $\mathbf{y}$ is, instead, imposed by an $\mathbf{h}$ automatically generated by the classification algorithm. The joint feature vector decomposes along the edges of $\mathbf{h}$:

$$\Phi(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \sum_{e\in\mathbf{h}} \phi(e).$$

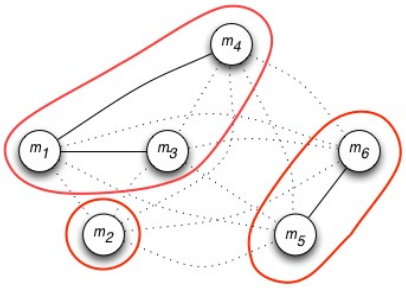
The learning procedure involves running Kruskal's algorithm for finding a maximum spanning forest of a graph containing all possible links between mentions. The resulting spanning forest, in which each connected component corresponds to a separate cluster (in Figure 1b clusters are circled), is a desired $\mathbf{h}$.

The LSVM implementation provided by the authors follows the *SVM^{struct}* API paradigm. In our

[2]It is a software package for implementing structural SVMs available at `http://svmlight.joachims.org/svm_struct.html`



(a)



(b)

Figure 1: Graphical models employed in structured learning algorithms

experiments, we analysed one of the LSVM specializations that is designed for CR. The inference by Kruskal's spanning algorithm is done on a fully-connected graph of mention pair relations. However, a large portion of mention pair links apparently do not convey significant information because they connect non-coreferring mentions, e.g., very distant mentions are very improbable to corefer. Thus, it has been a common practice in coreference research to adopt a preliminary strategy for mention pair filtering, e.g., Fernandes et al. (2012) preprocess the data by applying a set of linguistic filters (the so called sieves). This issue becomes crucial in the LSVM setting as Kruskal's algorithm includes sorting all document edges (this number is exponential in the number of mentions) by their weight. In our work, we intended to enable the LSVM implementation to operate on non-complete candidate graphs, i.e., whose edges have been filtered by some strategies.

4 Coreference Resolution with Latent Perceptron

The latent perceptron of Fernandes et al. (2012) is related to the earlier work of Yu and Joachims (2009) but they model CR as a spanning tree problem. They introduce document trees (Figure 1a), in which nodes represent mentions and edges – relations between them, plus an additional root node. The subtrees directly connected to the root node of such a tree form clusters. To obtain this tree, Edmonds' algorithm is run on a directed candidate graph of document mention relations. Such trees are implicit in data and hence are called latent. This modeling is incorporated into a latent perceptron framework in its loss-augmented formulation. It achieved the best results in the CoNLL 2012-Shared Task (Pradhan et al., 2012).

Edmonds' algorithm iterates over the tree nodes and chooses the best incoming edge (edge of maximum weight). By that means, the best antecedent is chosen for each mention (or no antecedent if the chosen edge starts in the root node). This strategy thereby fits the nature of the CR task very well.

5 Adapting *Latent SVM^{struct}* to filtered data

As mentioned before, we intend to enable the use of LSVM on filtered graphs, i.e., when some candidate edges between mention nodes are missing. The theoretical description of the algorithm does not impose any limitation on the use of partial graphs. However, the provided implementation requires fully-connected graphs. Indeed, a bare execution of LSVM on the partial data results into a low performance score (see Table 1).

In the implementation, each mention is assigned with an ID of the cluster it belongs to, which is chosen according to the rule $clusterID(m_i) = min_i\{m_i \cup \{m_j : \exists$ a positive edge between m_i and $m_j\}\}$, where m are the IDs of the mentions. Let us suppose that we have a cluster with 4 mentions $K = \{m_1, m_2, m_3, m_4\}$ (mentions receive an ID, corresponding to the order of their appearance in the document). If we are provided with all the edges then we surely obtain $\forall i = 1..4, clusterID(m_i) = m_1$. However, if an edge, e.g., (m_1, m_3), is missing, $clusterID(m_3) = m_2$ and it would differ from the cluster ID of the other coreferring mentions. Thus, we made the necessary modifications to the LSVM program code, which resolve the above problem by activating the following rule: $clusterID(m_i) = min\{m_i \cup \{m_j : \exists$ a positive route connecting m_i and $m_j\}\}$.

Another program issue requiring an adjustment is the construction of a gold spanning forest for the first iteration. In the original version of software, this is done by connecting consecutively the cluster edges. For the aforementioned cluster K, chain $\{(m_1, m_2), (m_2, m_3), (m_3, m_4)\}$ would

	All edges		Filtered edges	
Scorer Version	v4	v7	v4	v7
Original LSVM	60.22	56.56	53.15	46.67
Modified LSVM	60.22	56.56	60.31	57.18

(a) development set

	All edges		Filtered edges	
Scorer Version	v4	v7	v4	v7
Original LSVM	59.61	55.19	52.85	46.03
Modified LSVM	59.61	55.19	59.71	56.09

(b) test set

Table 1: Performance of the LSVM implementations on the English part of the CoNLL-2012 dataset.

be output. However, this is not a valid manner when instead of the entire graphs, some edges are filtered. Our modification therefore connects each mention m_i to $min\{m_j : m_j > m_i, \exists$ a positive edge between m_i and $m_j\}$.

Beside the other insignificant changes to the program code, our adjustments enabled us to train the LSVM on thoroughly filtered data while reaching basically the same performance as in the fully-connected case.

6 Experiments

In all our experiments, we used the English part of the corpus from the CoNLL 2012-Shared Task[3], which comprises 2,802, 343 and 348 documents for training, development and testing, respectively. We report our results in terms of the MELA score (Pradhan et al., 2012) computed using the versions 4 and 7 of the official CoNLL scorer. Our feature set is composed of *BART*[4](Versley et al., 2008) and some Fernandes et al. features.

Table 1(a) reports our experiments on the development set, using the original LSVM (Row 1) and our modified version enabling the use of filters (Row 2). The first column regards the use of a fully-connected coreference graph. The numbers confirm that we do not introduce any errors to the implementation since we obtain equal performance as with the original algorithm (v4 and v7 are different scorers). The results in the rightmost column are more interesting as they show that the original LSVM loses up to 10 absolute percent points whereas the modified version obtains practically the same results as when using unfiltered graphs. It should be noted that we use here only 3.94% of edges: this corresponds to a substantial speed-up of the learning and classification phases. Table 1(b) illustrates the same trend on the test set.

[3]http://conll.cemantix.org/2012/data.html

[4]http://bart-anaphora.org

	All edges		Filtered edges	
Scorer Version	v4	v7	v4	v7
Development	61.68	58.25	61.78	58.89
Test	61.21	57.64	61.23	57.90

Table 2: Accuracy of our implementation of the Latent Perceptron of Fernandes et al. (2012) on the English part of the CoNLL-2012 dataset.

In Table 2, we report the performance of our implementation of the modelling of Fernandes et al., showing that the perceptron model unexpectedly outperforms LSVM in all the settings. The main difference of the methods is that LSVM finds a global optimum[5], whereas perceptron simply finds a solution. We thus would expect higher accuracy from LSVM. However, LSVM uses graphs instead of trees along with a different spanning tree algorithm, i.e., Kruskal's vs. Edmond's used by the Latent Perceptron.

To shed some light on this question, we implemented the latent perceptron with the graph model and Kruskal's spanning algorithm as it is done in LSVM. Due to the time constraints, we could train this perceptron implementation only on a part of the filtered training set, constituted by 363 out of all $2,802$ documents. We obtained 58.79(v4) and 55.43(v7) on the development set. These results are lower than what we obtained with LSVM on the same data, i.e., 59.51(v4), 56.22(v7). Additionally, the same perceptron but using latent trees and Edmonds' algorithm scored 61.37(v4) and 58.33(v7). This suggests that Edmonds' spanning tree algorithm is superior to Kruskal's for CR and LSVM using it may outperform the latent perceptron.

7 Conclusions

We have performed a comparative analysis of the structured prediction frameworks for coreference resolution. Our experiments reveal that the graph modelling of Fernandes et al. and Edmonds' spanning algorithm seem to tackle the task more specifically. As a short-term future work, we intend to verify if LSVM benefits from using Edmonds' algorithm. We have also enabled the LSVM implementation to operate on partial graphs, which allows the framework to be combined with different filtering strategies and facilitates its comparison with other systems.

[5]Although, in latent methods, this is often not true as the data is not separable.

Acknowledgments

The research described in this paper has been partially supported by the EU FP7 grant #288024: LIMOSINE – Linguistically Motivated Semantic aggregation engiNes.

References

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jack R. Edmonds. 1967. Optimum branchings. *Journal of research of National Bureau of standards*, pages 233–240.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea, July. Association for Computational Linguistics.

Joseph B. Kruskal. 1956. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In *Proceedings of the American Mathematical Society, 7*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, page 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 104–, New York, NY, USA. ACM.

Yannick Versley, Simone Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: A modular toolkit for coreference resolution.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, New York, NY, USA. ACM.

10.12871/CLICIT2014140

Nominal Coercion in Space: Mass/Count Nouns and Distributional Semantics

Manuela Hürlimann
EM LCT/University of Trento
huerlimann.manuela@gmail.com

Raffaella Bernardi
Denis Paperno
University of Trento
{first.last}@unitn.it

Abstract

English Theoretical linguists analyse all nouns as either mass or count, but admit that noun meanings can be shifted from one class to the other and classify these shifts. We use distributional semantic models to check how the theoretical analysis of mass-count meaning shifts relates to the actual usage of the nouns.

Italiano *In linguistica i sostantivi inglesi sono divisi in numerabili e non numerabili. È però riconosciuto che il significato nominale può passare da una classe ad un'altra seguendo determinati tipi di "spostamenti". In questo lavoro, usiamo i modelli semantici distribuzionali per verificare se le teorie linguistiche sugli spostamenti del significato nominale abbiano riscontro nei dati.*

1 Introduction

It is generally assumed that if a mass (count) noun is used in a count (resp. mass) context, its meaning changes. Compare example (1), where *wine* is used in a mass context (as a bare singular; denoting a substance) to (2), where the use of the determiner *three* indicates a count usage, shifting its interpretation to *types of wine*.

(1) I like wine.

(2) Three wines grow in this region.

The same phenomenon can also be observed for count nouns: in example (3), *apple* is used in its more frequent count sense, while its bare usage in example (4) constitutes a mass usage with a slightly changed meaning — the focus is not on individual, whole apples as in the countable example, but on their material/substance.

(3) I bought five apples at the market.

(4) There is apple in the salad.

Data-based approaches to the mass/count phenomenon include Baldwin and Bond (2003), who classify nouns into five countability types based on lexico-syntactic features and Ryo Nagata et al. (2005), who use context words to distinguish between mass and count nouns.

Katz and Zamparelli (2012) were the first to study mass/count elasticity using distributional semantic models. First of all, they dispelled the view that there is a clear count/mass dichotomy: like in the examples above, many nouns which appear frequently in count contexts also appear frequently in mass contexts. Hence, rather than making a binary distinction (count vs. mass nouns), we should speak of *predominantly count* (resp., *predominantly mass*) nouns, i.e., nouns which occur more frequently in count (resp. mass) contexts than in mass (resp., count) contexts. Moreover, Katz and Zamparelli (2012) take pluralisation as a proxy for count usage and conjecture that for *predominantly count* nouns the similarity between singular and plural is higher than for *predominantly mass* nouns since the latter undergo a shift whereas the former do not. This conjecture finds quantitative support in their data – the 2-billion word ukWaC corpus.[1] We wonder whether other factors, such as polysemy, have an impact on this quantitative analysis and we further investigate nominal coercion by also considering the abstract vs. concrete dimension and polysemy.

Katz and Zamparelli (2012) notice that while plurals are invariably count, singulars can be a mixture of mass and count usages, and propose to use syntactic contexts to disambiguate mass and count usages in future studies.

We take up their suggestion and look at coercion using vector representations of mass vs. count us-

[1] wacky.sslmit.unibo.it/doku.php?id=corpora

ages.

According to the linguistic literature (Pelletier (1975)), instances of coercion fall into several shift classes. In this view, coerced nouns move towards a particular "destination":

- **Container shift:** Liquids (mass) are coerced into countable quantities contained in containers: "two beers, please!"
- **Kind shift:** Masses are coerced into a kind reading: "three wines grow in this region"
- **Food shift:** Animal nouns are coerced into a mass food meaning: "there was chicken in the salad"
- **Universal grinder:** Countables are coerced into a mass reading: "after the accident, there was dog all over the street"

We wonder whether these shift classes can be identified in the semantic space. Thus, we propose a simple experiment in which we assess whether the count usage vectors of typical mass nouns move towards (=become more similar to) these suggested destinations.

In sum, we address the following research questions: (1) Do nouns undergo noticeable shifts – and if so, what factors have an impact? (2) Can we interpret the destination of a shift in terms of standard shift classes?

2 Distributional Semantic Models

Distributional Semantic Models are based on the assumption that the meaning of a word can be captured by counting its co-occurrences in a corpus with other words in a given vocabulary. Hence, word meaning can be represented by a vector and semantic similarity between two words can be captured using the cosine similarity of the corresponding vectors Turney and Pantel (2010). The bigger the cosine similarity, the closer are the two words semantically.

Core Vector Space We collected co-occurrence statistics from the concatenation of ukWaC, a mid-2009 dump of the English Wikipedia, and the British National Corpus, a total of 2.8 billion words. For each target word, its co-occurrence with all context words in the same sentence was counted, with the top 20K most frequent content word lemmas being used as context items. We furthermore used Positive Pointwise Mutual Information as a weighting scheme, followed by dimensionality reduction (Singular Value Decomposition) to 400 dimensions. In this space, all usages of a noun are collapsed for building its vector. The model distinguishes, however, between singular and plural nouns (i.e., `cat-sg` and `cat-pl` are two different vectors). We consider those vectors as representing an average or "core" meaning across different usages.

Vector Space of Mass and Count Usages Mass and count usages of nouns were defined using the following determiners: *much*, *less* for mass usages, and *a*, *an*, *every*, *many*, *each*, *fewer*, cardinals, *more* + plural noun, *enough* + plural noun for count usages. In order to reduce noise due to parsing errors, determiners had to be adjacent to the noun and their part of speech tag had to be adjective (not adverb). Based on these syntactic patterns, co-occurrence values were collected for both usages and their final vector representation were then obtained by projection onto the core vector space.

3 Datasets

In order to understand whether polysemy and abstractness have an impact on Katz and Zamparelli (2012)'s results, we create a data set of singular and plular nouns. We expand on Katz and Zamparelli (2012)'s methodology by annotating these nouns with information on concreteness/abstractness and polysemy.

Secondly, in order to avoid side effects of noisy data and to overcome the limitations of the singular/plural nouns as a proxy for the mass/count distinction, we create a second data set filtered by noun frequency and use the vector representations of the disambiguated mass/count usages of the nouns.

3.1 Singular-Plural Data

This dataset contains a total of 3960 singular-plural noun pairs. Only nouns that occur in the corpus at least 10 times in either a mass or a count context were considered.

These nouns have been annotated with information about abstractness/concreteness and polysemy. We required nouns to be unambiguously annotated as either *abstraction.n.06* or *physical_entity.n.01* in WordNet.[2] Furthermore, for a

[2] wordnet.princeton.edu

more fine-grained measure of concreteness, we used the Ghent database Brysbaert et al. (2013) to assign a concreteness score (1=most abstract, 5=most concrete) to each noun. We used WordNet also to annotate polysemy, quantified as the number of different senses (synsets) for each noun.

3.2 Mass-Count Data

To overcome the ambiguity problems associated with the singular-plural data, we create an additional dataset of mass and count nouns and their *usage vectors*.

We use the output from the syntactic patterns of the singular-plural dataset above (see section 3.1) and take the intersection between the nouns that occur with count determiners and those that occur with mass determiners. We clean this list by excluding nouns which occur less than 10 times in a mass context, obtaining 2433 nouns.

4 Experiments

4.1 Exp. 1: Do nouns undergo shifts?

In this first experiment we use the vectors of the singular-plural dataset in order to verify the results by Katz and Zamparelli (2012) against our data and to furthermore check for effects of abstractness and polysemy. Our hypotheses are:

1. Mass nouns undergo greater singular-plural meaning shifts than count nouns.

2. The more abstract a noun (lower concreteness score), the greater its meaning shift between singular and plural.

3. Nouns with a higher degree of polysemy (greater number of synsets) show a greater singular-plural distance.

We then assess the correlations between these annotations and the singular-plural similarity using the cosine measure. In order to run the correlation analyses, we normalise the count and mass context frequencies, thus creating a continuous variable. We define an alternative measure, "massiness". For count context frequency $\mathtt{c}$ and mass context frequency $\mathtt{m}$, $massiness = \frac{m}{(m+c)}$. Massiness can take values between 0 and 1.

Table 1 shows the Pearson correlations between each of the annotations and the cosine similarity measure. All correlations are highly significant (p-values between 2.2e-16 and 6.40e-05).

	Pearson correlation with cosine
concreteness score	0.167
massiness	-0.225
synsets	-0.266

Table 1: Pearson correlation coefficient between annotated variables and cosine similarity.

While the correlations between the annotated variables and similarity scores are not large, they do reveal tendencies which intuitively make sense:

- **concreteness score:** The meaning of concrete nouns shifts less when pluralised.

- **massiness:** Nouns used more frequently in mass contexts undergo greater meaning shifts when pluralised.

- **synsets:** Smaller number of synsets (less polysemy) correlates with greater similarity. Nouns with more unambiguous meanings shift less when pluralised.

4.2 Exp. 2: Where does a shift take a noun?

An important aspect of nominal coercion is the destination of a coerced noun — since we found above that noun meanings indeed change, it would be interesting to investigate *how* they change, or, speaking in terms of the semantic space, *where* they are taken to by coercion.

Destinations of shift classes We look at the container and the kind shifts, which are the most intuitive and least controversial ones among those discussed in the linguistic literature. We take *beer*, *coffee* and *tea* as examples of mass nouns that undergo the container shift and *flour* and *wine* as examples of mass nouns that undergo the kind shift. We run a small-scale experiment in which we compare the cosine similarity of the mass and count usages of these nouns to another word taken as a potential destination of the shift. The results are reported in Table 2 — we can see that the count usage vectors are more similar to the expected destinations than the mass usage vectors, which is in accordance with the container and kind shift explanations.

How far does coercion take a noun? We conclude the analysis of the destination of nominal coercion by visualising the distance of usage vectors with respect to their "core" representation. We

Usage vector	destination	cosine
beer-n$_c$	pint-n	**0.674**
beer-n$_m$	pint-n	0.548
coffee-n$_c$	cup-n	**0.559**
coffee-n$_m$	cup-n	0.478
tea-n$_c$	cup-n	**0.577**
tea-n$_m$	cup-n	0.486
flour-n$_c$	variety-n	**0.267**
flour-n$_m$	variety-n	0.140
wine-n$_c$	variety-n	**0.470**
wine-n$_m$	variety-n	0.177

Table 2: Container vs. kind shifts.

generate a plot in which, for each noun, we put the cosine similarity of the mass usage vector to the "core" noun vector on the x-axis and the similarity of the count usage vector to the "core" noun on the y-axis (see Figure 1). It is evident that there is no strong relation between the two similarities, as indicated by the red fit line.

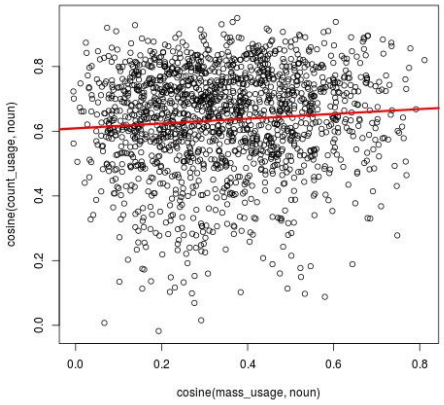


Figure 1: Similarity between mass usage and core noun plotted against similarity between count usage and core noun. Red line = linear fit.

Nouns in the lower left-hand corner (low-mass and low-count) are predominantly bare nouns; as they normally occur without determiners, their average meaning is not very similar to either the mass or the count usage. Words in the upper left-hand corner are nouns that are highly countable and do not seem to lend themselves much to mass usages. Contrary to the latter, words in the lower right-hand corner are nouns that are very "massy" and do not seem to be readily countable. The interesting cases (elastic nouns) are in the upper right-hand corner. For these nouns, both the mass and the count usage vectors are highly similar to the core noun vector. This corner seems to be where regular coercion, which is the subject of our study, lies. Many nouns in this corner shift from "abstract mental state" (mass) to "elements which elicit that state" (count), e.g. *love*, *fear*, *pleasure*. Similarly, *responsibility* shifts from a mental state to a list of concrete duties. Examples of nouns found in the four corners are reported in Table 3.

	Low-count	High-count
Low-mass	diving, dissension	framework, diet
High-mass	importance, distress	love, fear

Table 3: Contingency table: examples

To sum up, regular coercion turns out to only slightly modify the meaning of the noun, so that neither the mass nor the count meaning shifts too far from the core meaning.

5 Conclusions

We have seen how Distributional Semantics Models (DSMs) can be applied to investigate nominal coercion. DSMs can capture some aspects of mass/count noun meaning shifts, such as the fact that predominantly mass nouns undergo greater meaning shifts than predominantly count nouns when pluralised. We also find that abstractness and polysemy have an impact on singular-plural distance: abstract nouns and highly polysemous nouns have a greater singular-plural distance than concrete and monosemous nouns, respectively. Furthermore, our second experiment shows that coercion lies mainly in cases where both the mass and count usage vectors stay close to the averaged noun meaning. However, as our toy evaluation of clear cases of container and kind coercion shows, the direction of the shift can be differentiated based on usage vectors.

Acknowledgments

The first author was supported by the Erasmus Mundus European Masters Program in Language and Communication Technologies (EM LCT).

The other two authors were supported by COMPOSES (ERC 2011 Starting Independent Research Grant n. 283554).

We used the COMPOSES dissect toolkit (http://clic.cimec.unitn.it/composes/toolkit/) for our semantic space experiments.

We furthermore thank Roberto Zamparelli for sharing his huge knowledge of nominal coercion.

References

Baldwin, Timothy and Bond, Francis. 2003. "Learning the Countability of English Nouns from Corpus Data". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pp. 463–470.

Brysbaert, Marc; Warriner, Amy Beth and Kuperman, Victor. 2013. "Concreteness ratings for 40 thousand generally known English word lemmas". *Behavior research methods*, pp. 1–8.

Katz, Grahahm and Zamparelli, Roberto. 2012. "Quantifying Count/Mass Elasticity". *Proceedings of the 29th West Coast Conference on Formal Linguistics, ed. Jaehoon Choi et al.*, pp. 371-379. Somerville, MA: Cascadilla Proceedings Project.

Nagata, Ryo; Wakana, Takiro; Masui, Fumito; Kawai, Atsuo and Isu, Naoki. 2005. "Detecting Article Errors Based on the Mass Count Distinction". In: *Natural Language Processing IJCNLP 2005 Lecture Notes in Computer Science Volume 3651*, pp. 815–826.

Pelletier, F. J. 1975. "Non-singular reference". *Philosophia* 5, pp. 1-14.

Turney, Peter and Pantel, Patrick. 2010. "From frequency to meaning: Vector space models of semantics". *Journal of Artificial Intelligence Research* 37, pp. 141-188.

10.12871/CLICIT2014141

Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language

Claudio Iacobini
University of Salerno
ciacobini@unisa.it

Aurelio De Rosa
University of Salerno
aurelioderosa@gmail.com

Giovanna Schirato
University of Salerno
giovanna.schirato@gmail.com

Abstract

English. The realization of MIDIA (a balanced diachronic corpus of written Italian texts ranging from the XIII to the first half of the XX c.) has raised the issue of developing a strategy for PoS tagging able to properly analyze texts from different textual genres belonging to a broad span of the history of the Italian language. The paper briefly describes the MIDIA corpus; it focuses on the improvements to the contemporary Italian parameter file of the PoS tagging program Tree Tagger, made to adapt the software to the analysis of a textual basis characterized by strong morpho-syntactic and lexical variation; and, finally, it outlines the reasons and the advantages of the strategies adopted.

Italiano. *La realizzazione di MIDIA (un corpus diacronico bilanciato di testi scritti dell'italiano dal XIII alla prima metà del XX secolo) ha posto il problema di elaborare una strategia di PoS tagging capace di analizzare adeguatamente testi appartenenti a diversi generi testuali e che si estendono lungo un ampio arco temporale della storia della lingua italiana. Il paper, dopo una breve descrizione del corpus MIDIA, si focalizza sui cambiamenti apportati al file dei parametri dell'italiano contemporaneo per il programma di PoS tagging Tree-Tagger al fine di renderlo adeguato all'analisi di una base testuale caratterizzata da una forte variazione morfosintattica e lessicale, e evidenzia le motivazioni e i vantaggi delle strategie adottate.*

1 Introduction

The realization of MIDIA, a balanced diachronic corpus of Italian, raised the issue of the elaboration of a strategy of analysis of texts from different genres and time periods in the history of Italian. This temporal and textual diversity involves both a marked graphic, morphological and lexical variation in word forms, and differences in the ordering of the PoS. The program chosen for the PoS tagging is Tree Tagger (cf. Schmid 1994, 1995), and the parameter file, made of a lexicon and a training corpus, is the one developed by Baroni et al (2004) for contemporary Italian. The strategy we developed for the adjustement of the PoS tagging to different diachronic varieties has been to greatly increase the lexicon with a large amount of word forms belonging predominantly to Old Italian, and not to retrain the program with texts belonging to previous temporal stages. This solution turned out to be economical and effective: it has allowed a significant improvement of the correct assignment of PoS for texts both old and modern, with a success rate equal to or greater than 95% for the tested texts, and an optimal use of human resources.

2 MIDIA: a brief description

MIDIA (an acronym for Morfologia Italiana in Diacronia) is a balanced diachronic corpus of written Italian texts, fully annotated with the indication of the lemma and the part of speech. The corpus goes from the beginning of the thirteenth to the first half of the twentieth century.

Periodization is based on important linguistic, literary and cultural facts of Italian history. Five time periods have been distinguished: 1) 1200-1375 formation of Tuscan-centered Old Italian; 2) 1376-1532 affirmation of Italian outside Tuscany; 3) 1533-1691 standardization of Italian in the late Renaissance, Mannerist and Baroque periods; 4) 1692-1840 the birth of modern Italian: the age of Arcadia, the Enlightenment and Romanticism; 5) 1841-1947 the language of Italian political unification.

Texts belonging to seven genres have been collected: expositive prose; literary prose; normative and juridical prose; personal prose; scientific prose; poetry; spoken language mimesis. For each time period and genre 25 texts were selected. A section of 8000 tokens was extracted from each text; for a total of more than 7.5 million tokens.

The search tool we built, in the form of a web application, allows an easy extraction of the data, particularly devised for the study of word-formation in Italian from a diachronic point of view, but also usable for several other types of linguistic investigation. MIDIA can be queried for forms or lexemes also through the use of regular expressions, the search can be refined through the identification of word forms, lexemes or PoS that precede or follow the queried string, and through the use of metadata concerning time period, genre, author, and work.

Different types of outcome can be obtained. The default result shows the selected string in context (the value of 10 left and 10 right forms can be increased or decreased) together with the indication of PoS, lexeme, the metadata concerning author and work, and the ID of the file containing about 8,000 token texts from which the selected item is taken. Other outcomes consist of: distribution tables indicating the number of occurrences of the selected item distinguishing genres and periods; frequency lists showing the number of occurrences of the selected item divided in form, PoS and lemma; graphs and charts showing time evolution of the selected item according to author, genre, and period. All the types of outcome can be viewed online and downloaded in CSV format.

MIDIA is the outcome of the Prin project "The history of word-formation in Italian" funded by the Italian Ministry of Education University and Research. The corpus is freely available at the URL http://www.corpusmidia.unito.it/.

3 PoS tagging strategy for a diachronic corpus of the Italian language

The software we used to associate a part of speech to each word form of our corpus is Treetagger (cf. Schmid 1994, 1995). The application of the Tree Tagger software to a language involves the identification of a Tagset, the creation of a lexicon containing the a priori tag probabilities for each word, and a Tagged Corpus representing the (variety of the) language that is to be analyzed.

We started the automatic annotation with part-of-speech tags using the source files underlying the parameter file for contemporary Italian made by Baroni et al. (2004), which consists of a training corpus of about 115,000 tokens taken from the newspaper La Repubblica (years 1985-2000), and a lexicon which amounts to approximately 220,000 tokens (we thank Marco Baroni for his contribution to the realization of our project).

Our case presents special problems because of the variety of genres and the time span of the texts of the corpus (about PoS tagging of diachronic corpora, cf. Dipper et al. 2004, Martineau 2008, Sánchez-Marco et al. 2010, Stein 2008). We began to test the contemporary Italian TreeTagger (ContIt TT) on two literary prose texts of the first period (1200-1375) of our corpus (taken from Dante's *Vita Nuova* and Dino Compagni, *Cronica delle cose occorrenti ne' tempi suoi*) in order to figure out the problems that the program had with Old Italian texts. The results have been manually checked in order to find recurring mistakes and to think about some possible solutions for the improvement of PoS tagging.

The result of POS tagging on the two texts of the first period was then compared with that of a literary prose text of the most recent period (1841-1947) of our corpus: Italo Svevo, *La coscienza di Zeno*. As expected, the error rate of ContIt TT, fully satisfactory for modern texts (about 5%), was higher for Old Italian literary prose (about 13%). In addition, error analysis reveals that wrong assignments mainly concern PoS (exp. adjectives and verbs) of particular interest for the study of word-formation, for which the MIDIA corpus is especially conceived.

As is known, TreeTagger is a probabilistic PoS tagger that gives to each token of a text PoS and lemma information. The assignment of a particular PoS to each word form depends on the matching with a form present in the lexicon associated with the probabilities of co-occurrence of a PoS with other adjacent according to the information about PoS sequences obtained from the training corpus.

The strategy we adopted to cope with our diachronic corpus was to strongly enrich the contemporary Italian lexicon (that is, the list of forms with specification of PoS and lemma) and not to train it on a widened corpus to which were added Old Italian texts (cf. Gaeta et al. 2013). Our expectation was that PoS tagging of the diachronic corpus could be significantly improved even without adding to the training corpus ex-

amples of the typical syntactic patterns found in Old Italian texts.

The reason behind this decision is twofold. On the one hand we took a theoretical and methodological stance: we were confident that by adding more forms (especially those more typical of older texts) we could significantly improve the results of the analysis, i.e. to have a better "syntactic" analysis through more detailed word recognition. On the other hand we took a cautious position: since ContIt TT already had fairly good results also with Old Italian texts, we have preferred avoiding to alter the distribution of the sequence of PoS on which the program was set (by adding a training corpus made of early texts), especially considering that MIDIA corpus is made not only of texts belonging to Old Italian, but to the entire time span of the history of Italian.

Our expectation was that the recognition of word forms would significantly help the recognition of sentences, i.e. the recognition of sequences of PoS elements, and this was what happened (as we will show in section 4).

The enrichment of MIDIA Tree Tagger (MIDIA TT) lexicon results from the addition of about 230,000 word forms mainly dating from the XIV to the XVI c. (MIDIA TT lexicon actually counts about 550,000 forms).

For the implementation of the lexicon, in a first step we have made use of the available philological resources: word lists, lists of names, critical editions, glossaries and digital corpora (Corpus Taurinense TLIO); later, comparing the lexicon increased in this way with the set of forms used in the texts of the MIDIA corpus, we selected those absent from the lexicon, favoring forms with higher frequency and morphological variance, and we tagged them with a semiautomatic procedure according to the format required by Tree Tagger, paying particular attention to the homographs that would have troubled the recognition mechanisms of the program (for example, proper names were not included that would have generated ambiguity overlap with common names: *Prato, Potenza, Monaco, Fiume, Riga, Spine, Spira, Angelo, Norma, Nunzio, Leone,* etc.; with verbs: *Segna, Segni, Giura, Vendi*; or with numerals: *Cento*). For the same reason we have reduced the Tagset analyticity by suppressing the distinction between adjectives and pronouns for demonstratives, indefinites, numerals, possessives, interrogatives.

4 Checking the results of MIDIA PoS tagging and error analysis

In order to evaluate the performance of MIDIA TT, we have selected one text of literary prose for each of the time periods of the corpus, and for each text we prepared a gold standard PoS assignment through a thorough manual review revised and discussed within our research group.

These gold standards form the benchmark for the performance evaluation of the ContIt TT and MIDIA TT programs (the number of tokens manually checked for PoS assignment is 52,952).

Table 1: See appendix

Table 1 compares the number and the percentage of errors in ContIt TT and MIDIA TT PoS tagging for literary texts belonging to the five time periods. As may be noted, MIDIA TT has significantly better results than those of ContIT TT especially in the first periods; furthermore, we can notice that the result of MIDIA TT in period 1 is better than that of ContIt TT in period 5.

Tables 2 and 3 show some of the typical errors of ContIt TT (highlighted in bold) compared with MIDIA TT correct PoS tagging in texts belonging to the first period.

Table 2: See appendix

Table 3: See appendix

ContIt TT PoS tagging errors reported in bold in Table 2 are very likely to be attributed to the recognition of *ser* (antiquated form for 'mister', but similar in form to the verb *essere* 'to be') as a Noun, which results in the assignment of the form *dove* to the PoS WH instead of to Conjunction; the absence in ContIt lexicon of *giacea* and the proximity of this form to a proper noun (*Ciappelletto*) causes the erroneous tagging of this Verb to the adjectival class. Similarly, the form *allato* is recognized as a past participle (probably because of the final string), while *postoglisi* is not recognized as a past participle because of the combination of enclitic forms. MIDIA lexicon contains all these verb forms and allows the correct attribution of the PoS Conjunction to the word form *dove*, although in the lexicon this form corresponds to three different PoS (Noun, Adverb, and Conjunction).

In table 3 the ambiguity of *magnifico* (Noun, Adjective, and Verb) and the absence in ContIt lexicon of the word form *suggeritole* causes the

error in the assignment of PoS of these forms and of the adjacent word *quadretto*. From this brief error analysis, we may conclude that the failure to recognize word forms triggers a cascade effect of PoS assignment on nearby words, whereas a rich lexicon increases the possibility of a correct PoS assignment also for words that are not listed in the lexicon.

Table 4 shows the PoS with a higher percentage of errors in the text of the first period used as gold standard for PoS assignment (the column GS shows the expected number of tokens for each PoS; the left column of both ContIt TT and Midia TT shows the difference from GS, the right column the percentage of errors for each PoS assignment).

Table 4: See appendix

The errors in MIDIA TT are concentrated in clitics, auxiliary and modal verbs (which generally are still recognized as verbs). The nouns do not present serious problems either in MIDIA TT or in ContIt TT, while the latter has a high error rate in the adjectives, verbs and adverbs; the difficulty in recognizing the members of these PoS is probably due to their high graphic and morphological variation not accounted in ContIt lexicon. The main errors in the PoS tagging of Old Italian in MIDIA TT can be traced in part to the decision not to train MIDIA TT with texts of this period. The main differences that distinguish modern and contemporary Italian from Old Italian concern primarily the syntactic structure; among the syntactic differences, one of the most notable is the possibility to interpose nominal arguments between modal and auxiliary verbs and the main verb, and a greater freedom of clitic position (Renzi and Salvi, 2010; Dardano, 2013). The criterion of adding word forms to the lexicon cannot cope with these difficulties, while it has proved to be adequate for many other variation factors, such as lexical and morphological differences, and also the different positions of the main verbs or of the nominal constituents. The overall positive result on the texts of all the periods made us decide to maintain our choice. Moreover, the enriched lexicon can still provide a useful starting point for those just interested in the texts of Old Italian, who want to train a Tree Tagger parameter file specialized for these texts.

Table 5: See appendix

Table 5 compares auxiliaries, clitics and verbs PoS tagging in period 1 and 5. It shows that verb recognition is stable in the two periods for MIDIA TT, while the correct assignment of clitics and auxiliaries strongly improves in the most recent period for both MIDIA TT and ConIT TT. The good results in verb recognition already performed by MIDIA TT in period 1 may be attributed to the strong enrichment of the lexicon (cf. the high percentage of errors of Cont It TT in period 1), the differences in auxiliaries and clitics can be explained with changes in the syntactic order in the two periods of the Italian language under examination.

5 Conclusions

The strategy we devised to develop MIDIA PoS tagging for the analysis of texts belonging to different time periods and textual genres than that for which it was originally trained has proved to be successful and economical. Human resources have been concentrated on enriching the lexicon and on the review of automatic lexeme and PoS assignment.

Our results show that a larger lexicon improves the analysis also for words adjacent to those recognized by the matching with the word forms listed in the lexicon. This has some interesting consequences both on the strategies for text tagging and on the implementation of the program Tree Tagger for the analysis of texts with a great range of variation.

We plan to further enrich MIDIA lexicon by adding word forms from the corpus not yet listed in the lexicon.

References

Baroni Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi,. Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*.

Dardano Maurizio (ed.). 2013. *Sintassi dell'italiano antico*. Carocci, Roma.

Dipper Stefanie, Faulstich Lukas, Leser Ulf and Lüdeling Anke. 2004. Challenges in Modelling a Richly Annotated Diachronic corpus of German. *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, Lisbon, Portugal: 21-29.

Gaeta Livio, Claudio Iacobini, Davide Ricca, Marco Angster, Aurelio De Rosa, and Giovanna Schirato. 2013. MIDIA: a balanced diachronic corpus of Italian. *Conference held at 21st International Conference on Historical Linguistics (Oslo, 5-9 August 2013)*.

Martineau France. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus* [En ligne] 7: 136-155. URL : http://corpus.revues.org/1508

Renzi Lorenzo and Giampaolo Salvi. 2010. Italiano antico. In:. *Enciclopedia dell'italiano*. Roma, Istituto dell'Enciclopedia Italiana: 713-716.

Schmid Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

Sánchez-Marco Cristina, Boleda Gemma, Fontana Josep Maria, Domingo Judith. 2010. Annotation and representation of a diachronic corpus of Spanish. *Proceedings of the International Conference on Language Resources and Evaluation*, 17-23 May, Valletta, ELRA: 2713-2718.

Schmid Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland*.

Stein Achim. 2008. Syntactic Annotation of Old French Text Corpora. *Corpus* [En ligne] 7: 157-171. URL : http://corpus.revues.org/1510

Appendix

Period	ContIt TT		MIDIA TT	
1	1260	13.24%	478	5.02%
2	1117	9.87%	507	4.48%
3	1904	15.37%	493	3.98%
4	902	8.62%	272	2.60%
5	568	6.13%	220	2.38%
average		**10.86%**		**3.72%**

Table 1: Number and percentage of PoS tagging errors in the five time periods of the corpus MIDIA.

TEXT	ContIt TT	MIDIA TT
nella	ARTPRE	ARTPRE
camera	NOUN	NOUN
dove	**WH**	CON
ser	**VER:fin**	NOUN
Ciappelletto	NPR	NPR
giacea	**ADJ**	VER:fin
e	CON	CON
allato	**VER:ppast**	ADV
postoglisi	**NOUN**	VER:ppast:cli
a	PRE	PRE
sedere	VER:infi	VER:infi

Table 2: Error analysis (first period texts).

TEXT	ContIt TT	MIDIA TT
dipinse	VER:fin	VER:fin
un	ART	ART
magnifico	**NOUN**	ADJ
quadretto	**VER:fin**	NOUN
suggeritole	**NOUN**	VER:ppast:cli
dalla	ARTPRE	ARTPRE
mia	DET:poss	DET:poss
malattia	NOUN	NOUN

Table 3: Error analysis (first period texts).

POS	GS	ContIt TT		MIDIA TT	
ADJ	381	**166**	**43.6 %**	23	6.0 %
ADV	652	**132**	**20.3 %**	5	0.8 %
AUX	187	**78**	**41.7 %**	**61**	**32.6** %
CLI	287	**57**	**19.9 %**	**73**	**25.4** %
CON	565	41	7.3 %	10	1.8 %
DET	905	61	6.7 %	38	4.2 %
NOUN	1402	49	3.50 %	45	3.2 %
PRE	1080	79	7.31 %	1	0.1 %
PRO	542	43	7.9 %	1	0.2 %
VER	**1432**	**342**	**33.9** %	81	5.6 %
VER2	**134**	**46**	**34.3** %	**51**	**38.1** %

Table 4: PoS tagging errors (first period).

PoS	Period 1				
	GS	ContIT TT		MIDIA TT	
AUX	187	78	41.7%	61	32.6%
CLI	287	57	19.9%	73	25.4%
VER	1432	486	33.9%	81	5.6%
	Period 5				
	GS	ContIT TT		MIDIA TT	
AUX	213	6	2.8%	9	4.2%
CLI	342	49	14.3%	27	7.9%
VER	1476	151	10.2%	69	4.7%

Table 5: PoS tagging errors for auxiliaries, clitic and verbs in period 1 and 5.

10.12871/CLICIT2014142

Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion

Elisabetta Jezek
Università di Pavia
jezek@unipv.it

Laure Vieu
IRIT-CNRS - Université Toulouse III
vieu@irit.fr

Abstract

English In this paper we argue that the account of the notion of complex type based on copredication tests is problematic, because copredication is possible, albeit less frequent, also with expressions which exhibit polysemy due to coercion. We show through a distributional and lexico-syntactic pattern-based corpus analysis that the *variability* of copredication contexts is the key to distinguish complex types nouns from nouns subject to coercion.

Italiano *In questo contributo sosteniamo che il test di copredicazione utilizzato in letteratura per motivare l'esistenza di tipi complessi è problematico, in quanto la copredicazione è possibile, seppur con minor frequenza, anche con espressioni che esibiscono un comportamento polisemico a seguito di coercion. Attraverso una analisi distribuzionale che utilizza pattern lessico-sintattici mostriamo come la variabilità dei contesti di copredicazione è la chiave per distinguere nomi associati a tipi complessi da nomi soggetti a coercion.*

1 Introduction

Copredication can be defined as a "grammatical construction in which two predicates jointly apply to the same argument" (Asher 2011, 11). We focus here on copredications in which the two predicates select for incompatible types. An example is (1):

(1) *Lunch was delicious but took forever.*

where one predicate ('take forever') selects for the event sense of the argument *lunch* while the other ('delicious') selects for the food sense.

Polysemous expressions entering such copredication contexts are generally assumed to have a *complex type* (Pustejovsky 1995), that is, to lexically refer to entities "made up" of two (or more) components of a single type; it is thus assumed for example that *lunch* is of the complex type event • food.[1] Copredication as a defining criterion for linguistic expressions referring to complex types is, however, problematic, because copredication is possible, albeit less frequent, also with expressions which exhibit polysemy because of coercion, as in the case of the noun *sandwich* in such contexts as (2):

(2) *Sam grabbed and finished the sandwich in one minute.*

where the predicate *grab* selects for the simple type the noun *sandwich* is associated with (food), whereas *finish* coerces it to an event. The claim that the event sense exhibited by *sandwich* is coerced is supported by the low variability of event contexts in which *sandwich* appears (as opposed to *lunch*); see for example *"during lunch"* (780 hits for the Italian equivalent in our reference corpus, cf. section 3) vs. *"*during the sandwich"* (0 hits).

Our goal is therefore twofold: evaluate whether at the empirical level it is possible to distinguish, among nouns appearing in copredication contexts, between complex types and simple (or complex) types subject to coercion effects; and propose a method to extract complex type nouns from corpora, combining distributional and lexico-syntactic pattern-based analyses. Our working hypothesis is that lexicalized complex types appear in copredication patterns more systematically, and so that high variability of pair of predicates in copredication contexts is evidence of complex type nouns, while low variability points to simple (or complex) type nouns subject to coercion effects.

In the sections that follow, we will first raise the questions what counts as a copredication and what

[1] Dot/complex types have received different terminologies in the literature, particularly *nouns with facets* (Cruse 1995) and *dual aspect nouns* (Asher 2011).

copredication really tell us about the underlying semantics of the nouns that support it. Then, we will introduce the experiments we conducted so far to verify our hypothesis. Finally, we will draw some conclusions and point at the experiments we have planned as future work.

2 Copredication

2.1 What counts as a copredication?

In the literature, what exactly counts as a copredication is not clear. Typically, copredication has been restricted to classic coordinative constructions as in (3), where the adjective *voluminoso* 'bulky' selects for the physical sense of book, while *impegnativo* 'demanding' selects for the informational one.

(3) *È un libro voluminoso e impegnativo*.
'It is a bulky and demanding book'.

Research has shown, however, that copredication patterns based on coordination do not frequently mix different aspects but tend to predicate on a single aspect, as in (4), where both adjectives select for the same event aspect of *costruzione* 'construction' (Jezek and Melloni 2011):

(4) *La costruzione fu lenta e paziente*.
'The construction was slow and patient'.

Moreover, it has been claimed that constructions different from coordinative (or disjunctive) ones can be copredicative; for example, copredications with anaphoric pronouns (5)a, and structures where one of the predicates is located in a subordinative clause, as in (5)b and (5)c.

(5) a. *He paid the bill and threw it away*. (Asher 2011, 63).
b. *La construction, qui a commencé hier, sera très jolie* (Jacquey 2001, 155).
'The building, which started yesterday, will be very nice'.
c. *Una volta completata, la traduzione si può caricare in una sezione apposita del sito* (Jezek and Melloni 2011, 27).
'Once completed, the translation may be uploaded in a special section of the site'.

These copredication patterns may be disputable from both a structural and semantic point of view because they involve pronouns and coreference, and one could argue that pronominalization leaves room for phenomena such as bridging and associative anaphora.

In our work we focus on what we argue is a less disputable copredication pattern, namely [V [Det N Adj]]. This pattern is instantiated in contexts such as the following, where for example the predicate *bruciavano* selects for the physical aspect of *book*, whereas *controversi* selects for the informational one:

(6) *... bruciavano i libri controversi*.
'... they burned the controversial books'.

2.2 What does copredication really tell us?

As referenced above, it has also been noted that copredication may actually involve coercion (Asher and Pustejovsky 2006; corpus evidence in Pustejovsky and Jezek 2008). Consider:

(7) *Aprire il vino rosso con 30 minuti di anticipo*.
'Open the red wine 30 minutes in advance'.

In (7), *vino* 'wine' appears to denote both drink and container in the same context, due to the two predicates *rosso* 'red' and *aprire* 'open'. Despite the apparent polysemy, the noun *vino* is generally assumed to be lexically associated with a simple type (drink), and to license a sense extension to container in specific contexts only, as a coercion effect induced by the semantic requirements of the selecting predicate.

We claim that a single occurrence of a relevant copredication context is not enough to identify a complex type, and we conjecture that a *variety* of copredication contexts appearing with enough regularity might constitute evidence. Indeed, one can observe that *vino* 'wine' displays a limited variability, since it cannot be coerced into a container type by any predicate that would felicitously apply to *bottiglia* 'bottle', as shown by (8):

(8) **Ho rotto il vino rosso*.
'I broke the red wine'.

3 The experiment

We conducted a corpus-based study to assess the possibility to empirically distinguish between complex types and simple (or complex) types subject to coercion effects through the analysis of copredication contexts. The concrete goal of the experiment was, for a given complex type, to extract a list of candidate nouns that do appear in some copredication context, and compute the variability of copredication contexts to order these nouns. The hypothesis is that nouns shall be ordered from most likely being of the complex type at stake to most likely being of some other type but subject to coercion. We exploited the SketchEngine (Kilgarriff et al. 2014) tagged Italian corpus It-

TenTen10 (2,5 Gigawords) and its tools. The complex type chosen for this first experiment was information_object • physical_object of which 'book' is taken to be the prototype in the literature, and as detailed above, the copredication patterns used are of the form [V [Det N Adj]].

3.1 Predicate extraction

The copredication contexts of interest are those based on a transitive verb and an adjective that each select for a different type. The first step was therefore to pick four lists of predicates: transitive verbs selecting for information_object (Info) or physical_object (Phys) as object complements and adjectives that modify nouns of either type.

The starting point was a list of 10 seed nouns[2] considered as good examples of the complex type. We extracted from the corpus predicates applying to these seed nouns, that are frequent and shared enough: on the most frequent 200 verbs (V) and adjectives (A) in the collocational profiles (*WordSketches*) of each of these seed nouns, we performed 2-by-2 intersections and then union, which yielded 427 V and 388 A. We manually doubly classified them into Phys and Info, avoiding predicates (too) polysemic, generic, or subject to metaphorical uses. We thus gathered 65 VPhys, 53 VInfo, 18 APhys and 127 AInfo.

3.2 Candidate extraction

Using a manually selected subset of 6-14 frequent predicates of each category, a series of concordance built on the copredication pattern with all context pairs ⟨VPhys, AInfo⟩ and ⟨VInfo, APhys⟩ produced nouns occurring in these contexts. We then manually annotated 600+ randomly taken hits, checking for actual copredication with both aspects, thus extracting 97 different nouns. The 5 seed nouns not present among these 97 were added, obtaining 102 nouns, as candidates for the complex type Info • Phys. For the rest of the experiment, since the relevant copredications are rather sparse, we focussed on the 54 nouns with frequency above 200,000, and selected 28 (52%) ones, aiming at covering most of the various types appearing among these and including 7 seed nouns (marked * in the table).

[2] *articolo, diario, documento, etichetta, fumetto, giornale, lettera, libro, racconto, romanzo* ('article', 'diary', 'document', 'label', 'comic', 'newspaper', 'letter', 'book', 'short novel', 'novel')

3.3 Computing the copredication context variability

For all 28 nouns we extracted all occurrences of the [V [Det N Adj]] pattern, N fixed. The hits of each lexico-syntactic pattern are grouped by pairs ⟨V, A⟩ that we here call "copredication contexts" for this noun. We then extract the *relevant* contexts ⟨VPhys, AInfo⟩ and ⟨VInfo, APhys⟩ combining selected predicates in our four lists. The ratio of relevant contexts among all contexts is an indicator of the variability of Info • Phys copredication contexts for each noun, and this variability a sign of the conventionalisation of the lemma ability to jointly denote both Phys and Info referents.

The results, ordered from more variable to less variable, appear on Table 1, where **Hits** is the total number of hits of the lexico-syntactic pattern, **Cop. hits** are those hits with a relevant ⟨VPhys, AInfo⟩ or ⟨VInfo, APhys⟩ context, **Contexts** is the total number of ⟨V, A⟩ contexts, and **Cop. cont.** are the relevant ones. Ratios are in %.

Note that the hit ratio would yield a different order than the context ratio, since a single relevant context may have a large incidence. Indeed, with context ratio, the 7 seed nouns are ranked among the 10 first, while with the hit ratio, they would appear among the 14 first, and include at the very top *informazione* and *indicazione*, two nouns unlikely prototypes for the Info • Phys complex type.

4 Discussion

The copredication contexts extracted are sparse, and the ratio figures ordering the nouns are low (all below 3%). This might be due to the phenomenon of copredication across types being sparse, but obviously also because the 4 lists of predicates are by no means exhaustive. On the basis of a manual annotation of 200 (0,8%) hits on *libro*, the recall is estimated at 6%. A very high recall could not be reached without including polysemic or very generic predicates, thus lowering precision. Precision has been estimated for *libro*: 118 (86%) extracted copredication hits are indeed relevant cases. However, in the lower rows, precision drops: 9 (60%) for *volume* and even 0 for *fenomeno*, which means that if we had other means to screen the results, the ratio range would widen between top and bottom rows.

The method allows to distinguish four groups of lemmas (statistically significant partition, but finer-grained partitions could be drawn). At the

Lemma	Freq.	Hits	Cop. hits	Hit ratio	Contexts	Cop. cont.	Cont. ratio
lettera (letter)*	549552	13386	414	3.1	5513	130	2.4
giornale (newspaper)*	276139	6757	37	0.55	968	20	2.1
documento (document)*	547415	25615	313	1.2	11404	182	1.6
informazione (information)	1092596	68201	2635	3.9	18459	242	1.3
racconto (short novel)*	243777	7533	111	1.5	4418	56	1.3
capitolo (chapter)	218115	4982	60	1.2	2731	32	1.2
articolo (article)*	2458766	12885	104	0.81	6588	72	1.1
libro (book)*	968401	23958	137	0.57	10856	107	0.99
pagina (page)	716615	15850	111	0.70	8357	82	0.98
romanzo (novel)*	213778	7644	47	0.61	3844	35	0.91
testo (text)	528482	21080	108	0.51	9067	81	0.89
immagine (image)	641384	32097	256	0.80	19146	162	0.85
indicazione (indication)	279063	20831	651	3.1	6536	54	0.83
relazione (report)	744398	36274	467	1.3	15693	101	0.64
storia (story)	1505947	57074	235	0.41	21292	129	0.61
programma (program)	978951	39140	340	0.87	18029	103	0.57
parola (speech)	1087778	44619	139	0.31	16292	87	0.53
gioco (game)	637619	16815	60	0.36	8859	43	0.49
proposta (proposal)	716391	28007	149	0.53	12254	58	0.47
serie (series)	668564	12824	40	0.31	6872	31	0.45
dichiarazione (statement)	339720	13601	33	0.24	5817	25	0.43
fonte (source)	354620	20912	35	0.17	7692	33	0.43
riferimento (reference)	691282	18193	57	0.31	6705	27	0.40
ricerca (research)	1378351	25002	103	0.41	12228	46	0.38
carattere (character)	378986	45632	131	0.29	20504	70	0.34
volume (volume)	307808	6732	15	0.22	4445	15	0.34
pezzo (piece)	286093	13190	27	0.20	7201	23	0.32
prodotto (product)	837772	48285	72	0.15	20391	54	0.26
fenomeno (phenomenon)	342726	26876	20	0.074	11872	13	0.11

Table 1: Relevant copredication variability for 28 candidate Info • Phys nouns with high frequency

top, are those that arguably are prototypical examples of the complex type Info • Phys. Next comes a group of nouns with still classical examples of this dot-type, especially *libro*, as well as nouns of the simple type Info such as *informazione*. Since information objects generically depend on their physical realizations, coercion is readily available. What these data tell us is that the pure Info sense of *libro* (as in *il libro di Dante è stato tradotto in tante lingue* 'Dante's book has been translated in many languages') or *immagine* might prevail over their complex type sense. The next group gathers many nouns of a different complex type, Info • Event, such as speech act nouns, some of which, like *relazione*, do also have a lexicalized sense of document, while others, like *indicazione* and *dichiarazione*, are rather subject to coercion. The last group exhibits occasional coercion contexts, with the exception of *volume* which does have a standard Info • Phys sense but much less frequent than its spatial or sound quality sense.

We can therefore conclude that an experimental method to separate nouns of complex types from nouns subject to coercion appears possible. The proposed method constitutes the first attempt at semi-automatically extracting from corpus complex type nouns, something remaining elusive up to now. In addition, we learned that *letter* should be preferred over *book* as prototype of the complex type Info • Phys. In fact, this complex type is not the most straightforward since the dependence between the components of a dot object is not one-to-one. The case of Event • Food with *lunch* as prototype, in which there is such a tight symmetric dependence and no competition with separate simple senses, might prove easier to deal with. This will be tackled in a next experiment.

The predicate selection is a critical phase in the method proposed. It is difficult if not impossible to avoid polysemy and metaphorical uses, especially since the relevant copredications are sparse and we cannot rely only on highly specialized unfrequent predicates. In future work, we plan to experiment with fully automatic selection, exploiting distributional semantics methods. Dimension reduction through non-negative matrix factorization yields a possible interpretation of the dimensions in terms of "topics", which is confirmed by experiments (Van de Cruys et al. 2011). Building on this, we shall check whether "topics" for predicates correspond to selectional restrictions suitable to build our copredication patterns.

Acknowledgments

Thanks to Philippe Muller for help with programming issues and to the participants of the workshop on dot objects in May 2014 in Toulouse for feedback on early results of this work. We also acknowledge Tommaso Caselli and two anonymous reviewers for their useful comments.

Bibliography

N. Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge: Cambridge University Press.

N. Asher and J. Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6, 1–38.

D.A. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. Saint-Dizier and E. Viegas *Computational Lexical Semantics*, CUP, 33–49.

E. Jacquey. 2001. *Ambiguités Lexicales et Traitement Automatique des Langues: Modélisation de la Polysémie Logique et Application aux déverbaux d'action ambigus en Français*. Ph.D. Dissertation, Université de Nancy 2.

E. Jezek, and C. Melloni. 2011. Nominals, Polysemy and Co-predication. *Journal of Cognitive Science*, 12, 1–31.

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1–30. `http://www.sketchengine.co.uk/`

J. Pustejovsky. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

J. Pustejovsky, J. and E. Jezek. 2008. Semantic coercion in language: Beyond distributional analysis. Distributional Models of the Lexicon in Linguistics and Cognitive Science. Special Issue on *Italian Journal of Linguistics*, 20(1), 175–208.

T. Van de Cruys, T. Poibeau, and A. Korhonen. 2011. Latent vector weighting for word meaning in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

10.12871/CLICIT2014143

Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data

Fahad Khan
ILC CNR Pisa
fahad.khan@ilc.cnr.it

Francesca Frontini
ILC CNR Pisa
francesca.frontini@ilc.cnr.it

Abstract

English. This paper presents the ongoing project for the conversion and publication of the Italian lexicon Parole Simple Clips in linked open data, illustrating the chosen model, with a particular focus on the translation of the syntactic and semantic information pertaining verbs and their predicates.

Italiano. *Questo paper presenta il progetto in corso per la conversione e pubblicazione del lessico italiano Parole Simple Clips nel formato linked open data, descrivendo il modello adottato con particolare riferimento alla traduzione delle informazioni sintattico semantiche dei verbi e dei loro predicati.*

1 Introduction

The aim of the present paper is to describe the ongoing conversion of the semantic layer of the Parole Simple Clips (PSC) lexical resource into linked open data. We have previously presented the conversion of the nouns in PSC in (Del Gratta et al., 2013). In this paper we will continue this work by presenting the model we intend to use for converting the verbs.

In the next section we shall give a general background on the linguistic linked open data (LLOD) cloud and discuss the importance of putting lexical resources on the cloud. We also discuss the *lemon* model which we have chosen as the basis of the conversion of the PSC resource. In the following section we discuss PSC itself and give a brief overview of its structure.

Finally in the last section we will outline how we intend to proceed with the conversion of the PSC verbs, illustrating the proposed schema with an example.

2 Linguistic Linked Open Data

The term linked open data refers to the practice of publishing data online in a standardised format that makes the interlinking of distributed datasets more straightforward and so much more commonplace. Furthermore the modifier "open" in this context refers to the idea that the datasets in question should be free to be downloaded and used by the public.

Over the last few years data about the growing number of datasets published as linked open data, the so called linked open data cloud, has been presented in the form of a diagram in which each dataset is represented by a node and the links between each dataset by edges between the corresponding nodes.

The publishing of data as linked open data is based on principles first elucidated by Tim Berners Lee (Berners-Lee, 2006). These principles recommend the use of the resource description framework (RDF), a language that models data in terms of triples of resources. Each of the resources in a triple is named using a unique resource identifier (URI). An RDF triple can be regarded as representing data in the form of a subject-predicate-object statement.

The many advantages and benefits of the emerging linked open data paradigm are obvious from a scientific standpoint. By putting different resources on the linked open data cloud it becomes far easier to link them together with each other in ways which render single resources much more useful than before, it also makes them more accessible and usable, facilitaing their reuse in an open ended variety of contexts (as is the case with the linked data version of Wikipedia). Indeed, this fact has not been lost on the language resources community, and the specific part of the linked open data cloud diagram dealing with language resources and datasets now includes a wide array

of linguistic resources including translations of the current version of the Princeton wordnet and ItalWordnet into RDF (Assem et al., 2006; Gangemi et al., 2003), (Bartolini et al., 2013), as well as a number of important vocabularies for language resources.

The *lemon* model (McCrae et al., 2011) is currently one of the most popular rdf based models for enabling the publishing of lexical resources as linked open data on the web. Its original focus was on the addition of linguistic information to ontologies, but it has by now been used to translate numerous different kinds of lexical resources into RDF, including many different wordnets.

Because the intial focus was on enriching already existing ontologies with lingustic information, the lemon model makes a clear distinction between a lexicon and an ontology. The pairing of a lexicon and an ontology as a combined lexico-semantic resouce takes place via the interlinking of a RDF based lexicon with an RDF based ontology. This is done using so called sense objects which are pointed to by lexical entries and which then in turn point to the vocabulary items in an ontology.

3 Parole Simple Clips and the Generative Lexicon

Parole Simple Clips (PSC) is a large, multilayered Italian language lexicon, the result of work carried out within the framework of three successive European/national projects. The first two of these were the European projects PAROLE (Ruimy et al., 1998) and SIMPLE (Lenci et al., 2000a) which produced wide coverage lexicons for a number of different European languages, including Italian, and all of which were designed to a common set of guidelines. These lexicons are arranged into phonetic, morphological, syntactic and semantic layers (the semantic layers were actually added during the SIMPLE project, the other layers during the earlier PAROLE project). The last of the projects instrumental in the creation of PSC was CLIPS, an Italian national project, which had the aim of expanding upon the Italian Parole-Simple lexicon. In this paper we focus on the translation of the syntactic and semantic layers of PSC into RDF using the *lemon* model.

The construction of the semantic layer of PSC was heavily influenced by Generative Lexicon (GL) theory (Pustejovsky, 1991; Bel et al., 2000). GL theory posits a complex multi part structure for individual word senses, making provision for the encoding of information related to different, salient, dimensions of a lexical entry's meaning[1].

In GL a lexical entry contains information on the position of the lexical entry in a language wide type system its so called lexical type structure; a predicative argument structure; information on the event type of the entry, the event structure; as well as a data structure known as a qualia structure. This qualia structure presents four distinct, orthogonal aspects of a word's meaning in terms of which polysemy as well as the creative and the novel uses of words based on established meanings can be straightforwardly explained.

These four aspects or qualia roles contained in each lexical entry's qualia structure, can be defined as follows. **The formal quale**: this corresponds to the ontological isA relation; **the constitutive quale**: this encodes meronymic or partOf relationships between an entity and the entities of which it is composed; **the telic quale**: this encodes the purpose for which an entity is used; **the agentive quale**: this encodes the factors that were involved in an entity's coming into being.

3.1 The Structure of the Semantic Layer of PSC

The semantic layer of PSC builds upon this theoretical foundation by introducing the notion of an Extended Qualia Structure (Lenci et al., 2000a) according to which each of the four qualia roles are further elaborated by being broken down into more specific relations. This means that for example the constitutive relation is further elaborated by relations specifying whether a constitutive relation holds between two elements on the basis of location, group membership, etc; telic relations are specified in terms of purpose, classified with respect to direct and indirect telicity, etc.

In PSC this Extended Qualia Structure is represented as a relation that holds between semantic units or USems in the terminology of PSC. In addition to the extended qualia relations there are also a number of so called lexical relations organised into the following five clases SYNONYMY, POLYSEMY, ANTONYMY,

[1]This information is used to construct larger units of meaning through a compositional process in which, to use the slogan common in GL theory literature, the semantic load is more equally spread over all of the consituents of an utterance, rather than being largely focused on the verbs.

DERIVATION, METAPHOR.

PSC makes use of a language independent, 'upper' ontology that is also common to the PAROLE-SIMPLE lexicons for other European languages; this has been converted into an OWL ontology (Toral and Monachini, 2007) which we make use of in our translation. This ontology contains 153 so called semantic types which provide a higher level structuring of the circa 60k Italian language specific USems contained in PSC. In order to illustrate the levels of information available in PSC, we use the example of the verb "dare", to give.

The verbal lexical entry *dare* maps onto 3 different semantic units (USem): (i) USem7149dare as in "to give something to someone"; (ii) USem79492dare as in "to give the medicine to the patient" (make ingest); (iii) USem79493dare as in "the window faces the square".

In PSC, the first two USems map onto the same syntactic frame with 3 positions, representing subject, object and indirect object, all of which are noun phrases and the latter of which is introduced by the preposition "a". We also know that this frame selects the auxiliary "avere". The other USem uses a bivalent frame instead.

In the semantic layer, a mapping is defined between each USem and a predicate. This mapping is not one-to-one, as in some cases two senses may map onto one predicate[2].

The predicates are then linked to their argument structures, so for instance the predicate structure of USem7149dare has three arguments, the first has the role of **Agent** and selects ontological type *Human*, the second has the role of **Patient** and selects the ontological type *Concrete_entity*, the third has role **Beneficiary** and selects the ontological type *Human*. A linking is also available between the semantic and the syntactic structure; in this case the predicative structure and the syntactic frame of USem7149dare are linked by an isomorphic trivalent relation, which means that Position1 maps onto Argument1, Position2 maps onto Argument2, and Position3 maps onto Argument3.

Finally, each of the USems of *dare* linked to other USems in the lexicon by means of the complex network of relations of the Extended Qualia Structure, and is also linked to the Interlingual upper level SIMPLE ontology. So for instance USem7149dare has ontological type *Change_of_Possession* and is linked to USem3939cambiare ("to change") on the formal axis and to USemD6219privo ("deprived of") on the constitutive axis, the USem USemD6219privo being the resulting state of the USem7149dare. Lexical relations such as polysemy or derivation are also possible for verbs.

4 Converting PSC into linked Data with *lemon*

A detailed account of the challenges brought about by the translation of the PSC resource into RDF is presented in (Del Gratta et al., 2013). Here we will summarize that work and thus lay the ground for further discussion on the translation of the PSC verbs in the next section.

The main challenge that arose during the conversion of the PSC nouns related to how best to understand the status of the Usems, namely whether these were better viewed as *lemon* senses which could then in turn be understood as reified pairings of lexical entries with ontological vocabulary items; or whether PSC USems should instead be seen as elements in an ontological layer.

As mentioned above USems take part in lexical relations such as synonymy, polysemy and antonymy which in standard works are treated as relations between lexical senses[3]. On the other hand PSC USems also take part in (Extended Qualia Structure) relations that are arguably better classed as ontological relations holding between the referents of words rather than between their senses, e.g., produces, produced-by, used-for, is_a_follower_of, is_the_habit_of: at the very least it seems odd to say that the relation of synonymy and a relation specifying whether relations of one class "produce" members of another hold between the same kind of element.

In the end the considerations given above along with the fact that the lemon model makes such a clear distinction between lexicon and ontology led to the decision to duplicate the USems: once as *lemon* lexical senses, with lexical relations like synonymy holding between them, and in the second instance as ontological entities. These are then to be seen as an lower level of the already

[2]This is especially the case for reflexive verbs such as *incolonnarsi* ("to line up") vs their transitive counterparts ("to line something/one up"), that are represented as different senses and different syntactic frames, but have the same underlying argument structure.

[3]Although the aforementioned lexical relations can themselves bedefined differently in different sources.

existing SIMPLE OWL ontology.

4.1 The Verbs

The modelling of the PSC verbs in linked open data involves a number of challenges over and above those that arose during the modelling of the nouns. In particular it is important to represent information about both the syntactic frames and semantic predicates associated with verb senses[4]. In addition it is also desirable to have some kind of mapping between these two kinds of representation, so that the syntactic arguments of a verb frame can be mapped to the semantic arguments of the verb's semantic predicative representation.

One of the considerations that we have been most keenly aware of throughout the process of developing a model for the PSC verbs is that we are attempting to convert a legacy resource with a relatively long history and a well documented design that was developed through the collaboratio of a number of experts in the field.

We have therefore tried to remain as faithful as possible to the original intentions of the designers of PSC, while at the same time exploiting the advantages and opportunities offered up by the linked data model.

We present our proposal for verbs below. Once more we are working with the Italian verb *dare*.

```
:dare_1 a lemon:sense ;
        lemon:reference :USem7149dare ;
        psc:synBehavior frames:t-ind-xa ;

lmf:hasSemanticPredicate :PREDdare#1 ;
        psc:hasSynSemMapping
                    ssm:Isotrivalent .

:PREDdare#1 a lmf:SemanticPredicate ;
        lmf:hasArgument ARG0dare#1 ;
        lmf:hasArgument ARG1dare#1 ;
        lmf:hasArgument ARG2dare#1 .

:ARG0dare#1 a lmf:Argument ;
        a simple:ArgHuman .

:ARG1dare#1 a lmf:Argument ;
        a simple:Concrete_Entity .

:ARG2dare#1 a lmf:Argument ;
        a simple:ArgHuman .

:ARG2dare#1 lemon:marker :a .
```

The lexical entries point to their reified sense objects. In the example these are named dare_1, dare_2, whereas the USem ID is used to name the ontological counterpart of the original PSC USem, the reference object[5].

We use the psc prefix to refer to the name space main file containing the definitions of concepts and properties in the example.

Each lexical sense points to a lemon:frame by means of the psc:synBehavior property[6]. These frames are stored in a separate file, each frame in this file is an abstraction over many syntactic frames. So in the example the verb sense dare_1 is mapped to a frame t-ind-xa. This represents a transitive frame for a verb with both a direct and indirect object and which takes *avere* as an auxilliary verb.

The sense is also linked to a predicate object, which provides descriptions of the argument structure. We use the lmf property hasSemanticPredicate to link to an lmf SemanticPredicate PREDdare#1. The type selected by each argument of the predicatepoints back to the SIMPLE Ontology.

Finally the sense dare_1 is linked to ssm:Isotrivalent an object representing the mapping between the syntactic frame and the semantic predicate via the hasSynSemMapping property. We have created a file ssm that contains a number of these mappings as represented in the PSC specifications. The particular mapping object in question, Isotrivalent, represents the isomorphic trivalent relation mentioned above. Details on the best way of representing these mappings using OWL will be provided in the final paper.

5 Conclusion

In this paper we have presented our model for representing the PSC verbs using the lemon model. As we have stated above this is currently work in progress. In the final paper the link to the public dataset will be provided.

References

Mark Van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of wordnet to a standard rdf/owl representation. In *Proceedings of the Fifth International Conference on Language Resources and Eval-*

[4]It is also true that PSC nouns have predicative structure but this was ignored during the intial translation of PSC into linked data.

[5]Although for space reasons we are unable to show this here, in our model lexical relations are kept between senses, while ontological relations are implemented between usems qua ontological objects, as for nouns.

[6]In our model, this is a property of Senses rather than LexicalEntries.This is in keeping with the PSC specifications.

uation (LREC-2006), Genoa, Italy, May. European Language Resources Association (ELRA).

Roberto Bartolini, Riccardo Del Gratta, and Francesca Frontini. 2013. Towards the establishment of a linguistic linked data network for italian. In *2nd Workshop on Linked Data in Linguistics*, page 76.

Núria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. In *LREC* (DBL, 2000).

Tim Berners-Lee. 2006. Linked data. *W3C Design Issues*.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 111–125, Berlin, Heidelberg. Springer-Verlag.

N. Calzolari. 2008. Approaches towards a 'Lexical Web': the Role of Interoperability. In J. Webster, N. Ide, and A. Chengyu Fang, editors, *Proceedings of The First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 18–25.

Philipp Cimiano, John McCrae, Paul Buitelaar, and Elena Montiel-Ponsoda, 2012. *On the Role of Senses in the Ontology-Lexicon*.

2000. *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*. European Language Resources Association.

Riccardo Del Gratta, Francesca Frontini, Fahad Khan, and Monica Monachini. 2013. Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web Journal (Under Review)*.

Gil Francopoulo. 2013. *LMF - Lexical Markup Framework*. ISTE Ltd + John Wiley & sons, Inc, 1 edition.

Gil Francopulo, Romary Laurent, Monica Monachini, and Nicoletta Calzolari. 2006. Lexical markup framework (lmf iso-24613). In European Language Resources Association (ELRA), editor, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genova, IT.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer.

Yoshihiko Hayashi. 2011. Direct and indirect linking of lexical objects for evolving lexical linked data. In *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011)*, 10.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000a. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Alessandro Lenci, F. Busa, Nilda Ruimy, E. Gola, Monica Monachini, Nicoletta Calzolari, and Antonio Zampolli. 2000b. Simple linguistic specifications. Deliverable. In: LE-SIMPLE (LE4-8346), Deliverable D2.1 & D2.2. ILC and University of Pisa, Pisa, 404 pp. 2000.

Ernesto William De Luca, Martin Eul, and Andreas Nrnberger. 2007. Converting eurowordnet in owl and extending it with domain ontologies. In C. Kunze, L. Lemnitzer, and R. Osswald, editors, *Proceedings of the GLDV-2007 Workshop on Lexical-Semantic and Ontological Resources*, page 3948.

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.

James Pustejovsky. 1991. The generative lexicon. *Comput. Linguist.*, 17(4):409–441, dec.

N. Ruimy, O. Corazzari, E. Gola, A. Spanu, N. Calzolari, and A. Zampolli. 1998. The european le-parole project: The italian syntactic lexicon. In *Proceedings of the First International Conference on Language resources and Evaluation*, pages 241–248.

Antonio Toral and Monica Monachini. 2007. Simple-owl: a generative lexicon ontology for nlp and the semantic web. In *Workshop of Cooperative Construction of Linguistic Knowledge Bases, 10th Congress of Italian Association for Artificial Intelligence.*

10.12871/CLICIT2014144

A Preliminary Comparison of State-of-the-art Dependency Parsers on the Italian Stanford Dependency Treebank

Alberto Lavelli
FBK-irst
via Sommarive, 18 - Povo
I-38123 Trento (TN) - ITALY
lavelli@fbk.eu

Abstract

English. This paper reports the efforts involved in applying several state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank (ISDT). The aim of such efforts is twofold: first, to compare the performance and choose the parser to participate in the EVALITA 2014 task on dependency parsing; second, to investigate how simple it is to apply freely available state-of-the-art dependency parsers to a new language/treebank.

Italiano. *Questo articolo descrive le attività svolte per applicare vari analizzatori sintattici a dipendenza allo stato dell'arte all'Italian Stanford Dependency Treebank (ISDT). L'obiettivo di questi sforzi è duplice: in primo luogo, confrontare le prestazioni e scegliere il parser per partecipare al task EVALITA 2014 su dependency parsing; secondo, indagare quanto è facile applicare analizzatori sintattici a dipendenza liberamente disponibili a una nuova lingua / treebank.*

1 Introduction

Recently, there has been an increasing interest in dependency parsing, witnessed by the organisation of a number of shared tasks, e.g. Buchholz and Marsi (2006), Nivre et al. (2007). Concerning Italian, there have been tasks on dependency parsing in all the editions of the EVALITA evaluation campaign (Bosco et al., 2008; Bosco et al., 2009; Bosco and Mazzei, 2011; Bosco et al., 2014). In the 2014 edition, the task on dependency parsing exploits the Italian Stanford Dependency Treebank (ISDT), a new treebank featuring an annotation based on Stanford Dependencies (de Marneffe and Manning, 2008).

This paper reports the efforts involved in applying several state-of-the-art dependency parsers on ISDT. There are at least two motivations for such efforts. First, to compare the results and choose the parsers to participate in the EVALITA 2014 task on dependency parsing. Second, to investigate how simple it is to apply freely available state-of-the-art dependency parsers to a new language/treebank following the instructions available together with the code and possibly having a few interactions with the developers.

As in many other NLP fields, there are very few comparative articles when the performance of different parsers is compared. Most of the papers simply present the results of a newly proposed approach and compare them with the results reported in previous articles. In other cases, the papers are devoted to the application of the same tool to different languages/treebanks.

It is important to stress that the comparison concerns tools used more or less out of the box and that the results cannot be used to compare specific characteristics like: parsing algorithms, learning systems, ...

2 Parsers

The choice of the parsers used in this study started from the two we already applied at EVALITA 2011, i.e. MaltParser and the ensemble method described by Surdeanu and Manning (2010). We then identified a number of other dependency parsers that, in the last years, have shown state-of-the-art performance, that are freely available and with the possibility of training on new treebanks. The ones included in the study reported in this paper are the MATE dependency parsers, TurboParser, and ZPar.

We plan to include other dependency parsers in our study. We have not been able to exploit some of them because of different reasons: they are not yet available online, they lack documenta-

tion on how to train the parser on new treebanks, they have limitations in the encoding of texts (input texts only in ASCII and not in UTF-8), ...

MaltParser (Nivre et al., 2006) (version 1.8) implements the transition-based approach to dependency parsing, which has two essential components:

- A nondeterministic transition system for mapping sentences to dependency trees
- A classifier that predicts the next transition for every possible system configuration

Given these two components, dependency parsing can be performed as greedy deterministic search through the transition system, guided by the classifier. With this technique, it is possible to perform parsing in linear time for projective dependency trees and quadratic time for arbitrary (non-projective) trees (Nivre, 2008). MaltParser includes different built-in transition systems, different classifiers and techniques for recovering non-projective dependencies with strictly projective parsers.

The ensemble model made available by Mihai Surdeanu (Surdeanu and Manning, 2010)[1] implements a linear interpolation of several linear-time parsing models (all based on MaltParser). In particular, it combines five different variants of MaltParser (Nivre's arc-standard left-to-right, Nivre's arc-eager left-to-right, Covington's non projective left-to-right, Nivre's arc-standard right-to-left, Covington's non projective right-to-left) as base parsers.

The MATE tools[2] include both a graph-based parser (Bohnet, 2010) and a transition-based parser (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012). For the languages of the 2009 CoNLL Shared Task, the graph-based MATE parser reached accuracy scores similar or above the top performing systems with fast processing. The speed improvement is obtained with the use of Hash Kernels and parallel algorithms. The transition-based MATE parser is a model that takes into account complete structures as they become available to rescore the elements of a beam, combining the advantages of transition-based and graph-based approaches.

TurboParser (Martins et al., 2013)[3] (version 2.1) is a C++ package that implements graph-based dependency parsing exploiting third-order features.

ZPar (Zhang and Nivre, 2011) is a transition-based parser implemented in C++. ZPar supports multiple languages and multiple grammar formalisms. ZPar has been most heavily developed for Chinese and English, while it provides generic support for other languages. It leverages a global discriminative training and beam-search framework.

3 Data Set

The experiments reported in the paper are performed on the Italian Stanford Dependency Treebank (ISDT) (Bosco et al., 2013) version 2.0 released in the context of the EVALITA evaluation campaign on Dependency Parsing for Information Extraction (Bosco et al., 2014)[4]. There are three main novelties with respect to the previously available Italian treebanks: (i) the size of the dataset, which is much bigger than the resources used in the previous EVALITA campaigns; (ii) the annotation scheme, which is compliant to *de facto* standards at the level of both representation format (CoNLL) and adopted tagset (Stanford Dependency Scheme); (iii) its being defined with a specific view to supporting information extraction tasks, a feature inherited from the Stanford Dependency scheme.

The EVALITA task focuses on standard dependency parsing of Italian texts with evaluations aimed at testing the performance of parsing systems as well as their suitability to Information Extraction tasks.

The training set contains 7,414 sentences (158,561 tokens), the development set 564 sentences (12,014 tokens), and the test set 376 sentences (9,066 tokens).

4 Experiments

The level of interaction with the authors of the parsers varied. In two cases (ensemble, MaltParser), we have mainly exploited the experience gained in previous editions of EVALITA. In the case of the MATE parsers, we have had a few in-

[1] http://www.surdeanu.info/mihai/ensemble/

[2] https://code.google.com/p/mate-tools/

[3] http://www.ark.cs.cmu.edu/TurboParser/

[4] http://www.evalita.it/2014/tasks/dep_par4IE.

			collapsed and propagated		
	LAS		P	R	F_1
MATE stacking (TurboParser)	89.72		82.90	90.58	86.57
Ensemble (5 parsers)	89.72		82.64	90.34	86.32
ZPar	**89.53**		84.65	92.11	88.22
MATE stacking (transition-based)	89.02		82.09	89.77	85.76
TurboParser (model_type=full)	88.76		83.32	90.71	86.86
TurboParser (model_type=standard)	88.68		83.07	90.55	86.65
MATE graph-based	88.51		81.72	89.42	85.39
MATE transition-based	88.32		80.70	89.40	84.82
Ensemble (MaltParser v.1.8)	88.15		80.69	88.34	84.34
MaltParser (Covington non proj)	87.79		81.50	87.39	84.34
MaltParser (Nivre eager -PP head)	**87.53**		81.30	88.78	84.88
MaltParser (Nivre standard - MaltOptimizer)	86.35		81.17	89.04	84.92
Ensemble (MaltParser v.1.3)	86.27		78.57	86.28	82.24

Table 1: Results on the EVALITA 2014 development set without considering punctuation. The second column reports the results in term of Labeled Attachment Score (LAS). The score is in bold if the difference with the following line is statistically significant. The three columns on the right show the results in terms of Precision, Recall and F_1 for the collapsed and propagated relations.

teractions with the author who suggested the use of some undocumented options. In the case of TurboParser, we have simply used the parser as it is after reading the available documentation. Concerning ZPar, we have had a few interactions with the authors who helped solving some issues.

As for the ensemble, at the beginning we repeated what we had already done at EVALITA 2011 (Lavelli, 2011), i.e. using the ensemble as it is, simply exploiting the more accurate extended models for the base parsers. The results were unsatisfactory, because the ensemble is based on an old version of MaltParser (v.1.3) that performs worse than the current version (v.1.8). So we decided to apply the ensemble model both to the output produced by the current version of MaltParser and to the output produced by some of the parsers used in this study. In the latter case, we have used the output of the following 5 parsers: graph-based MATE parser, transition-based MATE parser, TurboParser (full model), MaltParser (Nivre's arc-eager, PP-head, left-to-right), and MaltParser (Nivre's arc-eager, PP-head, right-to-left).

Concerning MaltParser, in addition to using the best performing configurations at EVALITA 2011[5], we have used MaltOptimizer[6] (Ballesteros and Nivre, 2014) to identify the best configuration. According to MaltOptimizer, the best configuration is Nivre's arc-standard. However, we have obtained better results using the configurations used in EVALITA 2011. We are currently investigating this issue.

As for the MATE parsers, we have applied both the graph-based parser and the transition-based parser. Moreover, we have combined the graph-based parser with the output of another parser (both the transition-based parser and TurboParser) using stacking. Stacking is a technique of integrating two parsers at learning time[7], where one of the parser generates features for the other.

Concerning ZPar, the main difficulty was the fact that a lot of RAM is needed for processing long sentences (i.e., sentences with more than 100 tokens need 70 GB of RAM).

During the preparation of the participation to the task, the experiments were performed using the split provided by the organisers, i.e. training on the training set and testing using the development set.

When applying stacking, we have performed 10-fold cross validation of the first parser on the training set, using the resulting output to provide to the second parser the predictions used during learning. During parsing, the output of the first parser (trained on the whole training set and applied to the development set) has been provided to the second parser.

In Table 1 we report the parser results ranked according to decreasing Labeled Accuracy Score

[5]Nivre's arc-eager, PP-head, and Covington non projective.

[6]http://nil.fdi.ucm.es/maltoptimizer/

[7]Differently from what is done by the ensemble method described above where the combination takes place only at parsing time.

			collapsed and propagated		
	LAS		P	R	F_1
MATE stacking (transition-based)	87.67		79.14	88.14	83.40
Ensemble (5 parsers)	87.53		78.28	88.09	82.90
MATE stacking (TurboParser)	87.37		79.13	87.97	83.31
MATE transition-based	87.07		78.72	87.16	82.73
MATE graph-based	86.91		78.74	87.97	83.10
ZPar	86.79		80.30	88.93	84.39
TurboParser (model_type=full)	86.53		79.43	89.42	84.13
TurboParser (model_type=standard)	86.45		79.65	89.32	84.21
Ensemble (MaltParser v.1.8)	85.94		76.30	86.38	81.03
MaltParser (Nivre eager -PP head)	**85.82**		78.47	86.06	82.09
Ensemble (MaltParser v.1.3)	85.06		76.36	84.74	80.33
MaltParser (Covington non proj)	84.94		77.24	82.97	80.00
MaltParser (Nivre standard - MaltOptimizer)	84.44		76.53	86.99	81.43

Table 2: Results on the EVALITA 2014 test set without considering punctuation. The second column reports the results in term of Labeled Attachment Score (LAS). The score is in bold if the difference with the following line is statistically significant. The three columns on the right show the results in terms of Precision, Recall and F_1 for the collapsed and propagated relations.

(LAS), not considering punctuation. The score is in bold if the difference with the following line is statistically significant[8]. In the three columns on the right of the table the results for the collapsed and propagated relations are shown (both the conversion and the evaluation are performed using scripts provided by the organisers).

The ranking of the results according to LAS and according to Precision, Recall and F_1 are different. This made the choice of the parser for the participation difficult, given that the participants would have been ranked based on both measures.

According to the results on the development set, we decided to submit for the official evaluation three models: ZPar, MATE stacking (TurboParser), and the ensemble combining 5 of the best parsers. In this case, the training was performed using both the training and the development set. In Table 2. you may find the results of all the parsers used in this study (in italics those submitted to the official evaluation). Comparing Table 1 and Table 2 different rankings between parsers emerge. This calls for an analysis to understand the reasons of such difference. The results of a preliminary analysis and further details about our participation to the task are reported in Lavelli (2014).

The results obtained by the best system submitted to the official evaluation are: 87.89 (LAS), 81.89/90.45/85.95 (P/R/F_1). More details about the task and the results obtained by the participants are available in Bosco et al. (2014).

We are currently analysing the results shown above to understand how to further proceed in our investigation. A general preliminary consideration is that approaches that combine the results of different parsers perform better than those based on a single parser model, usually with the drawback of a bigger complexity.

5 Conclusions

In the paper we have reported on work in progress on the comparison between several state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank (ISDT).

In the near future, we plan to widen the scope of the comparison including more parsers.

Finally, we will perform an analysis of the results obtained by the different parsers considering not only their performance but also their behaviour in terms of speed, CPU load at training and parsing time, ease of use, licence agreement, ...

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923). We would like to thank the authors of the parsers for making them freely available. In particular, we would like to thank Bernd Bohnet, Joakim Nivre, Mihai Surdeanu, Yue Zhang, and Yijia Liu for kindly answering our questions on the practical application of their parsers and for providing useful suggestions.

[8] To compute the statistical significance of the differences between results, we have used MaltEval (Nilsson and Nivre, 2008)

References

Miguel Ballesteros and Joakim Nivre. 2014. MaltOptimizer: Fast and effective parser optimization. *Natural Language Engineering*, FirstView:1–27, 10.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both worlds – a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87, Avignon, France, April. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Cristina Bosco and Alessandro Mazzei. 2011. The EVALITA 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*, pages 24–25.

Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Giuseppe Attardi, Anna Corazza, Alberto Lavelli, Leonardo Lesmo, Giorgio Satta, and Maria Simi. 2008. Comparing Italian parsers on a common treebank: the EVALITA experience. In *Proceedings of LREC 2008*.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice DellOrletta, and Alessandro Lenci. 2009. Evalita09 parsing task: comparing dependency parsers and treebanks. In *Proceedings of EVALITA 2009*.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *Proceedings of EVALITA 2014*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August. Coling 2008 Organizing Committee.

Alberto Lavelli. 2011. An ensemble model for the EVALITA 2011 dependency parsing task. In *Working Notes of EVALITA 2011*.

Alberto Lavelli. 2014. Comparing state-of-the-art dependency parsers for the EVALITA 2014 dependency parsing task. In *Proceedings of EVALITA 2014*.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jens Nilsson and Joakim Nivre. 2008. MaltEval: an evaluation and visualization tool for dependency parsing. In *Proceedings of LREC 2008*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 2216–2219.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34:513–553.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652, Los Angeles, California, June. Association for Computational Linguistics.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.

10.12871/CLICIT2014145

SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations

Alessandro Lenci[1], Gianluca E. Lebani[1], Sara Castagnoli[2], Francesca Masini[2], Malvina Nissim[3]
[1]University of Pisa, Department of Philology, Literature, and Linguistics
[2]Alma Mater Studiorum, University of Bologna, LILEC
[3]Alma Mater Studiorum, University of Bologna, FICLIT
alessandro.lenci@ling.unipi.it,gianluca.lebani@for.unipi.it,
{s.castagnoli|francesca.masini|malvina.nissim}@unibo.it

Abstract

English. The paper presents SYMPAThy, a new approach to the extraction of Word Combinations. The approach is new in that it combines pattern-based (P-based) and syntax-based (S-based) methods in order to obtain an integrated and unified view of a lexeme's combinatory potential.

Italiano. *L'articolo presenta SYMPAThy, un nuovo metodo per l'estrazione di Combinazioni di Parole. L'originalità dell'approccio consiste nel combinare il metodo basato su sequenze di parti del discorso (P-based) e quello basato sulle dipendenze sintattiche (S-based) per arrivare a una visione integrata e unitaria del potenziale combinatorio di un lessema.*

1 Introduction: Word Combinations

The term Word Combinations (WOCs), as used here, broadly refers to the range of combinatory possibilities typically associated with a word.

On the one hand, it comprises so-called Multiword Expressions (MWEs), intended as a variety of recurrent word combinations that act as a single unit at some level of linguistic analysis (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008): they include phrasal lexemes, idioms, collocations, etc.

On the other hand, WOCs also include the preferred distributional interactions of a word (be it a verb, a noun or an adjective) with other lexical entries at a more abstract level, namely that of argument structure patterns, subcategorization frames, and selectional preferences. Therefore, WOCs include both the *normal* combinations of a word and their idiosyncratic *exploitations* (Hanks, 2013).

The *full combinatory potential* of a lexical entry can therefore be defined and observed at the level of syntactic dependencies and at the more constrained surface level. In both theory and practice, though, these two levels are often kept separate. Theoretically, argument structure is often perceived as a "regular" syntactic affair, whereas MWEs are characterised by "surprising properties not predicted by their component words" (Baldwin and Kim, 2010, 267). At the pratical level, in order to detect potentially different aspects of the combinatorics of a lexeme, different extraction methods are used – i.e. either a surface, pattern-based (**P-based**) method or a deeper, syntax-based (**S-based**) method – as their performance varies according to the different types of WOCs/MWEs (Sag et al., 2002; Evert and Krenn, 2005).

We argue that, in order to obtain a comprehensive picture of the combinatorial potential of a word and enhance extracting efficacy for WOCs, the P-based and S-based approaches should be combined. Thus, we extracted corpus data into a database where both P-based and S-based information is stored together and accessible at the same time. In this contribution we show its advantages. This methodology has been developed on Italian data, within the CombiNet[1] project, aimed at building an online resource for Italian WOCs.

2 Existing extraction methods

The automatic extraction of combinatory information at both the P-level and the S-level is usually carried out in a similar fashion: first, dependency or surface structures are automatically extracted from corpus data, and second, the extracted structures are ranked according to frequency and/or one or more association measures, in order to distinguish meaningful combinations from sequences of words that do not form any kind of relevant unit (Evert and Krenn, 2005; Ramisch et al., 2008; Villavicencio et al., 2007). Let us summarize pros and cons of both methods.

[1]http://combinet.humnet.unipi.it

2.1 P-based approach

P-based methods exploit shallow (POS-)patterns, and are often employed for extracting WOCs. The specification of POS-patterns is a necessary step to obtain a better set of candidate structures with respect to (adjacent) unspecified n-grams. However, despite any attempt to obtain a comprehensive list of language-appropriate patterns (Nissim et al., 2014), not every extracted combination is a WOC, even after association measures are applied. The string may be part of a larger WOC (see *stesso tempo* 'same time', which is a very frequent bigram in itself, but is in fact part of the larger *allo stesso tempo* 'at the same time'), or it may contain a WOC plus some extra element (e.g. *anno/i di crisi economica* 'year(s) of economic crisis', containing *crisi economica* 'economic crisis'). Overall, however, the P-based method yields satisfactory results for relatively fixed, adjacent, and short (2-4 words) WOCs (e.g. *alte sfere* 'high society').

Some WOCs, however, especially verbal ones[2], allow for higher degrees of syntactic flexibility (e.g. passivization, dislocation, variation/addition/dropping of a determiner, internal modification by means of adjectives/adverbs, etc.) (Villavicencio et al., 2007) and/or display a complexity which is difficult to capture without resorting to syntactic information. A collocation like *aprire una discussione* 'start a discussion', for instance, is syntagmatically non-fixed in a number of ways: the determiner can vary (*aprire una/la discussione* 'start a/the discussion'), the object can be modified (*aprire una lunga e difficile discussione* 'start a long and difficult discussion'), and passivization is allowed (*la discussione è stata aperta* 'the discussion was started'). This would require taking into account and specifying all possible variations a priori. Similarly, some idioms can be very difficult to capture with POS-patterns because of their length and complexity, which is hardly "generalizable" into meaningful POS sequences (e.g.: *dare un colpo al cerchio e uno alla botte* lit. give a blow to the ring and one to the barrel 'run with the hare and hunt with the hounds'). Last but not least, P-based approaches are not able to address more abstract combinatory information (e.g. argument structures) and are thus typically limited to MWEs.

[2]In Italian, verbal MWEs are less fixed than nominal ones (Voghera, 2004), even though variability is a thorny issue for nominal MWEs, too (Nissim and Zaninello, 2013).

2.2 S-based approach

S-based methods are based on dependency relations extracted from parsed corpora. They offer the possibility to extract co-occurrences of words in specific syntactic configurations (e.g. subject-verb, verb-object etc.) irrespective of their superficial realizations, i.e. generalizing over syntactic flexibility and interrupting material. S-based extraction methods thus have two major advantages. First, by moving away from surface forms, they can help account for the complexity and the syntactic variability that some WOCs – like the V+N combination *aprire una discussione* above – might exhibit. Second, by taking into account the dependency between elements, they minimise the risk of extracting unrelated words (Seretan et al., 2003). As a consequence, they are particularly useful to extract "abstract" structures such as lexical sets, i.e. lists of fillers in given slots (e.g. the most prototypical objects of a verb), argument structure patterns and subcategorization frames.

However, precisely because S-based methods abstract away from specific constructs and information (word order, morphosyntactic features, interrupting material, etc.), they do not consider how exactly words are combined. Thus, the regular phrase *gettare acqua su un fuoco* 'throw water on a fire' and the structurally similar idiom *gettare acqua sul fuoco* 'defuse' would be treated equally, on the basis of the combination of throw-water-fire.

Also, S-based approaches cannot distinguish frequent "regular" combinations (e.g. *gettare la sigaretta* 'throw the sigarette') from idiomatic combinations that have the very same syntactic structure (e.g. *gettare la spugna* lit. throw the sponge 'throw in the towel'). Statistical association measures alone are not able to discriminate between them as both *sigaretta* and *spugna* are likely to appear among the preferred fillers of the object slot of *gettare*.

3 SYMPAThy: A unified approach

P-based and S-based methods for WOC analysis are in fact highly complementary. In our view, the existing dualism does not reflect the fact that all these combinatory phenomena are interconnected with one another, and that there is a very intricate continuum that links fixed and flexible combinations, compositional and totally idiomatic ones.

In order to represent the full combinatory potential of lexemes, and in an attempt to disentan-

gle this continuum of WOCs, we propose to adopt a unified approach, whose theoretical premises lie in a constructionist view of the language architecture. In Construction Grammar, the basic unit of analysis is the Construction, intended as a conventionalized association of a form and a meaning that can vary in both complexity and schematicity (Fillmore et al., 1988; Goldberg, 2006; Hoffmann and Trousdale, 2013). Therefore, Constructions span from specific structures such as single words (Booij, 2010) to complex, abstract structures such as argument patterns (Goldberg, 1995), in what is known as the lexicon-syntax continuum, which comprises MWEs and other types of WOCs.

3.1 SYntactically Marked PATterns

We implemented this view in a distributional knowledge base, SYMPAThy (SYntactically Marked PATterns), built by extracting from a dependency-parsed corpus all the occurrences of a set of lemmas and processing them so as to obtain an integrated representation of the kinds of combinatorial information usually targeted in S-based and P-based methods, albeit separately.

The ultimate goal of our extraction algorithm is to filter and interpret the linguistic annotation provided by a pipeline of NLP tools and to represent it with a data format that allows for the simultaneous encoding of the following linguistic information, for any terminal node that depends on a given target lemma TL or on its direct governor:

- its lemma;
- its POS tag;
- its morphosyntactic features;
- its linear distance from the TL;
- the dependency path linking it to TL.

By building on an automatically annotated corpus, the actual implementation of the SYMPAThy extraction algorithm is largely dependent on the properties of the specific linguistic annotation tools exploited. Here we report examples extracted from a version of the "la Repubblica" corpus (Baroni et al., 2004) that has been POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency parsed with DeSR (Attardi and Dell'Orletta, 2009).

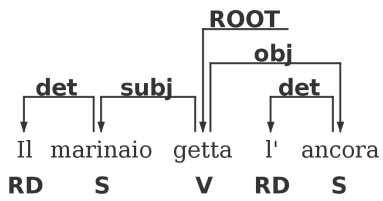
Figure 1 shows the different patterns that can be extracted from the sentence *il marinaio getta l'ancora* 'the sailor throws the anchor', for two different TLs: *gettare* 'throw' and *ancora* 'anchor'. In this representation, the terminal nodes are labeled with patterns of the form `lemma-pos|morphological features|distance_from_target`. For instance, the label `il-r|sm|-2` should be read as an instance of the singular masculine form (`sm`) of the lemma *il* 'the', that is an article (`r`) linearly placed two tokens on the left of TL[3].

The structural information encoded by our patterns, moreover, abstracts from the one-to-one dependency relations identified by the parser in order to build macro-constituents somehow reminiscent of the tree structure typical of phrase structure grammars. Such macro-constituents represent meaningful chunks of linguistic units, in which one element (the 'head', marked by a superscript ^{H}) is prominent with respect to the others. Non-head elements include intervening elements, like determiners, auxiliaries and quantifiers, whose presence is crucial to determine how fixed a linguistic construction is (and that is usually neglected in S-based approaches), and whose linear placement should be known a priori in a P-based perspective. This information is vital in distinguishing idioms, like *gettare acqua sul fuoco* (see Section 2.2), from otherwise identical compositional expressions like *gettare acqua su quel grande fuoco* ('throw water on that big fire').

Finally, the contrast between the two patterns reported in Figure 1 gives a measure of how much the SYMPAThy data representation format is target-dependent. On the one hand, both the syntactic annotation and the linear order are represented with respect to the TL: see the inverse OBJ-1 dependency in the *ancora*-based pattern, as well as the rationale of the indexing encoding the linear positions of terminal elements.

On the other hand, only the part of the sentence that is relevant to characterize the combinatorial behavior of the TL is extracted. In the preliminary work presented here, such a relevant portion includes all the constituents that are directly or indirectly governed by TL (e.g. the object of a verb together with the prepositional phrases modifying its nominal head), and the constituent that governs TL, thus encoding inverse relations like the OBJ-1 dependency in the lower pattern of Figure 1.

[3] For a description of the tagsets used to annotate the corpus, see: `http://medialab.di.unipi.it/wiki/Tanl_Tagsets`.



$[_{TGT}$ $[_{SUBJ}$ il-r|sm|-2 marinaio-s|sm|-1^{H}] **gettare-v|s3ip|$\mathbf{0}^{H}$**
$[_{OBJ}$ il-r|sn|1 ancora-s|sn|2^{H}]]

$[_{OBJ-1}$ gettare-v|s3ip|-2^{H} $[_{TGT}$ il-r|sn|-1 **ancora-s|sn|$\mathbf{0}^{H}$**]]

Figure 1: LEFT: dependency tree[4] for the sentence *il marinaio getta l'ancora* 'the sailor throws the anchor'; RIGHT: SYMPAThy patterns for the TLs *gettare* 'throw' (above) and *ancora* 'anchor' (below).

3.2 A sympathetic example

Here follows a small example showing how such a representation can be used to integrate S-based and P-based approaches. We extracted from our parsed version of the "la Repubblica" corpus all the SYMPAThy patterns featuring a transitive construction governed by the TL *gettare* 'throw'. In a S-based fashion, we ranked the nominal heads filling the object position and found that the most frequent object fillers of *gettare* are *spugna* 'sponge', *acqua* 'water' and *ombra* 'shadow'.

By taking into account the whole subcategorization frame in which these $<TL, obj>$ pairings occur, other interesting patterns emerge. When occurring with *acqua*, TL is often associated with a complement introduced by the preposition *su* and headed by the noun *fuoco* 'fire'. Another salient pattern displays TL with the object *ombra* and an indirect complement introduced by *su* and filled by a nominal head other than *fuoco*.

At the S-level only, it is difficult to guess what the status of these constructions is. Are they compositional or somehow fixed? If the latter, in which way and to what extent is their variation limited? The P-based side of the SYMPAThy data format comes in handy to address such issues. Here crucial pieces of information are the presence/absence of intervening material between TL and the heads of the governed constituents, how variable is the morphological behavior of the relevant lexical elements and to what extent they are free to be superficially realized with respect to TL.

By looking at this information, we can see that the strong association *gettare* + *obj:spugna* is due to the high frequency of the idiomatic expression *TL_la_spugna* 'throw in the towel'. Indeed, 98% of the patterns are linearly and morphologically fixed, with most of the remaining cases (1.7%) being superficial variations due to the presence of interrupting material, typically adverbs.

Cases with *acqua* in the object position present a more articulated picture. Half of them (53.5%) are instances of the rigid idiomatic expression *TL_acqua_sul_fuoco* 'defuse'. As for the remaining cases, even if there is a strong preference for realizing TL and the object one next to the other, with no morphological variation (84%), there is substantial variability in the number, type and filler of the indirect complement (36% of the remaining cases are instances of a subcategorization frame different from the simple transitive one).

When the object slot is filled by *ombra*, finally, the constructions appear to be freer. Even if there is a strong preference (40% of the cases) for the idiom *TL_(una|la)_ombra_su*, roughly meaning 'cast a shadow on', dimensions of variability include the presence/absence of a determiner, its type, and the optional presence of intervening tokens (e.g. adverbs/adjectives) between TL and the object.

Overall this brief example shows how P-based and S-based ideas can be used together to obtain a better description of the combinatoric behavior of lexemes, thus advocating for the usefulness of a resource like SYMPAThy that is able to bridge between the aforementioned approaches.

4 Conclusions

In this paper we presented SYMPAThy, a new method for the extraction of WOCs that exploits a variety of information typical of both P-based and S-based approaches. Although SYMPAThy was developed on Italian data, it can be adapted to other languages. In the future, we intend to exploit this combinatory base to model the gradient of schematicity/productivity and fixedness of combinations, in order to develop an "WOC-hood" indicator to classify the different types of WOCs on the basis of their distributional behavior.

[4]Plotted with DgAnnotator: `http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/`

Acknowledgments

This research was carried out within the CombiNet project (PRIN 2010-2011 *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, n. 20105B3HE8), coordinated by Raffaele Simone (Roma Tre University) and funded by the Italian Ministry of Education, University and Research (MIUR).

References

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL 2009*, pages 261–264.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of LREC 2004*, pages 1771–1774.

Geert Booij. 2010. *Construction morphology*. Oxford University Press, Oxford.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940.

Felice Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466. Special issue on Multiword Expression.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64(3):501–538.

Adele Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structures*. The University of Chicago Press, Chicago.

Adele Goldberg. 2006. *Constructions at work*. Oxford University Press, Oxford.

Stefan Th. Gries. 2008. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 3–25. John Benjamins, Amsterdam & Philadelphia.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Malvina Nissim and Andrea Zaninello. 2013. Modeling the internal variability of multiword expressions through a pattern-based method. *ACM Trans. Speech Lang. Process.*, 10(2):1–26.

Malvina Nissim, Sara Castagnoli, and Francesca Masini. 2014. Extracting mwes from italian corpora: A case study for refining the pos-pattern methodology. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 57–61.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the LREC Workshop MWE 2008*, pages 50–53.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of RANLP-03*, pages 424–431.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP-CoNLL 2007*, pages 1034–1043.

Miriam Voghera. 2004. Polirematiche. *Linguistica Pragensia*, 67(2):100–108.

10.12871/CLICIT2014146

Più l'ascolto e più *Mi piace!* Social media e Radio: uno studio preliminare del successo dei post

Eleonora Lisi
University of Rome "Tor Vergata"
eleonora.lisi.3@gmail.com

Emanuele Donati
Radio Dimensione Suono
e.donati@rds.it

Fabio Massimo Zanzotto
DII - University of Rome "Tor Vergata"
fabio.massimo.zanzotto@uniroma2.it

Abstract

English. Radio channels are fighting the final battle against other media. In this paper, we want to analyze how radio channels can exploit social networks to survive in this war. The final goal of the project is to find a strategy that radio commentators can use to propose successful posts in Facebook™. We will analyze a corpus of posts in order to correlate linguistic and stylistic features to the success of the post.

Italiano. *Le radio sono in un punto di non ritorno e combattono una strenua battaglia con i nuovi mezzi di comunicazione di massa.*
In questo articolo vogliamo analizzare come la radio possa sfruttare a proprio vantaggio i social networks. Nel particolare, vogliamo cercare di individuare una strategia utile agli speakers radiofonici per proporre dei post di successo in piattaforme quali Facebook™. Dunque, analizzeremo stilisticamente e linguisticamente un corpus di post scritti dagli speakers di una radio per correlare queste caratteristiche con il successo del post stesso in termini di visualizzazioni e di like.

1 Introduzione

La radio è stata introdotta in Italia come mezzo di comunicazione di massa nel 1924 e è stata la padrona dell'etere italiano fino a quando nel 1954 la televisione ha fatto il suo primo vagito. In realtà, sin dalle origini, la radio ha sempre dovuto combattere con mezzi generati da una tecnologia in evoluzione.

Con la televisione, i cui abbonati sono da subito cresciuti molto velocemente (Fonti Istat; Ortoleva & Scaramucci, 2003) la radio ha trovato un accordo. Importanti innovazioni tecnologiche hanno diversificato la radio dalla televisione negli anni '50 e '60. L'FM permetteva una moltiplicazione, a costo relativamente basso, delle stazioni emittenti, consentendo il superamento almeno parziale del modello "generalista" proprio della TV in favore di un'offerta più ampia e varia di programmi mirati; il transistor permise alla Radio di conquistare spazi al di fuori dell'ambiente domestico, mentre la nascita delle autoradio permise di seguire gli ascoltatori anche nei loro spostamenti quotidiani. Parallelamente a queste nuove tecnologie si sviluppò attorno alla Radio una nuova cultura giovanile, animata dal ritmo travolgente del Rock'n roll (metà anni Cinquanta) e ammaliata dal fascino della riproduzione su disco (Monteleone, 2011). Nel giro di quindici-venti anni dalle prime affermazioni del mezzo televisivo, la Radio aveva diversificato la propria offerta in termini di contenuti e palinsesti. La Radio riempie gli spazi temporali lasciati liberi dalla tv (ore mattutine e buona parte di quelle pomeridiane). La fruizione personale, mobile, relativamente distratta, divenne sottofondo e accompagnamento alle altre attività quotidiane, finendo per delineare un tratto caratteristico del rapporto con il pubblico, che si fece più intimo e profondo (Menduni, 2003). Questa nuova dimensione dell'ascolto lasciò in breve tempo intuire la possibilità di sfruttare un'antica ma grande risorsa che, quasi paradossalmente, avrebbe conferito alla Radio il volto di un mezzo innovativo: si tratta del cavo telefonico, un nuovo canale attraverso cui minimizzare le distanze con il pubblico e dare avvio all'era dell'interattività. Il 7 gennaio 1969 andò in onda alle 10:40 la prima puntata della trasmissione Radiofonica "Chiamate Roma 31-31". E nacque l'interazione con il pubblico.

In questi anni, la Radio è costretta a combattere contro un nuovo nemico che potrebbe diventare un suo alleato: il Web nella sua versione nuova dei Social Networks. Così come negli anni 50 e 60 la Radio si è reinventata iniziando l'interazione con il pubblico, così in questi anni la Radio potrebbe sfruttare i Social Networks per reinventarsi.

In questo articolo vogliamo analizzare come la radio possa sfruttare a proprio vantaggio i social

networks. Nel particolare, vogliamo cercare di proporre una strategia agli speaker radiofonici per proporre dei post di successo in piattaforme quali Facebook™. Dunque, analizzeremo stilisticamente e linguisticamente un corpus di post di speakers di una radio per correlare queste caratteristiche con il successo del post stesso in termini di visualizzazioni e di like. Da questo, cercheremo di derivare alcune linee guida per la scrittura di post di successo.

Il resto dell'articolo è organizzato come segue: la sezione 2 descrive il metodo di analisi stilistica, linguistica e contenutistica dei post. La sezione 3 analizza i risultati su un insieme di post di speakers di una radio.

2 Definizione di un post di successo attraverso l'analisi strutturale e linguistica dei contenuti

Per comprendere meglio il meccanismo di ibridazione tra Radio e Social Media può risultare molto utile analizzare da vicino le modalità e i prodotti della loro interazione. Questa analisi parte dallo studio di un campione di contenuti digitali (*post*) generati da esperti della comunicazione radiofonica sulla piattaforma social più nota al mondo: Facebook. Nello specifico si tratta di 220 post, pubblicati dagli speakers dell'emittente radiofonica RDS sulla pagina Facebook "RDS 100% grandi successi!". Facebook – Social Media per eccellenza – offre ai suoi iscritti la possibilità di tenere sotto controllo il livello di interattività generato di volta in volta dai contenuti pubblicati, grazie a una serie di strumenti utili a rilevare e monitorare i movimenti degli *ospiti* sulla propria pagina.

2.1 Gli indicatori di successo di un post

Nel condurre questa analisi sono stati presi in considerazione gli *indicatori* di successo (*insight*) più noti ai frequentatori della piattaforma, di seguito elencati secondo il diverso grado di coinvolgimento che ciascuno di essi implica: **numero di visualizzazioni; numero di *mi piace;* *e* numero di commenti; numero di condivisioni.**
I dati numerici relativi a ciascun *insight* sono stati di seguito inseriti in una tabella, al fianco del testo del post cui si riferivano.

2.2 Variabili strutturali dei post

In una seconda fase l'obiettivo è stato quello di capire se fosse possibile individuare una correlazione tra le costanti numeriche individuate e alcune variabili strutturali caratteristiche di ciascun post, così raggruppate: lunghezza del testo; presenza di immagini, foto o video, fascia oraria di pubblicazione, tipologia.
Lunghezza del testo Per calcolare l'incidenza della variabile *lunghezza del testo* è stato preso in considerazione il numero di battute di cui si componeva ciascun post. Sono state così individuate quattro classi di valori rispetto alla variabile *n. battute.*
Presenza di immagini, foto o video In uno spazio che cambia velocemente, ad attrarre la nostra attenzione sono spesso alcuni dettagli che si rivelano più immediati di altri nel trasmetterci informazioni e sensazioni. È il caso delle immagini, delle foto e dei video.
Fascia oraria di pubblicazione Così come accade nella definizione dei palinsesti, anche sui Social Networks la scelta di pubblicare contenuti in determinate fasce orarie, piuttosto che in altre, può rilevarsi più o meno proficua.
Tipologia Se lo scopo del messaggio che si vuole veicolare è, come in questo caso, quello di suscitare una particolare reazione nel destinatario, è importante capire quali contenuti possono attivare un comportamento in linea con il nostro scopo e quali invece possono produrre passività, indifferenza, assuefazione e quindi, effetti disfunzionali inutili, o peggio controproducenti.

3 Analisi di correlazione tra indicatori di successo e variabili strutturali

Come ben noto Internet e in modo particolare i Social Networks hanno modificato le nostre abitudini di lettura, portandoci, più o meno consapevolmente, a prediligere testi brevi e coincisi, coadiuvati da immagini d'effetto, o ancor meglio da foto e video. L'attenzione alla testualità su una piattaforma dinamica quale è Internet, può tuttavia divenire secondaria se non viene opportunamente correlata ad un'attenta valutazione dell'utenza distribuita nelle varie fasce orarie, la cui mancata osservanza potrebbe decretare il confino dei contenuti nell'oblio della memoria virtuale. Di seguito verrà illustrato il procedimento adottato nell'analisi di ciascuna variabile.

3.1 Lunghezza del testo

Per la lunghezza del testo abbiamo individuato 4 classi in funzione del numero di battute: 0*-50, 50-100, 100-200, e da 200 in su dove con zero si indicano i post con soltanto immagini, video o foto. Alla prima classe appartengono 24 *post* (10,9%

del totale), alla seconda 48 (21,81%), alla terza 102 (46.36%) e alla quarta 46 (20.9%). In seguito, per ciascuna classe di *post* è stato calcolato il valore medio del numero di *visualizzazioni*, di *mi piace*, di *commenti* e di *condivisioni*.

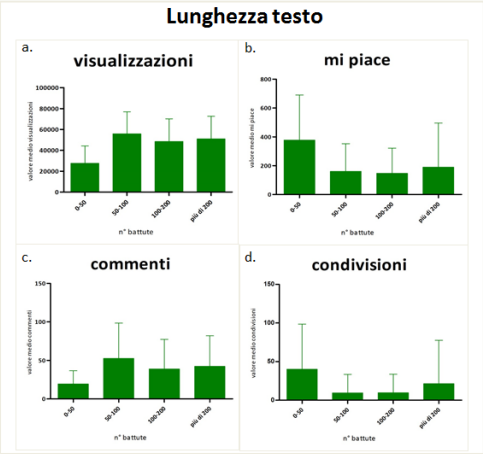


Figura 1. Confronto del valore medio di ciascun insight in ogni classe di valori riscontrata, relativamente al parametro n. battute

La Figura 1 mostra graficamente l'incidenza del parametro *lunghezza* del testo relativamente a ciascun tipo di feedback, sulla base del valore medio riscontrato per ogni classe di *post*. Il segmento verticale riportato all'apice di ogni barra indica la variazione standard all'interno di ciascuna distribuzione.

Osservazioni: Scrivere un *post* breve incide positivamente sul numero ***mi piace*** e di **condivisioni**. Non si riscontrano invece correlazioni significative tra la lunghezza del testo di un *post* e il numero di visualizzazioni e commenti.

3.2 Presenza di immagini, foto o video

Sul totale dei 220 *post* analizzati, 93 contengono immagini, foto o video (42,27% del totale).
La Figura 2 mostra la ripartizione dei *post* con (*with*) e senza (*without*) immagini, foto o video, all'interno di ciascuna classe di valori in cui sono stati precedentemente suddivisi i vari *hits*. Come si può notare, la presenza di immagini, foto o video non sembra influire sul numero di visualizzazioni e commenti, mentre incide in maniera positiva sul numero di *mi piace* e di condivisioni.

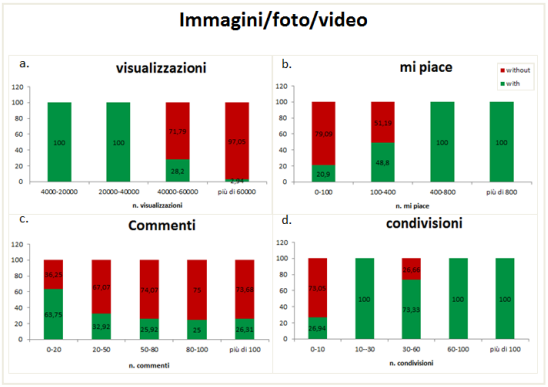


Figura 2. Confronto del valore medio di ciascun insight in ogni classe di valori riscontrata, relativamente al parametro immagini, foto o video

Osservazioni: Introdurre immagini, foto o video in un post fa aumentare il numero di ***mi piace*** e di **condivisioni**. La presenza della variabile non sembra incidere sul numero di visualizzazioni e commenti.

3.3 Fascia oraria di pubblicazione

Così come accade nella definizione dei palinsesti, anche sui Social Networks la scelta di pubblicare contenuti in determinate fasce orarie, piuttosto che in altre, può rilevarsi più o meno proficua. Sulla base dei dati raccolti, relativi all'orario di pubblicazione di ciascun post, individuiamo otto fasce orarie di pubblicazione (divise in gruppi di tre ore): 09:00-12:00; 12:00-15:00; 15:00-18:00; 18:00-21:00; 21:00-00:00; 00:00-03:00; 03:00-06:00; 06:00-09:00. Nella prima fascia oraria rientrano 45 *post*, nella seconda 42, nella terza 41, nella quarta 19, nella quinta 41, nella sesta 18, nella settima 12, nell'ottava 2 (quest'ultima è stata tralasciata in fase di analisi).
La Figura 3 mostra l'incidenza della variabile *fascia oraria* agisce relativamente a ciascun tipo di feedback, sulla base del valore medio riscontrato per ogni classe di *post.*
Osservazioni. La percentuale di *post* pubblicati è particolarmente nella fascia oraria **09:00-00:00** (fatta eccezione per la fascia oraria 18-21), mentre si abbassa notevolmente tra le 03:00 e le 09:00. Tenendo conto di questo dato e del valore della varianza molto alto in quasi tutti i casi, possiamo concludere che: (1) La scelta di pubblicare in diverse fasce orarie non incide (in questo specifico caso) in maniera significativa né sulle condivisioni, né sui *mi piace*, e tantomeno sui commenti; (2) la fascia oraria con più visualizzazioni sembrano essere quelle comprese tra le **09:00-15:00** e tra le **21:00-06:00**.

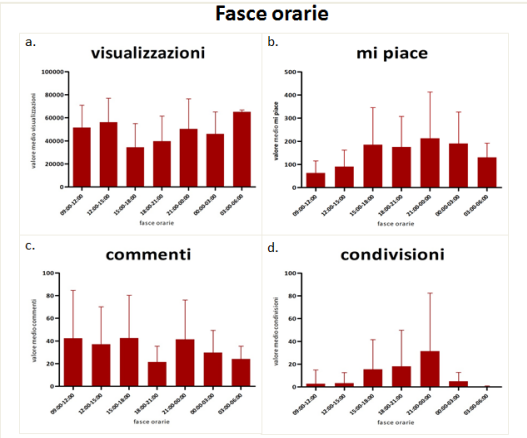


Figura 3. Confronto del valore medio di ciascun insight in ogni classe di valori riscontrata, relativamente al parametro fascia oraria

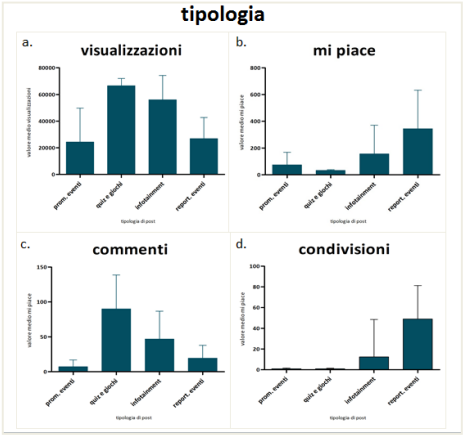


Figura 4. Confronto del valore medio di ciascun insight in ogni classe di valori riscontrata, relativamente al parametro tipologia

3.4 Tipologia

Tra i 220 post analizzati, sono state riscontrate quattro diverse tipologie a seconda del contenuto trattato: (1) Post di autopromozione o promozione di eventi; (2) Quiz e giochi; (3) Post d'intrattenimento o infotainment; (4)Reportage e descrizioni di eventi musicali

Alla prima tipologia appartengono 13 *post* (pari al 5,9% del totale) tra quelli analizzati; alla seconda 7 (il 3,18%); alla terza 155 (il 70,45%) e alla quarta 45 (il 20,45%). Anche in questo caso il valore medio del numero di visualizzazioni, mi piace, commenti e condivisioni per ciascuna classe di valori riscontrata, relativamente al parametro tipologia, dopo aver raggruppato in classi i *post*, si è proceduto calcolando per ciascuna classe il valore medio del numero di visualizzazioni (Mv), di *mi piace* (Mp), di commenti (Mcm) e di condivisioni (Mcd).

La Figura 4 mostra il valore medio di ciascun *hits* (e la corrispettiva variazione standard) rispetto alla variabile *tipologia* del *post*.

Osservazioni: I *post* che hanno come contenuto *quiz e giochi* e *intrattenimento e infotainment* fanno aumentare il numero di **visualizzazioni** e **commenti.** I *post* che contengono *reportage fotografici o descrizioni di eventi* producono (anche se con una variabilità abbastanza alta) un maggior numero di **condivisioni**.

4 Studi correlati

Questo studio preliminare è una base di partenza per reinventare il modello radiofonico e mostra come la radio possa sfruttare i recenti studi sulla diffusione virale dei post e delle informazioni come la viralità ed emozioni evocate da un contenuto (Berger, 2012), la viralità come fenomeno complesso descrivibile tramite molteplici indici (Guerini M. C., 2011), la viralità di citazioni da film (Danescu-Niculescu-Mizil, 2012), la viralità su twitter di contenuti linguistici (Tan, 2014), timing del post e rete sociale (Artzi, 2012), (Hong, 2011), la viralità e stile linguistico (Guerini M. A., 2012), la iralità dipendente da interazione testo e immagini (Khosla, Sarma, & Hamid, 2014) (Guerini M. J., 2013).

5 Conclusioni e sviluppi futuri

I risultati e le osservazioni ricavate da questo studio iniziale tendono in parte a confermare l'effettiva incidenza di alcuni parametri sulla riuscita di un post. In diversi casi però, i dati ricavati delineano scenari nuovi e inaspettati. Sono proprio risultati come questi a condurci verso una riflessione sulle sostanziali differenze tra la dimensione *on line* e quella *on air* della radio, come la mancanza (nel primo caso) di palinsesti e vincoli legati agli orari delle programmazioni e quindi alla routine degli appuntamenti quotidiani, nonché il particolare meccanismo a flusso, generato in primis dalla casualità e dall'imprevisto.

Lo studio presentato può essere un base per costruire un sistema predittivo in grado di prevedere se un post può avere successo come quelli usati per prevedere i rating dei film (Pang, Lee, & Vaithyanathan, 2002).

Bibliografia

Agosti, A. a. 2007. *Making readability indices readable.* LIWC. net, Austin, TX.

Artzi, Y. P. 2012. Predicting responses to microblog posts. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics.

Berger, J. a. 2012. What makes online content viral? *Journal of Marketing Research*, 192-205.

Danescu-Niculescu-Mizil, C. e. 2012. You had me at hello: How phrasing affects memorability. *50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.* Association for Computational Linguistics.

Guerini, M. A. 2012. *Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality?* ICWSM.

Guerini, M. C. 2011. *Exploring Text Virality in Social Networks.* ICWSM.

Guerini, M. J. 2013. Exploring image virality in google plus. *Social Computing (SocialCom), 2013 International Conference on. IEEE.*

Hong, L. O. 2011. Predicting popular messages in twitter. *20th international conference companion on World Wide Web.* ACM.

Istat. (2014). *Fruizione dei mass-media (giornali, tv, radio).* Retrieved Gennaio 18, 2014, from http://dati.istat.it

Khosla, A., Sarma, A. D., & Hamid, R. 2014. What makes an image popular? *23rd international conference on World wide web. International World Wide Web Conferences Steering Committee.*

Menduni, E. 2003. I pubblici della Radio. In M. Livolsi, *Il pubblico dei media. La ricerca nell'industria culturale* (pp. 151-166). Roma: Carocci.

Montefinese, M. e. 2013. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior research methods*, (pp. 1-17).

Monteleone, F. 2011. *Storia della Radio e della Televisione in Italia. Un secolo di costume, società e politica.* Marsilio.

Ortoleva, P., & Scaramucci, B. 2003. *Enciclopedia della Radio.* Milano: Garzanti.

Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing.*

Tan, C. L. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *52th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.* Association for Computational Linguistics.

Tonelli, S. K. 2012. Making readability indices readable. *First Workshop on Predicting and Improving Text Readability for target reader populations.* Association for Computational Linguistics.

10.12871/CLICIT2014147

Estimating Lexical Resources Impact in Text-to-Text Inference Tasks

Simone Magnolini
University of Brescia
FBK, Trento, Italy
magnolini@fbk.eu

Bernardo Magnini
FBK, Trento, Italy
magnini@fbk.eu

Abstract

English. This paper provides an empirical analysis of both the datasets and the lexical resources that are commonly used in text-to-text inference tasks (e.g. textual entailment, semantic similarity). According to the analysis, we define an index for the impact of a lexical resource, and we show that such index significantly correlates with the performance of a textual entailment system.

Italiano. *Questo articolo fornisce un'analisi empirica dei datasets e delle risorse lessicali comunemente usate per compiti di inferenza testo-a-testo (es., implicazione testuale, similaritá semantica). Come risultato definiamo un indice che misura l'impatto di una risorsa lessicale, e mostriamo che questo indice correla significativamente con le prestazioni di un sistema di implicazione testuale.*

1 Introduction

In the last decade text-to-text semantic inference has been a relevant topic in Computational Linguistics. Driven by the assumption that language understanding crucially depends on the ability to recognize semantic relations among portions of text, several text-to-text inference tasks have been proposed, including recognizing paraphrasing (Dolan and Brockett., 2005), recognizing textual entailment (RTE) (Dagan et al., 2005), and semantic similarity (Agirre et al., 2012). A common characteristic of such tasks is that the input are two portions of text, let's call them $Text1$ and $Text2$, and the output is a semantic relation between the two texts, possibly with a degree of confidence of the system. For instance, given the following text fragments:

Text1: *George Clooneys longest relationship ever might have been with a pig. The actor owned Max, a 300-pound pig.*
Text2: *Max is an animal.*

a system should be able to recognize that there is an "entailment" relation among $Text1$ and $Text2$.

While the task is very complex, requiring in principle to consider syntax, semantics and also pragmatics, current systems adopt rather simplified techniques, based on available linguistic resources. For instance, many RTE systems (Dagan et al., 2012) would attempt to take advantage of the fact that, according to WordNet, the word *animal* in $Text2$ is a hypernym of the word *pig* in $Text1$. A relevant aspect in text-to-text tasks is that datasets are usually composed of textual pairs for positive cases, where a certain relation occurs, and negative pairs, where a semantic relation doesn't appear. For instance, the following pair:

Text1: *John has a cat, named Felix, in his farm, it's a Maine Coon, it's the largest domesticated breed of cat.*
Text2: *Felix is the largest domesticated animal in John's farm.*

shows a case of "non-entailment".

In the paper we systematically investigate the relations between the distribution of lexical associations in textual entailment datasets and the system performance. As a result we define a "resource impact index" for a certain lexical resource with respect to a certain dataset, which indicates the capacity of the resource to discriminate between positive and negative pairs. We show that the "resource impact index" is homogeneous across several datasets and tasks, and that it correlates with the performance of the algorithm we chose in our

experiments.

2 Lexical resources and Text-to-Text Inferences

The role of lexical resources for recognizing text-to-text semantic relations (e.g. paraphrasing, textual entailment, textual similarity) has been under discussion since several years. This discussion is well reflected in the data reported by the RTE-5 "ablation tests" (Bentivogli et al., 2009), where the performance of a certain algorithm was measured removing one resource at time.

Challenge	T1/T2 Overlap (%)		
	YES	NO ENTAILMENT	
		Unknown	Contradiction
RTE - 1	68.64	64.12	
RTE - 2	70.63	63.32	
RTE - 3	69.62	55.54	
RTE - 4	68.95	57.36	67.97
RTE - 5	77.14	62.28	78.93

Table 1: Comparison among the structure of different RTE data-set (Bentivogli et al., 2009).

As an example, participants at the RTE evaluation reported that WordNet was useful (i.e. improved performance) 9 of the times, while 7 of the time it wasn't useful. As an initial explanation for such controversial behavior, Table 1, again extracted from (Bentivogli et al., 2009), suggests that the degree of word overlap among positive and negative pairs might be a key to understand the complexity of a text-to-text inference task, and, as a consequence, a key to interpret the system's performance. In this paper we extend this intuition, considering: (i) lexical associations (e.g. synonyms) other than word overlap, and (ii) datasets with different characteristics.

There are several factors which in principle can affect our experiments, and that we have carefully considered.

Resource. First, the impact of a resource depends on the quality of the resource itself. Lexical resources, particularly those that are automatically acquired, might include noisy data, which negatively affect performance. In addition, resources such as WordNet (Fellbaum, 1998) are particularly complex (i.e. dozen of different relations, deep taxonomic structure, fine grained sense distinctions) and their use needs tuning. We have selected lexical resources manually constructed, with a high degree of precision, and in the experiments we have used lexical relations separately, in order to keep under control their effect.

Inference Algorithm. Second, different algorithms may use different strategies to take advantage of resources. For instance, algorithms that calculate a distance or a similarity between $Text1$ and $Text2$ may assign different weights to a certain word association, on the basis on human intuitions (e.g. synonyms preserve entailment more than hypernyms). In our experiments we avoided as much as possible the use of settings not grounded on empirical evidences.

Dataset. Finally, datasets representing different inference phenomena, may manifest different behaviors with respect to the impact of a certain resource, specific for each inference type (e.g. entailment and semantic similarity). Although reaching a high level of generalization is limited by the existence itself of datasets, we have conducted experiments both on textual entailment and semantic similarity.

3 Resource Impact Index

In this Section we define the general model through which we estimate the impact of a lexical resource. The idea behind the model is quite simple: the impact of a resource on a dataset should be correlated to the capacity of the resource to discriminate positive pairs from negative pairs in the dataset. We measure this capacity in term of the number of *lexical alignments* that the resource can establish on positive and negative pairs, and then we calculate the difference among them (we call this measure the *resource impact differential - RID*). The smaller the RID, the smaller the impact of the resource on that dataset. In the following we provide a more precise definition of the model.

Dataset (D). A dataset is a set of text pairs $D = \{(T1, T2)\}$, with positive $(T1, T2)^p$ and negative $(T1, T2)^n$ pairs for a certain semantic relation (e.g. entailment, similarity).

Lexical Alignment (LexAl). We say that two tokens in a $(T1, T2)$ pair are aligned when there's some semantic association relation, including equality, between the two tokens. For instance, synonyms and morphological derivations are different types of lexical alignments.

Lexical Resource (LR). A Lexical Resource is a potential source of alignment among words. For instance, WordNet is a source for synonyms [1].

Resource Impact (RI). The impact of a resource LR on a data-set D is calculated as the number of lexical alignments returned by LR, normalized on the number of potential alignments for the data-set D. We use $|T1| * |T2|$ as potential alignments (Dagan et al., 2012, page 52), although there might be other options: $|T1| + |T2|$, $max(|T1|, |T2|)$, etc. RI ranges from 0, when no alignment is found, to 1, when all potential alignments are returned by LR.

$$RI_{(LR,D)} = \#LexAl/|T1| * |T2| \quad (1)$$

Resource Impact Differential (RID). The impact of a resource LR on a certain dataset D is given by the difference between the RI on positive pairs $(T1, T2)^p$ and on negative pairs $(T1, T2)^n$. A RID ranges from -1, when the RI is 0 for the entailed pairs and 1 for not entailed pairs, to 1, when the RI is 1 for entailed and 0 for not entailed pairs.

$$RID_{(LR,D)} = RI(T1, T2)^p - RI(T1, T2)^n \quad (2)$$

The RID measure isn't affected by the size of the dataset, because it's normalized on the maximum number of alignments. Finally, the coverage of the resource (i.e. the number of lexical alignments) is an upper of the bound of the RID (see 3), being the RID a difference.

$$\left|RID_{(LR,D)}\right| \leq \frac{\#LexAl}{|T1| \cdot |T2|} \quad (3)$$

4 Experiments

In this section we apply the model described in Section 3 to different datasets and resources, showing that the RID is highly correlated to the accuracy of a text-to-text inference algorithm.

Datasets. We use four different datasets in order to experiment different characteristics of text-to-text inferences. The RTE-3 dataset (Giampiccolo et al., 2007) for English has been used in the context of the Recognizing Textual Entailment shared tasks, it has been constructed mainly using application derived text fragments and it's balanced between positive and negative pairs (about 1600 in total). The Italian RTE-3 dataset is the translation of the English one. The RTE-5 dataset is similar to RTE-3, although Text-1 in pairs are usually much longer, which, in our terms, means that a higher number of alignments can be potentially generated by the same number of pairs. Finally the SICK dataset (Sentences Involving Compositional Knowldedge) (Marelli et al., 2014) has been recently used to highlight distributional properties, it isn't balanced (1299 positive and 3201 negative pairs), and $T1$ and $T2$, differently from RTE pairs, have similar length.

Sources for lexical alignments. We carried on experiments using four different sources of lexical alignments, whose use is quite diffused in the practice of text-to-test inference systems. The first source consists of a simple match among the lemmas in $T1$ and $T2$: if two lemmas are equal (case insensitive), then we count it as an alignment between $T1$ and $T2$. The second resource considers alignments due to the synonymy relation (e.g. *home* and *habitation*). The source is WordNet (Fellbaum, 1998), version 3.0 for English, and MultiWordNet (Pianta et al., 2002) for Italian. The third resource considers the hypernym relation (e.g. *dog* and *mammal*): as for synonymy we use WordNet. The last source of alignment are morphological derivations (e.g. *invention* and *invent*). For English derivations are covered again by WordNet, while for Italian we used MorphoDerivIT, a resource developed at FBK which has the same structure of CATVAR (Habash and Dorr, 2003) for English. Finally, in order to investigate the behavior of the RID in absence of any lexical alignment, we include a 0-Knowledge experimental baseline, where the system does not have access to any source of lexical alignment.

Algorithm. In order to verify our hypothesis that the RID index is correlated with the capacity of a system to correctly recognize textual entailment, we run all the experiments using EDITS (Negri et al., 2009) RTE based on calculating the Edit Distance between $T1$ and $T2$ in a pair. The algorithm calculate the minimum-weight series of edit operations (deletion, insertion and substitution) that transforms $T1$ into $T2$.The algorithm has an optimizer that decides the best cost for every edit op-

[1] In the paper we consider lexical resources that are supposed to provide similarity/compatibility alignments (e.g. synonyms). However, there might be resources (e.g. antonyms in WordNet) that are supposed to provide dissimilarity/opposition alignments. We'll investigate negative alignments in future work.

	RTE-3 eng		RTE-3 ita		RTE-5 eng		SICK eng	
	RID	Accuracy	RID	Accuracy	RID	Accuracy	RID	F1
0-Knowledge	0	0.537	0	0.543	0	0.533	0	0.005
Lemmas	87.164	0.617	84.594	0.641	36.169	0.6	523.342	0.347
Synonyms	-6.432	0.533	5.343	0.537	1.383	0.546	12.386	0.093
Hypernyms	-0.017	0.545	-1.790	0.543	7.969	0.556	48.665	0.221
Derivations	0.154	0.543	-0.024	0.536	2.830	0.545	-6.436	0
R correlation	0.996		0.991		0.985		0.851	

Table 2: Experimental results obtained on different datasets with different resources.

erations. The algorithm is normalized on the number of words of $T1$ and $T2$ after stop words are removed. As for linguistic processing, the Edit Distance algorithm needs tokenization, lemmatization and Part-of-Speech tagging (in order to access resources). We used TreeTagger (Schmid, 1995) for English and TextPro (Emanuele Pianta and Zanoli, 2008) for Italian. In addition we removed stop words, including some of the very common verbs. Finally, all the experiments have been conducted using the EXCITEMENT Open Platform (EOP) (Padó et al., 2014) (Magnini et al., 2014), a rich and modular open source software environment for textual inferences [2].

5 Results and Discussion

Table 2 reports the results of the experiments on the four datasets and the five sources of alignment (including the 0-Knowledge baseline) described in Section 4. For each resource we show the RID of the resource (given the very low values, the RID is shown multiplied by a 10^4 factor), and the accuracy achieved by the EDITS algorithm. The last row of the Table shows the Pearson correlation between the RID and the accuracy of the algorithm for each dataset, calculated as the mean of the correlations obtained for each resource on that dataset.

A first observation is that all RID values are very close to 0, indicating a low expected impact of the resources. Even the highest RID (i.e. 523.342 for lemmas on SICK), corresponds to a 5% of the potential impact of the resource. Negative RID values mean that the resource, somehow contrary to the expectation, produces more alignments for negative pairs than for positive (this is the case, for instance of synonyms on the English RTE-3). Alignment on lemmas is by far the resource with the best impact.

[2] http://hltfbk.github.io/Excitement-Open-Platform/

Finally, results fully confirm the initial hypothesis that the RID is correlated with the system performance; i.e. the accuracy for balanced datasets and the F1 for the unbalanced one. The Pearson correlation shows that R is close to 1 for all the RTE datasets (the slightly lower value on SICK reveals the different characteristics of the dataset), indicating that the RID is a very good predictor of the system performance, at least for the class of inference algorithms represented by EDITS. The low values for RID are also reflected in absolute low performance, showing again that when the system uses a low impact resource the accuracy is close to the baseline (i.e. the 0-Knowledge configuration).

6 Conclusion

We have proposed a method for estimating the impact of a lexical resource on the performance of a text-to-text semantic inference system. The starting point has been the definition of the RID index, which captures the intuition that in current datasets useful resources need to discriminate between positive and negative pairs. We have then shown that the RID index is highly correlated with the accuracy of the system for balanced datasets and with the F1 for the unbalanced one, a result that allows to use the RID as a reliable indicator of the impact of a resource.

As for future work, we intend to further generalize our current findings applying the same methodology to different text-to-text inference algorithms, starting from those already available in the EXCITEMENT Open Platform. We also want to conduct experiment on operation, like summing, with this index to describe to combined effect of different resources.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 385–393, Montréal, Canada.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognising textual entailment challenge. In *Proceedings of the TAC Workshop on Textual Entailment*, Gaithersburg, MD.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 177–190, Southampton, UK.

Ido Dagan, Dan Roth, and Fabio Massimo Zanzotto. 2012. *Recognizing Textual Entailment: Models and Applications*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005), Asia Federation of Natural Language Processing*.

Christian Girardi Emanuele Pianta and Roberto Zanoli. 2008. The textpro tool suite. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias 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/.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for english. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 17–23. Association for Computational Linguistics.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics, Demo papers*.

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 Nicoletta Calzolari (Conference Chair), 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'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Matteo Negri, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. 2009. Towards extensible textual entailment engines: the EDITS package. In *Proceeding of the Conference of the Italian Association for Artificial Intelligence*, pages 314–323, Reggio Emilia, Italy.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*. doi: 10.1017/S1351324913000351.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Developing an aligned multilingual database. In *Proc. 1st Intl Conference on Global WordNet*.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

10.12871/CLICIT2014148

Parting ways with the partitive view: a corpus-based account of the Italian particle "ne"

Alice Mariotti
University of Bologna
alice.mariotti11@gmail.com

Malvina Nissim
U. of Bologna / U. of Groningen
m.nissim@rug.nl

Abstract

English. The Italian clitic "ne" is traditionally described as a partitive particle. Through an annotation exercise leading to the creation of a 500 instance dataset, we show that the partitive feature isn't dominant, and the anaphoric properties of "ne", syntactically and semantically, are what we should focus on for a comprehensive picture of this particle, also in view of its computational treatment.

Italiano. *Il clitico "ne" è noto come 'ne partitivo'. Attraverso un esercizio di annotazione che ha condotto alla creazione di un corpus di 500 esempi, osserviamo che il tratto partitivo non è affatto dominante nell'uso del "ne", e che per avere un quadro completo di questo clitico è necessario concentrarsi sulle sue caratteristiche anaforiche, sia a livello sintattico che, specialmente, semantico, anche per lo sviluppo di sistemi di risoluzione automatica.*

1 Introduction and Background

The Italian particle "ne" is a clitic pronoun. Traditionally, linguistic accounts of "ne" focus on two of its aspects: its syntactic behaviour and its being a conveyor of partitive relations.

Syntactically, this particle has been studied extensively, especially in connection with unaccusative verbs (Belletti and Rizzi, 1981; Burzio, 1986; Sorace, 2000). In Russi's volume specifically dedicated to clitics, the chapter devoted to "ne" only focuses on the grammaticalisation process which brought the clitic to be incorporated in some verbs, causing it to lose its pronominal properties. It is referred to as "the 'partitive' ne" (Russi, 2008, p. 9).

In (Cordin, 2001), the clitic is described in detail, and shown to serve three main uses. It can be a partitive pronoun, usually followed by a quantifier, as in (1). It can be used purely anaphorically to refer to a previously introduced entity, such as "medicine" in (2). The third use is as a locative adverb, like in (3).[1]

(1) Quanti giocatori di quell'U17-U19 quest'anno o l'anno scorso hanno giocato minuti importanti in prima squadra? A me **ne** risultano *2 o 3*.
How many players of that U17-U19 [team] this year or last year have played important minutes in the first team? I think *2 or 3* [**of them**]

(2) Tu sai che la medicina fa bene e pretendi che il palato, pur sentendo**ne** *l'amaro*, continui a gustarla come se fosse dolce.
You know that the medicine is good for you, and you ask your palate to enjoy it as if it was sweet, in spite of tasting [**its**] *bitterness*.

(3) Me **ne** vado.
I'm leaving.

Note that for both partitive and non-partitive uses, in order to interpret the *ne*, the antecedent must be identified ("players of that U17-U19 [team]" in (1) and "medicine" for (2)). While there has been a recent effort to highlight the anaphoric properties of real occurrences of "ne" (Nissim and Perboni, 2008), there isn't as yet a comprehensive picture of this particle. In this paper, we contribute a series of annotation schemes that capture the anaphoric nature of "ne", and account for the different kinds of relations it establishes with its antecedent. We also contribute an annotated dataset that can be used for training automatic resolution systems, and that as of now provides us with a picture of this particle which is the most comprehensive to date.

2 Annotation schemes

Considering the examples above, we can see that the resolution of "ne" can be summarised as obeying the scheme in (4), where capturing the function of *ne* ($f(ne)$) is part of the resolution process. Figure 1 shows an example and its resolution.

(4) $predicate + [f(ne) + antecedent]$

[1] Unless otherwise specified, all examples are from "Paisà" (Lyding et al., 2014), a corpus of about 250 mio tokens of Italian web data. The *ne* is bold-faced, the antecedent is underlined, and the predicate is in italics. Note that *ne* is often used as an enclitic, such as in (2). This can be the case with any of the three uses described.

Example: "Sto pensando a un modo per staccare la spina, **ne** *ho veramente bisogno*."
'I'm thinking of a way to take I break, I really need to.'
$f(ne)$: "di + x" 'of + x'
antecedent: "staccare la spina" 'take a break' (lit. 'unplug')
predicate: "ho veramente bisogno" 'I really need to'

resolution: "ho veramente bisogno di staccare la spina" 'I really need to take a break'

Figure 1: Example of "ne" resolution components and procedure.

Table 1: Annotation scheme for the element "ne". The most used classes are highlighted in boldface.

ne	anaphoric	partitive	**type token**
		¬ partitive	**internal external prepobj**
		cataphoric vague	
	¬ anaphoric		

In order to account for all the entities involved in 4, we developed a set of annotation schemes that define three elements (ne, antecedent, and predicate) and their respective attributes. The schemes mainly build on on our own analysis of a random selection of corpus occurrences and on the only existing corpus-based work on "ne" (Nissim and Perboni, 2008), over which we introduce three substantial improvements:

(i) we distinguish between *type* and *token* for partitive uses to account for the difference between Example (7) and Example (8) below.

(ii) we add the values *internal*, *external*, and *prepobj* for non-partitive cases (see Section 2.1).

(iii) we mark explicit links between anaphor and antecedent, and anaphor and predicate.

Both (i) and (ii) are quite crucial conceptual distinctions, as we will see both in the scheme description as well as in the analysis of annotated data, while (iii) is important in the implementation of an automated resolution process. Additionally, the annotation is performed by means of a different, more appropriate annotation tool.

2.1 Scheme for *ne*

The scheme is summarised in Table 1. The primary branch for "ne" is its anaphoricity, which is a binary feature. All cases of non-anaphoricity are basically idiomatic uses, especially with pronominal verbs (see also Example 3). These cases won't be further specified in the annotation.

Differently, anaphoric occurrences are classified as one of four types: *partitive*, *non-partitive*, *cataphoric*, *vague*. The rare cataphoric cases (6) are annotated as such without additional features. The value vague is used when the instance is anaphoric but with an unclear or unspecified antecedent (5), with no further annotation.

(5) L'aggettivo puoi anche metterlo dopo il sostantivo, a questo modo potresti continuare: "l'alba rugiadosa **ne** *trae prestigio ed eleganza*".
You can even place the adjective after the noun, and so you could continue: "the dewy dawn *gains prestige and elegance* **from it**

(6) *Ce* **ne** *fossero ancora molti* di preti nello scautismo ...
I wish there still were many priests in the scouting movement ...

The main distinction is thus between partitive and non-partitive uses. Both values are then further detailed, and the resulting five categories – boldfaced in Table 1 – are the core of the scheme. Below we explain the opposition between *type* and *token* references for partitive cases, and the difference between *internal*, *external*, and *prepobj* for non-partitive cases.

partitive Consider Examples (7)-(8).

(7) **type** – Ho comprato quindici paste, e **ne** avrei prese *ancora*!
I bought fifteen *pastries*, and I would have bought even *more*!

(8) **token** – Ho comprato quindici paste, **ne** ho mangiate cinque.
I bought *fifteen pastries*, and I ate five [**of those**].

While in (8) the antecedent of "ne" is the whole NP "quindici paste" (fifteen pastries) and the predicate thus selects a subset ("cinque", five) of those

fifteen, the antecedent of "ne" in (7) is not a specific set of pastries, rather the class "pastries". Indeed, the predication is not about a portion or set of the aforementioned"fifteen pastries", but rather on instances of "pastries" in general. The *type* cases are akin to nominal ellipsis (Lobeck, 2006), and the type/token distinction is reflected in there being a direct or indirect anaphoric link.

The above are examples we made up for the sake of clarity: the contrast is explicit thanks to the occurrence of the same head noun. The opposition can anyway be observed in actual extracted data, too, and we report two cases in (9) and (10).

(9) **type** – non solo non risolve [...] i problemi che l'hanno scatenata, ma li aggrava e **ne** *crea di nuovi ancora più gravi*.

(10) **token** – È uno spettacolo grandioso costato 150 milioni di dollari; e probabilmente **ne** *incasserà il triplo o il quadruplo*.

non partitive We introduce three new distinctive values which capture both semantic and syntactic aspects. First, syntactically, we specify whether the function of "ne" is resolved within the predicate's argument structure, and it is thus annotated as a "prepositional object" (*prepobj*), as the example in Figure 1. Second, semantically, we distinguish between what we call "internal" and "external" references. "Internal" is a reference made – through the predicate of "ne" – to a feature which is already part of the antecedent, as in (12). Another case of "internal" is (2), where the bitter taste is an internal feature of the medicine. With "external" we mark references which introduce some feature – again, via the predicate – which is external to the whole represented by the antecedent, as in (11). This distinction, neglected in the literature, has semantic implications on the part-whole relation which gets established, or even created, between anaphor and antecedent.

(11) **external** – [...] il possesso dell'oggetto del nostro desiderio, posticipando**ne** *la soddisfazione*. [...] the possession of the object of our desire, procrastinating its satisfaction.

(12) **internal**[2] – [...] il possesso dell'oggetto del nostro desiderio, posticipando**ne** *l'intensità*. [...] the possession of the object of our desire, procrastinating its intensity.

[2]This example is made up on the basis of the *external* one above for easing the comparison.

2.2 Scheme for *antecedent*

The antecedent is what resolves the anaphoric interpretation of "ne". In the examples concerning the distinction between type and token, the antecedent is "paste" ('parties') in (7) and "quindici paste" ('fifteen pastries') in (8). While in both examples the antecedent is an NP, it is possible for a VP (as in Figure 1) and even a full sentence (S) to serve as antecedent. The annotators are asked to mark as *antecedent* the whole linguistic expression which they identify as the antecedent, and to assign some syntactic features to it. The annotation scheme is summarised in Table 2.

Table 2: Annotation scheme for antecedent

antecedent	NP	subject	modified ¬ modified
		object	modified ¬ modified
		other	
	VP S		

For each antecedent we thus specify its syntactic category, its grammatical role distinguishing just between *subject*, *object*, and any other role (*other*). For antecedents featuring as subject or object we also specify whether they have any sort of modification (adjectives, relative clause, and so on).

2.3 Scheme for *predicate*

The predicate of "ne" is what provides the completion to the interpretation of the anaphoric relation. In terms of annotation we specify only whether the predicate is a noun phrase or a verb phrase, and in the former case whether it is a modified NP or not.

3 Data selection and annotation

We collected 500 random occurrences from PAISÀ (Lyding et al., 2014), a web corpus of Italian which contains however good quality data. Instances of "ne" were extracted in a context of two preceding and one following sentence with respect to the matching sentence. In a given paragraph, then, possibly more than one occurrence of "ne" was included, but only one was occurrence per paragraph was highlighted to be annotated.

To perform the annotation we customised MMAX2, an XML-based annotation tool specifically devised to mark up coreference links (Müller

and Strube, 2006). We introduced two different links to connect each instance of "ne" to its predicate and to its antecedent. To ease the annotation process, these are visualised with different colours. A screenshot is given in Appendix A. We implemented the annotation categories described above, i.e. "ne", "antecedent", and "predicate", and made available, for each of them, all relevant attributes. MMAX2 lets developers create attributes in dependence of certain values assigned to other attributes so that, for instance, the attribute "class", whose values are "type" or "token", is activated only if the instance is annotated as "partitive". MMAX2 also lets annotate discontinuous material as part of the same entity, which came useful when annotating predicates and antecedents. The output is standoff XML.

The authors of this paper independently annotated the data, achieving a score of $K = .78$ on the classification of "ne". This is considerably lower than the agreement reported in (Nissim and Perboni, 2008), but the classification categories in our scheme are higher in number, and finer-grained.

4 Corpus Analysis

4.1 Distribution

Out of the 500 extracted instances, only two were not annotated because they were not actual occurrences of "ne". An overview of the distribution of annotated categories is given in Table 3.

Table 3: Distribution of categories in the dataset

anaphoric	partitive	token	19
		type	52
		total partitive	71
	¬partitive	external	107
		internal	33
		prepobj	127
		total ¬partitive	272
	cataphoric/vague		45
	total anaphoric		388
non-anaphoric			110
invalid			2
total			500

As we can see, only about one-fifth of the cases are non-anaphoric. Also, among anaphoric instances, we can observe that most commonly, "ne" is used in a purely anaphoric, *non-partitive* way, suggesting a behaviour different than the one described in the theoretical literature.

4.2 Anaphoric aspects

In terms of specific anaphoric relations, we observed both direct and indirect links. The observed combinations are summarised in Appendix B.

In coreference, we observe that the relationship between "ne" and its antecedent can be of all kinds apart from purely partitive, as it isn't a specific object that "ne" refers to, rather the class that the antecedent introduces (such as "pastries" in Example 7). Syntactically, the antecedent occurs as a direct object only in the partitive_token case, otherwise is always an indirect complement, usually introduced by "di" ('of') or "da" ('from').

Cases of indirect anaphora that we observed are of two sorts: (i) the reference to classes rather than objects, which is found with partitive *types*, and is syntactically akin to nominal ellipsis, as in (7); and (ii) instances of bridging (Clark, 1975; Kleiber, 1999). Within (i) we observe also cases of *other-anaphora* (Modjeska, 2002), such as (13) below.

(13) Possiamo tenere soltanto un versetto che ci accompagna però durante tutta la giornata e lo memorizziamo, lo ruminiamo, e domani **ne** prendiamo *un altro* [...]

We can only keep a verse with us during the day, and we memorise it, we think it over, and tomorrow we will take *another one*

Bridging is observed with cases of non-partitive external and especially internal features, as indeed bridging anaphors usually convey a (widely speaking) meronymic relation to their antecedent (e.g."intensità, 'intensity', in (12)). For the sake of space we cannot go into the details of the meronymic relations observed, but they have been classified according to (Cruse, 1985).

5 Conclusion

Actual corpus data, annotated thanks to the development of specific annotation schemes focused on the anaphoric potential of "ne", shows that the function of "ne" cannot be at all limited to a 'partitive' pronoun or as a test for certain syntactic types, as it is usually done in the theoretical literature. It also highlights several aspect of the anaphoric properties of "ne", both semantically and syntactically. We plan to exploit the dataset to develop an automatic resolution system.

References

A. Belletti and L. Rizzi. 1981. The syntax of ne: some implications. *The Linguistic Review*, 1:117–154.

L. Burzio. 1986. *Italian Syntax: A Government-Binding Approach*. Reidel, Dordrecht.

Herbert H. Clark. 1975. Bridging. In Roger Schank and Bonnie Nash-Webber, editors, *Theoretical Issues in Natural Language Processing*. The MIT Press, Cambridge, MA.

Patrizia Cordin. 2001. Il clitico "ne". In Lorenzo Renzi, Giampaolo Salvi, and Anna Cardinaletti, editors, *Grande grammatica italiana di consultazione dell'Italiano*, vol. I. Il Mulino, Bologna.

Alan D. Cruse. 1985. *Lexical Semantics*. Cambridge University Press, Cambridge.

G. Kleiber. 1999. Associative anaphora and part-whole relationship: the condition of alienation and the principle of ontological congruence. *Journal of Pragmatics*, pages 339–362.

Anne Lobeck. 2006. Ellipsis in DP. In Martin Everaert and Henk van Riemsdijk, editors, *The Blackwell Companion to Syntax*, volume 2, pages 145–173. Blackwell, Oxford.

V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, and V. Pirrelli. 2014. The PAISÀ corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WAC-9), in conjunction with EACL 2014*, Gothenburg, Sweden.

Natalia N. Modjeska. 2002. Lexical and grammatical role constraints in resolving other-anaphora. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Conference (DAARC 2002)*, pages 129–134, September.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun et al., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

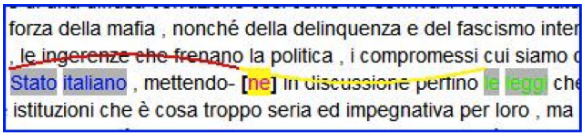
Malvina Nissim and Sara Perboni. 2008. The italian particle ne: Corpus construction and analysis. In Nicoletta Calzolari et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cinzia Russi. 2008. *Italian Clitics. An empirical study*, volume 193 of *Trends in Linguistics Monograph*. Walter de Gruyter, Berlin/New York.

A. Sorace. 2000. Gradients in auxiliary selection with intransitive verbs. *Language*, 76:859–890.

Appendix A: Screenshot of annotation in MMAX2

The instance of "ne" to be annotated is always enclosed in bold square brackets. Once annotated, the antecedent is in blue, the predicate in green. The red arc marks the link between anaphor and antecedent, while the yellow one links anaphor and predicate.



Appendix B: "Ne" types and anaphoric relations

Observed configurations of types of "ne" and anaphoric relations. For a description, please refer to Section 4.2.

		direct anaphora		
		d-obj	i-obj	encaps
partitive	type	–	–	–
	token	✓	–	–
¬partitive	prepobj	–	✓	✓
	external	–	✓	✓
	internal	–	✓	–

		indirect anaphora	
		nom ellipsis	bridging
partitive	type	✓	–
	token	✓	–
¬partitive	prepobj	–	–
	external	–	✓
	internal	–	✓

10.12871/CLICIT2014149

On the lexical coverage of some resources on Italian cooking recipes

Alessandro Mazzei
Università degli Studi di Torino
Corso Svizzera 185, 10149 Torino
mazzei@di.unito.it

Abstract

English. We describe an experiment designed to measure the lexical coverage of some resources over the Italian cooking recipes genre. First, we have built a small cooking recipe dataset; second, we have done a qualitative morpho-syntactic analysis of the dataset and third we have done a quantitative analysis of the lexical coverage of the dataset.

Italian. *Descriviamo un esperimento per valutare la copertura lessicale di alcune risorse sul genere delle ricette da cucina. Primo, abbiamo costruito un piccolo dataset di ricette. Secondo, ne abbiamo eseguito un'analisi qualitativa di sulla morfo-sintassi. Terzo, ne abbiamo eseguito un'analisi quantitativa della copertura lessicale.*

Introduction

The study reported in this paper is part of an applicative project in the field of nutrition. We are designing a software service for Diet Management (Fig. 1) that by using a smartphone allows one to retrieve, analyze and store the nutrition information about the courses. In our hypothetical scenario the interaction between the man and the food is mediated by an intelligent recommendation system that on the basis of various factors encourages or discourages the user to eat that specific course. The main factors that the system needs to account for are: (1) the diet that you intend to follow, (2) the food that have been eaten in the last days and, (3) the nutritional values of the ingredients of the course and its specific recipe. Crucially, in order to extract the complete salient nutrition information from a recipe, we need to analyze the sentences

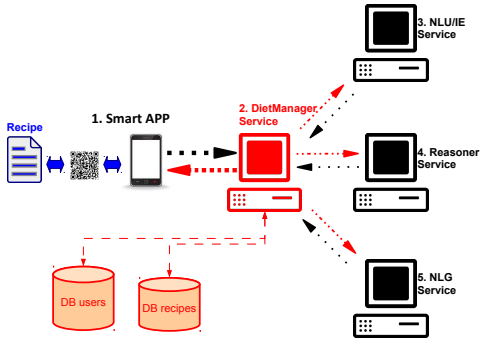


Figure 1: The architecture of the diet management system.

of the recipe. To allow the execution of this information extraction task we intend to use a syntactic parser together with a semantic interpreter. However, we intend to use both a *deep* interpreter, which have showed to have good performances in restricted domains (Fundel et al., 2007; Lesmo et al., 2013), as well as a *shallow* interpreter, that are most widely used in practical applications (Manning et al., 2008).

In order to optimize the information extraction task we need to evaluate the specific features of the applicative domain (Fisher, 2001; Jurafsky, 2014). In the next Sections we present a preliminary study towards the realization of our NLP system, i.e. a linguistic analysis of the Italian recipes domain.

1 Data set construction

The construction of a linguistic resource includes three main steps, i.e. collection, annotation and analysis of linguistic data. Since all these steps are very time-consuming, it is usual perform first off all tests on a preliminary small dataset. As a case study we selected three versions of the same recipe, that of the "caponata" (a Sicilian course consisting of cooked vegetables), respectively extracted from a WikiBook (210 tokens,

	Tok./Sent.	N/V	Con/Fun words
WikiBook	11.8	2.3	1.7
Cucchiaio	16.6	2.8	1.5
Cuochi	21.7	1.9	1.2

Table 1: The rates of the number of the tokens for sentences, of the number of nouns respect to verbs, of the number of the content words w.r.t function words.

15 sentences, *WikiBook* in the following) (Wikibooks, 2014), and from two famous Italian cooking books: "Il cucchiaio d'argento" (399 tokens, 23 sentences, *Cucchiaio* in the following) (D'Onofrio, 1997), and "Cuochi si diventa" (355 tokens, 16 sentences, *Cuochi* in the following) (Bay, 2003). The corpus obtained consists of 964 tokens corresponding to 54 sentences. The application of treebank development techniques to this very limited amount of data is only devoted to a preliminary qualitative evaluation of the feasibility of the information extraction task and the detection of the major difficulties that can be expected in the further development of our work. For what concern the collection of texts, the selection of data from three different books will have some impact on the further steps. In particular, this allows us to find different lexical choices in the three data sets, or different exploitation of specific linguistic constructions, such as passive versus active clauses, or different frequency of specific verbal forms, such as imperative versus present. Moreover, we can find different structures used to describe recipes or, in other words, different text styles within the cooking text genre.

In Table 1 we reported some statistics about the corpus. The number of tokens for sentence and the rate between content and function words reveal that *WikiBook* uses a simpler register with respect to the other sources. The style used by *Cuochi*, that is similar to a novel and does not follow the standard ingredients-methods template (Fisher, 2001), is revealed by the high number of tokens for sentence.[1]

2 Morpho-syntactic analysis

Following a typical strategy of semi-automatic annotation, i.e. automatic annotation followed by manual correction, for the annotation of our small dependency treebank we applied on the preliminary dataset two pipelines which integrate morphological and syntactic analysis, i.e. TULE (Turin University Linguistic Environment) (Lesmo, 2007) and DeSR (Dependency Shift Reduce) (Attardi, 2006). The exploitation of two different systems allows the comparison of the different outputs produced and the selection of the best one. Both TULE and DeSR have been tuned on a balanced corpus that does not contain recipes.

As far as the morphological analysis is concerned, first of all we have to observe that each error in the Part of Speech tagging (PoS, $1.7 - 3.2\%$ for TULE), such as the erroneous attribution of the grammatical category to a word (e.g. Verb rather than Noun), has an effect on the following analysis. For instance, it makes impossible to build a syntactic tree for some sentence or to recovery a meaning for some word in the semantic database. Because of this motivation, we started our error recovery process from the morphological annotation.

As far as the syntactic analysis is concerned, the performance of the parsers adopted are qualitatively similar even if the errors can vary. The problems more frequently detected are related to the sentence splitting which can be solved by using a pre-processing step. These problems are rare in *Cuochi* and mainly found in *Cucchiaio* or *WikiBook*, where the recipes are organized by a set of titles according to a sort of template (cf. (Fisher, 2001)), including e.g. the name of the recipe, "Ingredienti", "Ricetta", "Per saperne di più". This confirms that the selected books adopt a different style in the description of recipes also within the same text genre represented by cooking literature.

3 Lexical coverage experiment

With the aim to extract information from recipes (Maeta et al., 2014; Walter et al., 2011; Amélie Cordier, 2014; Shidochi et al., 2009; Haoran Xie and Lijuan Yu and Qing Li, 2010; Druck, 2013), a key issue regards the coverage of the lexicon. In order to extract the nutrition values from a specific recipe, we need to map the words contained into the recipe to a semantic organized repositories of lexical knowledge. A number of lexical resources are specialized on one specific domain while others resources are more general and, often, are automatically extracted from semi-structured

[1] For example, an newspaper section of the Turin University Treebank has $\sim$ 25 tokens for sentence (Bosco et al., 200).

resources (Hovy et al., 2013). In order to explore the automatic extraction of information from Italian recipes, we designed an experiment that uses both the types of resources.

In our experiment we have used 4 distinct Italian computational lexicons: 1 specialized lexicon, i.e. AGROVOC (FAO, 2014), and 3 general lexicons, i.e. MultiWordNet, BabelNet, UniversalWordnet. AGROVOC is a specialized lexicon, that is a controlled multi-language vocabulary, developed in collaboration with the FAO, covering a number of domains related to food, as nutrition, agriculture, environment, etc. It contains $40,000$ concepts organized in a hierarchies, that express lexical relations among concepts, as "narrow terms". Each concept is denoted by a number, and can be linked by different lexical items (terms) in different languages. AGROVOC is formalized as a RDF linked dataset but it is also available for download in various formats.[2] A notable feature of AGROVOC is the direct connection with other knowledge repository: in particular it is connected with DBpedia (Bizer et al., 2009), that often contains explicit annotation of the nutrition values.

MultiWordNet, BabelNet and UniversalWordnet are three general computational lexicons related to WordNet, that is a large lexical database of English (Miller, 1995; Fellbaum, 2005). Nouns, verbs, adjectives and adverbs are organized into sets of synonyms (synsets), each one denoting a distinct concept. Synsets are interlinked by means of semantic and lexical relations as ISA relation or hyperonymy relation. MultiWordNet is a lexical database in which an Italian WordNet is strictly aligned with WordNet[3]. The Italian synsets ($\sim 30,000$, that are linked by $\sim 40,000$ lemmas) are created in correspondence with the WordNet synsets and the semantic relations are imported from the corresponding English synsets (Pianta et al., 2002). BabelNet is a multilingual lexicalized ontology automatically created by linking Wikipedia to WordNet (Navigli and Ponzetto, 2012). The integration is obtained by an automatic mapping and by using statistical machine translation. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages, among them Italian. In this work we used BabelNet 1.1 consisting of 5 millions of concepts linked by 26 millions of word. UniversalWordNet is an automatically constructed multilingual lexical knowledge base based on WordNet (de Melo and Weikum, 2009). Combining different repositories of lexical knowledge (e.g. wikipedia), UniversalWordNet cosists of $1,500,000$ lemmas in over 200 languages. Note that the direct connections of UniversalWordNet and BabelNet towards wikipedia allows one to access to the nutrition values of foods since they are often represented in wikipedia.

In order to analyze and compare the possible use of these Italian lexical resources for information extraction, we performed a Named-Entity Recognition (NER) experiment. We introduced three semantic entities that are particularly relevant for the recipe analysis: FOOD, PREP (preparation), Q/D (quantity and devices). We mark with the FOOD label the words denoting food, e.g. melanzana (*aubergine*), pomodoro (*tomato*), sale (*salt*); we mark with PREP words denoting verbs that are involved with the preparation of a recipe, e.g. tagliare (*to cut*), miscelare (*to mix*), cuocere (*to cook*); we mark with Q/D words expressing quantities, e.g. minuti (*minutes*), grammi (*grams*) or denoting objects that are related with the recipe preparation, e.g. cucchiaio (*spoon*), coltello (*knife*). By using these three name entity categories, we annotated the three caponata recipes. In the columns "Tok." (tokens) of the Tables 2-3-4 are reported the number of words for each category.

We performed two distinct experiments for lexical coverage. The first experiment concerns AGROVOC, the second experiment concerns MultiWordNet, BabelNet and UniversalWordNet. In the first experiment we count the number of entities that can be retrieved by a straight search in AGROVOC for each name entity category: we search for the word form and for the corresponding lemma too. The columns AgrVoc-TP (true positives) of the Tables 2-3-4, report the number of retrieved tokens for each category, and the columns AgrVoc-rec report the corresponding coverage. In this experiment there are no "false positives", i.e. all the elements retrieved belongs to a meaningful categories (in other word precision is 100%). A first consideration regards the very low scores obtained on the PREP and Q/D categories. This fact could suggest that AGROVOC lexicon is not enough gen-

[2] In the experiment we used the SQL version of AGROVOC.

[3] MultiWordNet is natively aligned with WordNet 1.6. However we adopt a pivot table in order to use WordNet 3.0.

eral to be used for recipe analysis. A deeper analysis explains also the low score obtained on the FOOD category. Many of the terms are present in AGROVOC only in the plural form: for instance AGROVOC contains the entry "pomodori" (*tomatoes*) but does not contain "pomodoro" (*tomato*). Moreover, many food do not have a generic lexical entry: for instance AGROVOC contains the entry "peperoni verdi" (*green peppers*) but does not contain "peperoni" (*peppers*). However, the best scores for this experiment has been obtained on *WikiBook*, that is on the simplest recipe.

The second experiment, that involves MultiWordNet, BabelNet and UniversalWordNet, is more complex. We use a naive *Super-Sense Tagging algorithm* (NaiveSST) for the NER task. SST consists of annotating each significant entity in a text (nouns, verbs, adjectives and adverbs) within a general semantic tag belonging to the taxonomy defined by the WordNet lexicographer classes, that are called *super-senses* (Ciaramita, 2003). The lexicographer classes are 44 general semantic categories as "location", "food", "artifact", "plant", etc. The NaiveSST algorithm is very simple:

foreach *content word in the sentence* **do**
  Retrieve all the synsets corresponding to the word from MultiWordNet, BabelNet, UniversalWordNet
  foreach *super-sense of a synset* **do**
    if *the super-sense is food or plant or animal* **then**
      assign the label FOOD to the word
    end
    if *the super-sense is quantity or artifact* **then**
      assign the label Q/D to the word
    end
    if *the super-sense is creation or change or contact* **then**
      assign the label PREP to the word
    end
  end
end

Algorithm 1: The NaiveSST algorithm.

The columns NaiveSST-TP (true positives), NaiveSST-FP (false positives) of the Tables 2-3-4 report the number of correct/uncorrect labels for each category, while the NaiveSST-pre and NaiveSST-rec columns report the corresponding

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	37	23	62.2	35	5	87.5	94.6
PREP	19	1	5.3	15	8	65.2	79.0
Q/D	15	6	40.0	8	10	44.4	53.3
TOT.	71	30	42.3	58	23	71.6	81.7

Table 2: The results of the lexical semantic coverage experiment on the "WikiBook" version of the caponata recipe.

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	61	35	57.4	55	10	84.62	90.2
PREP	49	4	8.2	35	12	74.5	71.4
Q/D	31	1	3.2	27	10	73.0	87.1
TOT.	141	40	28.4	117	42	73.6	83.0

Table 3: The results of the lexical semantic coverage experiment on the "Cucchiaio d'argento" version of the caponata recipe.

precision and recall. In contrast with the previous experiment, the best scores here has been obtained on *Cuochi*. Indeed, the novel-style of *Cuochi* gives better results on the PoS tagging ($\sim 1.7\%$) and, as a consequence, on the correct lemmatization of the words. Also in this experiment the most difficult category is Q/D: this low value is related to the lemmatization process too. Often the lemmatizer is not able to recognize the correct lemma, e.g. "pentolino" (*small pot*) or "″" (*seconds*).

	Tok.	AgrVoc		NaiveSST			
		TP	rec%	TP	FP	pre%	rec%
FOOD	45	27	60.0	43	11	79.6	95.6
PREP	52	2	3.9	49	4	92.5	94.2
Q/D	43	3	7.0	32	26	55.2	74.4
TOT.	140	32	22.9	124	41	75.15	88.6

Table 4: The results of the lexical semantic coverage experiment on the "Cuochi si diventa" version of the caponata recipe.

Conclusions

In this paper we presented a preliminary study on cooking recipes in Italian. The qualitative analysis emphatizes the importance of the sentence splitter and of the PoS tagger for a correct morpho-syntactic analysis. From the quantitative lexical coverage analysis we can draw a number of speculations. First, there is a great linguistic variation among cookbooks. Second, general lexical resources outperform domain specific resources with respect to lexical coverage. Third, the lemmatization can improve the recall of the algorithm with respect to the lexical resource.

Acknowledgments

This work has been partially supported by the project MADiMAN, partially funded by Regione Piemonte, Innovation Hub for ICT, POR FESR 2007/2013 - Asse I - Attività I.1.3.
`http://di.unito.it/madiman`

We thank Cristina Bosco and Manuela Sanguinetti for their precious help in the linguistic analysis of the recipes.

References

Amélie Cordier. 2014. 4th Computer Cooking Contest, An event of ICCBR 2011. http://liris.cnrs.fr/ccc/ccc2011/doku.php.

G. Attardi. 2006. Experiments with a multilanguage non–projective dependency parser. In *Proceedings of the CoNLL-X '06*, New York City, New York.

Allan Bay. 2003. *Cuochi si diventa. Le ricette e i trucchi della buona cucina italiana di oggi*, volume 1. Feltrinelli.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semant.*, 7(3):154–165, September.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. 200. Building a treebank for italian: a data-driven annotation schema. In *In Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000*, pages 99–105.

Massimiliano Ciaramita. 2003. Supersense tagging of unknown nouns in wordnet. In *In EMNLP 2003*, pages 168–175.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

C. D'Onofrio. 1997. *Il Cucchiaio d'Argento*. Editoriale Domus. On-line version: http://www.cucchiaio.it/ricette/ricetta-caponata-classica .

Gregory Druck. 2013. Recipe Attribute Prediction using Review Text as Supervision. In *Cooking with Computers 2013, IJCAI workshop*.

FAO. 2014. AGROVOC Project. http://aims.fao.org/standards/agrovoc/.

Christiane Fellbaum. 2005. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford. Elsevier.

M. F. K. Fisher, 2001. *The Anatomy of a Recipe*, chapter 1, pages 13–24. Vintage.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, January.

Haoran Xie and Lijuan Yu and Qing Li. 2010. A Hybrid Semantic Item Model for Recipe Search by Example. In *ISM*, pages 254–259.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

D. Jurafsky. 2014. *The Language of Food: A Linguist Reads the Menu*. W. W. Norton.

Leonardo Lesmo, Alessandro Mazzei, Monica Palmirani, and Daniele P. Radicioni. 2013. TULSI: an NLP system for extracting legal modificatory provisions. *Artif. Intell. Law*, 21(2):139–172.

Leonardo Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. *Intelligenza artificiale*, 2:46–47.

Hirokuni Maeta, Shinsuke Mori, and Tetsuro Sasada. 2014. A framework for recipe text interpretation. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 553–558, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November.

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.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Yuka Shidochi, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase. 2009. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. In *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities*, CEA '09, pages 9–14, New York, NY, USA. ACM.

Kirstin Walter, Mirjam Minor, and Ralph Bergmann. 2011. Workflow extraction from cooking recipes. In Belen Diaz-Agudo and Amelie Cordier, editors, *Proceedings of the ICCBR 2011 Workshops*, pages 207–216.

Wikibooks. 2014. Wikibooks, manuali e libri di testo liberi: Libro di cucina - Ricette. it.wikibooks.org/wiki/Libro_di_cucina/Ricette/Caponata .

10.12871/CLICIT2014150

Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank

Anne-Lyse Minard
FBK, Trento, Italy
minard@fbk.eu

Alessandro Marchetti
FBK, Trento, Italy
alessandro.marchetti777@gmail.com

Manuela Speranza
FBK, Trento, Italy
manspera@fbk.eu

Abstract

English. In this paper we present ongoing work devoted to the extension of the Ita-TimeBank (Caselli et al., 2011) with event factuality annotation on top of TimeML annotation, where event factuality is represented on three main axes: time, polarity and certainty. We describe the annotation schema proposed for Italian and report on the results of our corpus analysis.

Italiano. *In questo articolo viene presentata un'estensione di Ita-TimeBank (Caselli et al., 2011), con l'annotazione della fattualità delle menzioni eventive già individuate secondo le specifiche di TimeML. La fattualità degli eventi è rappresentata attraverso tre dimensioni: tempo, polarità e certezza. Lo schema di annotazione proposto per l'italiano e l'analisi del corpus sono riportati e descritti.*

1 Introduction

In this work, we propose an annotation schema for factuality in Italian adapted from the schema for English developed in the NewsReader project[1] (Tonelli et al., 2014) and describe the annotation performed on top of event annotation in the Ita-TimeBank (Caselli et al., 2011). We aim at the creation of a reference corpus for training and testing a factuality recognizer for Italian.

The knowledge of the factual or non-factual nature of an event mentioned in a text is crucial for many applications (such as question answering, information extraction and temporal reasoning) because it allows us to recognize if an event refers to a real or to hypothetical situation, and enables us to assign it to its time of occurrence. In particular we are interested in the representation of information about a specific entity on a timeline, which enables easier access to related knowledge. The automatic creation of timelines requires the detection of situations and events in which target entities participate. To be able to place an event on a timeline, a system has to be able to select the events which happen or that are true at a certain point in time or in a time span. In a real context (such as the context of a newspaper article), the situations and events mentioned in texts can refer to real situations in the world, have no real counterpart, or have an uncertain nature.

The FactBank guidelines are the reference guidelines for factuality in English and FactBank is the reference corpus (Sauri and Pustejovsky, 2009). More recently other guidelines and resources have been developed (Wonsever et al., 2012; van Son et al., 2014), but, to the best of our knowledge, no resources exist for event factuality in Italian.

2 Related work

Several studies have been carried out on the representation of factuality information. In addition to the definition of annotation frameworks, these studies have been leading to the development of annotated corpora.

Our notion of event factuality is based on the notion of event as defined in the TimeML specifications (Pustejovsky et al., 2003a) and annotated in TimeBank (Pustejovsky et al., 2003b). *Event* is a cover term for situations that happen or occur, including predicates describing states or circumstances in which something obtains or holds true (Pustejovsky et al., 2003a).

Our main reference for factuality is FactBank (Sauri and Pustejovsky, 2009), where event factuality is defined as the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in a given

[1] http://www.newsreader-project.eu/

discourse.

van Son et al. (2014) propose an annotation schema inspired by FactBank. They add the distinction between past or present events and future events (temporality) to the FactBank schema. They then use three features (polarity, certainty and temporality) to annotate event factuality on top of the sentiment annotation in the MPQA corpus (Wiebe et al., 2005).

Wonsever et al. (2012) propose an event annotation schema based on TimeML for event factuality in Spanish texts. Factuality is annotated as a property of events that can have the following values: YES (factual), NO (non-factual), PROGRAMMED_FUTURE, NEGATED_FUTURE, POSSIBLE or INDEFINITE. Besides the factuality attribute they introduce an attribute to represent the semantic time of events, which can be different from the syntactic tense. In this way they duplicate both temporal information and polarity, as the factuality values include temporal and polarity information.

For Italian, to the best of our knowledge, there are no resources for factuality. The closest work to event factuality annotation that has been done is the annotation of attribution relations in a portion of the ISST corpus (Pareti and Prodanof, 2010). An attribution relation is the link between a source and what it expresses, and contains features providing information about the type of attitude and the factuality of the attribution. The focus of this annotation is on sources and their relations with events, while our work aims at describing factuality of events without explicitly annotating the relations between events and sources.

3 Annotation of factuality

As part of the NewsReader project, Tonelli et al. (2014) have defined guidelines for intra-document annotation at the semantic level, which provide an annotation schema of factuality for English based on TimeML annotation and the annotation framework proposed by van Son et al. (2014).

Following this annotation schema, we propose guidelines for event factuality annotation in Italian where we represent factuality by means of three attributes associated to event mentions: certainty, time, and polarity.

Certainty. We define the certainty attribute as how certain the source is about an event, with the following three values: `certain`, `possible`, `probable`. Modals and modal adverbs are typical markers of both probable (e.g. *essere probabile* - *be likely*) and possible (e.g. *potere* - *may*, *can*) events. The `underspecified` value is used for events for which it is not possible to assign a certainty value. In example (1) the event *portare* is `possible` due to the presence of *potere*. Certainty is determined according to the main source, which can be the utterer (in cases of direct speech, indirect speech or reported speech) or the author of the news. In (2) the source used to determine the certainty of *detto* is the writer and for *giocato* it is *Gianluca Nuzzo*. In both cases the source is certain about the event.

(1) *L'aumento delle tasse potrebbe* ***portare*** *nelle casse più di 500.000 euro.* [The tax increase could **bring** in more than 500,000 euros.]

(2) *"Durante l'ultimo mese ho* ***giocato*** *pochissimo", ha* ***detto*** *Gianluca Nuzzo.* ["During the last month I **played** very little, said Gian Luca Nuzzo".]

Time. The time attribute specifies the time an event took place or will take place. Its values are `non future` (for present and past events), `future` (for events that will take place), and `underspecified` (used for general events and when the time of an event cannot be determined). In the case o reported speech, the value of the time attribute is related to the time of utterance and not to the time of writing (i.e. when the utterance is reported).

Polarity. The polarity attribute captures if an event is affirmed or negated and, consequently, it can be either `positive` or `negative`; when there is not enough information available to detect the polarity of an event, it is `underspecified`.

Special cases. The *special_cases* layer is needed in order to make a distinction between hypothetical events in conditionals that do not refer to the real world and general statements that are not anchored in time, among others. This annotation can have the attribute COND_ID_CLAUSE if the event is in the "if clause" of the condition, COND_MAIN_CLAUSE if it is in the main clause, GEN for a general statement or NONE otherwise.

Factuality value. Combining the three attributes certainty, time and polarity, and taking into account the special case layer, we can determine whether the term considered refers to a fac-

tual, a counterfactual or a non factual event.

We can say that an expression refers to a FACTUAL event if it is annotated as certainty `certain`, time `non future`, and polarity `positive`, while it refers to a COUNTERFACTUAL event (i.e. an event which did not take place) if it annotated as certainty `certain`, time `non future`, and polarity `negative`. In any other combination of annotation, the event referred by the term can be considered NON FACTUAL, either because it refers to a future event, or because it is not certain (`possible` or `probable`) if the event will happen or not.

The *special cases* layer changes the status of the factuality value FACTUAL to a NON FACTUAL value, i.e. an event annotated as FACTUAL will be considered as NON FACTUAL when part of a conditional construction or of a general statement.

4 The corpus

The Ita-TimeBank is a language resource manually annotated with temporal and event information (Caselli et al., 2011). It consists of two corpora, the CELCT corpus and the ILC corpus, that have been developed in parallel following the It-TimeML annotation scheme, an adaptation to Italian of the TimeML annotation scheme (Pustejovsky et al., 2003a). The CELCT corpus, created within the LiveMemories project[2], consists of news stories taken from the Italian Content Annotation Bank (I-CAB)[3] (Magnini et al., 2006), which in turn consists of 525 news articles from the local newspaper "L'Adige"[4]. The ILC corpus is composed of 171 newspaper stories collected from the Italian Syntactic-Semantic Treebank, the PAROLE corpus, and the web.

From the Ita-TimeBank, which was first released for the EVENTI task at EVALITA 2014[5], we selected a subset of news stories to be annotated with factuality. The subset consists of 170 documents taken from the CELCT corpus and contains 10,205 events.

We annotated factuality values on top of the TimeML annotation. The TimeML specifications consider as *events* predicates describing situations that happen or occur, together with predicates describing states and circumstances. Each event is classified into one of the following TimeML classes: REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE and STATE.

In the corpus, within the 10,205 event mentions, there are 6,300 verbs, 3,526 nouns, 352 adjectives and 27 prepositions. The distribution among TimeML classes is the following: 5,292 OCCURRENCE, 2,352 STATE, 900 I_ACTION, 864 I_STATE, 439 REPORTING, 258 ASPECTUAL and 100 PERCEPTION.

With respect to the TimeML annotation, we do not annotate factuality for events of the class STATE because we do not consider it relevant for "circumstances in which something obtains or holds true" (Pustejovsky et al., 2003a). Likewise we do not annotate factuality for events of the class I_STATE because we use them to determine the certainty of their eventive argument (e.g. *sperare - hope*).

The annotation of factuality has been done for 6,989 events from 170 articles by using the CELCT Annotation Tool (Lenzi et al., 2012).

5 Results

In the following section, we report on the inter-annotator agreement and then we present a first analysis of the annotated corpus.

5.1 Inter-Annotator agreement

We have computed the agreement between two annotators on the four factuality attributes assigned to 92 events. For the agreement score we used accuracy and we computed it as the number of matching attribute values divided by the number of events. For each of the four attributes we obtained good agreement, with accuracy values over 0.91.

A study of the annotations on which we found disagreement shows that the problem stems from the `underspecified` values for time, polarity and certainty attributes. The `underspecified` value is used when it is not possible to assign another value to an attribute by using information available in the text. More precise rules should be defined in order to help annotators decide if they can use the `underspecified` value or not.

5.2 Corpus analysis

Factuality attributes have been annotated on top of 4,114 verbal events and 2,870 nominal events, for a total of 6,989 events.

[2] `http://www.livememories.org`
[3] `http://ontotext.fbk.eu/icab.html`
[4] `http://www.ladige.it/`
[5] `http://www.evalita.it/2014/tasks/eventi`

	event classes					*news topics*				
	IACT	REP	PER	OCC	ASP	Trento	Sport	Economy	Culture	News
# events	900	439	100	5,292	258	3,084	886	735	684	1,600
Factual (%)	65.2	**84.5**	66.0	69.0	65.5	68.2	**71.1**	66.4	62.9	**74.6**
Counterfactual (%)	3.8	2.7	8.0	3.8	1.6	4.5	4.4	1.4	2.5	3.5
Future - certain (%)	9.0	2.5	6.0	10.9	**21.3**	9.5	14.0	**16.9**	**16.5**	4.8
Future - uncertain (%)	**14.2**	6.6	12.0	8.9	6.6	11.6	8.5	2.4	**13.6**	7.1
Non future - uncertain (%)	2.6	0.9	2	1.8	1.9	2.7	0.8	0.5	0.3	2.1

Table 1: Corpus statistics: correlation of event factuality with event classes and news topics.

We combined the values of certainty, polarity and relative time attributes of events in order to obtain their factuality value. The factuality values were then studied in comparison with event parts-of-speech, TimeML event classes and news topics. In Table 1, we report the statistics on event factuality in the corpus.

As expected, in newspaper articles the majority of events mentioned are FACTUAL. We observed that there is a higher proportion of nominal FACTUAL events (73.8%) than verbal FACTUAL events (66.1%). On the contrary, `uncertain` events are mainly verbs.

The relation between TimeML event classes and factuality values was studied in order to determine their correlation. Some expected phenomena were observed, in particular that REPORTING events[6] are mainly FACTUAL (84.5%) because they are often used to introduce reported speech and that events of the class ASPECTUAL[7] contain a high proportion of `future` events, mainly `certain`. Considering the events of the class I_ACTION[8] it can be noted that the proportion of `uncertain` events (17%) is higher than in other classes.

The distribution of the factuality value of events in the Ita-TimeBank was also studied according to the topic of each news article considered. The news of the CELCT corpus are categorized in 5 topics: news stories, local news, economy, culture and sport.

The main distinction we observed is between cultural news and all the other kinds of news. Cultural news contains a lower proportion of FACTUAL events (62.9%) and a higher proportion of `future` events (30.1%) than the other categories of news articles, while around 14% of the event mentions in cultural news were annotated as `uncertain`. Indeed cultural news contains both reports about past cultural events and announcement of future events. On the contrary, in news stories there is a high proportion of factual events and very few future events.

6 Conclusion

In this paper we have presented an annotation schema of event factuality in Italian and the annotation task done on the Ita-TimeBank. In our schema, factuality information is represented by three attributes: time of the event, polarity of the statement and certainty of the source about the event.

We have selected from the Ita-TimeBank 170 documents containing 10,205 events and we have annotated them following the proposed annotation schema. The annotated corpus is freely available for non commercial purposes from `https://hlt.fbk.eu/technologies/fact-ita-bank`.

The resource has been used to develop a system based on machine learning for the automatic identification of factuality in Italian. The tool has been evaluated on a test dataset and obtained 76.6% accuracy, i.e. the system identified the right value of the three attributes in 76.6% of the events. This system will be integrated in the TextPro tool suite (Pianta et al., 2008).

Acknowledgments

This research was funded by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project.

[6] "REPORTING events describe the action of a person or an organization declaring something, narrating an event, informing about an event, etc." (Pustejovsky et al., 2003a)

[7] ASPECTUAL events "code information on a particular phase or aspect in the description of another event" (Caselli et al., 2011)

[8] "I_ACTION events describe an action or situation which introduces another event as its argument" (Pustejovsky et al., 2003a)

References

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *LREC*, pages 333–338.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*.

Silvia Pareti and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation, LREC10*.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34.

James Pustejovsky, Patrick Hanks, Roser Saur, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March.

Roser Sauri and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Sara Tonelli, Rachele Sprugnoli, and Manuela Speranza. 2014. NewsReader Guidelines for Annotation at Document Level, Extension of Deliverable D3.1. In *Technical Report NWR-2014-2*.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, pages 162–210.

Dina Wonsever, Aiala Ros, Marisa Malcuori, Guillermo Moncecchi, and Alan Descoins. 2012. Event Annotation Schemes and Event Recognition in Spanish Texts. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 7182 of *Lecture Notes in Computer Science*, pages 206–218. Springer.

10.12871/CLICIT2014151

An English-Italian MWE dictionary

Johanna Monti
Dipartimento di Scienze
Umanistiche e Sociali
Università degli Studi di Sassari
Via Roma 151 - Sassari
[jmonti@uniss.it]

Abstract

English. The translation of Multiword Expressions (MWEs) requires the knowledge of the correct equivalent in the target language which is hardly ever the result of a literal translation. This paper is based on the assumption that the proper treatment of MWEs in Natural Language Processing (NLP) applications and in particular in Machine Translation and Translation technologies calls for a computational approach which must be, at least partially, knowledge-based, and in particular should be grounded on an explicit linguistic description of MWEs, both using an electronic dictionary and a set of rules. The hypothesis is that a linguistic approach can complement probabilistic methodologies to help identify and translate MWEs correctly since hand-crafted and linguistically-motivated resources, in the form of electronic dictionaries and local grammars, obtain accurate and reliable results for NLP purposes. The methodology adopted for this research work is based on (i) Nooj, an NLP environment which allows the development and testing of the linguistic resources, (ii) an electronic English-Italian MWE dictionary, (iii) a set of local grammars. The dictionary mainly consists of English phrasal verbs, support verb constructions, idiomatic expressions and collocations together with their translation in Italian and contains different types of MWE POS patterns.

Italiano. *La traduzione delle polirematiche richiede la conoscenza del corretto equivalente nella lingua di arrivo che raramente è il risultato di una traduzione letterale. Questo contributo si basa sul presupposto che il corretto trattamento delle polirematiche in applicazioni di Trattamento Automatico del Linguaggio (TAL) ed in particolare di Traduzione Automatica e nelle tecnologie per la traduzione, più in generale, richiede un approccio computazionale che deve essere, almeno in parte, basato su dati linguistici, ed in particolare su una descrizione linguistica esplicita delle polirematiche, mediante l'uso di un dizionario macchina ed un insieme di regole. L'ipotesi è che un approccio linguistico può integrare le metodologie statistico-probabilistiche per una corretta identificazione e traduzione delle polirematiche, poiché risorse linguistiche quali dizionari macchina e grammatiche locali ottengono risultati accurati per gli scopi del TAL. La metodologia adottata per questa ricerca si basa su (i) Nooj, un ambiente TAL che permette lo sviluppo e la sperimentazione di risorse linguistiche, (ii) un dizionario macchina Inglese-Italiano di polirematiche, (iii) un insieme di grammatiche locali. Il dizionario è costituito principalmente da verbi frasali, verbi supporto, espressioni idiomatiche e collocazioni inglesi e contiene diversi tipi di modelli di polirematiche nonché la loro traduzione in lingua italiana.*

1 Introduction

This paper presents a bilingual dictionary of MWEs from English to Italian. MWEs are a complex linguistic phenomenon, ranging from lexical units with a relatively high degree of internal variability to expressions that are frozen or semi-frozen. They are very frequent and productive word groups both in everyday languages and in languages for special purposes and are the result of human creativity which is not ruled by algorithmic processes, but by very complex processes which are not fully representable in a machine code since they are driven by flexibility and intuition. Their interpretation and translation sometimes present unex-

pected obstacles mainly because of inherent ambiguities, structural and lexical asymmetries between languages and, finally, cultural differences.

The identification, interpretation and translation of MWEs still represent open challenges, both from a theoretical and a practical point of view, in the field of Machine Translation and Translation technologies.

Empirical approaches bring interesting complementary robustness-oriented solutions but taken alone, they can hardly cope with this complex linguistic phenomenon for various reasons. For instance, statistical approaches fail to identify and process non high-frequent MWEs in texts or, on the contrary, they are not able to recognise strings of words as single meaning units, even if they are very frequent.

Furthermore, MWEs change continuously both in number and in internal structure with idiosyncratic morphological, syntactic, semantic, pragmatic and translational behaviours.

The main assumption of this paper is that the proper treatment of MWEs in NLP applications calls for a computational approach which must be, at least partially, knowledge-based, and in particular should be grounded on an explicit linguistic description of MWEs, both using a dictionary and a set of rules.

The methodology adopted for this research work is based on: (I) Nooj an NLP environment which allows the development and testing of the linguistic resources, (ii) an electronic English-Italian (E-I) MWE dictionary, based on an accurate linguistic description that accounts for different types of MWEs and their semantic properties by means of well-defined steps: identification, interpretation, disambiguation and finally application, (iii) a set of local grammars.

2 Related work

The current theoretical work on this topic deals with different formalisms and techniques relevant for MWE processing in MT as well as other translation applications such as automatic recognition of MWEs in a monolingual or bilingual setting, alignment and paraphrasing methodologies, development, features and usefulness of handcrafted monolingual and bilingual linguistic resources and grammars and the use of MWEs in Statistical Machine Translation (SMT) domain adaptation, as well as empirical work concerning their modelling accuracy and descriptive adequacy across various language pairs.

The importance of the correct processing of MWEs in MT and Computer-aided translation (CAT) tools has been stressed by several authors. Thurmair (2004) underlines how translating MWEs word-by-word destroys their original meanings. Villavicenzio et al. (2005) underline how MT systems must recognise MWEs in order to preserve meaning and produce accurate translations. Váradi (2006) highlights how MWEs significantly contribute to the robustness of MT systems since they reduce ambiguity in word-for-word MT matching and proposes the use of local grammars to capture the productive regularity of MWEs. Hurskainen (2008) states that the main translation problems in MT are linked to MWEs. Rayson et al. (2010) underline the need for a deeper understanding of the structural and semantic properties of MWEs in order to develop more efficient algorithms.

Different solutions have been proposed in order to guarantee proper handling of MWEs in an MT process. Diaconescu (2004) stresses the difficulties of MWE processing in MT and proposes a method based on Generative Dependency Grammars with features. Lambert & Banchs (2006) suggest a strategy for identifying and using MWEs in SMT, based on grouping bilingual MWEs before performing statistical alignment. Moszczyński (2010) explores the potential benefits of creating specialised MWE lexica for translation and localisation applications.

Recently, increasing attention has been paid to MWE processing in MT and translation technologies and one of the latest initiatives in this research area is the MUMTTT workshop series specifically devoted to“ Multiword Units in Machine Translation and Translation Technology” (Monti & al. 2013). Finally, experiments in incorporating MWEs information in SMT have been carried out by Parra et al. (2014), who add compound lists to training sets in SMT, Kordoni & Simova (2014), who integrate phrasal verb information in a phrase-based SMT system, and finally Cholakov & Kordoni (2014), who use a linguistically informed method for integrating phrasal verbs into SMT systems. Automatic and manual evaluations of the results of these experiments show improvements in MT quality.

3 NooJ: an NLP environment for the development and testing of MWE linguistic resources

NooJ is a freeware linguistic-engineering development platform used to develop large-coverage formalised descriptions of natural languages and apply them to large corpora, in real time (Silberztein, 2002).
The knowledge bases used by this tool are: electronic dictionaries (simple words, MWEs and frozen expressions) and grammars represented by organised sets of graphs to formalise various linguistic aspects such as semi-frozen phenomena (local grammars), syntax (grammars for phrases and full sentences) and semantics (named entity recognition, transformational analysis). NooJ's linguistic engine includes several computational devices used both to formalise linguistic phenomena and parse texts such as: (i) Recursive Transition Networks (RTNs), (ii) Enhanced Recursive Transition Networks (ERTNs), (iii) Regular Expressions (RegExs) and finally (IV) Context-Free Grammars (CFGs in general).
NooJ is a tool that is particularly suitable for processing different types of MWEs and several experiments have already been carried out in this area: for instance, Machonis (2007 and 2008), Anastasiadis, Papadopoulou & Gavriilidou (2011), Aoughlis (2011). These are only a few examples of the various analysis performed in the last few years on MWE using NooJ as an NLP development and testing environment.

4 The Dictionary of English-Italian MWEs

The translation of MWEs requires the knowledge of the correct equivalent in the target language which is hardly ever the result of a literal translation. Given their arbitrariness, MT and Translation technologies have to rely on the availability of ready solutions in the source and target language in order to perform an accurate translation process.
The English-Italian MWE dictionary is the result of a contrastive English-Italian analysis of continuous and discontinuous MWEs with different degrees of variability of co-occurrence among words and different syntactic structures, carried out during the development and testing of the English-Italian language pair for Logos, a rule-based MT system, and subsequently further developed in the framework of the Lexicon-Grammar (LG) formalism (Monti, 2012).
The dictionary is based on the LG approach to MWEs (Gross, 1986), where these complex and varied linguistic phenomena are described according to a flat structure composed of the POS tags of the MWE elements and their sequence. Furthermore, according to this approach it is possible to distinguish fixed MWEs and MWEs that allow syntactic variations, such as the insertion of other elements or the variation of one or more elements. Green et al. (2011) adopt a similar approach for the MWE description and show the usefulness of this model for several NLP tasks in which MWE pre-grouping has improved accuracy.
The E-I MWE dictionary contains over 10,000 entries and is used to represent and recognise various types of MWEs. Each entry of the dictionary is given a coherent linguistic description consisting of: (i) the grammatical category for each constituent of the MWE: noun (N), Verb (V), adjective (A), preposition (PREP), determiner (DET), adverb (ADV), conjunction (CONJ); (ii) one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to nominalise them), preceded by the tag +FLX; (iii) one or more syntactic properties (e.g. "+transitive" or +N0VN1PREPN2); (iv) one or more semantic properties (e.g. distributional classes such as "+Human", domain classes such as "+Politics"); (v) the translation into Italian.
The dictionary contains different types of MWE POS patterns. The main part of the dictionary consists of English phrasal verbs, support verb constructions, idiomatic expressions and collocations together with their Italian translations.

Intransitive Verbs:

[VIntrans+ADJ]
lie,V+FLX=LIE+JM+FXC+Intrans+ADJ="flat"+IT ="sdraiarsi"

[VIntrans+PART]
bear,V+FLX=BEAR+JM+FXC+Intrans+PART= "down"+IT="avanzare"

[VIntrans+PART+PREP+N2]
break,V+FLX=SPEAK+JM+FXC+Intrans+ PART="off"+PREP="from"+N2="work"+IT "interrompere il lavoro"

[VIntrans+PART+PREP+Ving]
break,V+FLX=SPEAK+JM+FXC+Intrans+ PART= "off"+PREP="from"+VG+IT="smettere di Vinf"

[VIntrans+PREP+N2]
account,V+FLX=ASK+JM+FXC+Intrans+PREP

=“for”+N2+IT=“spiegare N2”

Transitive Verbs:

[VTrans+N1]
advance,V+FLX=LIVE+JM+FXC+Trans+N1= “reason”+IT=“esporre N1”

[VTrans+ADJ+N1]
break,V+FLX=SPEAK+JM+FXC+Trans+N1+ ADJ= "free”+IT=“liberare N1”

[VTrans+PART+N1]
bring,V+FLX=BRING+JM+FXC+Trans+PART= “up”+N1=“question”+IT=“sollevare N1(problema)”

[VTrans+PART+N1+PREP+N2]
bring,V+FLX=BRING+JM+FXC+Trans+PART=“back”+N1+PREP=“from”+N2=“memory”+IT= “richiamare a N2(mente)“

[VTrans +N1+PREP+N2]
break,V+FLX=SPEAK+JM+FXC+Trans+N1= news”+PREP=“to”+N2Hum+IT=“comunicare N1 a N2”

[VTrans+N1+PREP+Ving]
bar,V+FLX=ADMIT+JM+FXC+Trans+N1Hum+PREP=“from”+VG+IT=“impedire a N1 di Vinf”

5 Grammars

Syntactic or semantic grammars (.nog files) are used to recognise and annotate expressions in texts, e.g. to tag noun phrases, certain syntactic constructs or idiomatic expressions, extract certain expressions (name of companies, expressions of dates, addresses, etc.), or disambiguate words by filtering out some lexical or syntactic annotations in the text.
These grammars recognise different types of MWEs, such as frozen and semi-frozen units, and are particularly useful with discontinuous MWEs (Machonis, 2008 and Silberztein, 2008).
It is possible: (i) to identify MWEs of different types in texts by means of specific local grammars, (ii) annotate texts with the corresponding translations of the identified MWEs, (iii) export the annotated texts in XML.
Annotated texts can be used in this way for instance for SMT training purposes.
Once texts are annotated, they can be exported as XML files, like in the following example:

He <EXPV TYPE="JM" IT="rinunciare a"> abandons</EXPV> the <EXPN IT="appello"> appeal</EXPN>.

He <EXPV TYPE="JM" IT="rinunciare a">abandons</EXPV> the <EXPN IT="speranza"> hope</EXPN>.

He <EXPV TYPE="JM" IT="acquisire ">acquires</EXPV> a <EXPN IT="conoscenza"> knowledge</EXPN> of the specific domain.

6 Future work

For future work, we plan to further investigate MWEs in particular with respect to cross-linguistic asymmetries and translational equivalences.
Our final goal is to integrate MWE treatment in either data-driven or hybrid approaches to MT in order to achieve high quality translation by combining probabilistic and linguistic information.
However, to achieve this goal, we must devise efficient strategies for representing deep attributes and semantic properties for MWEs in a cross-linguistic perspective.

7 Conclusion

In conclusion, the focus of this research for the coming years will be to improve the results obtained so far and to extend the research work to provide a more comprehensive methodology for MWE processing in MT and translation technologies, taking into account not only the analysis phase but also the generation one.
This experiment provides, on the one hand, an investigation of a broad variety of combinations of MWE types and an exemplification of their behaviour in texts extracted from different corpora and, on the other hand, a representation method that foresees the interaction of an electronic dictionary and a set of local grammars to efficiently handle different types of MWEs and their properties in MT as well as in other types of NLP applications.
This research work has therefore produced two main results in the field of MWE processing so far:

- the development of a first version of an English-Italian electronic dictionary, specifically devoted to different MWEs types,
- the analysis of a first set of specific MWE structures from a semanto-syntactic point of view and the development of local grammars for the identification of continuous and discontinuous MWEs in the form of FST/FSA.

References

Anastasiadis, M., Papadopoulou, L., & Gavrilidou, Z. 2011. Processing Greek frozen expressions with Nooj. K. Vučković, B. Bekavac, & M. Silberztein (eds). *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik)*. Newcastle: Cambridge Scholars Publishing.

Aoughlis, F. 2011. A French-English MT system for Computer Science Compound Words. K. Vučković, B. Bekavac, & M. Silberztein (eds). *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik)*. Newcastle: Cambridge Scholars Publishing.

Cholakov, K. & Kordoni, V. 2014. Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014*, Association for Computational Linguistics: 196-201. http://aclweb.org/anthology/D14-1024

Diaconescu, S. 2004. Multiword Expression Translation Using Generative Dependency Grammar. *Advances in Natural Language Processing 4th International Conference, EsTAL 2004, October 20-22*, Alicante, Spain: 243-254.

Green, S., de Marneffe, M. C., Bauer, J., & Manning, C. D. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: 725-735.

Gross, M. 1986. Lexicon-Grammar: the representation of compound words. *Proceedings of COLING '86*. Bonn: University of Bonn, http://acl.ldc.upenn.edu/C/C86/C86-1001.pdf.

Hurskainen, A. 2008. *Multiword Expressions and Machine Translation.* Technical Reports. Language Technology Report No 1.

Kordoni, V. & Simova I. 2014. Multiword expressions in Machine Translation. *LREC 2014 Proceedings*.

Lambert, P., & Banchs, R. 2006. Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation. *Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context.* Trento, Italy.

Machonis, P. A. 2007. Look this up and try it out: an original approach to parsing phrasal verbs. *Actes du 26 Colloque international Lexique Grammaire, Bonifacio 2-6 octobre 2007.*

Machonis, P. A. 2008. NooJ: a practical method for Parsing Phrasal Verbs. *Proceedings of the 2007 International NooJ Conference.* Newcastle: Cambridge Scholars Publishing: 149-161.

Monti, J. 2012. *Multi-word Unit Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation* – PhD dissertation in Computational Linguistics- Università degli Studi di Salerno.

Monti J, Mitkov R, Corpas Pastor G, Seretan V (eds). 2013. *MT Summit workshop proceedings for: Multiword Units in Machine Translation and Translation Technologies (Organised at the 14th Machine Translation Summit)*. CH-4123 Allschwil: The European Association for Machine Translation.

Moszczyński, R. 2007. A Practical Classification of Multiword Expressions. *ACL '07 Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop.* Association for Computational Linguistics.

Parra Escartín, C., Peitz, S., and Ney, H. 2014. German Compounds and Statistical Machine Translation. Can they get along? *EACL 2014, Tenth Workshop on Multiword Expressions (MWE 2014)*, Gothenburg, Sweden, April 2014: 48-56.

Rayson, P., Piao, S., Sharoff, S., Evert, S., & Villada Moirón, B. 2010. Multiword expressions: hard going or plain sailing? *Journal of Language Resources and Evaluation. Lang Resources & Evaluation 44*: 1-5.

Silberztein, M. 2002. *NooJ Manual*. Available for download at: www.nooj4nlp.net.

Silberztein, M. 2008. Complex Annotations with NooJ. X. *Proceedings of the 2007 International NooJ Conference*, Jun 2007, Barcelone, Spain. Cambridge Scholars Publishing. <hal-00498042>

Thurmair, G. 2004. Multilingual content processing. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.

Váradi, T. 2006. Multiword Units in an MT Lexicon. *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts* Trento, Italy: Association for Computational Linguistics: 73-78.

Villavicencio, A. B. 2005. Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Journal of Computer Speech and Language Processing, 19(4)*: 365-377.

10.12871/CLICIT2014152

ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment

Giovanni Moretti, Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive 18 Trento
moretti@fbk.eu
satonelli@fbk.eu

Stefano Menini, Rachele Sprugnoli
Fondazione Bruno Kessler and
University of Trento
menini@fbk.eu
sprugnoli@fbk.eu

Abstract

English. This work presents ALCIDE (*Analysis of Language and Content In a Digital Environment*), a new platform for Historical Content Analysis. Our aim is to improve Digital Humanities studies integrating methodologies taken from human language technology and an easily understandable data structure representation. ALCIDE provides a wide collection of tools that go beyond simple metadata indexing, implementing functions of textual analysis such as named entity recognition, key-concept extraction, lemma and string-based search and geo-tagging.

Italiano. *Questo articolo presenta ALCIDE (Analysis of Language and Content In a Digital Environment), una nuova piattaforma per l'analisi di documenti storici. Il nostro obiettivo è quello di migliorare la ricerca nell'ambito dell' Informatica Umanistica integrando metodologie mutuate dalle tecnologie del linguaggio con la rappresentazione intuitiva di strutture dati complesse. ALCIDE offre una vasta gamma di strumenti per l'analisi testuale che vanno oltre la semplice indicizzazione dei metadati: ad esempio, il riconoscimento di nomi propri di entità, estrazione di concetti, ricerca basata su lemmi e stringhe, geo-tagging.*

1 Introduction

In this paper we present ALCIDE (*Analysis of Language and Content In a Digital Environment*), a new platform for Historical Content Analysis. Our aim is to improve Digital Humanities studies implementing both methodologies taken from human language technology and an easily understandable data structure representation. ALCIDE provides a wide collection of tools that go beyond text indexing, implementing functions of textual analysis such as: named entities recognition (e.g. identification of names of persons and locations within texts, key-concept extraction, textual search and geotagging). Every function and information provided by ALCIDE is time bounded and all query functions are related to this feature; the leitmotif of the portal can be summarized as: "All I want to know related to a time period".

Our work aims at providing a flexible tool combining automatic semantic analysis and manual annotation tailored to the temporal dimension of documents. The ALCIDE platform currently supports corpus analysis of English and Italian documents.

2 Related Works

Recently, several projects for the textual analysis of documents in the field of the Humanities have been presented: some of them focus only on temporal reasoning, e.g. Topotime[1](Grossner and Meeks, 2014) based on meta-data, whereas others perform word frequency analysis without a full exploitation of Natural Language Processing (NLP) techniques and temporal information, e.g. WordSeer[2] (Muralidharan and Hearst, 2013) and VOYANT (Rockwell, 2003)(Rockwell et al., 2010). Similarly to ALCIDE, WMATRIX (Rayson, 2008) is based on an automatic part-of-speech (Garside, 1987) and a semantic tagger (Rayson et al., 2004) for English to extract multiwords expressions, lemma variants and key concepts. With the only exception of key concept clouds, however, WMATRIX does not provide graphical visualizations of extracted data.

[1] http://kgeographer.com/wp/topotime/
[2] http://wordseer.berkeley.edu

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<xml archive="I.pdf">
    <file id="1">
        <head>
            <url>I.pdf_doc_number_1</url>
            <date>1901-01-01</date>
            <title>Title of the document</title>
            <description>Description of the document</description>
            <location>Geographic location</location>
            <abstract>Content of the abstract</abstract>
            <context/>
        </head>
        <content>Text of the document in italian</content>
        <original_text>Text of the document not in italian
          - if present<original_text>
    </file>
```

Figure 1: Sample of the XML input document

3 Data Preparation

ALCIDE allows the user to upload and perform analyses on any kind of documents, on condition that the documents are structured in XML according to a specific rule set.

3.1 XML Format

To fully exploit all the features of ALCIDE, the XML format must contain information about the title, the date, the location and the other information displayed in Fig. 1. A single XML file can contain multiple documents identified by a unique id. This allows users to upload an entire corpus at once.

The data can be easily imported into a database structure and given as input to NLP tools. In case the documents are available in pdf format, they need to be converted first into XML using, for instance, the JPedal PDF Java Library[3].

3.2 Data Processing

Once the documents are converted into XML and uploaded in the platform, the imported XML data is processed by TextPro[4]. TextPro is a NLP suite for Italian and English developed at Fondazione Bruno Kessler. It provides a pipeline of modules for tokenization, sentence splitting, morphological analysis, Part-of-Speech tagging, lemmatization, multi-word recognition, keyword extraction, chunking and named entity recognition(Pianta et al., 2008).

Taking the text contained in the XML file as input, TextPro returns the output of the analysis in a tabular format, with one token (and relative information) per line. When possible, TextPro modules have been tailored to the historical domain, for instance the keyword extractor and the named entity recognizer. However, we cannot expect the overall performance to be the same as for news data, on which the system was trained. The Italian POS-tagger, for instance, reached 0.98 accuracy on contemporary news stories (Pianta and Zanoli, 2007) and 0.95 on a sample of Alcide De Gasperi's writings (around 9,000 tokens written between 1906 and 1911).

3.3 Lemma Indexes

From the TextPro output, three different temporal indexes of lemmas are automatically created by ALCIDE, one for nouns, one for verbs and one for adjectives along with a timestamp. Indexing the lemma allows the portal to retrieve every document containing a certain word regardless of its declination.

3.4 Database Structure

The database structure is the core of ALCIDE and all the data presented in the graphical interface are accessible by using a query system. Data are provided both by the XML files and the TextPro analyses. The database is able to perform a large number of different queries in order to obtain the analyses requested by the user. Examples of possible queries are: the extraction of the documents published in a particular time span, in a specific city or containing a specific person name or key concept in a certain period of time. The database approach grants a good performance in case of multiple access and offers the possibility to easily update the data. Figure 2 shows that certain categories in the database such as countries or key concept can be used to group a set of documents. The database is able to relate any category to each other and then extend a category with the properties of the other related object.

4 Platform Functionalities

All processes presented in the previous Section are performed once. After the data is loaded and automatically processed, the following functionalities can be accessed through the web-based platform.

4.1 Geographical Distribution

The platform displays the geographical distribution of the documents (place of publication) and allows the user to extract all the documents related to a specific area (country or town) in a particular

[3] http://sourceforge.net/projects/jpedal/

[4] http://textpro.fbk.eu

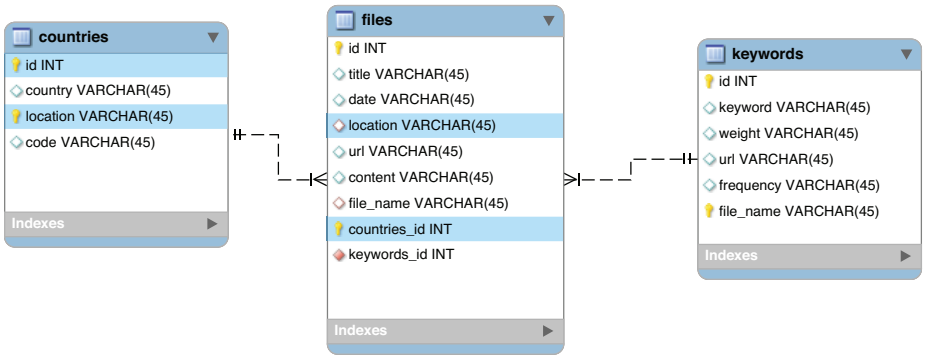


Figure 2: A simplified db graph of the structure

time span. To display data about the locations, the platform uses the Google GeoChart library[5].

4.2 Named Entity Recognition

The automatic extraction of person, location and organization names rely on the EntityPro (Pianta et al., 2008) module of TextPro. The module was originally trained on contemporary newspaper stories, on which it reached a performance of 92.12 F1 for Persons and 85.54 for GPEs. However, since the same tool obtained respectively 75.75 and 86.23 F1 on historical data (a set of Alcide De Gasperi's writings) a domain-specific adaptation was necessary. This was carried out by compiling black and white lists of common proper names for the period of interest and exploiting the tool in-built filtering functionality.

The data obtained is displayed together with the documents to highlight the most relevant persons in the text. It is also possible to query the system in order to obtain all the documents related to a specific entity or visualize in a graph the relevance of an entity over time.

4.3 Keyword Extraction

Keyword extraction is provided by the KX module embedded into the TextPro Suite. KX is a system for key-phrase extraction (both single and multi-word expressions) which exploits basic linguistic annotation combined with simple statistical measures to select a list of weighted keywords from a document (Pianta and Tonelli, 2010). KX was initially developed to work on news, patent documents and scientific articles. However, since ALCIDE is typically meant to deal with historical corpora, we tailored key-words extraction to the historians' requirement giving a higher rank to abstract concepts. This is done by boosting the relevance of concepts with a specific ending (e.g. '-ism', '-ty' in English and '-ismo', '-itidine' in Italian) usually expressing an abstract meaning. We also gave higher priority to generic key-concepts by boosting those expressed by single words.

Similarly to Named Entities, documents are displayed together with their most relevant keywords. Moreover, the portal allows the user to query the keywords characterizing a selected time span, the documents related to a specific keyword and the relevance of a keyword over the time.

4.4 Advanced Search Functions

One of the features we are interested in is to perform an efficient search of words or group of words in the whole collection of documents. The platform offers two main text search options. The first one is a full text search that gives the possibility to search for the match of one or more specific strings in a text. The second function performs a lemma based search, that looks for documents containing a specific verb, noun, or adjective in all its forms giving a lemma in input (e.g. searching for the verb *fight* the engine retrieves all the document containing *fight*, *fighting*, *fought*, etc).

Both the search functions give the possibility to perform the query in documents issued in a specific time span and to display in a graph the trend of the target term usage over time.

5 Graphical Interface

The graphical interface was developed to represent all previously mentioned data in an intuitive visualization framework. The interface provides the

[5] https://developers.google.com/chart/interactive/docs/gallery/geochart

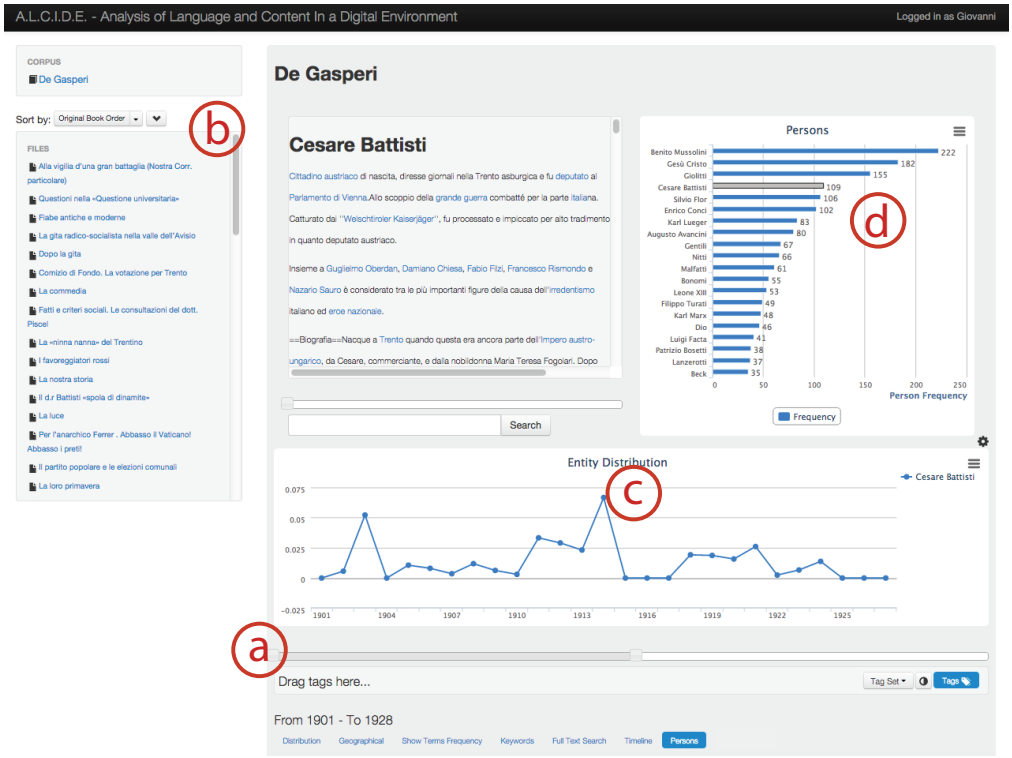


Figure 3: Sample of Graphical Interface

possibility to easily change the time span to which a search is referred. This feature is implemented through a horizontal slider (Fig.3 a) to modify the upper and the lower bound of a certain time period. The list of the retrieved documents (Fig.3 b) is always visible and accessible in the main view. In order to graphically represent the trend of an analysis (e.g. the document distribution or the number of mentions for a person) we use a line chart. All the nodes in the graph (Fig.3 c) can be used to query the system and retrieve the corresponding documents. The ranked list of keywords and entities is graphically represented by a horizontal bar chart (Fig.3 d) and are sorted by relevance to be easily identified by the user, as presented also in previous works (Few, 2013). All the bars displayed in the chart can be used to perform additional analyses by filtering and retrieving the corresponding data, for instance to get all the documents containing a particular concept in a specific time span.

Expert users can customize the set of meta-data associated with the corpus (e.g. speech transcription, propaganda materials, etc) and manually assign them to the documents. The added tags are stored in the database and can be further used to perform new queries on the collection.

6 Conclusions and Future Works

In this paper we described the general workflow and specific characteristics of the ALCIDE platform.

In the future, we aim to improve the efficiency of current functionalities and to add new ones such as (i) identification of temporal expressions and events (and the extraction of relations between them), (ii) distributional semantic analysis (i.e. quantification and categorization of semantic similarities between linguistic elements) and (iii) sentiment analysis on statements and key-concepts.

ALCIDE is already online but it is password protected. When the implementation stage will be more advanced, we will make it freely accessible and users will be allowed to upload their corpora in Excel, XML or TEI format and explore them with the platform. For the moment a video of ALCIDE demo is available at `http://dh.fbk.eu/projects/alcide-analysis-language-and-content-digital-environment`.

Acknowledgments

We would like to thank Christian Girardi for providing support in integrating and customizing TextPro.

References

Stephen Few. 2013. Data visualization for human perception. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*

Roger Garside. 1987. The claws word-tagging system. In *The computational analysis of English*. Longman, London.

Karl Grossner and Elijah Meeks. 2014. Topotime: Representing historical temporality. In *Proceedings of DH2014, Lusanne*. Alliance of Digital Humanities Organizations.

Aditi Muralidharan and Marti A Hearst. 2013. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing*, 28(2):283–295.

Emanuele Pianta and Sara Tonelli. 2010. Kx: A flexible system for keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 170–173. Association for Computational Linguistics.

Emanuele Pianta and Roberto Zanoli. 2007. Tagpro: A system for italian pos tagging based on svm. *Intelligenza Artificiale*, 4(2):8–9.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*.

Paul Rayson, Dawn Archer, Scott Piao, and AM McEnery. 2004. The ucrel semantic analysis system.

Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Geoffrey Rockwell, Stéfan G Sinclair, Stan Ruecker, and Peter Organisciak. 2010. Ubiquitous text analysis. *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, 2(1).

Geoffrey Rockwell. 2003. What is text analysis, really? *Literary and linguistic computing*, 18(2):209–219.

10.12871/CLICIT2014153

Inner Speech, Dialogue Text and Collaborative Learning in Virtual Learning Communities

Stefanos Nikiforos **Katia Lida Kermanidis**
Department of Informatics, Ionian University,
7, Tsirigoti Sq., 49100, Corfu, Greece
{c13niki;kerman}@ionio.gr

Abstract

English. Virtual Learning Communities offer new opportunities in education and set new challenges in Computer Supported Collaborative Learning. In this study, a detailed linguistic analysis in the discourse among the class members is proposed in five distinct test case scenarios, in order to detect whether a Virtual Class is a community or not. Communities are of particular importance as they provide benefits to students and effectively improve knowledge perception. This analysis is focused on two axes: inner speech and collaborative learning as they both are basic features of a community.

Italiano. *Le comunità di apprendimento virtuale offrono nuove opportunità nel campo dell'istruzione e propongono nuove sfide nella Supported Collaborative Learning. In questo lavoro viene proposta, in cinque scenari distinti di prova, un'analisi linguistica dettagliata del discorso instaurato tra i membri di una classe. L'analisi è volta a rilevare se una classe virtuale sia o no una comunità. Le comunità sono di particolare importanza in quanto forniscono benefici per gli studenti e sono un modo efficace di migliorare la percezione della conoscenza. Questa analisi è focalizzata su due assi: il discorso interiore e l'apprendimento collaborativo in quanto entrambi sono caratteristiche fondamentali di una comunità.*

1 Introduction

Virtual Learning Communities (VLCs) constitute an aspect of particular importance for Computer Supported Collaborative Learning (CSCL). The stronger the sense of community is, the more effectively is learning perceived, resulting in less isolation and greater satisfaction (Rovai, 2002; Daniel et al, 2003; Innes, 2007). Strong feelings of community provide benefits to students by increasing 1) the commitment to group goals, 2) collaboration among them and 3) motivation to learn (Rovai, 2002). Virtual Classes (VCs) are frequently created and embodied in the learning procedure (Dillenbourg and Fischer, 2007). Nevertheless there are questions arising: *Is every VC always a community as well? How can we detect the existence of a community? What are its idiosyncratic properties?* Sharing of knowledge within a community is achieved through shared codes and language (Daniel et al, 2003; Stahl, 2000; Innes, 2007). Language is not only a communication tool; it also serves knowledge and information exchange (Dillenbourg and Fischer, 2007; Knipfer et al, 2009; Daniel et al, 2003; Bielaczyc and Collins, 1999). Communication and dialogue are in a privileged position in the learning process due to the assumption that knowledge is socially constructed (Innes, 2007).

Collaborative learning (CL) is strongly associated with *communities* as it occurs when individuals are *actively* engaged in a *community* in which learning takes place through collaborative efforts (Stahl et al, 2006). This active engagement is achieved through public discussion, which is a central way for a community to expand its knowledge (Bielaczyc and Collins, 1999). Developing an understanding of how meaning is collaboratively constructed, preserved, and re-learned through the media of *language* in group interaction, is a challenge for CL theory (Daniel et al, 2003;Wells, 2002; Warschauer, 1997; Koschmann, 1999). *Inner speech (IS)* is an esoteric mental language, usually not outerly expressed, having an idiosyncratic syntax. When outerly expressed, its structure consists of apparent lack of cohesion, extensive fragmentation and abbreviation compared to the outer (formal) language used in most everyday interactions. Clauses keep only the predicate and its accompanying words, while the subject and its dependents are omitted. This does not lead to misunderstandings if the thoughts of the individuals are in accordance (they form a community). The more identical the thoughts of the individuals are, the less linguistic cues are used (Vygotsky, 2008; Socolov, 1972).

Various works using discourse analysis have been presented in the CSCL field: some of them focus on the role of dialogue (Wells, 2002), oth-

ers examine the relationship between teachers and students (Blau et al., 1998; Veermans and Cesareni, 2005), while others focus on the type of the language used (Maness, 2008; Innes, 2007), on knowledge building (Zhang et al., 2007), or on the scripts addressed (Kollar et al., 2005). Spanger et al. (2009) analyzed a corpus of referring expressions targeting to develop algorithms for generating expressions in a situated collaboration. Other studies use machine learning techniques in order to build automated classifiers of affect in chat logs (Brooks, 2013). Rovai (2002), examined the relationship between the sense of community and cognitive learning in an online educational environment. Daniel et al. (2003) explored how the notions of social capital and trust can be extended in virtual communities.

Unlike these studies, the proposed approach, for the first time to the authors' knowledge, takes into account the correlation between community properties and both inner speech and collaborative learning features (Bielaczyc and Collins, 1999) by applying linguistic analysis to the discourse among class members as a means for community detection. To this end, the discourse of four different types of VCs is analyzed and compared against non-conversational language use.

2 Inner speech linguistic analysis model

In a community, under certain conditions, the specific features of inner speech appear in outer (surface) speech (Socolov, 1972). The stronger the presence of inner speech, the more confident we are of the existence of a community. The stronger the specific mental action of inner speech is, the clearer the peculiarities of its syntax structure appear (Vygotsky, 2008; Wiley, 2006). A linguistic analysis based on the following features is therefore proposed (Appendix A).

In inner speech there is a common code for communication among the communicating parties (Emerson, 1983) transforming the language genre and style, and making it more specific (*IS1,IS2,IS3)* (Vygotsky, 2008; Wiley, 2006). The main feature of inner speech is ellipticity (Vygotsky, 2008). The *informal clauses* (Maness, 2008; Pérez-Sabater, 2012), the clauses having *no verb*, the semantically abbreviated clauses being *elliptical* in meaning, the reduced use of *subordination* [1] and of *prepositional phrases* and the average number of words in the clauses (Wiley, 2006) are features of ellipticity in the language. *Punctuation* is likely to be sparse as well (Brooks et al., 2013; Pérez-Sabater, 2012; Mannes, 2008). The *word types* used, are another indicator of inner speech. In inner speech, use of *adverbs* is not so essential, due to the common/mutual understanding (Emerson, 1983). Absence of a*djectives* makes the language elliptical, ambiguous and general. In inner speech "adjectives and other modifiers can usually be dispensed with"(Wiley, 2006). Use of *Greeklish* (informal written language, typing Greek words with Latin letters), *informal words* (shortened and simplified word forms, idioms, diminutives) and *emoticons* indicate informal communication, a basic feature of inner speech, (Brooks et al., 2013; Pérez-Sabater, 2012; Mannes, 2008).

In inner speech, where common/mutual understanding exists, the message is definite and clear to the receiver (Emerson, 1983; Mairesse et al, 2007). Therefore the use of indefinite articles will be limited, while definite articles are likely to constitute the majority. Using additional terms *(IS13,14,15,16),* is essential for achieving formal communication, but not necessary for inner speech. *Abbreviation* is a core feature of inner speech (Vygotsky, 2008; Socolov, 1972; Wiley, 2006). *Metaphors* are powerful for creating and exchanging rich sets of meaning (Daniel et al, 2003). Use of abbreviation and metaphors require a prior common understanding between the sender and the receiver, indicating inner speech. In contrast, use of *similes* indicates a necessity for additional information. So, their *absence* is an indicator for inner speech. *IS-20:*The percentage of distinct words in the discourse within a community is usually restricted (Vygotsky, 2008). Therefore, the *vocabulary richness* is poor (Wiley, 2006; Mairesse et al, 2007).

3 Collaborative learning linguistic analysis model

Collaboration is considered to be the most important shared characteristic in VLCs (Daniel et al, 2003). Analysis of the discourse, among the members of a class, focused on specific characteristics (Appendix B), can provide us with index marks of collaborative learning (CL).

Use of *verbs in the 1st person plural form* constitutes an indicator of team action or knowledge that has been produced collaboratively (Mc Millan and Chavis, 1986). Emotion is an elementary characteristic of the discourse within

[1] In case subordinate clauses are used as an object, they are not to be taken into account, because they are essential for the meaning of the sentence.

a community (Mc Millan and Chavis, 1986; Brooks et al., 2013) and is directly related to inner speech as well (Wiley, 2006). Emotion is distinguished between *positive* and *negative*. In the case of CL, the majority of the emotional words will express positive emotion, as there is strong correlation between the members' positive experience and the community bond (Mc Millan and Chavis, 1986; Mairesse et al, 2007). Community members feel the need to reward their partners for their effort (Bielaczyc and Collins, 1999; Mc Millan and Chavis, 1986; Mairesse et al, 2007). *Clauses of negation* (containing negative words: *no, not, don't*) are likely to be *less* frequent as collaboration *increases* (Mairesse et al, 2007). *Clauses of reason:* their use shows that a member of a team respects his team (he is proposing something, without giving orders). Use of *familiarity words* indicates the intimacy among the members of a team which has been transformed into a community (Mairesse et al, 2007). In a VC where the students do not know each other before the creation of the class, this metric is a strong indication of the existence of a community. Use of *inclusive words* (like *together, team, company, community*) and *social words* (like *friend, colleague, mate*) offer an index of a feeling of membership (Mairesse et al, 2007) and provide an index mark for the existence of a community (Mairesse et al, 2007). Using *pronouns in the 1st person plural form* indicates the sense of belonging to a team, the co-construction of knowledge and the feeling of sharing with others (Mc Millan and Chavis, 1986). The average number of 1st person pronouns to the total number of pronouns *(CA12)* and to the total number of personal and possessive pronouns *(CA13) is therefore counted*.

4 Case studies and Results

The five different learning communities used as case studies in this work are described in this section. *Virtual class 1 (VC1)* was created between an elementary school (ES:20 students, ages 11-12) and a high school (HS:20 students, ages 12-13) located in two different towns in Greece. The target of that project was the collaboration between the two classes in order to create a wiki about the location they live in. Students were divided into working groups of two or three. The teachers had a supporting and inspirational role and tried to minimize their involvement. Wikispaces was the collaborative platform used. During the project students were exchanging communication messages via a special web page. Discourse in VC1 is divided in two subgroups *(VC1.1, VC1.2)* for the needs of the analysis. *VC1.1* contains the discourse among the team members after having completed their task. Students expressed their impressions and feelings for the already completed project. In this case, there was *no problem* to be solved and the students chatted in a more free frame. *VC1.2* pertains to the discourse among the team members during the project. *Virtual Class 2 (VC2)* was created between two elementary schools (ES1 and ES2: 20 students each, ages 11-12) located in different towns in Greece. ES1 students were the same ones described in VC1. Designing of this project was the same as in VC1. The two main differences that have to be mentioned were: i) the difference between the educative level of the students in VC1 which does not exist in this VC, and ii) the previous experience for the ES1 students gained through their participation in VC1. *Virtual class 3 (VC3):*A real class was transformed into a virtual one through running a project using online collaborative tools. The target of the project was the creation of presentations for a national holiday. The students were the same of ES1 that joined in the two aforementioned VCs. Students were divided in groups of two or three. *Teachers had an active instructive role*. The selected environment was Google Drive. Two files were created in order to create a collaborative platform: one presentation file and one document file for the necessities of the communication among the group members. *Student's essay texts (ST):* The results of the conversational analysis (usually informal-Brooks, 2013; Bielaczyc and Collins, 1999) in the aforementioned VCs are compared against non-conversational language use, in order to detect differences. For this reason, students' essay texts (ST) were used in the analysis. These texts are narrative and they were written by the students of ES1 that took part in the VCs. They were written within the linguistics course in their school throughout the same school year when the case studies took place, by 7 different students (4 boys and 3 girls) out of a total of 20 in the class. They contain 3.577 words and 666 clauses, while VC1.1 had 210 and 52, VC1.2 had 453 and 106, VC2 had 471 and 102 and VC3 had 704 and 147 respectively. In the analysis these essay texts were treated as a single corpus.

Appendices C and D show the results for the two linguistic analysis models (percentage values for all aforementioned features). Statistical sig-

nificance testing (two tailed independent t-test - Roussos and Tsaousis, 2006) was applied to detect differences between every VC and the ST. Bold indicates significance at p<,05 level, italics at p<,02 level and asterisks at p<,01 level.

5 Discussion

VCs examined in this study were transformed into communities, providing students with the benefits of the community membership. In VC3 which was a priori a community as the students had already been working as a team for seven years (from kindergarten till the 6th grade), the community existence was confirmed. Comparison between the VCs and the ST reveals that there are statistically significant differences in the language used. In VCs the language was mainly informal, elliptical in meaning and abbreviated (the basic features of inner speech). The students of these VCs collaborated enough and had the membership feeling. The active instructive role of the teachers affects the language and makes it more formal. There are differences in the language use between problem-based and non-problem based projects. The existence of a common code and the mutual understanding in communities was confirmed. Existence of emotion among community members and their positive attitude was confirmed as well.

6 Conclusion

Applying linguistic analysis to the discourse among the members of a VC can provide us with useful results. Combining the result of the two categories (inner speech and collaboration) we can get strong indications of community existence. Furthermore, results of the analysis can help us improve the design of the VCs. However there is room for future research, e.g. applying this model and evaluating it on a larger corpus and different case studies.

References

Bielaczyc Katerine and Collins Allan. 1999. Learning communities in classrooms: a reconceptualization of educational practice. In C. M. Reigeluth (Ed.): *Instructional design theories and models*, vol. II, Mahwah NJ: Lawrence Erlbaum Associates.

Blau R. Susan, Hall John, and Strauss Tracy. 1998. Exploring the Tutor/Client Conversation: A Linguistic Analysis. *The Writing Center Journal*, Volume 19, Number 1, Fall/Winter 1998.

Brooks Michael, Kuksenok Katie, Torkildson Megan K., Perry Daniel, Robinson John J., Scott Taylor J., Anicello Ona , Zukowski Ariana, Harris Paul and Aragon Cecilia R. 2013. Statistical Affect Detection in Collaborative Chat. *CSCW '13 Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 317-328.

Daniel Ben, Schwier A. Richard and McCalla Gordon. 2003. Social Capital in Virtual Learning Communities and Distributed Communities of Practice. *Canadian Journal of Learning and Technology / La revue canadienne de l'apprentissage et de la technologie*, [S.l.], Oct. 2003.

Dillenbourg P. and Fischer F. 2007. Basics of Computer-Supported Collaborative Learning. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 21, pp. 111-130.

Emerson, Caryl. 1983. The Outer Word and Inner Speech: Bakhtin, Vygotsky, and the Internalization of Language. *Critical Inquiry*, 10 (2):245, pp. 245-264.

Innes B. Robert. 2007. Dialogic Communication in Collaborative Problem Solving Groups. *International Journal for the Scholarship of Teaching and Learning*, 1(1).

Knipfer Kristin, Mayr Eva, Zahn Carmen, Schwan Stephan and Hesse Friedrich W. 2009. Computer support for knowledge communication in science exhibitions: Novel perspectives from research on collaborative learning. *Educational Research Review*, Volume 4, Issue 3, pp. 196-209.

Kollar Ingo, Fischer Frank and Slotta James D. 2005. Internal and external collaboration scripts in web-based science learning at schools. *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years! (CSCL '05), International Society of the Learning Sciences*, pp. 331-340.

Koschmann Timothy. 1999. Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. *CSCL '99 Proceedings of the 1999 conference on Computer support for collaborative learning*, Article No. 38.

Mairesse Francois, Walker A. Marilyn, Mehl R. Matthias, and Moore K. Roger. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, 2007, pp. 457-500

Maness M. Jack. 2008. A Linguistic Analysis of Chat Reference Conversations with 18–24 Year-Old College Students. *The Journal of Academic Librarianship*, Vol. 34, Iss. 1, January 2008, pp. 31-38.

McMillan W. David and Chavis George M. David. 1986. Sense of Community: A Definition and Theory. *Journal of Community Psychology*, Volume 14.

Pérez-Sabater Carmen. 2012. The Linguistics of Social Networking: A Study of Writing Conventions on Facebook. *Linguistik online*, 56, 6/2012.

Roussos L. Petros and Tsaousis Giannis. 2006. *Applied Statistics in Social Sciences*. Ellinika Grammata, Athens. (In Greek)

Rovai P. Alfred. 2002. Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. *The Internet and Higher Education*, Volume 5, Issue 4, pp. 319-332.

Socolov N. A. 1972. *Inner Speech and Thought*. Plenum, New York, pp. 46-122.

Spanger Phillip, Masaaki Yasuhara, Ryu Iida and Takenobu Tokunaga. 2009. A Japanese corpus of referring expressions used in a situated collaboration task. *Proceedings of the 12th European Workshop on Natural Language Generation*, pp. 110-113.

Stahl Gerry. 2000. A Model of Collaborative Knowledge-Building. In B. Fishman and S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences*, (pp. 70-77), Mahwah, NJ: Erlbaum.

Stahl G., Koschmann T. and Suthers D. 2006. Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409-426), Cambridge, UK: Cambridge University Press.

Veermans Marjaana and Cesareni Donatella. 2005. The nature of the discourse in web-based Collaborative Learning Environments: Case studies from four different countries. *Computers & Education*, Vol.45, Iss.3, November 2005, pp. 316–336.

Vygotsky S. Lev. 2008. *Thought and Language*. Gnosi, Athens, pp. 378-436. (In Greek)

Warschauer Mark. 1997. Computer-Mediated Collaborative Learning: Theory and Practice. *The Modern Language Journal*, Volume 81, Issue 4, pp. 470–481, Winter 1997.

Wells Gordon. 2002. The Role of Dialogue in Activity Theory. *Mind, Culture, and Activity*, Vol.9, Iss.1, pp. 43-66.

Wiley Norbert. 2006. Inner Speech as a Language: A Saussurean Inquiry. *Journal for the Theory of Social Behaviour*, Volume 36, Issue 3, pp. 319–341, September 2006.

Zhang Jianwei, Scardamalia Marlene, Lamon Mary, Messina Richard and Reeve Richard. 2007. Socio-cognitive dynamics of knowledge building in the work of 9- and 10-year-olds. *Educational Technology Research and Development*, April 2007, Vol. 55, Iss. 2, pp. 117-145.

Appendices

IS1	Omission of Subjects	IS2	Omission of Conjunction
IS3	Informal clauses	IS4	Omission of verbs
IS5	Elliptical clauses	IS6	Words per clause
IS7	Words per period	IS8a	Parenthesis
IS8b	Commas	IS8c	Question marks
IS8d	Dots	IS8e	Exclamation Marks
IS8f	Full stops	IS8g	Punctuation (total)
IS9a	Adverbs	IS9b	Adjectives
IS9c	Greeklish	IS9d	Informal words
IS9e	Emoticons	IS10a	Advs of place
IS10b	Advs of time	IS10c	Advs of manner
IS10d	Advs of certainty	IS10e	Quantitative advs
IS10f	Interrogative advs	IS10g	Relative advs
IS10h	Viewpoint & commenting advs	IS11a	Subordinate clauses
IS11b	Prepositional phrases	IS12a	Definite articles/total
IS12b	indefin articles/total	IS12c	Articles/total words
IS12d	Articles/periods	IS13	Apposition
IS14	Epexegesis	IS15	Additional terms in genitive case
IS16	Additional terms in accusative case	IS17	Abbreviations
IS18	Metaphors	IS19	Similes
IS20	Word variety		

Appendix A. Inner Speech Analysis summary

CA1	Verbs in 1st plural person	CA2	Emotional clauses
CA3	Rewarding clauses	CA4	Clauses of negation
CA5	Clauses of reason	CA6	Familiarity words
CA7	Inclusive words	CA8	Social words
CA9	Emotional words	CA10	Positive emotion
CA11	Negative emotion	CA12	Use of 1st person plural pronouns
CA13	Use of 1st person plural pronouns		

Appendix B. Collaboration Analysis summary

Feature id	*VC 1.1*	*VC 1.2*	*VC 2*	*VC 3*	*ST*
CA-1	**0,76***	**0,43***	**0,31***	**0,03**	0,16
CA-2	**0,33***	***0,21***	***0,22***	0,10	0,05
CA-3	0,02	**0,16***	***0,16***	0,06	0
CA-4	**0***	0,03	0,07	0,06	0,05
CA-5	**0,29***	0,02	0,01	0,01	0,02
CA-6	**0,33***	**0,05***	**0,02**	**0,04**	0
CA-7	0,01	0	0	0	0
CA-8	**0,06**	0	0,01	0,02	0
CA-9	**0,18***	**0,10***	**0,07**	0,04	0,02
CA-10	**1,00***	0,55	**0,94**	0,68	0,65
CA-11	0*	0,45	**0,06**	0,32	0,35
CA-12	**0,33**	0,44	**0,68***	0	0,08
CA-13	0,34	**0,47**	**0,70***	0	0,10

Appendix C. Results for collaborative linguistic analysis model

(Bold indicates significance at p<,05 level, italics at p<,02 level and asterisks at p<,01 level)

Feature id	*VC 1.1*	*VC 1.2*	*VC 2*	*VC 3*	*ST*
IS-1	**0,88**	**0,92***	**0,85**	0,51	0,67
IS-2	**0**	0,02	0,12	0,14	0,03
IS-3	***0,23***	**0,75***	**0,58***	**0,69***	0
IS-4	0,04	**0,21***	**0,25***	**0,14**	0,01
IS-5	0,12	**0,57***	**0,34***	**0,61***	0,04
IS-6	**4,04***	**4,27***	**4,62***	4,79	5,37
IS-7	12,35	**7,08***	**7,03***	**8,09***	12,82
IS-8.a	0	0	0,03	0,01	0
IS-8.b	**0,02***	**0,08***	**0,02***	**0,14**	0,18
IS-8.c	0,02	0,03	0,20	**0,07**	0
IS-8.d	0,02	0,03	0,02	0,02	0
IS-8.e	0,52	**0,28***	0,15	0,06	0,02
IS-8.f	**0,13***	**0,26***	**0,24***	0,25	0,40
IS-8.g	0,71	0,68	0,65	0,54	0,61
IS-9.a	0,12	0,07	0,06	0,06	0,06
IS-9.b	**0,01***	0,07	***0,04***	0,08	0,09
IS-9.c	0	0	0,13	0,01	0
IS-9.d	0,02	**0,06***	**0,07***	***0,08***	0
IS-9.e	0	0	0	0	0
IS-10.a	0,16	0,18	0,33	0,26	0,16
IS-10.b	**0***	0,12	**0***	**0,09**	0,34
IS-10.c	0,28	0,24	**0,07***	0,26	0,23
IS-10.d	0	0	0	0	0
IS-10.e	***0,56***	0,39	0,30	0,33	0,27
IS-10.f	0	0	0	0	0
IS-10.g	0	0,03	0	0	0,01
IS-10.h	0	0,03	***0,30***	0,05	0
IS-11.a	0,29	**0,12**	0,14	**0,07***	0,23
IS-11.b	0,25	***0,21***	0,25	0,29	0,36
IS-12.a	**1,00***	**0,98**	**0,97***	0,94	0,88
IS-12.b	**0***	**0,02**	**0,03***	0,06	0,12
IS-12.c	***0,09***	0,12	**0,12***	0,16	0,15
IS-12.d	**1,12**	**0,84***	**0,87***	**1,26***	1,90
IS-13	0	0	0	0	0
IS-14	0,02	0	0,01	0,01	0,01
IS-15	**0***	**0,01***	**0***	0,02	0,03
IS-16	**0***	**0***	**0***	**0***	0,01
IS-17	0	0	0,02	0,04	0
IS-18	**0***	**0***	0,04	0,12	0,04
IS-19	***0***	***0***	***0***	***0***	0,02
IS-20	0,34	0,39	0,36	0,38	*0,50*
IS-20	*Average of VCs:* **0,37***				*0,50*

Appendix D. Results for IS linguistic analysis model

(Bold indicates significance at p<,05 level, italics at p<,02 level and asterisks at p<,01 level)

10.12871/CLICIT2014154

Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare.

Maria Palmerini
Cedat 85
m.palmerini@cedat85.com

Renata Savy
DipSUm / Lab.L.A. Università di Salerno
rsavy@unisa.it

Abstract

Italiano. *Il lavoro presenta i risultati di un lavoro di classificazione e analisi linguistica degli errori di un sistema di riconoscimento automatico (ASR), prodotto da Cedat'85. Si tratta della prima fase di una ricerca volta alla messa a punto di strategie di riduzione dell'errore.*

English. *The research project aims to analyze and evaluate the errors generated by Cedat 85's automatic speech recognition system (ASR), in order to develop new strategies for error reduction. The first phase of the project, which is explored in this paper, consists of a linguistic annotation, classification and analysis of errors.*

1 Introduzione

Il progetto di ricerca è nato da una collaborazione fra l'Università di Salerno e Cedat 85, azienda leader in Italia nel settore del trattamento automatico del parlato. Lo scopo del progetto è una valutazione accurata degli errori prodotti da un sistema di trascrizione automatica del parlato (ASR), passati al setaccio di una più fine analisi linguistica e successiva metadatazione.
La stima più utilizzata del *word error rate* (WER) di un sistema ASR è calcolata in maniera automatica e si basa sull'analisi di una trascrizione manuale (allineata al segnale) e la relativa trascrizione ottenuta dal sistema ASR. Su questo confronto vengono individuate le parole errate (*Substitutions*), quelle mancanti (*Deletetions*) e quelle erroneamente inserite (*Insertions*) nonché le parole totali (N) per una valutazione:

$$WER = \frac{(S+D+I)x100}{N}$$

Questa stima non entra nel merito della causa né della rilevanza dell'errore, costituendo piuttosto un riferimento di massima per una valutazione grossolana di un sistema ASR, senza alcuna indicazione sulla sua reale utilità e adeguatezza, né sulle possibilità di intervento e miglioramento.
Gran parte dei sistemi ASR di ultima generazione, che lavorano su parlato spontaneo, utilizzano tecnologie ed algoritmi che possono sfruttare al meglio l'enorme potenza di calcolo attualmente disponibile, ma differiscono in modo rilevante nella scelta dei parametri, dei passi intermedi, nei criteri di selezione dei candidati più probabili, negli strumenti per il trattamento dei dati di addestramento. Un criterio 'qualitativo', oltre che quantitativo, di valutazione degli errori si rende necessario per un adeguamento del sistema all'ambiente di riferimento, e per l'indicazione su eventuali interventi migliorativi.
Studi recenti, sia di ambito tecnologico che linguistico e psicolinguistico, indicano correlazioni tra errori e frequenza nel vocabolario o nell'uso delle parole, velocità d'eloquio, ambiguità (omofonia) e confondibilità acustica (coppie minime e sub-minime). Mancano tuttavia studi sistematici che prendano in considerazione la correlazione con classi morfo-lessicali, strutture fonologiche e sillabiche, sequenze sintagmatiche, ordine dei costituenti e soprattutto, fattori prosodici.
In questo contributo presentiamo una prima parte dei risultati di una ricerca più ampia sul peso di questi fattori, soffermandoci sui criteri della classificazione linguistica dei dati e sulle correlazioni ottenute tra presenza (e tipo) di errore e categorie fono-morfologiche e morfo-sintattiche.

2 Corpus e metodo di analisi

Cedat 85 ha messo a disposizione un corpus di registrazioni audio (che chiameremo *test set*, v. §2.2) con relative trascrizioni manuali e trascrizioni prodotte automaticamente dal proprio sistema ASR. Su questi dati è stato calcolato il *word error rate* (WER) in modo automatico, grazie al tool Sclite, componente del pacchetto *Speech Recognition Scoring Toolkit* (SCTK) realizzato dal *National Institute of Standards and Technology* (NIST).
Sono inoltre stati messi a disposizione il *phone set* e il dizionario utilizzati dal sistema ASR.

2.1 Il sistema ASR

Il sistema per il riconoscimento automatico del parlato continuo di Cedat 85 è un sistema di ultima generazione, *speaker independent* (che quindi non richiede addestramento specifico sulla singola voce), basato su modelli statistici di tipo markoviano[1]. Nel sistema ASR analizzato la decodifica del parlato avviene grazie a due moduli che interagiscono fra loro: un 'modello acustico', deputato al riconoscimento dei suoni significativi all'interno del segnale, e un 'modello di linguaggio', cui spetta l'individuazione di parole singole (unigrammi) e sequenze di parole (bigrammi e trigrammi). Entrambi i moduli si basano su un dizionario (lessicale e fonologico). I modelli acustici per la lingua italiana sono stati addestrati su centinaia di ore di parlato proveniente da vari ambienti sia microfonici, sia telefonici. Sono stati messi a punto diversi modelli di linguaggio, dal politico al televisivo, dalle lezioni universitarie al giudiziario.

2.2 Il test set

Il *test set* sottoposto ad analisi è suddiviso in 4 *subset* appartenenti a 4 diversi domini; 3 di tipo microfonico (politico, televisivo, giudiziario) e uno di tipo telefonico (sms vocali e telefonate di call center). I subset microfonici ammontano a circa 25min. di parlato ognuno, mentre il subset telefonico è composto da 109 messaggi vocali e 20 min. circa di interazioni di call center.
Su tale *test set* è stato calcolato il WER, suddiviso nelle tre categorie di errori: *Insertion* (I), *Deletion* (D), *Substitution* (S).

2.3 Metodo di classificazione

L'indagine è stata svolta in 3 fasi. Nella fase preliminare le categorie del WER sono state scorporate sui 4 diversi domini.
Nella seconda fase si è proceduto alla catalogazione degli errori per ogni dominio secondo il sistema di metadatazione linguistica (descritto in §3). L'analisi uditiva è stata corredata da una minima osservazione spettrografica. Per ciascuna stringa è stato effettuato il confronto puntuale tra le due trascrizioni per ogni item marcato da errore; l'etichettatura ha riguardato sempre l'elemento del *Reference text* (trascrizione manuale), fatta eccezione per i casi di 'inserzione' in cui è stato marcato l'elemento inserito dal sistema automatico. A valle dell'etichettatura, sono stati scorporati dal WER tutti i casi di 'falso errore', attribuibili a incomprensione o refusi del trascrittore umano. Il calcolo delle correlazioni riguarda quindi il corpus 'epurato'.
Infine, in una terza fase è stato effettuato un *PoS-tagging* di tutti i testi di riferimento dei 4 subset, allo scopo di 'pesare' i dati delle correlazioni individuate tra errore e categorie lessicali e ricavare indicazioni più puntuali e impiegabili per future ottimizzazioni del modello.

3 Il sistema di annotazione

Il modello di annotazione linguistica è stato progettato dal Laboratorio di Linguistica Applicata dell'Università di Salerno, mettendo a punto un sistema di metadatazione che prende in esame diverse caratteristiche. Schematicamente possiamo distinguere tra tre tipi di categorizzazione: 1) lessicale (*Pos*), ulteriormente articolata al suo interno; 2) 'morfologica' (implicata esclusivamente per alcune PoS); 3) 'fonetico-fonologica'. Di seguito si presenta l'elenco delle categorie del modello e relativi valori che ognuna può assumere. Tutte le etichette si riferiscono alle parole grafiche (unigrammi) considerate dal sistema.

Error Type: indica il tipo di errore secondo il sistema di misurazione automatica; può assumere i valori di I(nsertion), D(eletion), S(ubstitution).

Error Category: indica la categoria lessicale della parola oggetto dell'errore; assume i valori di Noun (N), Verb (V), Adjective (Adj), Adverb (Adv), FunctionWord (FW) and Other (O); quest'ultima categoria marca fenomeni di disfluenza, ripetizioni, false partenze e simili.

Error Subcategory: prevede una sottocategorizzazione sintattico-semantica delle PoS maggiori e una capillare descrizione delle parole funzionali, delle esitazioni e altri fenomeni (*marcatori discorsivi, false partenze, autocorrezioni, ripetizioni, pause piene, lapsus, errate pronunce*).

Verb + Clitics: assume valore 'True' (T) nel caso in cui il target dell'errore sia una forma verbale con clitico pronominale (es: *dimmi*).

Derivate: indica se il target dell'errore in questione è una parola derivata, e quindi presenta affissazione; i valori possibili per questo campo sono 'P','S' e 'P+S'.

Position: riferisce la posizione di Avverbio rispetto a Verbo e Aggettivo rispetto a Nome; assume valori 'Pre' e 'Post'.

Morphological Complexity: indica il grado di composizione morfologica della parola target secondo una 'scala di morfo-complessità' calcolata partendo dal lessema-base e aggiungendo +1 per ogni nuovo morfema, ad esempio:

industria	1
industri-ale	2
industri-al-izzare	3
industri-al-izza-zione	4
de-industri-al-izza-zione	5

[1] Il sistema è attualmente impiegato in numerose applicazioni e servizi già commercializzati da Cedat 85.

Phonological Length: indica la lunghezza in fonemi del target di errore, basata sulla trascrizione fonologica del vocabolario di riferimento.

Syllabic Length: indica la lunghezza in sillabe fonologiche del target di errore.

Accentual Type: indica il tipo accentuale del target di errore: tronco, piano, sdrucciolo, bisdrucciolo.

Omophones: indica la possibile esistenza di omofoni per la parola target; assume valori booleani (t/f).

Minimal Pairs: indica la possibile esistenza di coppie minime con la parola target; assume valori booleiani (t/f).

Alcune delle categorie sopra elencate presentano evidenti correlazioni in partenza: la presenza di clitico pronominale sul verbo o di affissazione, ad esempio, implica complessità morfologica e può comportare maggiore lunghezza fonologica e sillabica, nonché influenzare il tipo accentuale. Ciononostante, ogni parametro è stato valutato separatamente, per poter *a posteriori* verificare la concomitanza di più fattori critici.

4 Primi risultati

In questa prima analisi dei risultati riportiamo solo le correlazioni rivelatesi significative e soprattutto adeguate ad avanzare ipotesi utili per indirizzare le indagini successive. I valori nelle tabelle si intendono come percentuali sul totale degli errori del corpus di controllo.

La prima verifica linguistica riguarda la distribuzione dell'errore nelle diverse categorie lessicali, che mostra una situazione omogenea, diversa solo per il dominio telefonico.

	N	V	ADJ	ADV	FW	O
politico	11,2	11,6	5,1	3,8	**29,3**	**38,7**
media	15,8	18,7	2,5	3,2	**25,7**	**34,2**
giustizia	7,7	17,7	2,6	3,5	**33,2**	**35,2**
telefonico	17,6	21,4	3,6	8,1	**33,8**	15,3

Tabella 1. Distribuzione di Error category nei 4 subset.

I dati in tab.1[2] evidenziano una pesante concentrazione dell'errore per la classe delle parole funzionali (FW) e delle produzioni disfluenti (O), oscillante tra il 30 e 38%. Tra le parti variabili del discorso sono scarsamente affetti da errore aggettivi e avverbi (fatta eccezione per il corpus telefonico), mentre una percentuale leggermente più alta si registra nella classe dei e, per i corpora TV e Telefonico, anche per la classe dei nomi.

I successivi dati significativi ci sembra riguardino la correlazione tra percentuale di errore e complessità morfologica, sillabica e fonologica (le ultime valutate in termini di 'lunghezza'). Le tabb.2 e 3 riportano in dettaglio i dati delle prime due categorie (mentre è più difficile riassumere i dati sulla lunghezza fonologica, altamente variabile e disomogenea):

	0	1	2	3	4	5
politico	**39,1**	**38,7**	21,4	0,8	-	-
media	**29,6**	**51,8**	14,1	3,9	0,7	-
giustizia	**35,2**	**34,3**	**24,4**	5,4	0,6	-
telefonico	10,3	**42,2**	**38,0**	9,2	0,2	0,2

Tabella 2. Distribuzione del WER nella categoria Morpho_complex dei 4 subset (con valore di morfocomplessità 0 sono state indicate le esitazioni e i fenomeni di disfluenza)**.**

Appare netta, dunque, un'elevata concentrazione di errori per le parole a bassa complessità morfologica (0-2), mentre quasi nulla per parole con valore di complessità morfologica superiore a 5.

	1	2	3	4	5	6
politico	10,2	**33,9**	**28,8**	13,6	10,2	3,4
media	11,1	**35,6**	17,8	15,6	**20,0**	-
giustizia	8,3	**33,3**	**38,9**	8,3	8,3	-
telefonico	14,3	**43,9**	**26,5**	9,2	4,08	2

Tabella 3. Distribuzione del WER nella categoria Syllabic length dei 4 subset.

In ultimo, sembra emergere una tendenza dell'errore (con poche eccezioni) a diminuire in modo direttamente proporzionale all'aumentare della lunghezza della parola: le parole bi- e trisillabiche concentrano, in media, oltre il 30% di errori per tutti i corpora; solo le parole monosillabiche contrastano questa tendenza generale. I dati sulla lunghezza fonologica indicano più affette da errore le parole costituite da 1 a 5 fonemi (fin oltre il 60% per quelle monofonematiche).

L'errore, dunque, si concentra sulle parole di lunghezza medio-bassa e a ridotta complessità morfologica, per ridursi poi in modo significativo nelle parole più complesse e più lunghe. Le due categorie PoS maggiormente affette da errore di riconoscimento (FW e O) sono, infatti, anche quelle che correlano con bassi o nulli valori di complessità morfologica e numero di fonemi.

Un ulteriore conteggio si rende però necessario per valutare il peso e l'incidenza del WER sulle diverse categorie lessicali. In tabella 4 riportiamo i dati di frequenza delle diverse PoS rispetto all'intero corpus, mentre in tabella 5 le percentuali di errore ricalcolate su questo insieme:

	N	V	ADJ	ADV	FW	O
politico	**23,3**	14,9	10,3	7,5	**35,8**	8,2
media	**28,5**	15,9	9,2	6,5	**36,7**	3,1
giustizia	**20,3**	**20,3**	7,0	9,8	**36,3**	6,2
telefonico	**21,7**	19,5	6,9	13,1	**31,2**	7,7

Tabella 4. Dati del PoS tagging su tutte le parole dei 4 subset.

	N	V	ADJ	ADV	FW	O
politico	7,3	11,8	7,6	7,7	12,4	**73**
media	4.0	8,5	2,0	3,5	5,1	**82,3**
giustizia	5,5	12,4	5,2	5,0	13,0	**83,3**
telefonico	**27,3**	**38,8**	**17,5**	**20,7**	**36,4**	**66,2**

Tabella 5. Incidenza dell'errore ricalcolata sul totale delle parole del corpus divise in categorie.

[2] Le tendenze regolari sono segnalate in grassetto, mentre le celle ombreggiate evidenziano dati in controtendenza.

Le PoS maggiormente affette da errore (FW e O, tab.1) hanno distribuzione frequenziale molto diversa nel corpus (tab.4): le prime, com'era prevedibile, mostrano un alto numero di occorrenze (con frequenza >30%, direttamente seguite dai Nomi); le seconde, invece, sono poco frequenti rispetto al totale delle parole del *test set* (solo il 3-8%). Ne deriva che l'incidenza dell'errore (tabella 5) è molto più significativa nel secondo caso, raggiungendo livelli anche molto maggiori dei 2/3 degli items (tra il 66 e l'83% del totale).

5 Considerazioni preliminari

Sebbene i risultati sopra esposti rappresentino un'elaborazione parziale dei dati dell'analisi del WER condotta nella ricerca, essi consentono di avanzare alcune considerazioni preliminari a future e più approfondite valutazioni.

In primo luogo, volendo misurare globalmente l'efficienza del sistema di trascrizione basato su ASR, occorre interpretare i dati inclusi in tabella 5, che mostrano percentuali di errore basse o trascurabili, comprese tra il 2% e il 13%, equamente suddivise per tutte le PoS. Fa eccezione il dominio 'telefonico' (per il quale v.oltre). Se una buona parte del WER complessivo (>25%) incide sulla categoria delle FW di un testo (tab.1), è pur vero che essa ha valori di frequenza altissimi che normalizzano l'incidenza dei mancati riconoscimenti del sistema, rendendola comparabile ad altre PoS, nonostante la loro minore complessità morfologica ed estensione fonologica.

Questo dato è d'altronde coerente col funzionamento del sistema ASR, nel quale agiscono, compensandosi, il modello acustico, che riconosce con maggiore accuratezza parole dotate di maggior 'corpo fonico', e il modello di linguaggio, che fornisce miglior supporto sulle stringhe di parole più ricorrenti, riuscendo ad integrare il riconoscimento di parole grammaticali dove l'informazione acustica è più carente (anche per fenomeni di coarticolazione e ipoarticolazione).

Una valutazione diversa va riservata ai Nomi, che mostrano un comportamento parzialmente oscillante: concentrano, infatti, percentuali variabili del WER (tab.1), anche se la loro incidenza appare normalizzata nel rapporto tra loro frequenza assoluta (22-28% sull'intero corpus) e i casi di mancato riconoscimento (tra il 4 e il 7%). In ogni caso, come classe aperta, essi sono in genere meno prevedibili e maggiormente specifici rispetto a ciascun dominio: richiedono pertanto una massiccia 'personalizzazione' del vocabolario (implementazione effettuata con addestramento sullo specifico dominio), più semplice su alcuni domini a lessico meno variabile (politico e giudiziario), più aleatoria su domini più liberi.

Risulta così che un'incidenza davvero significativa del WER si ottiene unicamente nella classe etichettata come O(ther) che racchiude in genere fenomeni di disfluenza del parlato costituiti da espressioni non lessicali, esitazioni, parole interrotte o mal pronunciate; elementi non inclusi nel vocabolario né considerati nel modello di linguaggio e quindi soggetti a errori di riconoscimento quasi per *default.* Va considerata, inoltre, l'alta variabilità delle possibili forme che essi assumono nella trascrizione ortografica manuale, dove è inevitabile un elevato tasso di interpretazione e resa grafica soggettiva, in mancanza di un modello di trascrizione standardizzato. Dal confronto tra queste rese variabili e il tentativo del sistema ASR di associarle ad entrate del vocabolario acusticamente più 'vicine' deriva l'alto tasso di WER ad esse associato (>35% del WER complessivo, >66% sul totale delle occorrenze).

A parte quest'ultimo dato, dunque, l'errore non sembra essere correlato significativamente a particolari categorie lessicali, quanto piuttosto all'estensione e al 'corpo' delle parole: unità lessicali più estese, infatti, contengono maggiori informazioni acustiche e devono competere con un minor numero di candidati simili.

6 Conclusioni e sviluppi successivi

A valle di questa preliminare fase di analisi ci sembra si possa azzardare una prima conclusione importante: la valutazione quantitativa del *word error rate* sovrastima le falle di riconoscimento di un sistema ASR. La metadatazione linguistica effettuata e la successiva valutazione qualitativa normalizza i dati del WER e reindirizza la maggior quota verso fenomeni non lessicali, imprevedibili quanto poco significativi per la misura dell'efficienza del sistema. In quest'ambito, oltretutto, l'indecisione e la confusione di resa grafica sono pressoché pari per la trascrizione automatica e quella manuale. Ciò nonostante, il peso degli errati riconoscimenti di questi segmenti può essere ridotto adottando uno schema di annotazione più fine, sia in termini di norme più salde per i trascrittori, sia come modello per il sistema ASR. Ci limitiamo infine a ipotizzare che alcuni secondari interventi sul *phone set*, l'arricchimento del vocabolario con le varianti fonetiche possibili, e un migliore trattamento dei fenomeni prosodici potrebbero migliorare di qualche grado le prestazioni del sistema.

References

Daniel Jurafsky, James H. Martin. 2009. *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, second edition, New Jersey, Pearson, Prentice Hall.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003). Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. *IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, US Virgin Islands.

Patti Price. 1990. Evaluation of Spoken Language System: the ATIS domain. *Proceedings of DARPA Speech and Natural Language Workshop*, Hidden Valley, PA.

10.12871/CLICIT2014155

"Il Piave mormorava…": Recognizing Locations and other Named Entities in Italian Texts on the Great War

Lucia Passaro
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica,
University of Pisa (Italy)
lucia.passaro@for.unipi.it

Alessandro Lenci
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica,
University of Pisa (Italy)
alessandro.lenci@ling.unipi.it

Abstract

English. Increasing amounts of sources about World War I (WWI) are nowadays available in digital form. In this paper, we illustrate the automatic creation of a NE-annotated domain corpus used to adapt an existing NER to Italian WWI texts. We discuss the annotation of the training and test corpus and provide results of the system evaluation.

Italiano. *Negli ultimi anni, si sono resi disponibili in formato digitale un numero sempre maggiore di materiali riguardanti la Prima Guerra Mondiale. In questo lavoro illustriamo la creazione automatica di un corpus di addestramento per adattare un NER esistente a testi italiani sulla Prima Guerra Mondiale e presentiamo i risultati della valutazione del nostro sistema addestrato sul nuovo corpus.*

1 Introduction

Increasing amounts of sources about World War I (WWI) are nowadays available in digital form. The centenary of the Great War is also going to foster this trend, with new historical sources being digitized. This wealth of digital documents offers us an unprecedented possibility to achieve a multidimensional and multiperspectival insight on war events, understanding how soldiers and citizens of different countries and social conditions experienced and described the events in which they were involved together, albeit on opposite fronts and with different roles. Grasping this unique opportunity however calls for advanced methods for the automatic semantic analysis of digital historical sources. The application of NLP methods and tools to historical texts is indeed attracting growing interest and raises interesting and highly challenging research issues (Piotrowsky 2012).

The research presented in this paper is part of a larger project dealing with the digitization and computational analysis of Italian War Bulletins of the First World War (for details see Boschetti et al. 2014). In particular, we focus here on the domain and language adaptation of a Named Entity Recognizer (NER) for Italian. As a byproduct of this project, we illustrate the automatic creation of a NE-annotated domain corpus used to adapt the NER to the WWI texts.

War bulletins (WBs) were issued by the Italian Comando Supremo "Supreme Headquarters" during WWI and WWII as the official daily report about the military operations of the Italian armed forces. They are plenty of Named Entities, mostly geographical locations, often referring to small places unmarked in normal geographic maps or with their name changed during the last century because of geopolitical events, hence hardly attested in any gazetteer.

To accomplish the Named Entity Recognition task, several approaches have been proposed such as Rule Based Systems (Grover et al., 2008; Mikheev et al., 1999a; Mikheev et al., 1999b), Machine Learning based (Alex et al., 2006; Finkel et al., 2005; Hachey et al., 2005; Nissim et al., 2004, including HMM, Maximum Entropy, Decision Tree, Support Vector Machines and Conditional Random Field) and hybrid approaches (Srihari et al., 2001). We used a Machine Learning approach to recognize NEs.

Rule-based systems usually give good results, but require long development time by expert linguists. Machine learning techniques, on the contrary, use a collection of annotated documents for training the classifiers. Therefore the development time moves from the definition of rules to the preparation of annotated corpora.

The problems of the NER in WWI bulletins are larger than those encountered in modern texts. The language used in such texts is early

20th century Italian, which is quite different from contemporary Italian in many respects and belongs to the technical and military jargon. These texts are therefore difficult to analyze using available Italian resources for NER, typically based on contemporary, standard Italian. Grover et al. (2008) describe the main problems encountered by NER systems on historical texts. They evaluated a rule-based NER system for person and place names on two sets of British Parliamentary records from the 17th and 19th centuries. One of the most important issues they had to deal was the gap between archaic and contemporary language.

This paper is structured as follows: In section 2, we present the CoLingLab NER and in section 3 we describe its adaptation to WWI texts.

2 The CoLingLab NER

The CoLingLab NER is a NER for Italian developed with the Stanford CoreNLP NER (Finkel et al., 2005). The Stanford NER, also known as CRFClassifier, is a Java implementation of Named Entity Recognizer (NER) available for download under the GNU General Public License.

The classification algorithm used in Stanford NER is a Conditional Random Field (CRF) as in Lafferty et al. (2001). This model represents the state of the art in sequence modeling, allowing both a discriminative training, and a calculation of a flow of probability for the entire sequence.

The CoLingLab NER was trained on I-CAB (Italian Content Annotation Treebank), a corpus of Italian news, annotated with semantic information at different levels: Temporal Expressions, Named Entities, relations between entities (Magnini et al., 2006). I-CAB is composed of 525 news documents taken from the local newspaper 'L'Adige'. (Time span: September-October, 2004). The NE classes annotated in this corpus are: Locations (LOC), Geo-Political Entities (GPE), Organizations (ORG) and Persons (PER).

Entity	P	R	F1	TP	FP	FN
B-GPE	0.828	0.765	0.795	870	181	267
B-LOC	0.767	0.295	0.426	46	14	110
B-ORG	0.726	0.65	0.684	834	315	455
B-PER	0.881	0.82	0.85	1892	255	413
I-GPE	0.73	0.583	0.649	84	31	60
I-LOC	0.833	0.366	0.508	30	6	52
I-ORG	0.556	0.442	0.493	192	153	242
I-PER	0.835	0.862	0.848	891	176	143
MicroAVG	0.811	0.735	0.771	4839	1131	1742

Table 1– CoLingLab NER trained and tested on I-CAB.

Table 1 reports the performance of the CoLingLab NER and Table 2 compares it with other state-of-the-art NER systems for Italian in EVALITA 2011.[1]

Participant	FB1	Precision	Recall
FBK_Alam_ro1	63.56	65.55	**61.69**
UniPi_SimiDeiRossi_ro1	58.19	65.90	52.09
UniPi_SimiDeiRossi_ro2	52.15	54.83	49.72
CoLingLab	**65.66**	**76.96**	59.76
BASELINE	44.93	38.84	53.28

Table 2 – Comparison between the CoLingLab NER and the 3 top models in Evalita (2011)

3 Adapting the NER to WWI bulletins

We use as test corpus (WB1) the Italian WBs of WWI. These texts come from the digitization of bulletins published in "I Bollettini della Guerra 1915-1918", preface by Benito Mussolini, Milano, Alpes, 1923 (pages VIII + 596).

To speed up the creation of the gold standard annotated corpus, these texts were first tagged semi-automatically with the existing NER, and then manually checked by an annotator to fix the incorrect tags and to add missing annotations.

The tagset consists of five entity classes, with begin-internal notation: Locations (LOC; e.g., *Monte Bianco*), Persons (PER; e.g., *Brandino Brandini*), Military Organizations (MIL; e.g., *Brigata Sassari,* Sassari Brigade), Ships (SHP; e.g., *Czepel*), Airplanes (PLN; e.g., *Aviatik*). The final test corpus consists of 1361 bulletins, covering the period from May 24, 1915 up to November 11, 1918 (1282 days).

In particular, the corpus is composed of 189,783 tokens, with the following NE distribution: 24 PER (13 B-PER – 11 I-PER); 19,171 LOC (12,542 B-LOC – 6,629 I-LOC); 38 SHP (33 B-SHP – 5 I-SHP); 1,249 MIL (615 B-MIL – 634 I-MIL); 54 PLN (52 B-PLN – 2 I-PLN). The corpus was automatically POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009) using Support Vector Machines as learning algorithm.

In the following sections, we describe 3 experiments. First, we annotated the WBs using the existing CoLingLab NER trained on I-CAB. Then, we combined the I-CAB resource with WB2, a new domain NE-annotated corpus creat-

[1]The presented results have been produced according to the 'open' modality, therefore, with the possibility of using any type of supplementary data.

ed *ad hoc* from a digitized version of WBs of WWII. In the last experiment, we annotated WB1 with the NER trained on WB2 only. Tables 4, 5 and 6 report the following values to evaluate the performance of the various models: precision (P), recall (R), F1-Measure (F1), true positives (TP), false positives (FP), false negatives (FN), and microaveraged total scores (MicroAVG).

3.1 Features

In all the three experiments we used the same feature sets: morphological and orthographical features, information about the word shape, the part-of-speech tag, named entity tag, and contextual features.

In particular, we trained the models with the following types of features:

Word features. We used two different features: the first one considered next and previous words. For example, in the expression *"capitano [David Frazer]"* (captain David Frazer), the presence of the word "captain" helps to determine that the following words belong to the class PERSON. The second one considered a window of 6 words (3 preceding and 3 following the target word). It is useful to deal with cases like "*capitano di corvetta [Bandino Bandini]* PER" (Lieutenant Commander [Bandino Bandini] PER).

Orthographic features. We considered "word shape" features such as spelling, capital letters, presence of non–alphabetical characters etc.

Linguistic features. We used the word position in the sentence (numeric attribute), the lemma and the PoStag (nominal attribute).

Terms. We employed complex terms as features to train the model. Terms have been extracted with EXTra (Passaro et al., 2014). For example, the expression *"capitano di corvetta"* (Lieutenant Commander) is recognized by the system as a single item.

It is worth stressing that no information from gazetteers was used in the experiments reported below. It is clear that the system could be extended using lists of names of people, military groups, places, planes, and ships taken by several sources.

3.2 Experiment 1

In this experiment we tagged WWI texts using the CoLingLab NER (see Section 2) trained on a modified version of I-CAB in which we merged Locations with Geopolitical Entities, and we mapped I-CAB's Organizations into Military Organizations. Table 3 shows the mapping between I-CAB NEs and WBs NEs, which produced the following distribution of NEs: 10,487 PER (6,955, B-PER – 3,532 I-PER); 5,636 LOC (4,474 B-LOC – 1,162 I-LOC); 8,304 MIL (4,947 B-MIL – 3,357 I-MIL).

I-CAB		WWII-Bulletins	
B-LOC	LOC	LOC	B-LOC
I-LOC			
B-GPE	GPE		I-LOC
I-GPE			
B-ORG	ORG	MIL	B-MIL
I-ORG			I-MIL
B-PER	PER	PER	B-PER
I-PER			I-PER
-	-	SHP	B-SHP
-			I-SHP
-	-	PLN	B-PLN
-			I-PLN

Table 3– Mapping I-CAB and WB2 classes

Table 4 shows the results obtained using this mapped version of I-CAB as training corpus and the bulletins of WWI as test:

Entity	P	R	F1	TP	FP	FN
B-LOC	0.879	0.425	0.573	5327	732	7210
B-MIL	0.056	0.111	0.075	68	1142	541
B-PER	0.005	0.692	0.01	9	1747	4
I-LOC	0.827	0.433	0.568	2116	442	2771
I-MIL	0.077	0.093	0.084	33	395	323
I-PER	0.006	0.37	0.0118	3	498	5
Micro AVG	0.604	0.408	0.487	7556	4956	10950

Table 4–Annotation results using mapped I-CAB

In this experiment, the CoLingLab NER did not achieve good results on WB1. In fact, we can notice a significant decrease in the system ability to identify all kinds of NEs. This is due to the huge difference between the training and the test corpus, both in the language (modern Italian and generalist in I-CAB, archaic and military in WB1) and in the distribution of NEs, which in WB1 are strongly biased towards Locations.

3.3 Experiment 2

Given the unsatisfactory results obtained by annotating WB1 with a NER trained on a corpus from modern standard Italian, we have retrained the classifier using texts more similar to the test corpus.

Since the process of building annotated corpora is very expensive, we created a new automatically annotated training corpus (WB2) in a very fast way. We started from an html version of World War II Bulletins freely available[2] on the

[2]http://www.alieuomini.it/pagine/dettaglio/bollettini_di_guerra

Web, which includes an index containing different classes of NEs attested in the bulletins. The WW II bulletins were automatically downloaded and cleaned of html tags. The NE index was projected on WB2 to create a new training corpus, which was linguistically annotated with the same tools used for WB1. WB2 consists of 1,201 bulletins covering the time span from June 12th 1940 to September 8th 1943 (typically a bulletin per day), for a total of 211,778 tokens. WB2 is annotated with the same five classes as WB1, i.e. PER, LOC, MIL, PLN, and SHP. The class LOC includes both geo-political entities (e.g. *Italia*) and real locations (e.g. *Monte Bianco*), because such distinction was not marked in the original resource we used for the automatic construction of WB2.

We made a first experiment on this dataset using 10-fold cross-validation, obtaining a F1-Measure ~95%. This good result encouraged us to use WB2 as a gold standard to annotate WB1.

The model in the second experiment has been trained on the combination of I-CAB and WB2. Therefore, we mapped I-CAB's classes to WB2 classes as described in Table 3. The results obtained in this experiment are shown in Table 5. The combined corpora allowed us to increase the performances by 19%. It is worth noticing the significant improvement on Locations. This means that the new corpus provides the NER with much more evidence to identify this class. However, this improvement did not affect the recognition of PER and MIL. In these cases, in fact, we can observe a great number of false positives surely due to fact that I-CAB is very biased towards this class. Moreover, some semantic classes are not recognized because of the dearth of examples in the training data.

Entity	P	R	F1	TP	FP	FN
B-LOC	0.886	0.649	0.75	8141	1044	4396
B-MIL	0.174	0.186	0.18	113	537	496
B-PER	0.016	0.846	0.031	11	695	2
I-LOC	0.846	0.579	0.688	2831	517	2056
I-MIL	0.226	0.216	0.221	77	264	279
I-PER	0.02	0.625	0.038	5	250	3
Totals	0.772	0.604	0.678	11178	3307	7328

Table 5 – Annotation results using I-CAB + WB2

3.4 Experiment 3

In the last experiment, we trained our NER only on the WB2 corpus. This has the advantage of containing texts temporally and thematically closer to WB1, and a more balanced proportion of entity types. Results are presented in Table 6. For the sake of comparison with the previous experiments, we only provide a report for Locations, Persons and Military Organizations, leaving aside the identification of the SHP and PLN classes.

Entity	P	R	F1	TP	FP	FN
B-LOC	0.816	0.82	0.818	10279	2312	2258
B-MIL	0.474	0.074	0.128	45	50	564
B-PER	0.151	0.615	0.242	8	45	5
I-LOC	0.783	0.687	0.732	3359	929	1528
I-MIL	0.34	0.0899	0.144	32	57	324
I-PER	0.098	0.625	0.169	5	46	3
Totals	0,8	0.746	0.772	13728	3439	4682

Table 6 – Annotation results using WB2

The global scores obtained in this third experiment are higher than those in the second one, with a much lower amount of FPs per Persons and Military Organizations.

3.5 Discussion

Analyzing the results of the three experiments, the adapted NER performs better for Location names. This may reflect the sparsity of the data in the other classes.

It should be noticed that in the experiments 1 and 2, the number of false positives for persons and military organizations is very high. This seems to be a direct consequence of the different distribution of the observations in I-CAB compared to WBs.

Unsurprisingly, the best performing model is the one that has been entirely domain-tuned.

We are confident that new lexicons and gazetteers could help us to improve the identification of Locations and other Named Entities.

4 Conclusion

Location names play an important role in historical texts, especially in those - like WBs - describing the unfolding of military operations.

In this paper, we presented the results of adapting an Italian NER to Italian texts about WWI through the automatic creation of a new NE-annotated corpus of WBs. The adapted NER shows a significantly increased ability to identify Locations.

In the near feature, we aim at processing other types of texts about the Great War (e.g., letters, diaries and newspapers) as part of a more general project of information extraction and text mining of war memories.

References

Attardi, G., Dell'Orletta, F., Simi, M., Turian, J. (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In Proceedings of EVALITA 2009, Reggio Emilia, Italy.

Bartalesi Lenzi, V., Speranza, M., and Sprugnoli, R. (2011). EVALITA 2011: Description and Results of the Named Entity Recognition on Transcribed Broadcast News Task. In Working Notes of EVALITA 2011, 24-25th January 2012, Rome, Italy.

Boschetti F., Cimino A., Dell'Orletta F:,Lebani G., Passaro L., Picchi P., Venturi G., Montemagni S., LenciA. (2014). Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II. In Proceedings of the LREC 2014 Workshop on "Language resources and technologies for processing and linking historical documents and archives- Deploying Linked Open Data in Cultural Heritage", Reykjawik, Iceland.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy.

Finkel J.R., Grenager T. and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

Grover C., Givon S., Tobin R. and Ball J. (2008). Named Entity Recognition for Digitised Historical Texts. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco.

Hachey B., Alex B., and Becker M. (2005). Investigating the effects of selective sampling on the annotation task. In Proceedings of the 9th Conference on Computational Natural Language Learning.

Lafferty J., McCallum A., and Pereira F.(2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th ICML. Morgan Kaufmann, San Francisco, CA.

Magnini B., Pianta E., Speranza M., Bartalesi Lenzi V. and Sprugnoli V. (2011).ITALIAN CONTENT ANNOTATION BANK (I-CAB): Named Entities.

Nissim M., Matheson C. and Reid J. (2004). Recognising geographical entities in Scottish historical documents. In Proceedings of the Workshop on Geographic Information Retrieval, SIGIR ACM 2004.

Mikheev A., Grover C. and Moens M. (1999a). XML tools and architecture for named entity recognition. Journal of Markup Languages: Theory and Practice, 1(3).

Mikheev A., Grover C. and Moens M. (1999b). Named entity recognition without gazetteers. In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99).

Passaro, L., Lebani, G.E. and Lenci A. (2014). Extracting terms with EXTra. submitted.

Piotrowsky M. (2012).Natural Language Processing for Historical Texts, Morgan & Claypool.

Srihari R., Niu C., and Li W., (2001). A hybrid approach for named entity and sub-type tagging. In Proceedings of the 6th Applied Natural Language Processing Conference, pages 247–254, Seattle.

10.12871/CLICIT2014156

The Importance of Being *sum*. Network Analysis of a Latin Dependency Treebank

Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
marco.passarotti@unicatt.it

Abstract

English. Network theory provides a suitable framework to model the structure of language as a complex system. Based on a network built from a Latin dependency treebank, this paper applies methods for network analysis to show the key role of the verb *sum* (*to be*) in the overall structure of the network.

Italiano. *La teoria dei grafi fornisce un valido supporto alla modellizzazione strutturale del sistema linguistico. Basandosi su un network costruito a partire da una treebank a dipendenze del latino, l'articolo applica diversi metodi di analisi dei grafi, mostrando l'importanza del ruolo rivestito dal verbo* sum *(*essere*) nella struttura complessiva del network.*

1 Introduction

Considering language as a complex system with deep relations between its components is a widespread approach in contemporary linguistics (Briscoe, 1998; Lamb, 1998; Steels, 2000; Hudson, 2007). Such a view implies that language features complex network structures at all its levels of analysis (phonetic, morphological, lexical, syntactic, semantic).

Network theory provides a suitable framework to model the structure of linguistic systems from such a perspective. Network theory is the study of elements, called *vertices* or *nodes*, and their connections, called *edges* or *links*. A complex network is a (un)directed graph $G(V, E)$ which is given by a set of vertices V and a set of edges E (Ferrer i Cancho, 2010).

Vertices and edges can represent different things in networks. In a language network, the vertices can be different linguistic units (for instance, words), while the edges can represent different kinds of relations holding between these units (for instance, syntactic relations).

So far, all the network-based studies in linguistics have concerned modern and living languages (Mehler, 2008a). However, times are mature enough for extending such approach also to the study of ancient languages. Indeed, the last years have seen a large growth of language resources for ancient languages. Among these resources are syntactically annotated corpora (treebanks), which provide essential information for building syntactic language networks.

2 From a Dependency Treebank to a Syntactic Dependency Network

For the purpose of the present study, we use the *Index Thomisticus* Treebank, a Medieval Latin dependency treebank based on the works of Thomas Aquinas (IT-TB; http://itreebank.marginalia.it; Passarotti, 2011). Presently, the IT-TB includes around 200,000 nodes in approximately 11,000 sentences. For homogeneity reasons, in this work we consider the subset of the IT-TB that features the in-line

annotation of the text of the *Summa contra Gentiles* (entire first book and chapters 1-65 of the second one) for a total of 110,224 nodes.

Automatic data cleaning was performed before building the network, by excluding punctuation marks, function words and elliptical dependency relations from the input data. Then, the method developed by Ferrer i Cancho et alii (2004) was applied to build the network.

According to this method, a dependency relation appearing in the treebank is converted into an edge in the network. The vertices of the network are lemmas. Two lemmas are linked if they appear at least once in a modifier-head relation (dependency) in the treebank.

Then a syntactic dependency network is constructed by accumulating sentence structures from the treebank. The treebank is parsed sentence by sentence and new vertices are added to the network. When a vertex is already present in the network, more links are added to it.

The result is a syntactic dependency network containing all lemmas and all dependency relations of the treebank. All connections between particular lemmas are counted, which means that the graph reflects the frequency of connections. The network is an emergent property of sentence structures (Ferrer i Cancho, 2005; Ferrer i Cancho et al., 2004), while the structure of a single sentence is a subgraph of the global network (Bollobás, 1998).

The free software Cytoscape was used for network creation and computing (Shannon et al., 2003; Saito et al., 2012).

Figure 2 presents the syntactic dependency network of the subset of the IT-TB used in this work. Vertices and edges are arranged according to the Edge-weighted Spring Embedded layout setting provided by Cytoscape (Kohl et al., 2011). Edges are weighted by frequency, the most central relations in the network being those most frequent in the treebank.

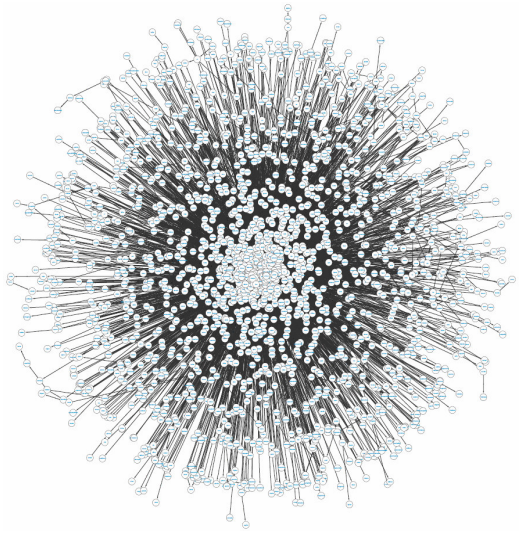


Figure 1. The network of the IT-TB

The drawing in figure 1 is messy and not much informative. In order to both analyze and categorize the network, we use a number of topological indices that are able to unravel fundamental properties of the network that are hidden to the eye.

3 Topological Indices

Most complex networks are characterized by highly heterogeneous distributions (Newman, 2005a). This property means that there are many vertices having a few connections and a few vertices with a disproportionately large number of connections. The most connected vertices in a network are called *hubs* (Albert & Barabási, 2002; Newman, 2003).

In network analysis, the centrality of a vertex is a topological index that measures its relative importance within a graph. We use two measures of centrality ('betweenness' and 'closeness') to calculate the importance of a vertex in a syntactic dependency network, i.e. to find hubs in the network. The higher are betweenness and closeness centralities of a vertex, the more important the vertex is in the network.

The *betweenness centrality* of a vertex *v, g(v)*, is a measure of the number of minimum distance (or "shortest") paths running through *v* (Ferrer i Cancho et al., 2004).

Closeness centrality. In a network, the length of the shortest paths between all pairs of vertices is a natural distance metric. The "farness" of a vertex *s* is the sum of its distances to all other vertices, and its "closeness" is the inverse of the farness (Sabidussi, 1966). Thus, the more central a vertex is, the lower is its total distance to all other vertices. Closeness centrality is a measure of how long it takes to spread information from *s* to all other vertices sequentially in the network (Newman, 2005b; Wuchty & Stadler, 2003).

Further, we use the following topological indices in order to categorize a syntactic dependency network by evaluating its complexity (Mehler, 2008b).

The so-called *degree* of a vertex *s* is the number of different relations holding between *s* and other vertices in the network. The *average degree* $\varepsilon(G)$=*edges/vertices* of a graph *G* is the proportion of edges with respect to the number of vertices.

Clustering coefficient is the probability that two vertices that are neighbours of a given vertex are neighbours of each other (Solé et al., 2010). In other words, it is a measure of the relative frequency of triangles in a network.

Average path length. Path length is defined as the average minimal distance between any pair of vertices (Solé et al., 2010). The average path length *d* is defined as the average shortest distance between any pair of vertices in a network.

Together with the clustering coefficient, the average path length of a graph *G* constitutes the 'small-world model' of Watts & Strogatz (1998), which has proved to be an appropriate model for many types of networks (like, for instance, biological and social ones). If a network has a high clustering coefficient but also a very short average path length in comparison to random graphs with the same number of vertices, it is a small-world network.

4 Hubs in the IT-TB Network

For each vertex in the IT-TB network, we calculated its betweenness and closeness centralities using the Cytoscape app CytoNCA (http://apps.cytoscape.org/apps/cytonca).

Table 1 presents the rates of the centrality measures of the first five lemmas in the IT-TB network ranked by betweenness centrality. The table reports also the degree for each lemma.

R.	Lemma	Betw. C.	Clos. C.	Deg.
1	*sum* (*to be*)	1793719.9	0.2822	1095
2	*dico* (*to say*)	324728.16	0.2558	401
3	*possum* (*can*)	307137.8	0.2581	464
4	*habeo* (*to have*)	214495.38	0.2535	351
5	*facio* (*to make*)	146891.89	0.2507	289

Table 1. Results on centrality measures

Although some lemmas are differently ranked according to different centrality measures (for instance, *dico* is second by betweenness centrality, but it is third by both closeness centrality and degree), *sum* remains always first. This shows that *sum* is the "most hub" among the hubs of the IT-TB network.

Hubs are the key components of the complexity of a network. They support high efficiency of network traversal, but, just because of their important role in the web, their loss heavily impact the performance of the whole system (Jeong et al., 2002). If the most highly connected vertices are removed, the network properties change radically and the network breaks into fragments, sometimes even promoting a system's collapse (Albert & Barabási, 2000).

Following its status of most hub vertex in the IT-TB network, we removed the vertex of *sum* and of all its direct neighbours from the network. Further, we removed all those vertices that become isolated in the network after such a removal is applied (i.e. those with degree = 0; in total: 702 vertices). Figure 2 presents the subnetwork that results from these modifications.

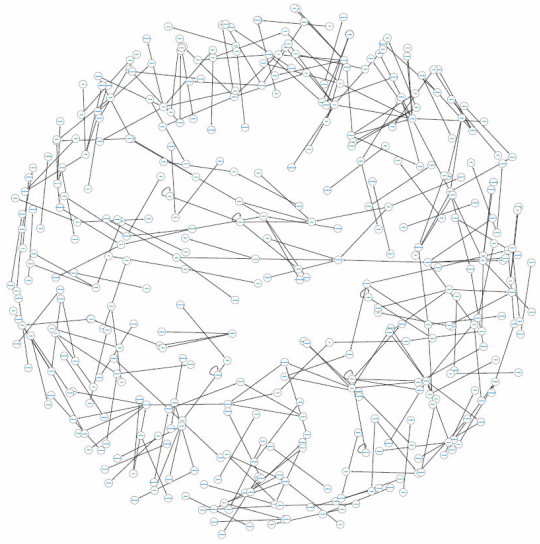


Figure 2. The IT-TB no-*sum* subnetwork

The counterpart of the subnetwork in figure 2 is the subnetwork formed only by the vertex of *sum* and its direct neighbours (figure 3).

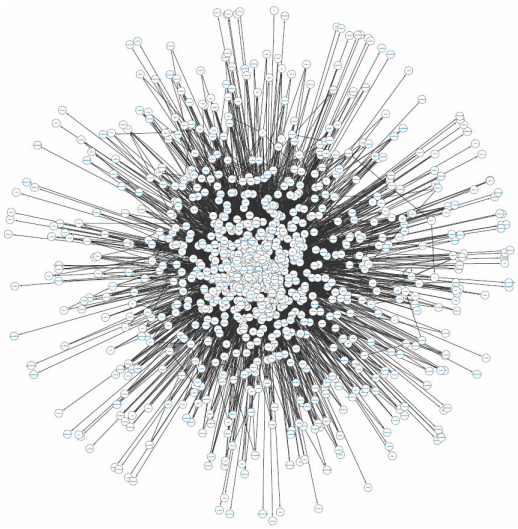


Figure 3. The IT-TB *sum*-only subnetwork

While figure 2 shows that removing the vertex of *sum* and those of its direct neighbours makes the network lose its connecting core, figure 3 presents a very much connected subnetwork.

In order to evaluate the role of *sum* in the network beyond the graphical layout of the subnetworks, we calculated the above mentioned topological indices of the full network of the IT-TB (1) and of the subnetworks reported respectively in figures 2 (2) and 3 (3). Table 2 presents the results.

	1	**2**	**3**
N. of vertices	2,198	398	1,098
N. of edges	19,031	301	15,486
Average degree	8.6583	0.7562	14.1038
Average path length	3.108	1.4883	2.5242
Clustering coefficient	0.247	0.081	0.352

Table 2. Results on topological indices

From the rates reported in table 2 it turns out that the subnetwork 2 is less small-world than 1 and 3, i.e. 2 is less connected and more fragmented than 1 and 3. This is shown by the clustering coefficient, which is dramatically lower in 2 than in 1 and 3. Although the average path length of 2 is shorter than 1 and 3, this is motivated by the much lower number of vertices in 2 than in 1 and 3, and not by the more small-worldness of 2. This is more clear if we look at the relation between the number of edges and the number of vertices in the networks. While in 1 and 3, the edges are much more than the vertices, in 2 the opposite holds, thus leading to much different average degrees.

The subnetwork 3 is even more small-world than 1. 3 is smaller than 1, as it results from removing a number of vertices from 1. This is why the average path length of 3 is shorter than 1. However, both the average degree and the clustering coefficient of 3 are higher than 1. It is worth noting that 3 includes, alone, half of the total of the vertices occurring in 1 and around 75% of the edges of 1: this shows that the vertex of *sum* is directly connected to half the vertices of the network and these connections cover most of those that occur in the IT-TB network.

5 Conclusion

While the most widespread tools for querying and analyzing treebanks give results in terms of lists of words or sequences of trees, network analysis permits a synoptic view of all the relations that hold between the words in a treebank. This makes network analysis a powerful method to fully exploit the structural information provided by a treebank, for a better understanding of the properties of language as a complex system with interconnnected elements.

References

R. Albert, H. Jeong and A.L. Barabási. 2000. Error and attack tolerance of complex networks. *Nature*, 406: 378-382.

R. Albert and A.L. Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74: 47-97.

B. Bollobás. 1998. *Modern Graph Theory*. Vol. 184 of *Graduate Texts in Mathematics*. Springer, New York.

T. Briscoe. 1998. Language as a Complex Adaptive System: Coevolution of Language and of the Language Acquisition Device. P. Coppen, H. van Halteren and L. Teunissen (eds.), *Proceedings of Eighth Computational Linguistics in the Netherlands Conference*. Rodopi, Amsterdam, 3-40.

R. Ferrer i Cancho, R.V. Solé and R. Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review* E69, 051915(8).

R. Ferrer i Cancho. 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. V. Levickij and G. Altmman (eds.), *Problems of quantitative linguistics*, 60-75.

R. Ferrer i Cancho. 2010. Network theory. P.C. Hogan (ed.), *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press, Cambridge, 555-557.

R. Hudson. 2007. *Language Networks. The New Word Grammar*. Oxford University Press, Oxford.

H. Jeong, S.P. Mason, A.L. Barabási and Z.N. Oltvai. 2002. Lethality and centrality in protein networks. *Nature*, 411: 41-42.

M. Kohl, S. Wiese and B. Warscheid. 2011. Cytoscape: software for visualization and analysis of biological networks. *Methods in Molecular Biology*, 696: 291-303.

S.M. Lamb. 1998. *Pathways of the Brain. The Neurocognitive Basis of Language*. John Benjamins, Amsterdam.

A. Mehler. 2008a. Large text networks as an object of corpus linguistic studies. A. Lüdeling and K. Merja (eds.), *Corpus Linguistics. An International Handbook of the Science of Language and Society*. De Gruyter, Berlin - New York, 328-382.

A. Mehler. 2008b. Structural similarities of complex networks: A computational model by example of Wiki graphs. *Applied Artificial Intelligence*, 22: 619-683.

M.E.J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*, 45.2: 167-256.

M.E.J. Newman. 2005a. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46: 323-351.

M.E.J. Newman. 2005b. A measure of betweenness centrality based on random walks. *Social Networks*, 27: 39-54.

M. Passarotti. 2011. Language Resources. The State of the Art of Latin and the *Index Thomisticus* Treebank Project. M.S. Ortola (ed.), *Corpus anciens et Bases de données. «ALIENTO. Échanges sapientiels en Méditerranée», N°2*. Presses universitaires de Nancy, Nancy, 301-320.

G. Sabidussi. 1966. The centrality index of a graph. *Psychometrika*, 31: 581-603.

R. Saito, M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, S. Lotia, A.R. Pico, G.D. Bader and T. Ideker. 2012. A travel guide to Cytoscape plugins. *Naure Publishing Group*, 9(11): 1069-76.

P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11): 2498-504.

R.V. Solé, B. Corominas-Murtra, S. Valverde and L. Steels. 2010. Language networks: Their structure, function, and evolution. *Complexity*, 15(6): 20-26.

L. Steels. 2000. Language as a Complex Adaptive System. M. Schoenauer (ed.), *Proceedings of PPSN VI, Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 17-26.

D.J. Watts and S.H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393: 440-442.

S. Wuchty and P.F. Stadler. 2003. Centers of complex networks. *Journal of Theoretical Biology*, 223(1): 45-53.

10.12871/CLICIT2014157

I-ChatbIT: an Intelligent Chatbot for the Italian Language

Arianna Pipitone and **Vincenzo Cannella** and **Roberto Pirrone**
DICGIM - Dipartimento di Ingegneria Chimica, Gestionale, Informatica e Meccanica
University of Palermo
{*arianna.pipitone, vincenzo.cannella26, roberto.pirrone*}*@unipa.it*

Abstract

English. A novel chatbot architecture for the Italian language is presented that is aimed at implementing cognitive understanding of the query by locating its correspondent subgraph in the agent's KB by means of a graph matching strategy purposely devised. The FCG engine is used for producing replies starting from the semantic poles extracted from the candidate answers' subgraphs. The system implements a suitable disambiguation strategy for selecting the correct answer by analyzing the commonsense knowledge related to the adverbs in the query that is embedded in the lexical constructions of the adverbs themselves as a proper set of features. The whole system is presented, and a complete example is reported throughout the paper.

Italiano. *Si presenta una nuova architettura di chatbot per l'italiano che implementa una forma di comprensione di natura cognitiva della query individuando il corrispondente stottografo nella base di conoscenza dell'agente con tecniche di graph matching definite appositamente. Il sistema FCG usato per la produzione a partire dai poli semantici estratti da tutti i sottografi coandidati alla risposta. Il sistrema effettua una disambigazione a partire dalla conoscenza di senso comun sugli avverbi che codificata come un apposito insieme di caratteristiche all'interno delle relative costruzioni lessicali. Si presenta l'intera architettura e viene svolto un intero esempio di funzionamento.*

1 Introduction

In recent years the Question-Answering systems (QAs) have been improved by the integration with Natural Language Processing (NLP) techniques, which make them able to interact with humans in a *dynamic way*: the production of answers is more sophisticated than the classical chatterbots, where some sentence templates are pre-loaded and linked to the specific questions.

In this paper we propose a new methodology that integrates the chatterbot technology with the Cognitive Linguistics (CL) (Langacker, 1987) principles, with the aim of developing a QA system that is able to harvest a linguistic knowledge from its inner KB, and use it for composing answers dynamically. Grammatical templates and structures tailored to the Italian language that are constructions of the Construction Grammar (CxG) (Goldberg, 1995) and a linguistic Italian source of verbs have been developed purposely, and used for the NL production. The result of the methodology implementation is I-ChatbIT, an Italian chatbot that is intelligent not only for the dynamic nature of the answers, but in the sense of *cognitive understanding* and *production* of NL sentences. Cognitive understanding of the NL query is achieved by placing it in the system's KB, which represents the conceptualization of the world as it has been perceived by the agent. The outcome of such a process is the generation of what we call the *meaning activation* subgraph in the KB. Browsing this subgraph, the system elaborates and detects the content of the answer, that is next grammatically composed through the linguistic base. The FCG engine is then used as the key component for producing the answer. Summarily, the work reports the modeling of the two tasks outlined above.

The paper is arranged as follow: in the next section the most popular chatbots are shown, devoting particular attention to the Italian ones. Section 3 describes the implemented methodology explaining

in detail a practical example. Finally, conclusions and future works are discussed in section 4.

2 The Italian Chatbots

There are no many Italian chatbots in literature. We refer to the most recent and widespread ones. QUASAR (Soriano et al., 2005) uses simple pattern matching rules for the answer extraction and it splits the Italian among the provided language. Eloisa and Romana (available at `http://www.eloisa.it/`) are the most recent Italian chatbots, the former speaking on generic arguments (as sports, politics and so on), the latter specifically for history and folklore of Rome city. Both have a basic form of intelligence because they learn new contents during the conversation, even if no learning algorithms have been made mentioned by the authors. Among cognitive QA systems, the best known cognitive technology is Watson (Ferrucci, 2012) from IBM, which was specifically developed for answering questions at the Jeopardy quiz show. The core is the UIMA (Ferrucci and Lally, 2004) framework on which the whole system is implemented. However, this system does not provide Italian language by now. Finally there are many virtual assistants developed for the Italian, but neither of them uses cognitive approaches. The base technology is using controlled NL and pattern matching; however these systems act on specific and restricted tasks as the services provided by telephonic companies, booking flights and so on.

3 Building I-ChatbIT

Figure 1 shows the I-ChatbIT architecture; the main modules are the *Meaning Activator* and the *Answer Composer*, which are connected to the Knowledge Base (KB) and to the linguistic base (composed by our *Italian Verbs Source* (IVS) and MultiWordnet (Pianta et al., 2002) (MWn)). The whole system is managed by the *Controller*, which acts as the user interface too. The KB contains the inner domain representation owned by the system. We used OWL ontologies for such a component. The KB can be replaced so the system is domain independent. MWn and the IVS form the linguistic base of the system. We are currently expanding the IVS to cover the other parts of speech, and it will become the only Italian dictionary of the system. In this phase MWn is used for retrieving parts of speech other than verbs. The Meaning Activator (MA) implements the meaning-activation process through a graph similarity search between the *query-graph* (a graph representation of the query) and the *conceptual-graph* (a graph representation of the KB): the result of search is a set of subgraphs that correspond to placing the query in the KB. Browsing such subgraphs some facts are detected, and they are the candidates for composing the answer. All the candidate facts are inputted to the Answer Composer (AC) that generates grammatical constructions, and filters them according to the linguistic information that is needed for context disambiguation. Filtered constructions are finally plunged and produced by the Fluid Construction Grammar (FCG) engine (Steels and de Beule, 2006). If the answer is not exhaustive for the user, the Controller involves the Semantic Annotator described in (Pipitone and Pirrone, 2012) that retrieves external contents; such contents are re-elaborated as facts by the AC and the process is iterated. Each component is next carefully described.

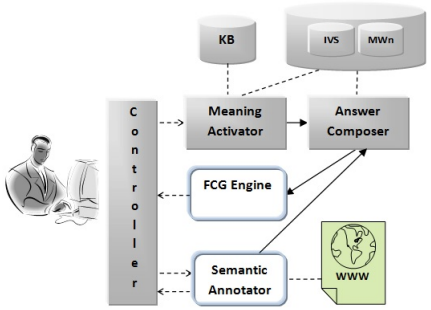


Figure 1: I-ChatbIT Architecture.

3.1 The Meaning Activator

The strategy adopted for implementing cognitive understanding in the MA relies on applying the Graph Edit Distance (GED) method (Zeng et al., 2009) between the query-graph Q and the conceptual-graph C, so that their GED is no larger than a distance threshold τ. In particular, the *query-graph* is the triple $Q = (N_q, E_q, L_q)$ where the nodes set N_q contains the macro-syntactic roles of the NL query, parsed by the Freeling parser (Padr and Stanilovsky, 2012). These nodes are sequentially connected reflecting their position in the query. The labels set L_q are labels nodes, and correspond to the tokens of the query outputted by the parser. For example, the query-graph for the question *"Dov'è nato il famoso Giotto?"* is shown in figure 2. The *conceptual-graph* is the 4-tuple $C = (N_c, E_c, L_c, \sigma)$ where the nodes set

$N_c = C_n \cup R_n$ is the union set of the set C_n containing the concepts in the KB, and the set R_n that contains relations. An edge in E_c connects only a concept-node to a relation-node if the concept is either the domain or the range for the relation itself. The edge is labeled with a progressive number for tracing the entities involved in the relation. σ is a label function $\sigma : N_c \rightarrow L_c$ that associates to each node $n_c \in N_c$ a list of strings $l_c \in L_c$ that are obtained by querying the linguistic sources on-the-fly, as it is next described. An example of conceptual graph is shown in figure 2. For GED computation, we refer to the two following parameters:

- a *similarity measure* between nodes, that is the Jaro–Winkler distance (Winkler, 1990) between the labels associated to them as described in 3.2.1;
- a *graph edit distance* ged between subgraphs, that represents the number of primitive graph edit operations to make them isomorphic. There are six primitive edit operations (Zeng et al., 2009): node insertion and deletion, node label substitution, edge insertion and deletion, and edge label substitution. For our purposes, the above constraints for connecting nodes make label substitution useless, so we refer only to the remaining four operations.

Given Q, C and a distance threshold τ, the problem is *to find a set of subgraphs* $I = \{I_i\}$ *with* $I_i = (N_i, E_i, L_i) \subset C$ *so that* I_i *and* Q *are isomorphic for a number of primitive edit operations* $ged <= \tau$. $M_{act} \equiv \bigcup_i I_i$ corresponds to the meaning activation area of the query. Considering that Q is a linear sequence of nodes and edges, an isomorphism in C will be a sequence too. Threshold τ is necessary for avoiding that the query is sparse in the KB. The τ value has been fixed arbitrarily to 10. The strategy computes the isomorphisms applying the *k*-AT algorithm (Wang et al., 2012), which defines a *q-gram* as a tree consisting of a vertex *v* and the paths starting at *v* with length no longer than q. In our case, the vertexes are nodes from M_{act}, the *k*-AT has been customized for using only four edit operations as explained before.

Once the isomorphisms are detected, MA probes the KB for retrieving connected facts, for example it adds nodes that are either the domain or the range of some relation node if they are not yet included in the subgraphs, or retrieves adjacent triples to the nodes involved in the isomorphisms. In our example there are two isomorphisms $I_1 = \{Giotto - datanascita\}$ and $I_2 = \{Giotto - luogonascita\}$, and ged is equal to 5 for both of them. They are candidates as possible answers. The AC will provide the correct disambiguation between them. If no disambiguation is possible, the answer is composed by the conjunction of them and results in an expanded sentence.

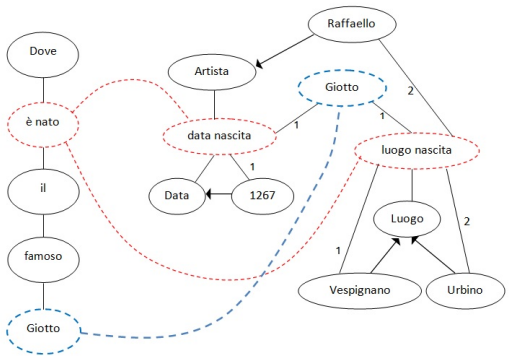


Figure 2: An example of Q and C, joint through the Jaro–Winkler distance, and the computation of the related isomorphisms.

3.2 The Answer Composer

Once the M_{act} subgraphs set is detected by MA, the correct NL sentence has to be composed. For this purpose, we use the FCG engine where system puts the linguistic information about the domain according to the FCG formalism, that is the CxG. Lexical and grammatical descriptions of the domain terms must be represented as *Constructions*, that are form-meaning couples. Form pole contains the syntactic features of terms, while the semantic one contains meaning. Lexical constructions are related to a single word, and conjunctions of them generate grammatical constructions. FCG uses the same set of constructions for both parsing and production, by iterating merging and unification processes on the involved poles (syntatic poles in parsing, semantic ones in production). In this phase, the FCG engine of the system contains lexical constructions for the Italian adverbs and articles, that were manually created. For adverbs, the features embed some commonsense knowledge about them: for example, the lexical construction for the adverb "dove" stores some features like "luogo", "posto", "destra", "sinistra" and so on. Query parsing allows obtaining the semantic poles related to the query, so the probing strategy performed by MA is necessary for retriev-

ing others facts from KB and composing the M_{act} related to the answer. For this reason we use FCG only in production: once the M_{act} is fed to the FCG engine, it unifies the related semantic poles to the correspondent lexical and grammatical constructions, and produces the answer. The opposite way is not possible because we would need to map all possible subgraphs in the KB as facts in the FCG, with a consequent combinatorial explosion.

3.2.1 Filling FCG through linguistic sources

Lexical and grammatical constructions about the domain form the linguistic base of the system and are generated by querying the KB and the linguistic sources (IVS and MWn). In particular, the KB concepts and relations labels are retrieved, and tokenized according to the algorithm described in (Pipitone et al., 2013), that models the cognitive task of reading. As a consequence I-ChatbIT learns the KB content. The system queries either IVS or MWn according to the stem of the label. In case of a verb stem verb, IVS provides all the related information, which includes the related argument structures (Goldberg et al., 2004) and synonyms, as it shown in next section. In the all other cases, the system refers to MWn, and it retrieves synonyms, hypernyms, hyponyms for each label along with the verbal information of the verb included in the definition. The lexical and grammatical constructions of all these terms are generated as described by some of the authors in (Pipitone and Pirrone, 2012) where terms that refer to the same nodes are considered synonymic constructions.

3.2.2 Answer production and disambiguation

FCG contains constructions tailored on the KB. When KB subgraphs are put to the AC, it builds the correspondent meaning poles, and the related constructions fire; all of them are candidates for being used in production. At this point AC applies the *disambiguation process*. Adverb tokens in the query are parsed by the FCG engine, and their corresponding lexical constructions fire. Disambiguation chooses the subgraph that has a link to the adverbial features stored in the construction. In our example, the candidates facts from MA are the subgraphs {Giotto - data nascita - 1267 - is_a Data} and {Giotto - luogo nascita - Luogo}; the lexical construction of the adverb "dove" allows selecting the second subgraph. If the query were "Quando è nato il famoso Giotto?" the first subgraph would be selected using the commonsense knowledge stored in the related lexical construction ("ora", "tempo","data", and so on).

3.3 The Italian Verbs Source

The IVS contains approximately five thousands verbs, classified into distinct groups. They represent the most common verbs usually used in a common conversation in Italian. All inflexions of each verb have been stored and annotated. The storage adopts a compressed description of verbs. Each inflextion is derived by combining the root of the verb with the corresponding suffix. Suitable rules choose the proper inflexion on the basis of tense, person, number and gender are used to choose the proper inflexion. Verbs have been grouped on the basis of their suffix class, according to the base rules of Italian grammar. A finer grouping has been made according to the origin of the verb. This choice allows a more compact description of the verbs' conjiugations. Irregular verbs have been treated using ad hoc rules for producing their inflections.
Each tense is described by a construction, containing tense, person, number, and gender as its features. Each compound form is described by a single construction, and not as combination of other constructions. All possible active, passive, and reflexive forms have been stored. All verbs have been classified as transitive, intransitive and semitransitive. This information is stored into each verb construction too. Finally, each verb is joined to a list of possible synonymies and analogies.

4 Conclusions and future works

A novel chatbot architecture for the Italian language has been presented that is aimed at implementing cognitive understanding of the query by locating its correspondent subgraph in the agent's KB by means of a GED strategy based on the k-AT algorithm, and the Jaro–Winkler distance. The FCG engine is used for producing replies starting from the semantic poles extracted from the candidate answers' subgraphs. The system implements a suitable disambiguation strategy for selecting the correct answer by analyzing the commonsense knowledge related to the adverbs in the query that is embedded in the lexical constructions of the adverbs themselves as a proper set of features. Future works are aimed at completing the IVS, and using explicit commonsense knowledge inside the KB for fine disambiguation. Finally, the graph matching strategy will be further tuned.

References

David Ferrucci and Adam Lally. 2004. Uima: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, September.

David A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3):1.

Adele E. Goldberg, Devin M. Casenhiser, and Nitya Sethuraman. 2004. Learning Argument Structure Generalizations. *Cognitive Linguistics*, 15(3):289–316.

A. E. Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford University Press, Stanford, CA. Vol 1, 1987(Hardcover), 1999(Paperback).

Llus Padr 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.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Arianna Pipitone and Roberto Pirrone. 2012. Cognitive linguistics as the underlying framework for semantic annotation. In *ICSC*, pages 52–59. IEEE Computer Society.

Arianna Pipitone, Maria Carmela Campisi, and Roberto Pirrone. 2013. An a* based semantic tokenizer for increasing the performance of semantic applications. In *ICSC*, pages 393–394.

Jos Manuel Gmez Soriano, Davide Buscaldi, Empar Bisbal Asensi, Paolo Rosso, and Emilio Sanchis Arnal. 2005. Quasar: The question answering system of the universidad politcnica de valencia. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Mller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 439–448. Springer.

Luc Steels and Joachim de Beule. 2006. A (very) brief introduction to fluid construction grammar. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, ScaNaLU '06, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guoren Wang, Bin Wang, Xiaochun Yang, and Ge Yu. 2012. Efficiently indexing large sparse graphs for similarity search. *IEEE Trans. on Knowl. and Data Eng.*, 24(3):440–451, March.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.

Zhiping Zeng, Anthony K. H. Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. 2009. Comparing stars: On approximating graph edit distance. *PVLDB*, 2(1):25–36.

10.12871/CLICIT2014158

Two-dimensional Wordlikeness Effects in Lexical Organisation

Vito Pirrelli
ILC-CNR Pisa
vito.pirrelli@ilc.cnr.it

Claudia Marzi
ILC-CNR Pisa
claudia.marzi@ilc.cnr.it

Marcello Ferro
ILC-CNR Pisa
marcello.ferro@ilc.cnr.it

Abstract

English. The main focus of research on wordlikeness has been on how serial processing strategies affect perception of similarity and, ultimately, the global network of associative relations among words in the mental lexicon. Comparatively little effort has been put so far, however, into an analysis of the reverse relationship: namely, how global organisation effects influence the speakers' perception of word similarity and of words' internal structure. In this paper, we explore the relationship between the two dimensions of wordlikeness (the "syntagmatic" and the "paradigmatic" one), to suggest that the same set of principles of memory organisation can account for both dimensions.

Italiano. *Gran parte dei lavori sulla nozione di "familiarità lessicale" ha analizzato come le strategie di elaborazione seriale influenzino la percezione della similarità all'interno della rete di relazioni formali nel lessico mentale. Poca attenzione è stata tuttavia dedicata finora a come queste relazioni globali influenzino la percezione della similarità lessicale. L'articolo esplora questa interconnessione tra relazioni sintagmatiche e paradigmatiche, attribuendola a un insieme omogeneo di principi per l'organizzazione della memoria seriale.*

1 Introduction

The language faculty requires the fundamental ability to retain sequences of symbolic items, access them in recognition and production, find similarities and differences among them, and assess their degree of typicality (or WORDLIKENESS) with respect to other words in the lexicon. In particular, perception of formal redundancy appears to be a crucial precondition to morphology induction, epitomised by the so-called WORD ALIGNMENT problem. The problem arises whenever one has to identify recurrence of the same pattern at different positions in time, e.g. *book* in *handbook*, or *mach* in both German *macht* and *gemacht*. Clearly, no "conjunctive" letter coding scheme (e.g., Coltheart et al. 2001; Harm & Seidenberg 1999; McClelland & Rumelhart 1981; Perry et al. 2007; Plaut et al. 1996), which requires that the representation of each symbol in a string be anchored to its position, would account for such an ability. In Davis' (2010) SPATIAL ENCODING, the identity of the letter is described as a Gaussian activity function whose max value is centred on the letter's actual position, enforcing a form of fuzzy matching, common to other models disjunctively encoding a symbol and its position (Grainger & van Heuven 2003; Henson 1998; Page & Norris 1998, among others).

The role of specific within-word letter positions interacts with short-term LEXICAL BUFFERING and LEXICALITY effects. Recalling a stored representation requires that all symbols forming that representation are simultaneously activated and sustained in working memory, waiting to be serially retrieved. Buffering accounts for the comparative difficulty in recalling long words: more concurrently-activated nodes are easier to be confused, missed or jumbled than fewer nodes are. Notably, more frequent words are less likely to be confused than low-frequency words, since long-term entrenchment improves performance of immediate serial recall in working memory (Baddeley 1964; Gathercole et al. 1991).

Serial (or syntagmatic) accounts of local ordering effects in word processing are often complemented by evidence of another, more global (or paradigmatic) dimension of word perception, based on the observation that, in the normal course of processing a word, other non-target neighbouring words become active. In the word recognition literature, there is substantial agreement on the inhibitory role of lexical neighbours (Goldinger et al. 1989; Luce & Pisoni 1998; Luce et al. 1990). Other things being equal, target words with a large number of neighbours take more time to be recognised and repeated, as they suffer from their neighbours' competition in lexical buffering. This is particularly true when

the target word is low-frequency. Nonetheless, there is contrasting evidence that dense neighbourhoods may speed up word reading time rather than delaying it (Huntsman & Lima 2002), and that high-entropy word families make their members more readily accessible than low-entropy families (Baayen et al. 2006).

Marzi et al. (2014) provide clear computational evidence of interactive effects of paradigm regularity and type/token lexical frequency on the acquisition of German verb inflection. Token frequency plays a paramount role in item-based learning, with highly frequent words being acquired at comparatively earlier stages than low-frequency words. Morphological regularity, on the other hand, has an impact on paradigm acquisition, regular paradigms being learned, on average, within a shorter time span than fully or partially irregular paradigms. Finally, frequency distribution of paradigmatically-related words significantly interacts with morphological regularity. Acquisition of regular paradigms depends less heavily on item-based storage and is thus less affected by differences in frequency distributions of paradigm members. Conversely, irregular paradigms are less prone to be generalised through information spreading and their acquisition mainly relies on itemised storage, thus being more strongly affected by the frequency distribution of paradigm members and by frequency-based competition, both intra- and inter-paradigmatically.

We suggest that compounded evidence of wordlikeness and paradigm frequency effects can be accounted for within a unitary computational model of lexical memory. We provide here preliminary evidence in this direction, by looking at the way a specific, neuro-biologically inspired computational model of lexical memories, Temporal Self-Organising Maps (TSOMs), accounts for such effects.

2 TSOMs

TSOMs are a variant of classical Kohonen's SOMs (Kohonen 2001), augmented with re-entrant Hebbian connections defined over a temporal layer encoding probabilistic expectations upon immediately upcoming stimuli (Koutnik 2007; Ferro et al. 2010; Pirrelli et al. 2011; Marzi et al. 2012a, 2012b). TSOMs consist of a network of memory nodes simultaneously responding to time-bound stimuli with varying levels of activation (Fig. 1). Through learning, nodes acquire selective sensitivity to input stimuli, i.e. they respond more strongly to a particular class of stimuli than to others. Selective sensitivity is based on both nature of the stimulus (through *what* connections), and stimulus timing (through *when* connections) (see Fig. 1). Accordingly, more nodes tend to be recruited to respond to the same symbol, each node being more sensitive to a specific occurrence of the symbol in context.

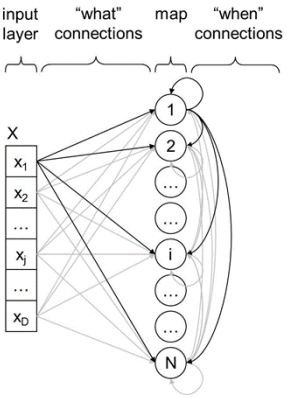


Figure 1: Outline architecture of a TSOM.

TSOMs can be trained on word forms as time-bound sequences of symbols by presenting each symbol on the input layer one at a time. The pattern of node activation prompted by each symbol is eventually integrated into a word memory trace, whose top-most activated nodes are named Best Matching Units (BMUs). Co-activation of the same BMUs by different input words reflects the extent to which the map perceives formal redundancies between words. We contend that perception of wordlikeness and morphological structure has to do with information sharing and co-activation levels between word forms.

2.1 Activation and co-activation effects

Two quantitative correlates have been suggested to account for effects of human perception of wordlikeness: N-GRAM PROBABILITY DENSITY (the likelihood that a word form results from concatenation of sublexical chunks of *n* length), and LEXICAL DENSITY (the number of word forms in the lexicon that are similar to a specific target word) (Bailey & Hahn 2001).

The two measures are highly correlated and thus easy to be confounded in measuring their independent effects on lexical tasks (Bard 1990). Bailey and Hahn (2001) propose to define *n*-gram probability densities in terms of the geometric mean of the product of the independent probabilities of bigram and trigram types extracted from the lexicon. In addition, following Luce and Pisoni (1998), the lexical neighbourhood of a target word can be defined as the set of word

forms obtained from the target by substitution, deletion or insertion of one symbol.

With a view to establishing functional correlates between the behaviour of a TSOM and evidence of probability and neighbourhood density effects on word processing, we trained 10 instances of a TSOM on 700 uniformly-distributed Italian verb forms, belonging to a fixed set of 14 cells of the 50 most frequent verb paradigms in the Italian Tree Bank (Montemagni et al. 2003). We tested the 10 map instances on the task of RECALLING [1] four data sets: (i) the original TRAINING SET; (ii) a set of 50 TEST WORDS sampled from the same 50 paradigms of the original training set; (iii) an additional set of novel Italian verb paradigms which were not part of the original training (hereafter NOVEL WORDS); iv) a set of German verb forms (or Italian NON-WORDS).

On training and test words, accuracy is respectively 99.2% and 96.4%. On novel words, recall (44.4%) significantly correlates with both node ACTIVATION STRENGTH (r=0.471, p<.00001), i.e. the per-word mean activation level of BMUs in the words' memory traces, and between-node CONNECTION STRENGTH (r=0.506, p<.00001), i.e. the per-word mean strength of the temporal connections between consecutives BMUs. Equally significant but lower correlations with activation and connection strengths are found for recall scores on non-words (12.4%): r=0.335, p<.00001 and r=0.367, p<.00001. We observe that recall scores somewhat reflect a word familiarity gradient, ranging from known words (training set) and known word stems with novel inflections (test words) to novel paradigms (novel words) and non-words. In particular, the gradient reflects the extent to which a map has developed expectations on incoming words, which in turn are encoded as weights on temporal connections. Both connection and activation strength thus capture probabilistic knowledge of Italian orthotactic constraints.

In fact, *n*-gram probability does not explain recall scores entirely. Forward probabilities account for degrees of entrenchment in integrated memory traces but they say nothing about co-activation of other formally-related words. This information has to do with neighbourhood density and is controlled by the degree of global lexical co-activation by an input word, i.e. by the extent to which the word memory trace reverberates with all other memory traces in the lexicon (Fig. 4). Note that both test and novel words exhibit comparatively high levels of global co-activation, in contrast with non-words, whose degree of paradigmatic wordlikeness is consistently poorer (p<.00001).

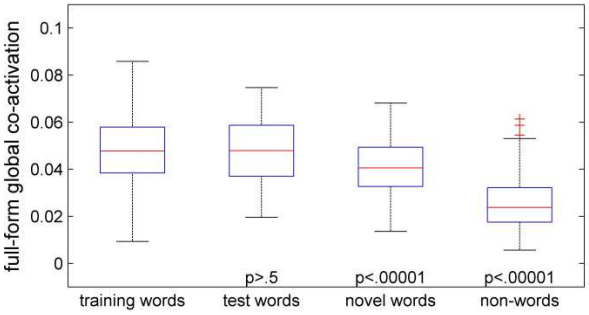


Figure 4: Per-word global co-activation.

We explain this overall effect by looking at differential values of activation strengths for stems and suffixes in Fig. 5. Here, Italian novel words score more highly on suffixes than on stems. As expected, they are recalled consistently more poorly, but their degree of perceived familiarity is due to their fitting Italian recurrent morphological patterns.

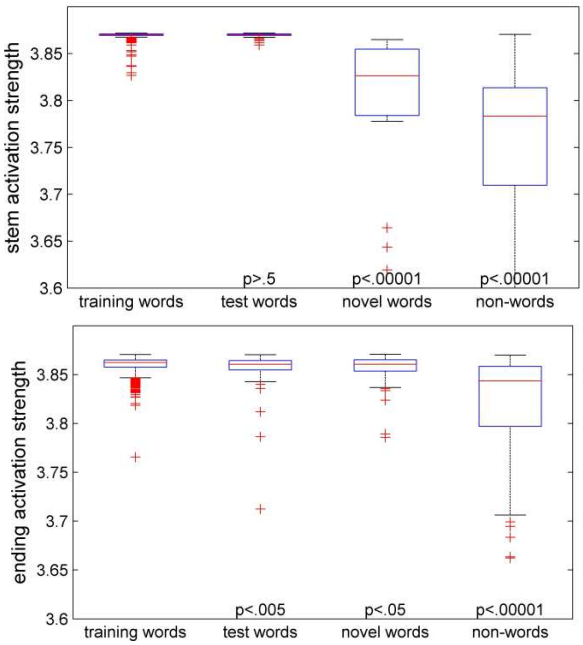


Figure 5: Per-stem (top panel) and per-ending (bottom panel) activation strength.

2.2 Frequency effects

In section 1, we overviewed contrasting evidence of inhibitory and facilitatory effects of neighbourhood density and neighbourhood frequency in different word processing tasks. To test Baayen and colleagues' claim that large, evenly-distributed word families facilitate accessibility of their own members, we assessed, for each Ital-

[1] Recall simulates the process of retrieving a sequence of letters from an integrated word memory trace. A word is recalled accurately if the map retrieves all its symbols in the correct left-to-right order.

ian word in our training set, the level of confusability of its memory trace on the map in the recall task. A word memory trace contains, over and above target BMUs, also nodes that are associated with concurrently activated neighbours. By increasingly filtering out nodes with lower activation levels in the trace, we can make it easier for the map to reinstate the appropriate sequence of target nodes by eliminating spurious competitors. Fig. 6 shows the box plot distribution of the mean filtering level for classes of words having up to 2 neighbours, between 3 and 12 neighbours, and more than 12 neighbours.

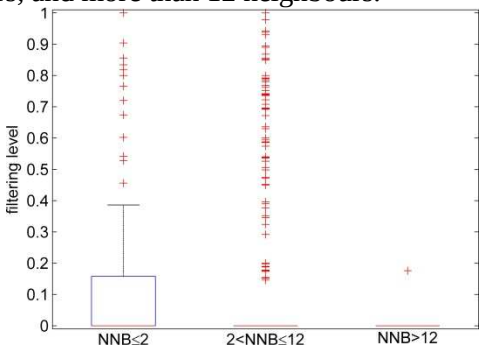


Figure 6: Filtering levels on word memory traces for serial recall, for three neighbourhood-density bins.

In TSOMs, words with sparser neighbours require more filtering to be recalled correctly from their memory traces. This is due to the facilitatory effect of having more words that consistently activate the same sequences of nodes. Fewer neighbours weaken this effect, making it more difficult to recover the right sequence of nodes from a word memory trace. This greater difficulty is reflected by larger filtering levels in Fig. 6.

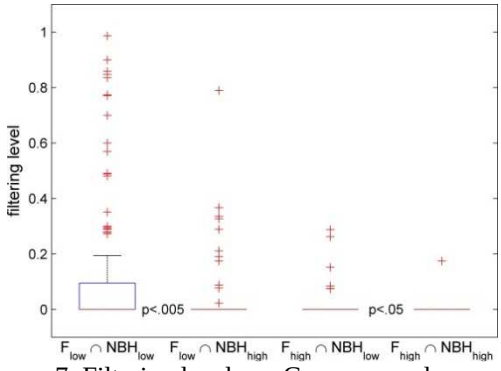


Figure 7: Filtering levels on German word memory traces for serial recall, for four classes of word-frequency by neighbourhood-entropy bins.

However, neighbours are not always helpful. If a word to be recalled is associated with high-frequency neighbours, these neighbours tend to strongly interfere with recall, eventually leading the map astray. The lower the frequency of the target word is, the more prone to interference from competing neighbours it will be, as shown in Fig. 7 for German verbs, where low-frequency words in low-entropy neighbourhoods ($F_{low} \cap NBH_{low}$) appear to require a significantly higher level of filtering than words in high-entropy neighbourhoods do.

3 Concluding remarks

Wordlikeness is a fundamental determinant of lexical organisation and access. Two quantitative measures of wordlikeness, namely *n*-gram probability and neighbourhood density, relate to important dimensions of lexical organisation: the syntagmatic (or horizontal) dimension, which controls the level of predictability and entrenchment of a serial memory trace, and the paradigmatic (or vertical) dimension, which controls the number of neighbours that are co-activated by the target word. The two dimensions are nicely captured by TSOMs, allowing the investigation of their dynamic interaction.

In accessing and recalling a target word, a large pool of neighbours can be an advantage, since they tend to support patterns of activation that are shared by the target word. However, their help may turn out to interfere with recall, if the connection strength of one or more neighbours is overwhelmingly higher than that of the target. Deeply entrenched friends eventually become competitors. This dynamic establishes a nice connection with paradigm acquisition, where a uniform distribution of paradigm members is helpful in spreading morphotactic information and speed up acquisition, and paradigmatically-related forms in skewed distributions compete with one another (Marzi et al. 2014). We argue that both neighbourhood and morphological effects are the result of predictive (syntagmatic) activation and competitive (paradigmatic) co-activation of parallel processing nodes in densely interconnected networks.

As a final qualification, our experiments illustrate the dynamic of activation and storage of letter strings, with no information about morphological content. They provide evidence of the first access stages of early lexical processing, where strategies of automatic segmentation are sensitive to possibly apparent morphological information (Post et al. 2008). Nonetheless, our data suggest that perception of wordlikeness and morphological structure can be accounted for by a common pool of principles governing the organisation of long-term memories for time series.

References

Baddeley, A. D. (1964). Immediate memory and the "perception" of letter sequences. *Quarterly Journal of Experimental Psychology*, 16, 364–367.

Gathercole, S. E., C. Willis, H. Emslie & A.D. Baddeley (1991). The influence of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics*, 12, 349–367.

Bard, E.G. (1990). Competition, lateral inhibition, and frequency: Comments on the chapters of Frauenfelder and Peeters, Marslen-Wilson, and others. In G. T. M. Altmann (ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, 185-210. MIT Press.

Bailey, T. M., & U. Hahn (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4), 568-591.

Baayen, R., L. Feldman & R. Schreuder (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53, 496–512.

Coltheart, M., K. Rastle, C. Perry, R. Langdon & J. Ziegler (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, 117.3, pp. 713-758.

Ferro, M., D. Ognibene, G. Pezzulo & V. Pirrelli (2010). Reading as active sensing: a computational model of gaze planning in word recognition. *Frontiers in Neurorobotics*, 4(6), 1-16.

Goldinger, S. D., P. Luce, & D. Pisoni (1989). Priming lexical neighbours of spoken words: Effects of competition and inhibition. *Journal of Memory & Language*, 28, 501-518.

Grainger, J. & W. van Heuven (2003). Modeling letter position coding in printed word perception. *The mental lexicon*, 1-24. New York, Nova Science.

Harm, M.W. & M.S. Seidenberg (1999). Phonology, Reading Acquisition and Dyslexia: Insights from Connectionist Models. *Psychological Review*, 106(3), 491-528.

Henson, R.N. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, 36, 73-137.

Huntsman, L.A. & S.D. Lima (2002). Orthographic Neighbors and Visual Word Recognition. *Journal of Psycholinguistic Research*, 31, 289-306.

Kohonen, T. (2001). Self-Organizing Maps. Heidelberg, Springer-Verlag.

Koutnik, J. (2007). Inductive Modelling of Temporal Sequences by Means of Self-organization. In Proceeding of International Workshop on Inductive Modelling (IWIM 2007), Prague, 269-277.

Luce, P. & D. Pisoni (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing*, 19(1), 1-36.

Luce, P., D. Pisoni & S. D.Goldinger (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, 122-147. Cambridge, MA: MIT Press.

Marzi C., M. Ferro & V. Pirrelli. (2012a). Prediction and Generalisation in Word Processing and Storage. In 8thMediterranean Morphology Meeting Proceedings on Morphology and the architecture of the grammar, 113-130.

Marzi C., M. Ferro & V. Pirrelli. (2012b). Word alignment and paradigm induction. *Lingue e Linguaggio* XI, 2. 251-274. Bologna: Il Mulino.

Marzi C., M. Ferro & V. Pirrelli. (2014). Morphological structure through lexical parsability. *Lingue e Linguaggio*, 13(2), forthcoming.

McClelland, J.L. & D.E. Rumelhart (1981). An interactive activation model of context effects in letter perception: Part 1. An account of Basic Findings. *Psychological Review*, 88, 375-407.

Montemagni, S. *et al.* (2003). Building the Italian syntactic-semantic treebank. In Abeillè, A. (ed.) *Building and Using Parsed Corpora*, 189–210. Dordrecht Kluwer.

Page, M.P.A. & D. Norris (1998). The primacy model: a new model of immediate serial recall. *Psychological Review*, 105, pp. 761-781.

Perry, C., J.C. Ziegler & M. Zorzi (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, 114(2), 273-315.

Pirrelli, V., M. Ferro & B. Calderone (2011). Learning paradigms in time and space. Computational evidence from Romance languages. In Maiden, M., J.C. Smith, M. Goldbach & M.O. Hinzelin (eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, 135-157. Oxford, Oxford University Press.

Plaut, D.C., J.L. McClelland, M.S. Seidenberg & K. Patterson (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56-115.

Post, B., W. Marslen-Wilson, W., B. Randall & L.K. Tyler (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition* 109 1-17.

10.12871/CLICIT2014159

Toward Disambiguating Typed Predicate-Argument Structures for Italian

Octavian Popescu, Ngoc Phuoc An Vo, Anna Feltracco, Elisabetta Jezek and Bernardo Magnini
University of Pavia - jezek@unipv.it
FBK, Trento, Italy - magnini|popescu|ngoc|feltracco@fbk.eu

Abstract

English. We report a word sense disambiguation experiment on Italian verbs where both the sense inventory and the training data are derived from T-PAS, a lexical resource of typed predicate-argument structures grounded on corpora. We present a probabilistic model for sense disambiguation that exploits the semantic features associated to each argument position of a verb.

Italiano. *Questo lavoro riporta un esperimento di disambiguazione per verbi italiani, in cui sia la lista dei sensi che i dati di addestramento sono derivati da T-PAS, una risorsa che contiene strutture argomentali tipizzate ricavate da corpora. Presentiamo un modello probabilistico per la disambiguazione che utilizza informazioni semantiche associate a ciascuna posizione argomentale del verbo.*

1 Introduction

Word Sense Disambiguation (WSD) (see (Agirre and Edmonds, 2006) for a comprehensive survey of the topic) is a task in Computational Linguistics where a system has to automatically select the correct sense of a target word in context, given a list of possible senses for it. For instance, given the target word *chair* in the context of the sentence *The cat is on the chair*, and given two possible senses for the word, let's call them *chair_as_forniture* and *chair_as_human*, a WSD system should be able to select the first sense as the appropriate one. An important aspect of WSD is that its complexity is affected by the ambiguity (i.e. the number of senses) of the words to be disambiguated. This has led in the past to discussing various characteristics of available sense repositories (e.g. WordNet, Fellbaum 1998), including the nature and the number of sense distinctions, particularly with respect to the application goals of WSD.

In this paper we address Word Sense Disambiguation of Italian verbs. Differently form previous work on WSD for Italian (Bertagna et al. 2007), where the sense repository was ItalWordNet (Roventini et al. 2003), in our experiments we use verb senses derived from T-PAS, a repository of Typed Predicate Argument Structures for Italian acquired from corpora. There are two benefits of this choice: (i) word sense distinctions are now grounded on actual sense occurrences in corpora, this way ensuring a natural selection with respect of sense granularity; (ii) as in T-PAS for each verb sense a number of sentences are collected, there is no further need to annotate data for training and testing, avoiding the issue of re-interpreting sense distinctions by different people.

The paper is organized as follows. Section 2 introduces T-PAS, including the methodology for its acquisition. Section 3 presents the probabilistic model that we have used for verb disambiguation and Section 4 reports on experimental results.

2 The T-PAS resource

T-PAS (Jezek et al. 2014) is a repository of Typed Predicate Argument Structures (T-PAS) for Italian acquired from corpora by manual clustering of distributional information about Italian verbs, freely available under a Creative Common Attribution 3.0 license [1]. T-PAS are corpus-derived verb patterns with specification of the expected semantic type (ST) for each argument slot, such as [[Human]] guida [[Vehicle]]. T-PAS is the first resource for Italian in which semantic selection properties and sense-in context distinctions of verbal predicates are characterized fully on empirical ground. In the

[1] tpas.fbk.eu.

resource, the acquisition of T-PAS is totally corpus-driven. We discover the most salient verbal patterns using a lexicographic procedure called Corpus Pattern Analysis (CPA, Hanks 2004), which relies on the analysis of co-occurrence statistics of syntactic slots in concrete examples found in corpora.

Important reference points for the T-PAS project are FrameNet (Ruppenhofer et al. 2010), and VerbNet (Kipper-Schuler 2005) and PDEV (Hanks and Pustejovksy 2005), a pattern dictionary of English verbs which is the main product of the CPA procedure applied to English. As for Italian, a complementary project is LexIt (Lenci et al. 2012), a resource providing automatically acquired distributional information about verbs, adjectives and nouns.

T-PAS is being developed at the Dept. of Humanities of the University of Pavia, in collaboration with the Human Language Technology group of Fondazione Bruno Kessler (FBK), Trento, and the technical support of the Faculty of Informatics at Masaryk University in Brno (CZ). The first release contains 1000 analyzed average polysemy verbs, selected on the basis of random extraction of 1000 lemmas out of the total set of fundamental lemmas of Sabatini Coletti 2008, according to the following proportions: 10 % 2-sense verbs, 60 % 3-5-sense verbs, 30 % 6-11-sense verbs.

The resource consists of three components: a repository of corpus-derived T-PAS linked to lexical units (verbs); an inventory of about 230 corpus-derived semantic classes for nouns, relevant for disambiguation of the verb in context; a corpus of sentences that instantiate T-PAS, tagged with lexical unit (verb) and pattern number. The reference corpus is a reduced version of ItWAC (Baroni & Kilgarriff, 2006).

As referenced above, T-PAS specifies the expected semantic type (ST) for each argument slot in the structure; in ST annotation, the analyst employs a shallow list of semantic type labels ([[Human]], [[Artifact]], [[Event]], ecc.) which was obtained by applying the CPA procedure to the analysis of concordances for ca 1500 English and Italian verbs.

Pattern acquisition and ST tagging involves the following steps:

- choose a target verb and create a sample of 250 concordances in the corpus;
- while browsing the corpus lines, identify the variety of relevant syntagmatic structures corresponding to the minimal contexts where all words are disambiguated;
- identify the typing constraint of each argument slot of the structure by inspecting the lexical set of fillers: such constraints are crucial to distinguish among the different senses of the target verb in context. Each semantic class of fillers corresponds to a category from the inventory the analyst is provided with. If none of the existing ones captures the selectional properties of the predicate, the analyst can propose a new ST or list a lexical set, in case no generalization can be done;
- when the structures and the typing constraints are identified, registration of the patterns in the Resource using the Pattern Editor (see Fig. 1.) Each pattern has a unique identification number, and a description of its sense, expressed in the form of an implicature linked to the typing constrains of the pattern, for example the T-PAS in Fig. 1. has the implicature [[Human]] *legge* [[Document]] *con grande interesse* (read with high interest):

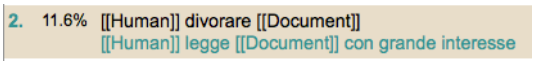


Fig 1. Selected pattern for verb *divorare*

- assignment of the 250 instances of the sample to the corresponding patterns, as shown in Fig. 2:



Fig 2. Sample annotation for pattern 2 of *divorare (devour)* - SketchEngine

In this phase, the analyst annotates the corpus line by assigning it the same number associated with the pattern.

3 Disambiguation Method

In this section, we present a disambiguation method for corpus patterns and apply it to the task of verb disambiguation with respect to the T-PAS resource. The method is based on identifying the important elements of a pattern which are disambiguating the verb in the text. The importance of such elements

is evaluated according to their effect on the sense of the verb, expressed as a relationship between the senses of the words inside a pattern. It has been noted that the relationship between verb meaning and semantic types is constrained, such that the context matched by a pattern is the sufficiently minimal context for disambiguation. This relationship, called chain clarifying relationship (CCR), is instrumental in doing pattern matching as well as in finding new patterns. (Popescu & Magnini 2007, Popescu 2012).

From a practical point of view, the probability of occurrence of a word and the probability of the verb are independent given the probability of the semantic type. As such, the CCR is very efficient in dealing with sparseness problem. This observation has a big positive impact on the disambiguation system, because it directly addresses two issues:1) the necessity of large quantity of training alleviating the data sparseness problem (Popescu 2007, Popescu 2013) and 2) the overfitting of probabilities, with important consequences for the disambiguation of less frequent cases (Popescu et. al 2007). The method divides the vocabulary in congruence classes generated by CCR for each verb and we build a classifier accordingly (Popescu 2013 and Popescu et al. 2014). To this end, we carry out an automatic analysis of the patterns at the training phase, which allows us to compute a confusion matrix for each verb pattern number and congruence class. The exact procedure is presented below.

We introduce here a probabilistic system which does partial pattern matching in text on the basis of individual probabilities which can be learned form training. Matching a corpus pattern against a verbal phrase involves labelling the heads of the constituents with semantic features and the verb with a pattern number. We build a probabilistic model in which we compute the probability in Equation (1).

$$p(t_0, t_1, t_2, t_3, .., t_n, w_1, w_2, w_n) \quad (1)$$

where t_0 is the pattern number, t_i is the semantic type of the word w_i, which is the head of the ith constituent, with i from 1 to n. For a given sentence we choose the most probable labeling, Equation (2)

$$p(t_0^c, t_1^c, t_2^c, t_3^c, .., t_n^c, w_1^c, w_2^c, w_n^c) = \arg\max_{t_i} p(t_0, \mathrm{w}_n) \quad (2)$$

On the basis of the relationship existing between the senses of the fillers of the corpus pattern given by CCR, and the fact that the patterns have a regular language structure, we learn for each verb its discriminative patterns with semantic types. Using the chain formula, and grouping the terms conveniently, Equation (1) becomes Equation (3).

$$\begin{aligned} & p(t_0, t_1, t_2, t_3, .., t_n, w_1, w_2, w_n) = p(t_0)p(w_1|t_0)... \\ & ... \; p(t_n|t_0, w_1, t_1, w_2, ..., t_{n-1}, w_n) \\ & \simeq \mathrm{p(t_0)}p(w_1|t_0)pt_1|t_0, w_1)p(w_2|t_0)p(t_2|t_0, w_2) \; ... \\ & ... \; p(t_n|t_0, w_n) \\ & \simeq \mathrm{p(t_0)}p(w_1|t_0)p(t_1|t_0)p(t_1|w_1)p(w_2|t_0)p(t_2|t_0)p(t_2|w_2) \; ... \\ & ... \; p(t_n|t_0)p(t_n|t_0)p(t_n|w_n) \end{aligned} \quad (3)$$

The quantities on the right hand side are computed as follows:

- $p(t_0)$ is the probability of a certain pattern. This probability is estimated from the training corpus, via ratio of frequencies.

- $p(w_i|t_0)$ is the probability of a certain word to be the head of a constituent. We used the Italian dependency parser MaltParser (Lavelli et al. 2009) for determining the head of the syntactic constituents and their dependency relationships. However, we allow for any content word to fulfill the role of subject, object or prepositional object with a certain probability. This probability is set a priori on the basis of the parser's probability error and the distance between the word and the verb.

- $p(t_i|t_0, w_i)$ is the probability that a certain word at a certain position in the pattern carries a specific semantic type. This probability is equated to $p(t_i|w_i)p(t_i|t_0)$, assuming independence between the verb sense and the word given the semantic type. The first of the two later probabilities is extracted from Semcor (Miller et al. 1993, Pianta el al. 2002) and Lin distance (Lin 1998), considering the minimal distance between a word and a semantic type. The second probability is computed at the training phase considering the frequency of a semantic type inside the pattern.

The probabilities may be affected by the way the training corpus is compiled. It is assumed that the examples have been drawn randomly from a large, domain independent corpus. We call the resulting model the CF_CCR, from chain formula with CCR.

4 Experiments and Results

We performed the following experiment: we have considered all the verbs present in T-PAS at this

System	*Attribute*	*Macro Average*
5libSVM	5 words	67.871
10libSVM	10 words	65.556
CF_CCR	syn-sem	71.497

Table 1: Direct global evaluation

moment. We have excluded the mono pattern verbs, as in this case there is no ambiguity, and we are left with 916 verbs. For each verb we split the T-PAS examples into train and test considering 80% and 20% respectively. We trained two SVM bag of words models considering a window of 5 and 10 tokens around the target verb. We used the WEKA libsvm library and we compared the results against the model presented in the previous section. The global results, macro average, are presented in Table 1. We report the precision here, corresponding to the true positive, as we annotated all examples.

The model 10libSVM performed worse than the other two, probably due to the noise introduced. On the other hand, it is surprising how well the 5 window model performed. We reckon that this is because of the fact that most of the time the direct object is within 5 words distance from the verb and the majority of the T-PAS patterns consider the direct object as the main distinctive feature, and the set of words occurring in the T-PAS examples is relatively small. Therefore the probability of seeing the same word in test and train is big.

We considered to investigate more the distribution of the results. For this, we considered the better performing bag of word system, namely five words window system, 5libSVM against CF_CCR.

The variation of precision is actually large. It ranges from below 10% to 90%.The number of verbs which are disambiguated with a precision bigger than 60% represents the large majority with 72% of the verbs. This suggests, that instead of macro average, a more indicative analysis could be carried out by distinguishing between precision on verbs with different number of patterns.

We looked into the influence of the number of patterns for the accuracy of the systems. As expected, the bigger the number of patterns the less precise is the system. The extremities are the ones that have an accelerated rate of growth. For example, the precision over 90% and under 10% goes from the biggest (lowest) coverage for 2 patterns, to lowest (bigger) for 9 patterns. The behaviour of CF_CCR is somewhat different from 5libSVM, the CF_CCR is able to identify more senses, thus achieves a better precision for verbs with more than 6 patterns, than 5libSVM does. In Table 2 we present comparatively how many times the system makes a less than 50% accurate prediction for patterns that have a frequency greater than 10%. As we can see, the CF_CCR system is between 6% to 20% percent better than 5libSVM in coping with these cases, proving that combining syntactic and semantic information reduces the overfitting. The fact that the absolute number decreased also with the number of patterns is due mainly to the fact that also the number of examples decreases drastically.

No. Patterns	*5libSVM*	*CF_CCR*
2	57	53
3	118	109
4	126	114
5	112	98
6	85	77
7	50	44
8	29	23
9	28	21

Table 2: Errors on patterns with frequency $>10\%$

5 Conclusion

We have presented a word sense disambiguation system for Italian verbs, whose senses have been derived from T-PAS, a lexical resource that we have recently developed. This is the first work (we hope that many others will follow) attempting to use T-PAS for a NLP task. The WSD system takes advantage of the T-PAS structure, particularly the presence of semantic types for each verbal argument position. Results, although preliminary, show a very good precision.

As for the future, we intend to consolidate the disambiguation methodology and we aim at a more detailed annotation of the sentence argument, corresponding to the internal structure of verb patterns. We plan to extend the analysis of the relationship between senses of the different positions in a pattern in order to implement a metrics based on tree and also to substitute the role of the parser with an independent pattern matching system. The probabilistic model presented in Section 3 can be extended in order to determine also the probability that a certain word is a syntactic head.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923) and by the BCROCE project of the autonomous province of Trento.

Reference

E. Agirre and P. Edmonds, 2006. *Word sense disambiguation: algorithms and applications*, Springer.

M. Baroni and A. Kilgarriff. 2006. Large Linguistically-Processed Web Corpora for Multiple Languages. In *EACL 2006 Proceedings*, pp. 87-90.

F. Bertagna, A. Toral and N. Calzolari, 2007. *EVALITA 2007: THE ALL-WORDS WSD TASK*, in *Proceedings of Evalita 2007*, Rome.

C. Fellbaum, 1998. *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

P. Hanks. 2004. Corpus Pattern Analysis. In G. Williams and S. Vessier (eds.). *Proceedings of the XI EURALEX International Congress*. Lorient, France (July 6-10, 2004), pp. 87-98.

P. Hanks and J. Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, 10 (2).

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini and O. Popescu. 2014. T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari et al. (eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland (May 26-31, 2014), Paris: European Language Resources Association (ELRA), 890-895.

K. Kipper-Schuler. 2005. *VerbNet: A broad coverage, comprehensive verb lexicon*. Ph.D. Thesis. University of Pennsylvania, USA.

A. Lenci, G. Lapesa and G. Bonansinga. 2012. LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC12), Istanbul (May 23-25, 2012), pp. 3712-3718.

A. Lavelli, J. Hall, J. Nilsson and J. Nivre. 2009. MaltParser at the EVALITA 2009 Dependency Parsing Task. In *Proceedings of EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*, Reggio Emilia, Italy.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Global WordNet Conference*, Mysore

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, Madison, Wisconsin, USA.

O. Popescu and B. Magnini. 2007. Sense Discriminative Patterns for Word Sense Disambiguation. In *Proceedings of the SCAR Workshop 2007, NODALIDA*, Tartu.

O. Popescu, S. Tonelli, Sara and E. Pianta. 2007. IRST-BP: Preposition Disambiguation based on Chain Clarifying Relationships Contexts. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague.

O. Popescu. Building a Resource of Patterns Using Semantic Types. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC-12), Istanbul.

O. Popescu. 2013. Learning Corpus Pattern with Finite State Automata. In *Proceedings of the ICSC 2013*, Postadam.

O. Popescu, M. Palmer, P. Hanks.2014. In *Mapping CPA Patterns onto OntoNotes Senses*. LREC 2014: 882-889.

A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, A. Zampolli, 2003. *ItalWordnet: Buiding a Large Semantic Database for the Automatic Treatment of Italian.* In *Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa (Linguistica Computazionale a Pisa), Linguistica Computazionale, Special Issue, Vol. XVI-XIX, pag. 745-791*.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson and J. Scheffczyk. 2010. *FrameNet II: Extended theory and practice*. International Computer Science Institute, University of Berkeley. (Manuscript, Version of September 14, 2010).

F. Sabatini and V. Coletti 2007. *Il Sabatini-Coletti. Dizionario della lingua italiana 2008*, Milano, Rizzoli-Larousse.

10.12871/CLICIT2014160

Il corpus Speaky

Fabio Poroli, Massimiliano Todisco, Michele Cornacchia,
Cristina Delogu, Andrea Paoloni, Mauro Falcone
Fondazione Ugo Bordoni
Viale del Policlinico 147 – 00161 Roma
{ fporoli,mtodisco,mcornacchia,cdelogu,apaoloni,mfalcone @fub.it }

Abstract

Italiano. In questo lavoro presentiamo un corpus di dialogo uomo-macchina acquisito nell'ambito del progetto SpeakyAcutattile con la tecnica del Mago di Oz. Il corpus contiene più di 60 ore di registrazione di audio, trascritto ortograficamente, e di video. Si presta in particolar modo all'analisi della gestione del turno e della risoluzione degli errori. La simulazione del sistema con il Mago di Oz è stata orientata a una produzione di dialogo vocale senza vincoli da parte del soggetto, sia a livello di gestione del turno, sia a livello di composizione della frase.

English. *In this paper we describe a corpus of man-machine dialogue achieved in the context of SpeakyAcutattile project by the Wizard-of-Oz technique. The corpus consists of more than 60 hours of audio, orthographically transcribed, and video recording. It is particularly suited for the analysis of both turn managing and errors recovering. The system simulation by Wizard-of Oz has been oriented to support a restrictions-free vocal production by subjects, whether for turn managing or for input string composition. .*

1 Introduzione

In questo lavoro presentiamo un corpus di dialogo uomo-macchina acquisito nell'ambito del progetto Speaky Acutattile, una piattaforma digitale per la domotica pensata per il sostegno all'utenza debole (anziani, non vedenti, ecc.), in cui la Fondazione Ugo Bordoni ha introdotto un'interfaccia utente basata sul riconoscimento della voce (Poroli et al., 2013). La piattaforma è stata progettata per fornire agli utenti uno strumento semplificato per la gestione degli elettrodomestici e degli altri dispositivi multimediali presenti in casa (televisione, stereo, etc.), ma anche per l'accesso in rete ai molti servizi di pubblica utilità, come l'assistenza sanitaria, i pagamenti online, le prenotazioni, l'acquisto di titoli di viaggio, ecc. Per la raccolta dati è stata utilizzata la tecnica del Mago di Oz (Fraser and Gilbert, 1991; Dahlback et al., 1993). La tecnica, sebbene richieda maggiori attenzioni e risorse rispetto ad altre strategie di elicitazione del parlato, viene comunemente collocata fra i sistemi più affidabili per la prototipazione di interfacce vocali *user-oriented* e la raccolta dati sulle modalità di interazione con gli utenti.
Eccettuati alcuni vizi strutturali legati al contesto sperimentale (come ad esempio, il minor coinvolgimento del soggetto rispetto all'utente reale), la rilevanza di un *corpus* di dialogo uomo-macchina raccolto con tale metodo viene determinata dalla definizione di alcuni parametri che fissano *a priori* il comportamento del *Mago*, di fatto rendendolo da parte dell'utente il più possibile assimilabile ad una macchina (*machine-like)*.
In questo lavoro è stato inoltre applicato un modello di simulazione di sistema a iniziativa mista (Allen et al., 2001) con grammatiche "*frame-and-slot*" (Bobrow et al., 1977), comprensivo del protocollo di comportamento del dialogo.

La tecnica del Mago di Oz ha consentito pertanto di elaborare le grammatiche di comprensione del dialogo con alcune varianti, verificando nel contempo le reazioni dei soggetti di fronte a un sistema che appariva come reale e non costringeva a percorsi di interazione obbligati per la risoluzione dei compiti.

2 Allestimento dell'acquisizione

2.1 Ambiente sperimentale e soggetti

L'acquisizione dei dati sperimentali è stata condotta nel laboratorio di usabilità del Ministero dello Sviluppo Economico a Roma. Il laboratorio era formato da due stanze, separate da una finestra con specchio riflettente a una via. Analoghe sessioni di registrazione sono state realizzate anche nelle città di Palermo, Torino e Padova , con il Mago di Oz connesso in remoto per il controllo dell'interazione utente.
Ogni soggetto veniva accompagnato e fatto sedere a un tavolo su cui si trovava una lista riepilogativa dei compiti da svolgere. Uno sperimenta-

tore coordinava l'accoglienza, compilava la liberatoria di *privacy* per la sessione, forniva le istruzioni di base e assistenza su richiesta anche durante la fase attiva dell'interlocuzione tra utente e Mago.
Il soggetto, nel caso di appartenenza alla classe Anziani, poteva usufruire di feedback informativi su uno schermo 42" (a distanza di 3m circa) che visualizzava in un angolo un avatar umanoide parlante (Figura 1) denominato Lucia (Cosi et al., 2003). Un ambiente associabile al dominio coinvolto e al compito da svolgere completava il *setting* grafico delle videate, per esempio un menu di prodotti da acquistare fra quelli menzionati nel compito o i canali TV preselezionati un una lista di preferenze.



Figura 1: Schermata di lavoro di Speaky-WOz (lato utente)

Il Progetto Speaky Acutattile ha sviluppato dunque l'idea di una piattaforma digitale avanzata per la domotica, costituita da più moduli o dispositivi polifunzionali integrabili, conforme agli standard vigenti e con interfaccia semplice controllata per mezzo della voce.

Il programma di Progetto ha richiesto nello specifico che i servizi fossero rivolti a un'utenza diversamente abile non-vedente (o ipo-vedente) e agli anziani in *digital divide*, cioè persone con età nell'intervallo 65-80 anni, di media scolarizzazione e non dotati di competenze informatiche di base. Per ognuna delle quattro città partecipanti hanno partecipato 20 soggetti (bipartiti per genere M/F, con istruzione medio-bassa e senza conoscenze pregresse in materia di ICT), di cui tipicamente 10 anziani e 10 non-vedenti, per una totale complessivo sul territorio nazionale di 80 individui (oltre a 9 soggetti utilizzati nel *pretest*).

La Tabella 1 riassume le caratteristiche delle due classi utenza.

Città	Soggetti	M/F	Età Media	DS Età	SMB	NO ICT
Roma AN	10	1,0	66,7	6,4	X	X
Roma NV	10	0,7	64,1	16,6	X	X
Padova AN	10	0,3	72,0	5,3	X	X
Padova NV	10	0,9	56,1	16,2	X	X
Palermo AN	10	1,0	69,0	14,0	X	X
Palermo NV	10	0,4	50,8	20,8	X	X
Torino AN	10	1,0	70,1	4,0	X	X
Torino NV	10	1,5	53,3	11,2	X	X

Tabella 1: Utenza sperimentale (Legenda: AN=Anziani, NV=Non-Vedenti, SMB=Scolarizzazione Medio-Bassa, NO ICT=nessuna esperienza ICT pregressa)

2.2 Compiti

Sono stati redatti in totale 48 compiti: ogni soggetto ha svolto circa 20 compiti diversi, composti ognuno da 2-3 attività connesse tra loro. I compiti sono stati progettati in conformità delle caratteristiche del modulo di comprensione del futuro sistema Speaky, secondo il modello *frame-and-slot*: ogni sotto-compito prevedeva perciò un certo numero di variabili da fornire al sistema (di cui alcune obbligatorie e altre facoltative) per il completamento dell'attività. Le istruzioni ai soggetti sono state impartite in due momenti o fasi:
- all'accoglienza con una descrizione a voce del compito da svolgere, ai fini della contestualizzazione degli obiettivi da raggiungere;
- durante il compito, quando il soggetto poteva consultare un promemoria riepilogativo delle richieste all'esecuzione del compito (Tabella 2).

Descrizione	**Descrizione estesa**	**Variabili**
Impostare gli orari per l'assunzione di alcuni medicinali.	Il soggetto deve dare il nome del medicinale, la quantità, l'orario d'assunzione ed eventualmente il giorno.	(S1) nome, (S2) quantità, (S3) orario, (S4) giorni della settimana.

Tabella 2: Esempio di promemoria riepilogativo di un compito

2.3 Frasi del Mago di Oz verso i soggetti

Per ogni compito è stato predisposto un insieme di frasi predefinite (Tabella 3) e dipendenti dal dominio (*domain-dependent*), che il Mago inviava ai soggetti di volta in volta, in consonanza con gli obiettivi generali e l'occorrenza specifica dell'azione richiesta.

Compito	Sub	Fase	Out	Tipo	TESTO DA INVIARE
1	1	1	1	1	Ciao! Come posso aiutarti?
1	1	2	1	4	Non riesco a comprendere, puoi ripetere?
1	1	3	1	6	Sono aperte le finestre del salotto e della cucina, le altre sono chiuse.
1	1	3	2	6	Nel salotto la finestra è aperta.
1	1	3	3	6	La finestra della cucina è aperta.
1	1	3	4	6	In bagno la finestra è chiusa.
1	1	3	5	6	La finestra della camera da letto è chiusa.
1	2	1	1	1	Ti serve altro?
1	2	1	2	2	Se vuoi posso chiuderle o aprirle.
1	2	1	3	2	Vuoi chiuderne o aprirne qualcuna?
1	2	2	1	3	Vuoi aprire la finestra del bagno?
1	2	3	1	6	Ho chiuso le finestre del salotto e della cucina.
1	2	3	2	6	Ho aperto la finestra della camera da letto
1	2	3	3	6	Ho chiuso le finestre di salotto e cucina, e aperto la camera da letto
1	2	3	4	6	La finestra del bagno è già chiusa.
1	3	1	1	1	Posso esserti ancora utile?
1	3	2	1	3	L'antifurto non è impostato.
1	3	2	2	3	Vuoi che l'antifurto si attivi quando esci di casa o impostare un orario?
1	3	3	1	6	L'antifurto si attiverà quando esci di casa.
1	3	3	2	6	L'antifurto si attiverà all'ora impostata.

Tabella 3: Numerazione delle risposte

Ogni insieme di compiti è diviso rispettivamente in sotto-compiti, fase del dialogo e tipologia delle frasi, a partire dalla sintesi del dialogo pratico (Allen et al., 2000) proposta da Alexandersson et al. (1997). Come illustrato nella Tabella 3, la prima colonna definisce il compito, la seconda il sotto-compito, la terza la fase del dialogo (1 = apertura, 2 = negoziazione, 3 = chiusura) mentre la quarta il tipo di frasi (1 = apertura generica, 2 = apertura guidata, 3 = richiesta di completamento, 4 = richiesta di ripetizione, 5 = richiesta di conferma errata, 6 = di completamento). Il set generico, *domain-independent*, è invece uguale per ogni compito e comprende le frasi il cui uso è esteso a ogni interazione, come i feedback di accordo, i saluti e le risposte a richieste fuori dominio. Durante l'interazione il Mago usa un'interfaccia grafica, vedi Figura 2, per la selezione manuale dei testi audio da inviare agli altoparlanti del sistema collocati nella stanza utenti del laboratorio.



Figura 2: interfaccia grafica del Mago

2.3 Svolgimento dell'interazione

Ogni dialogo inizia con una frase di attivazione del parlante, a cui segue una risposta del tipo "*How may I help you*?" (Gorin et al., 1997), con cui viene lasciata l'iniziativa al parlante per indicare l'attività da svolgere e, potenzialmente, per organizzarne la risoluzione in un solo turno. La fase di negoziazione, collocata tra l'apertura del compito e il suo completamento, è caratterizzata da diversi tipi di frasi: richieste di completamento, richieste di riformulazione, richieste di conferma. Successivamente all'apertura del compito, l'iniziativa passa al parlante, la cui frase può o meno includere tutte le informazioni necessarie; nel caso non vi siano tutte le informazioni necessarie, l'iniziativa torna al mago, il cui compito è elicitare i dati mancanti con frasi di completamento predisposti per coprire ogni caso possibile di assenza di informazioni. La fase di negoziazione prevede anche errori simulati tramite richieste di ripetizione e/o di conferma. Anche in questo caso è stato rispettato per gran parte un protocollo definito a priori: ogni compito prevedeva, infatti, l'uso di una richiesta di ripetizione (ex: «Non ho capito, puoi ripetere?») e di una richiesta di conferma errata, scritta appositamente per ogni compito, da usare coerentemente con le informazioni presenti nella frase del parlante. A seguito del completamento della prima attività, mancando un'eventuale apertura di quella successiva da parte dell'utente (entro tre secondi), è compito del Mago indirizzare il dialogo verso il secondo sotto-compito con una richiesta di apertura generica («Ti serve altro?»).
Per la gestione del dialogo è stato usato un modello a iniziativa mista. Ad esempio, a fronte di un richiesta di conferma errata, il parlante può, infatti, correggere egli stesso l'informazione direttamente nel turno successivo a quello del Mago (es. W: «Vuoi avere informazioni sui treni da Roma a Torino?» – U: «No, da Roma a Mila-

no»); allo stesso modo può prendere il turno (e l'iniziativa) subito dopo la chiusura dell'attività per aprire l'attività successiva. In assenza di un'apertura, il Mago imposta l'avvio di una seconda attività dopo 2-3 secondi di silenzio.

2.4 Descrizione del corpus

Il *corpus* (disponibile in formato audio, video e testuale) è costituito dalle registrazioni delle 80 sessioni di interazione con il sistema simulato, condotte con altrettanti utenti. Ogni sessione comprende circa 20 [1] dialoghi pratici tra il soggetto e il sistema simulato, oltre alle istruzioni iniziali fornite dallo sperimentatore al soggetto e le brevi interazioni tra un dialogo e l'altro. La durata media di ogni sessione è stata di 43 minuti, per un totale di più di 60 ore di registrazione.
Il segnale vocale utile pronunciato dai soggetti è stimabile in circa 16 ore (circa il 25% del registrato disponibile). Tale segnale è stato acquisito da cinque diversi canali a Roma, per tutte le altre città si hanno solo due canali: microfonico e da ripresa video frontale. La tabella 4 mostra i formati utilizzati per tutti i dispositivi e le relative dimensioni dei file per soggetto.

DISPOSITIVO	FORMATO	DIMENSIONI x SOGGETTO
Radiomicrofono Sennheiser XSW 12	PCM wav 48kHz @24bit mono	~ 600 MB
Telefono cellulare Samsung Galaxy SII	PCM wav 32kHz @16bit mono	~ 200 MB
Array microfonico Microsoft Kinect	PCM wav 16kHz @16bit mono	~ 100 MB
Video front ZOOM Q3HD	MPEG-4 1280x720 @25fps PCM wav 48kHz @24bit stereo	~ 3 GB
Video back ZOOM Q3HD	MPEG-4 1280x720 @25fps PCM wav 48kHz @24bit stereo	~ 3 GB

Tabella 4: Dispositivi e formati di acquisizione

Il *corpus* è disponibile anche in formato testuale, trascritto a partire dalla registrazione effettuata tramite il radiomicrofono. Al momento non sono state presi in considerazione le analisi dei dati video (che riprendono i movimenti e le espressioni del soggetto da due diverse angolazioni). Il *corpus* testuale è stato sincronizzato alle tracce audio tramite il software Transcriber 1.5.1 (Barras et al., 2000). Considerato l'allineamento del testo con i file audio, che consente un rapido recupero dei segmenti di dialogo, la trascrizione è stata di tipo ortografico, organizzata per turni. Sono tuttavia stati annotati fenomeni dialogici tipici, come pause, pause piene, false partenze e sovrapposizioni.

3 Ulteriori considerazioni sul corpus

Il *corpus* si presta particolarmente a studi sulla gestione del turno e dell'iniziativa, e sulla gestione degli errori. Tali analisi, oltre a darci informazioni su alcune meccaniche dialogiche di una particolare situazione comunicativa (il dialogo uomo-macchina), possono costituire un utile supporto conoscitivo per integrare le grammatiche di comprensione e le architetture del gestore di dialogo. Ovviamente, il *corpus* raccolto presenta alcuni limiti su altri livelli di analisi linguistica. Infatti, l'utenza principale del sistema, composta da anziani e non vedenti, ha reso necessaria la presenza di uno sperimentatore nella stanza del soggetto e l'uso di un foglio riepilogativo delle attività, variabili che potevano condizionare le scelte lessicali e morfologiche da parte dei soggetti. Da un punto di vista applicativo, tale condizionamento non è un problema: l'ampliamento del dizionario e delle possibili situazioni nel singolo turno di dialogo andranno certamente implementati in una fase successiva del progetto, con dati ottenuti dall'uso reale del sistema reale. Al contrario, il comportamento degli utenti nelle situazioni d'errore e in relazione alla gestione del turno sembra essere meno sensibile al contesto sperimentale, e fornisce valide informazioni per la progettazione del sistema, sia nell'ambito del progetto Speaky, sia, più in generale, per lo studio dell'interazione uomo-macchina.

4 Conclusioni e future attività

La tecnica del Mago di Oz ci ha permesso di ottenere un *corpus* controllato su alcuni aspetti dell'interazione che forniscono indicazioni per l'architettura del sistema di dialogo. I dati attuali verranno integrati con l'acquisizione di un nuovo *corpus* in cui il Mago di Oz "umano" verrà sostituito dal prototipo del sistema, a fronte dello stesso tipo di utenza sperimentale e degli stessi scenari d'uso, allo scopo di ottenere dati confrontabili con gli attuali, sia per migliorare le prestazioni del sistema, sia per ottenere preziosi informazioni sulla tecnica del Mago di Oz.
Le politiche di distribuzione del database saranno definite al termine del progetto (giugno 2015), e auspicabilmente saranno di gratuità per attività di ricerca, ovviamente previo accordo NDA (*Non Disclosure Agreement*).

[1] Variazione dovuta alla presenza o meno dell'ultimo compito sul controllo delle funzioni domotiche interattive (p.e. regolazione altezza delle tapparelle).

Bibliografia

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. *Dialogue Acts in VERBMOBIL-2*. Verbmobil-Report, 204.

James F. Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6:1-16

James F. Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2001. Towards Conversational Human-Computer Interaction. *AI Magazine*, 22 (4):27-37

Claude Barras, Edouard Geoffrois, Zhibiao Wu and Mark Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication (special issue on Speech Annotation and Corpus Tools)*, 33 (1–2).

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry S. Thompson and Terry Winograd. 1977. GUS, A frame driven dialog system. *Artificial Intelligence*, 8:155–173

Piero Cosi, Andrea Fusaro, Graziano Tisato. 2003. LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. *Proceedings of Eurospeech 2003*, Geneve, Switzerland.

Nils Dahlback, Arne Jonsson and Lars Ahrenberg. 1993. Wizard of Oz studies–why and how. *Knowledge-based systems*, 6(4):258–266

Allen L. Gorin, Giuseppe Riccardi and Jeremy H. Wright. 1997. How may I Help You?. *Speech Communication*, 23:113-127

Norman Fraser and Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language*, 5(1): 81–99

Fabio Poroli, Cristina Delogu, Mauro Falcone, Andrea Paoloni, Massimiliano Todisco. 2013. Prime indagini su un corpus di dialogo uomo-macchina raccolto nell'ambito del Progetto SpeakyAcutattile. *Atti del IX Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce*, Venezia, Italy.

Fabio Poroli, Andrea Paoloni, Massimiliano Todisco. 2014 (in corso di stampa). Gestione degli errori in un corpus di dialogo uomo-macchina: strategie di riformulazione. *Atti del X Convegno Nazionale AISV – Associazione Italiana di Scienze della Voce*, Torino, Italy.

10.12871/CLICIT2014161

Converting the parallel treebank ParTUT in Universal Stanford Dependencies

Manuela Sanguinetti
Dipartimento di Informatica
Università di Torino (Italy)
Corso Svizzera, 185, 10149 Torino
manuela.sanguinetti@unito.it

Cristina Bosco
Dipartimento di Informatica
Università di Torino (Italy)
Corso Svizzera, 185, 10149 Torino
cristina.bosco@unito.it

Abstract

English. Assuming the increased need of language resources encoded with shared representation formats, the paper describes a project for the conversion of the multilingual parallel treebank ParTUT in the *de facto* standard of the Stanford Dependencies (SD) representation. More specifically, it reports the conversion process, currently implemented as a prototype, into the Universal SD format, more oriented to a cross-linguistic perspective and, therefore, more suitable for the purpose of our resource.

Italiano. *Considerando la crescente necessità di risorse linguistiche codificate in formati ampiamente condivisi, l'articolo presenta un progetto per la conversione di una risorsa multilingue annotata a livello sintattico nel formato, considerato uno standard de facto, delle Stanford Dependencies (SD). Più precisamente l'articolo descrive il processo di conversione, di cui è attualmente sviluppato un prototipo, nelle Universal Stanford Dependencies, una versione delle SD maggiormente orientata a una prospettiva inter-linguistica e, per questo, particolarmente adatta agli scopi della nostra risorsa.*

1 Introduction

The increasing need to use language resources for the development and training of automatic systems goes hand in hand with the opportunity to make such resources available and accessible. This opportunity, however, is often precluded by the use of different formats for encoding linguistic content. Such differences may be dictated by several factors that, in the specific case of syntactically annotated corpora, or treebanks, may include the choice of constituency *v*s dependency-based paradigm, the specific morphological and syntactical features of the language at issue, or the end use the resource has been designed for. This variety of formats makes it more difficult the reuse of these resources in different contexts.

In the case of parsing, and of treebanks, a few steps towards the spread of formats that could be easily shared by the community has led, also thanks to the efforts devoted to the organization of evaluation campaigns, to the use of what have then become *de facto* standards. This is the case, for example, of the Penn Treebank format for constituency paradigms (Mitchell et al., 1993).
Within the framework of dependency-based representations, a new format has recently gained increasing success, i.e. that of the Stanford Typed Dependencies. The emergence of this format is attested by several projects on the conversion and harmonization of treebanks into this representation format (Bosco et al., 2013; Haverinen et al., 2013; McDonald et al., 2013; Tsarfaty, 2013; Rosa et al., 2014).

The project described in this paper is part of these ones and concerns in particular the conversion into the Stanford Dependencies of a multilingual parallel treebank for Italian, English and French called ParTUT. The next section will provide a brief description of ParTUT and its native format, along with that of the Universal Stanford Dependencies, while Section 3 will be devoted to the description of the conversion process, with some observations on its implications in the future development of ParTUT.

2 Data set

In this section, we provide an overview of ParTUT and of the two annotation formats at issue, focusing on their design principles and peculiarities.

2.1 The ParTUT parallel treebank

ParTUT[1] is a parallel treebank for Italian, English and French, designed as a multilingual development of the Italian Turin University Treebank (TUT)[2] (Bosco, 2001), which is also the reference treebank for the past parsing tracks of Evalita, the evaluation campaign for Italian NLP tools[3].

The whole treebank currently comprises an overall amount of 148,000 tokens, with approximately 2,200 sentences in the Italian and English sections, and 1,050 sentences for French[4].
ParTUT has been developed by applying the same strategy, i.e. automatic annotation followed by manual correction, and tool exploited in the Italian TUT project, i.e. the Turin University Linguistic Environment (TULE) (Lesmo, 2007; Lesmo, 2009), first developed for Italian and then extended for the other languages of ParTUT (Bosco et al., 2012). Moreover, one of the main developments of the treebank is also the creation of a system for the automatic alignment of parallel sentences taking explicitly into account the syntactic annotation that is included in these sentences (Sanguinetti and Bosco, 2012; Sanguinetti et al., 2013; Sanguinetti et al., 2014).

2.2 The TUT representation format

The treebank is annotated in a dependency-based formalism, partially inspired by the Word Grammar (Hudson, 1990), in particular for what concerns the head selection criteria for determiners and prepositions (that are considered as governors of the nominal and prepositional groups respectively). Other typical features of TUT and ParTUT trees are the use of null elements and the explicit representation of the predicate-argument structure not only for verbs but also for nouns and adjectives.

For what concerns the dependency labels, they were conceived as composed of two components[5] according to the following pattern:

morphoSyntactic–functionalSyntactic.

The main (and mandatory) feature is the second one, specifying the syntactic function of the node in relation to its governor, i.e. whether the node is an argument (ARG), a modifier (MOD) or a more specialized kind of argument (e.g. OBJ or SUBJ) or modifier (e.g. RMOD for restrictive modifiers and APPOSITION for the others) or something else (e.g. COORD or SEPARATOR). This component can be preceded by another one that specifies the morphological category *a*) of the governing item, in case of arguments (e.g. PREP-ARG for the argument of a Preposition), *b*) of the dependent, in case of modifiers (e.g. PREP-RMOD for a prepositional restrictive modifier). In some cases, the subcategory type of this additional component is also included (after a '+' sign), as in DET+DEF-ARG, which should be read as argument of a definite Determiner.

TUT aims at being as linguistically accurate as possible, providing a large number of labels for each of these two components, which can be easily combined together to express the specificity of a large variety of syntactic relations. It thus results in a high flexibility of the format that allowed its application to languages different from the original one (that is Italian).

2.3 The Stanford Typed Dependencies

The Stanford Dependencies (SD) representation (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe and Manning, 2008; de Marneffe et al., 2013) was originally developed for English syntax to provide a scheme that could be easy to use in practical NLP tasks, like Information Extraction. This led to the choice of a format that was theory-neutral as regards the specific grammar, and of a set of widely recognized grammatical relations. Indeed, one of the key features of SD representation, throughout the different versions proposed, is namely the trade-off between linguistic fidelity and readability, which is probably the main factor that determined its usability, and, finally, its success.

Recently, a new version of the SD scheme has been proposed, i.e. the Universal Stanford Dependencies (USD)[6], a revised set of relations more oriented to provide a uniform and consistent structural representation across languages of different linguistic typologies (de Marneffe et al., 2014).

[1] See `http://www.di.unito.it/~tutreeb/partut.html`

[2] `http://www.di.unito.it/~tutreeb`

[3] `http://www.evalita.it/`

[4] The resource is under constant development, and the French part of the newest texts recently added to the collection is yet to be analyzed and included.

[5] In the Italian TUT there is also a third one (omitted here and in the current ParTUT annotation) concerning the *semantic role* of the dependent with respect to its governor.

[6] `http://universaldependencies.github.io/docs/`

By virtue of this claim, more emphasis is put on the *lexicalist hypothesis*, that ultimately favors the relations between content words, with the aim of properly dealing with compounding and rich morphology. This affects, among the other things, the treatment of prepositions, which – rather than mediate between the modified word and the modifier – are now attached as dependents of the latter.
Furthermore, in order to allow the proper recognition of language-specific phenomena, USD representation also opens to possible extensions by adding new grammatical relations as subtypes of the existing ones. This flexibility in the labeling scheme is a valuable feature that USD has in common with the TUT format.

In light of these observations, in this conversion project we opted for the USD representation scheme as the target format.

3 Converting ParTUT

In this section, we describe the current, preliminary, stage of this project. This phase mainly consists in a qualitative comparison of the two formats at hand, drafting a basic mapping scheme between the two relation taxonomies and highlighting the main factors that could impact – both positively and negatively – the conversion process, currently implemented as a prototype.

Mapping scheme As expected, we encountered only 13 cases of 1:1 correspondences between the items of the two relation sets, although, conversely, in relatively few cases (9) a counterpart could not be found either in the source or the target format. The remaining ones entailed a multiple correspondence either 1:*n* or *m*:1. A small selection of such cases, based on the 15 most commonly used relations in ParTUT, is proposed in Table 1.

Preliminary observations The conversion from TUT to USD seems to be especially feasible because of the high flexibility of the involved schemes and their openness to cross-linguistic applications. Furthermore, we can benefit from the fact that we are moving from a source format with a high level of detail to a target format that is more underspecified[7].

[7]TUT scheme comprises an overall amount of 11 *morphoSyntactic* and 27 *functionalSyntactic* features (not to mention their subtypes) that can be combined together, while USD taxonomy includes 42 grammatical relations (which is a further reduction in number, with respect to the previous SD versions).

TUT	USD	H.m.
VERB-SUBJ	*nsubj, csubj*	Y
VERB-OBJ	*dobj, xcomp*	N
VERB-SUBJ/ VERB-INDCOMPL-AGENT	—	—
VERB-OBJ/VERB-SUBJ	*nsubjpass*	N
PREP-ARG	*case*	Y
DET+DEF-ARG	*det, poss*	Y
DEF+INDEF-ARG	*det*	Y
CONJ-ARG	*mark, xcomp*	Y
PREP-RMOD	*case*	Y
ADJC+QUALIF-RMOD	*amod*	N
COORD2ND+BASE	*conj, cc*	Y
COORD+BASE	*cc*	N
END	*punct*	N
SEPARATOR	*punct*	N
TOP-VERB	*root*	N

Table 1: A mapping scheme between the 15 most used syntactic relations in ParTUT and their counterparts in USD. The third column reports whether there is a (either direct or complex) head movement (H.m.) when transforming TUT representation into USD.

English-particular relations, for example, can be easily mapped onto the ones used in ParTUT, and, except for one specific case (that of verb particles), can also be applied to Italian as well as French constructions. Such cases are, respectively, *a*) temporal modifiers expressed with a NP; *b*) predeterminers; *c*) words preceding a conjunction; *d*) possessives.

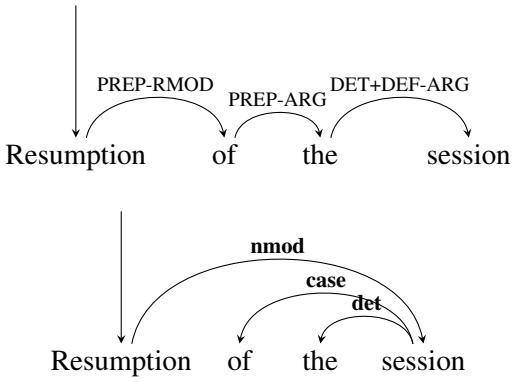
TUT	USD	H.m.
PARTICLE* (*English only)	*prt*	N
NOUN-RMOD-TIME	*tmod*	Y
PDET-RMOD	*predet*	N
COORDANTEC	*preconj*	N
DET+DEF-ARG	*poss*	Y

Table 2: English-particular relations in USD that can be mapped onto the ones used in ParTUT. Unless stated otherwise, all the relations reported in the table can also be applied to Italian and French.

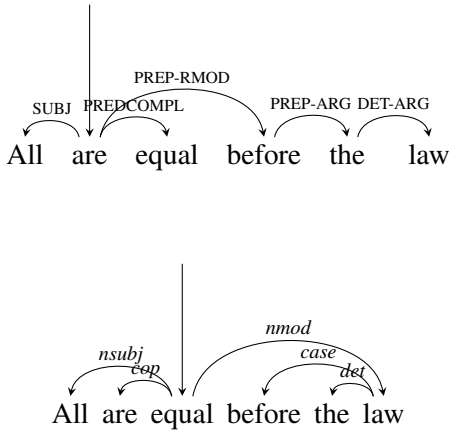
However, as briefly introduced in Section 2.3, the choice to establish meaningful syntactic links between content words not only characterizes this version of SD with respect to the previous ones,

but it also marks a clear boundary with the TUT representation. This aspect entails two basic types of conversion procedures in case of non-direct correspondences, that mainly concern the head selection criteria, and that can be summarized as follows:

- a direct head swapping, where conversion is carried out by a simple inversion of head and dependent roles, as in the case of determiners and prepositions (see below two parallel examples of TUT, in the upper part, and USD, in the lower one):



- a complex transformation that may involve the whole subtree. This is the case, for example, of copulative verbs, that are annotated as heads in ParTUT, and as dependents – together with the subject itself – of the predicative complement in USD (see below).



On the other hand, a more semantically-oriented representation has its benefits as well, especially when dealing with parallel texts in different languages annotated according to the same scheme[8]. This proves useful for translation purposes, which is one of the main goal ParTUT has been conceived for, since it could make it easier the identification of translational correspondences, both manually and automatically, and it constitutes therefore a meaningful step for the further development of the resource as a whole.

Implemented conversion The implementation of the converter is driven by the mapping scheme and observations mentioned above. Each single relation is classified according to different perspectives, including e.g. granularity and mapping cardinality. Adequate procedures are developed to deal with the transformations necessary to the conversion for each relation class. Some procedures, e.g. those implementing a complex restructuring rather than a simple relation renaming, exploit not only the syntactic knowledge but also PoS tagging associated to dependency nodes.
The output of the conversion is made available in different notations known in literature: besides the typical bracketed notation of SD, the converted version will be also released in CoNLL-U[9] and using the Universal PoS tagset proposed by Petrov et al. (2012)

4 Conclusion

In this paper, we briefly described the ongoing project of conversion of a multilingual parallel treebank from its native representation format, i.e. TUT, into the Universal Stanford Dependencies. The main advantages of such attempt lie in the opportunity to release the parallel resource in a widely recognized annotation format that opens its usability to a number of NLP tasks, and in a resulting representation of parallel syntactic structures that are more uniform and, therefore, easier to put in correspondence. Conversion, however, is not a straightforward process, and a number of issues are yet to be tackled in order to obtain a converted version that is fully compliant with the target format. The next steps of this work will focus in particular on such issues.

[8] Although recent works (Schwartz et al., 2012) seem to point to the fact that while content word-based schemes are more readable and "interlingually" comparable, they are harder to learn by machines; this is, in fact, an aspect we intend to verify in the validation phase of the converted resource, by using it as training set for a statistical parser, as also described in Simi et al. (2014).

[9] http://universaldependencies.github.io/docs/format.html

References

Cristina Bosco. 2001. Grammatical relation's system in treebank annotation. In *Proceedings of Student Research Workshop of Joint ACL/EACL Meeting*, pp. 1–6.

Cristina Bosco and Alessandro Mazzei. 2012. The EVALITA Dependency Parsing Task: From 2007 to 2011. In B. Magnini, F. Cutugno, M. Falcone and E. Pianta (Eds.), *Evaluation of Natural Language and Speech Tools for Italian*, pp. 1–12.

Cristina Bosco and Manuela Sanguinetti and Leonardo Lesmo 2012. The Parallel-TUT: a multilingual and multiformat parallel treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), pp. 1932–1938.

Cristina Bosco, Simonetta Montemagni and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 61–69.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), pp. 449–454.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependencies representation. In *Coling2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, CrossParser 08*, pp. 1–8 .

Marie-Catherine de Marneffe and Christopher D. Manning 2008. Stanford Typed Dependencies manual (Revised for the Stanford Parser v. 3.3 in December 2013). `http://nlp.stanford.edu/software/dependencies_manual.pdf`.

Marie-Catherine de Marneffe and Miriam Connor and Natalia Silveira and Samuel R. Bowman and Timothy Dozat and Christopher D. Manning. 2013. More Constructions, More Genres: Extending Stanford Dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics* (DepLing 2013), pp. 187–196.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen,Filip Ginter, Joakim Nivre and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), pp. 4585–4592.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski and Filip Ginter 2013. Building the essential resources for Finnish: the Turku Dependency Treebank. In *Language Resources and Evaluation*, 48:3, pp. 494–531.

Richard Hudson. 1990. *Word Grammar*. Basil Blackwell, Oxford and New York.

Leonardo Lesmo 2007. The rule-based parser of the NLP group of the University of Torino. In *Intelligenza artificiale*, IV:2, pp. 46–47 .

Leonardo Lesmo. 2009. The Turin University Parser at Evalita 2009. In *Proceedings of Evalita '09*, Reggio Emilia, Italy.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of 51st annual meeting of the Association for Computational Linguistics* (ACL'13), pp. 92-97 .

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*, 19:2.

Slav Petrov, Dipanjan Das and Ryan McDonald 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12).

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman and Zdeněk Žabokrtský. 2014. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pp. 2334–2341.

Manuela Sanguinetti and Cristina Bosco. 2012. Translational divergences and their alignment in a parallel treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories* (TLT11), pp. 169–180.

Manuela Sanguinetti, Cristina Bosco and Leonardo Lesmo 2013. Dependency and Constituency in Translation Shift Analysis. In *Proceedings of the 2nd Conference on Dependency Linguistics* (DepLing'13), pp. 282-291 .

Manuela Sanguinetti, Cristina Bosco and Loredana Cupi. 2014. Exploiting *catenae* in a parallel treebank alignment. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC'14), pp. 1824-1831.

Roy Schwartz and Omri Abend and Ari Rappoport. 2012. Learnability-Based Syntactic Annotation Design. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING'12), pp. 2405–2422.

Maria Simi, Cristina Bosco and Simonetta Montemagni 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of*

the 9th International Conference on Language Resources and Evaluation, (LREC'14), pp. 83–90.

Reut Tsarfaty 2013. A unified morpho-syntactic scheme of Stanford Dependencies. In *Proceedings of 51st annual meeting of the Association for Computational Linguistics* (ACL'13), pp. 578–584.

10.12871/CLICIT2014162

Developing corpora and tools for sentiment analysis: the experience of the University of Turin group

Manuela Sanguinetti, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, Leonardo Allisio, Valeria Mussa and Cristina Bosco
Dipartimento di Informatica
Università di Torino
{msanguin,sulis,patti,ruffo,bosco@di.unito.it},
{leonardo.allisio,valeria.mussa@studenti.unito.it}

Abstract

English. The paper describes the ongoing experience at the University of Turin in developing linguistic resources and tools for sentiment analysis of social media. We describe in particular the development of Senti-TUT, a human annotated corpus of Italian Tweets including labels for sentiment polarity and irony, which has been recently exploited within the SENTIment POLarity Classification shared task at Evalita 2014. Furthermore, we report about our ongoing work on the Felicittà web-based platform for estimating happiness in Italian cities, which provides visualization techniques to interactively explore the results of sentiment analysis performed over Italian geotagged Tweets.

Italiano. *L'articolo presenta l'esperienza fatta presso l'Università di Torino nello sviluppo di risorse linguistiche e strumenti per la sentiment analysis di social media. In particolare, viene descritto Senti-TUT, un corpus di Tweet in Italiano, che include annotazioni relative alla polarità del sentiment e alla presenza di ironia, utilizzato nell'ambito del task di SENTIment POLarity Classification di Evalita 2014. Inoltre viene presentato il lavoro su Felicittà, una piattaforma Web per la stima della felicità nelle città italiane, che fornisce diverse modalità di visualizzazione del grado di felicità che emerge da un'analisi del sentiment su messaggi Twitter geolocalizzati in Italiano.*

1 Introduction

Several efforts are currently devoted to automatically mining opinions and sentiments from natural language, e.g. in social media posts, news and reviews about commercial products. This task entails a deep understanding of the explicit and implicit information conveyed by the language, and most of the approaches applied refer to annotated corpora and adequate tools for their analysis.

In this paper, we will describe the experiences carried on at the Computer Science Department of the University of Turin in the development of corpora and tools for Sentiment Analysis and Opinion Mining (SA&OM) during the last few years. These experiences grew and are still growing in a scenario where an heterogeneous group of researchers featured by skills varying from computational linguistics, sociology, visualization techniques, big data analysis and ontologies cooperates. Both the annotation applied in the developed corpora and the tools for analyzing and displaying data analysis depend in fact on a cross-fertilization of different research areas and on the expertise gained by the group members in their respective research fields. The projects we will describe are currently oriented to the investigation of aspects of data analysis that can be observed in such a particular perspective, e.g. figurative language or disagreement deep analysis, rather than to the achievement of high scores in the application of classifiers and statistical tools.

The paper is organized as follows. The next section provides an overview on the annotated corpus Senti-TUT, which includes two main datasets: TW-NEWS (political domain) and TW-FELICITTA (generic collection), while Section 3 describes the main uses of Senti-TUT and the Felicittà application context.

2 Annotating corpora for SA&OM

The experience on human annotation of social media data for SA&OM mainly refers to the Senti-TUT corpus of Italian Tweets, featured by different stages of development (Gianti et al., 2012; Bosco et al., 2013; Bosco et al., 2014). We have

relied on our skills in building linguistic resources, such as TUT[1].

Tweets have been annotated at the message level. Among the main goals we pursued in the annotation of this corpus, there is the study of irony, a specific phenomenon which can affect SA&OM systems performances (Riloff et al., 2013; Reyes et al., 2012; Reyes et al., 2013; Hao and Veale, 2010; González-Ibáñez et al., 2011; Davidov et al., 2011; Maynard and Greenwood, 2014; Rosenthal et al., 2014). To deal with this issue, we extended a traditional polarity-based framework with a new dimension which explicitly accounts for irony. According to literature, boundaries in meaning between different types of irony are fuzzy (Gibbs and Colston, 2007) and this could be an argument in favor of annotation approaches where different types of irony are not distinguished, as the one adopted in Senti-TUT. We thus designed and applied to the collected data an annotation oriented to the description of Tweet polarity, which is suitable for high level tasks, such as classifying the polarity of a given text. The annotation scheme included the traditional labels for distinguishing among positive, negative or neutral sentiment. Moreover, we introduced the labels HUM, to mark the intention of the author of the post to express irony or sarcasm, and MIXED, to mark the presence of more than one sentiment within a Tweet[2]. Summarizing, our tagset includes:

POS	positive
NEG	negative
NONE	neutral (no sentiment expressed)
MIXED	mixed (POS and NEG both)
HUM	ironic
UN	unintelligible

Having a distinguished tag for irony did not prevent us from reconsidering these Tweets at a later stage, and force their classification according to traditional annotation schemes for the sentiment analysis task, i.e. applying a positive or negative polarity label, e.g. to measure how an automatic traditional sentiment classifier can be wrong, as we did in (Bosco et al., 2013). Similarly, identifying Tweets containing mixed sentiment can be useful in order to measure how the phenomenon impacts the performances of sentiment classifiers. Moreover, having distinguished tags for irony and mixed sentiment can be helpful to a better development of the corpora, in order to increase the inter-annotator agreement, since cases, that typically can be source of disagreement on the polarity valence, are recognized and labeled separately.

2.1 The Senti-TUT core

The first stage of development of the Senti-TUT project [3] led to the results described in (Bosco et al., 2013; Gianti et al., 2012). The major aims of the project are the development of a resource missing for Italian, and the study of a particular linguistic device: irony. This motivated the selection of data domain and source, i.e. politics and Twitter: Tweets expressing political opinions contain extensive use of irony. The corpus developed at this stage includes a dataset called TW-NEWS, composed of $3,288$ posts collected in the time frame between October 2012 and February 2013 and that focuses on the past Monti's government in Italy. They where collected and filtered, relying on the Blogmeter social media monitoring platform[4]. For each post in TW-NEWS, we collected in the first phase two independent annotations. The inter-annotator agreement calculated at this stage, according to the Cohen's κ score, was $\kappa = 0.65$ (Artstein and Poesio, 2008). The second step entailed the collection of cases when the annotators disagreed (about 25% of data). A third annotator thus attempted to solve the disagreement or discarded the inherently disagreement cases (around 2% of the data). This is motivated by the need of datasets that can be sufficiently unambiguous to be useful for training of classifiers and automatic tools. A second dataset, called TW-SPINO and composed of $1,159$ messages from the Twitter section of Spinoza [5] (a very popular Italian blog of posts with sharp satire on politics) has been collected in order to extend the size of the set of ironic Tweets tagged as HUM.

2.2 The TW-FELICITTA corpus

The TW-FELICITTA corpus (Bosco et al., 2014) can be seen as a further extension of Senti–TUT, mainly developed to validate the approach applied

[1] http://www.di.unito.it/~tutreeb

[2] About the MIXED label see also the gold standard presented in (Saif et al., 2013)

[3] http://www.di.unito.it/~tutreeb/sentiTUT.html

[4] http://www.blogmeter.eu

[5] http://www.spinoza.it

in the Felicittà project (see Section 3). The $1,500$ Italian Tweets here collected were randomly extracted from those collected by Twitter API, paying attention at avoiding geographic and temporal bias.

TW-FELICITTA corpus is a general-purpose resource. This means that data are not filtered in some way, but are more representative of the Twitter language and topics in general. The absence of a specific domain context made the interpretation and annotation of the posts more difficult. The annotation process involved four human annotators. We collected not less than three independent annotations for each Tweet according to the annotation scheme described above and relying on a set of shared guidelines. The inter–annotator agreement achieved was 0.51 (Fleiss, 1971). Hypothesizing that the 'soft disagreement' (i.e. disagreement occurring when we detect two agreeing and one disagreeing tags) was at least in part motivated by annotators biases or errors, after a further discussion of the guidelines, we applied a fourth independent annotation to the Tweets in soft disagreement. The resulting final corpus consists of $1,235$ Tweets with agreed annotation and 265 Tweets with disagreed annotation.

Table1 presents an overview of the distribution of tags (UN excluded) referring to the three annotated datasets currently included in Senti-TUT.

label	News	Felicittà	Spino
POS	513	338	-
NEG	788	368	-
NONE	1.026	260	-
MIXED	235	39	-
HUM	726	168	1.159

Table 1: Distribution of Senti-TUT tags in TW-NEWS, TW-FELICITTA and TW-SPINO.

The development of TW-FELICITTA also provided the basis for reflecting on the need of a framework to capture and analyze the nature of the disagreement (i.e. Tweets on which the disagreement reflects semantic ambiguity in the target instances and provides useful information). Hypothesizing that the analysis of the disagreement should be considered as a starting point for a deeper understanding of the task to be automated in our sentiment engine (in tune with the arguments in (Inel et al., 2014)), we investigated the use of different measures to analyze the following complementary aspects: the *subjectivity of each sentiment label* and the *subjectiveness of the involved annotators*.

Agreement analysis For what concerns the detection of the *subjectivity of the sentiment labels* in our annotation scheme, we hypothesized that when a sentiment label is more involved in the occurrence of disagreement, this is because it is more difficult to annotate, as its meaning is less shared among the annotators and there is a larger range of subjectivity in its interpretation. In order to estimate the subjectivity degree of each label *L*, we calculated the percentage of cases where *L* produced an agreement or disagreement among annotators. Table 2 shows how much a label has been used in percentage to contribute to the definition of an agreed or disagreed annotation of the Tweets.

label	agreement	disagreement
POS	26.3	14.4
NEG	29.2	17.8
NONE	21.8	23.5
MIXED	3.3	8.8
HUM	11.9	13.0
UN	7.6	22.5

Table 2: A measure of subjectivity of tags annotated in TW-FELICITTA

It should be observed, in particular, that while POS and NEG labels seem to have a higher reference to the agreement, for UN and MIXED the opposite situation happens.

Assuming a perspective oriented to the single annotators and referring to all the annotated tags, as above, we also measured the *subjectiveness* of each *annotator involved in the task* according to the variation in the exploitation of the labels. For each label *L*, starting from the total amount of times when *L* has been annotated, we calculated the average usage of the label. Then we calculated the deviation with respect to the average and we observed how this varies among the annotators. The deviation with respect to the average usage of the label is maximum for the MIXED and UN tags, and minimum for POS and NEG, showing that the annotators are more confident in exploit-

ing the latter tags (Table 3).

label	total	avg	dev. +	dev. -
NEG	1,592	398	15.32%	14.82%
POS	1,421	355.25	6.68%	5.13%
NONE	1,281	320.25	24.90%	16.31%
HUM	700	175	28.57%	31.42%
UN	569	142	73.94%	35.21%
MIXED	237	59.25	46.83%	80.18%

Table 3: A measure of variation among the exploitation of the labels in TW-FELICITTA.

3 Exploitation of Senti-TUT and ongoing applications of SA on Italian tweets

Irony and emotion detection A preliminary corpus-based analysis of phenomena connected to irony, in particular polarity reversing and frequency of emotions, is reported in (Bosco et al., 2013) and involved the Tweets tagged by HUM in TW-NEWS and TW-SPINO. We applied rule-based automatic classification techniques in (Bolioli et al., 2013) to annotate ironic Tweets according to seven categories: Ekman's basic emotions (*anger, disgust, fear, joy, sadness, surprise*) plus *love*. These emotions were expressed in 20% of the dataset and distributed differently in the corpora. What emerged was that irony was often used in conjunction with a seemingly positive statement to reflect a negative one (rarely the other way around). This is in accordance with theoretical accounts (Gibbs and Colston, 2007), reporting that expressing a positive attitude in a negative mode is rare and harder for humans to process, as compared to expressing a negative attitude in a positive mode.

Felicittà Felicittà (Allisio et al., 2013) is an online platform for estimating happiness in the Italian cities, which daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets, in order to enable the summarization of SA outcomes. The system automatically analyzes posts and classifies them according to traditional polarity labels according to a pipeline which performs a shallow analysis of Tweets and applies a lexicon-based approach looking for the word polarity in WordNetAffect (Strapparava and Valitutti, 2004). At the current stage of the project, we are investigating both the visualization techniques and data aggregation, also in order to compare, in future works, results extracted from Twitter about specific topics to those extracted from other sources like demographic reports.

SENTIPOLC For what concerns the exploitation of the Senti–TUT corpus, a further step is related to its use within the new shared task on sentiment analysis in Twitter (SENTIment POLarity Classification – SENTIPOLC[6]), as part of Evalita 2014[7]. SENTIPOLC represents a valuable forum to validate the data and to compare our experience to that of both the participants and colleagues co-organizing the task from University of Bologna, University of Groningen, Universitat Politècnica de València and the industry partner Blogmeter (CELI). (Basile et al., 2014). The main focus is on detecting sentiment polarity in Twitter messages as in SemEval 2013 - Task 2 (Nakov et al., 2013), but a pilot task on irony detection has been also proposed. The datasets released include Twitter posts from TW–NEWS and TW–FELICITTA annotated by the Turin & Blogmeter teams, and other posts collected and annotated by Bologna, Groningen and València teams (Basile and Nissim, 2013).

4 Conclusion

The paper describes the experiences done at the University of Turin on topics related to SA&OM, with a special focus on the main directions we are following. The first one is the development of annotated corpora for Italian that can be exploited both in automatic systems' training, in evaluation fora, and in investigating the nature of the data itself, also by a detailed analysis of the disagreement occurring in the datasets. The second direction, which is exemplified by ongoing work on the Felicittà platform, consists in the development of applications of SA on social media in the social and behavioral sciences field, where SA techniques can contribute to interpret the degree of well-being of a country (Mitchell et al., 2013; Quercia et al., 2012), with a special focus on displaying the results generated by the analysis in a graphic form that can be easily readable also for non-expert users.

[6] http://www.di.unito.it/~tutreeb/sentipolc-evalita14/

[7] http://www.evalita.it

Acknowledgments

We acknowledge Ing. Sergio Rabellino, who leads the ICT team of the Computer Science Department at the University of Turin, for the support in the development of a web platform for Twitter annotated data release based on RESTful technology.

References

Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, and Giancarlo Ruffo. 2013. Felicittà: Visualizing and estimating happiness in Italian cities from geotagged Tweets. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096, pages 95–106. CEUR-WS.org.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy.

Andrea Bolioli, Federica Salamino, and Veronica Porzionato. 2013. Social media monitoring in real life with blogmeter platform. In *ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 156–163. CEUR-WS.org.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti, and Emilio Sulis. 2014. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicittà. In B. Schuller, P. Buitelaar, L. Devillers, C. Pelachaud, T. Declerck, A. Batliner, P. Rosso, and S. Gaines, editors, *Proc. of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. European Language Resources Association.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2011. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the CONLL'11*, pages 107–116, Portland, Oregon (USA).

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3@LREC'12)*, pages 1–7, Istanbul, Turkey.

Raymond W Gibbs and Herbert L. Colston, editors. 2007. *Irony in Language and Thought*. Lawrence Erlbaum Associates, New York.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yanfen Hao and Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.*, 20(4):635–650, November.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandecic, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *Proceedings of ISWC 2014*, volume 8797 of *Lecture Notes in Computer Science*, pages 486–504. Springer International Publishing.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. 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'14)*, Reykjavik, Iceland, may. European Language Resources Association.

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on*

Semantic Evaluation (SemEval 2013), pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 965–968, New York, NY, USA. ACM.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowledge Engineering*, 74:1–12.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proc. of EMNLP*, pages 704–714. ACL.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Hassan Saif, Miriam Fernandez, He Yulan, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 9–21. CEUR-WS.org.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA.

10.12871/CLICIT2014163

Unsupervised Antonym-Synonym Discrimination in Vector Space

Enrico Santus[*], Qin Lu[*], Alessandro Lenci[§], Chu-Ren Huang[*]

[*] The Hong Kong Polytechnic University, Hong Kong
e.santus@connect.polyu.hk, {qin.lu, churen.huang}@polyu.edu.hk

[§] University of Pisa, Italy
alessandro.lenci@ling.unipi.it

Abstract

English. Automatic detection of antonymy is an important task in Natural Language Processing (NLP). However, currently, there is no effective measure to discriminate antonyms from synonyms because they share many common features. In this paper, we introduce *APAnt*, a new Average-Precision-based measure for the unsupervised identification of antonymy using Distributional Semantic Models (DSMs). *APAnt* makes use of Average Precision to estimate the extent and salience of the intersection among the most descriptive contexts of two target words. Evaluation shows that the proposed method is able to distinguish antonyms and synonyms with high accuracy, outperforming a baseline model implementing the *co-occurrence hypothesis*.

Italiano. *Sebbene l'identificazione automatica di antonimi sia un compito fondamentale del Natural Language Processing (NLP), ad oggi non esistono sistemi soddisfacenti per risolvere questo problema. Gli antonimi, infatti, condividono molte caratteristiche con i sinonimi, e vengono spesso confusi con essi. In questo articolo introduciamo APAnt, una misura basata sull'Average Precision (AP) per l'identificazione automatica degli antonimi nei Modelli Distribuzionali (DSMs). APAnt fa uso dell'AP per stimare il grado e la rilevanza dell'intersezione tra i contesti più descrittivi di due parole target. I risultati dimostrano che APAnt è in grado di distinguere gli antonimi dai sinonimi con elevata precisione, superando la baseline basata sull'ipotesi della co-occorrenza.*

1 Introduction

Antonymy is one of the fundamental relations shaping the organization of the semantic lexicon and its identification is very challenging for computational models (Mohammad et al., 2008). Yet, antonymy is essential for many Natural Language Processing (NLP) applications, such as Machine Translation (MT), Sentiment Analysis (SA) and Information Retrieval (IR) (Roth and Schulte im Walde, 2014; Mohammad et al., 2013).

As well as for other semantic relations, computational lexicons and thesauri explicitly encoding antonymy already exist. Although such resources are often used to support the above mentioned NLP tasks, they have low coverage and many scholars have shown their limits: Mohammad et al. (2013), for example, have noticed that "more than 90% of the contrasting pairs in GRE closest-to-opposite questions are not listed as opposites in WordNet".

The automatic identification of semantic relations is a core task in computational semantics. Distributional Semantic Models (DSMs) have often been used for their well known ability to identify semantically similar lexemes using corpus-derived co-occurrences encoded as distributional vectors (Santus et al., 2014a; Baroni and Lenci, 2010; Turney and Pantel, 2010; Padó and Lapata, 2007; Sahlgren, 2006). These models are based on the *Distributional Hypothesis* (Harris, 1954) and represent lexical semantic similarity in function of distributional similarity, which can be measured by *vector cosine* (Turney and Pantel, 2010). However, these models are characterized by a major shortcoming. That is, they are not able to discriminate among different kinds of semantic relations linking distributionally similar lexemes. For instance, the nearest neighbors of *castle* in the vector space typically include hypernyms like *building*, co-hyponyms like *house*, meronyms like *brick*, antonyms like *shack*, together with other semantically related words. While impressive results have been achieved in the automatic

identification of synonymy (Baroni and Lenci, 2010; Padó and Lapata, 2007), methods for the identification of hypernymy (Santus et al., 2014a; Lenci and Benotto, 2012) and antonymy (Roth and Schulte im Walde, 2014; Mohammad et al., 2013) still need much work to achieve satisfying precision and coverage (Turney, 2008; Mohammad et al., 2008). This is the reason why semi-supervised pattern-based approaches have often been preferred to purely unsupervised DSMs (Pantel and Pennacchiotti, 2006; Hearst, 1992)

In this paper, we introduce a new Average-Precision-based distributional measures that is able to successfully discriminate antonyms from synonyms, outperforming a baseline implementing the *co-occurrence hypothesis,* formulated by Charles and Miller in 1989 and confirmed in other studies, such as those of Justeson and Katz (1991) and Fellbaum (1995).

2 Defining Semantic Opposition

People do not always agree on classifying word-pairs as antonyms (Mohammad et al., 2013), confirming that antonymy classification is indeed a difficult task, even for native speakers of a language. Antonymy is in fact a complex relation and opposites can be of different types, making this class hard to define (Cruse, 1986).

Over the years, many scholars from different disciplines have tried to contribute to its definition. Though, they are yet to reach any conclusive agreement. Kempson (1977) defines opposites as word-pairs with a "binary incompatible relation", such that the presence of one meaning entails the absence of the other. In this sense, *giant* and *dwarf* are good opposites, while *giant* and *person* are not. Cruse (1986) points out the paradox of simultaneous similarity and difference between the antonyms, claiming that opposites are indeed similar in every dimension of meaning except in a specific one (e.g. both *giant* and *dwarf* refer to a person, with a head, two legs and two feet, but their size is different).

In our work, we aim to distinguish antonyms from synonyms. Therefore we will adopt the word "antonym" in its broader sense.

3 Related Works

Most of the work about the automatic antonymy identification is based on the *co-occurrence hypothesis*, proposed by Charles and Miller (1989), who have noticed that antonyms co-occur in the same sentence more often than expected by chance (Justeson and Katz, 1991; Fellbaum, 1995).

Other automatic methods include pattern based approaches (Schulte im Walde and Köper, 2013; Lobanova et al., 2010; Turney, 2008; Pantel and Pennacchiotti, 2006; Lin et al., 2003), which rely on specific patterns to distinguish antonymy-related pairs from others. Pattern based methods, however, are mostly semi-supervised. Moreover they require a large amount of data and suffer from low recall, because they can be applied only to frequently occurring words, which are the only ones likely to fit into the given patterns.

Mohammad et al. (2013) have used an analogical method based on a given set of contrasting words to identify and classify different kinds of opposites by hypothesizing that for every opposing pair of words, A and B, there is at least another opposing pair, C and D, such that A is similar to C and B is similar to D. Their approach outperforms other measures, but still is not completely unsupervised and it relies on thesauri, which are manually created resources.

More recently, Roth and Schulte im Walde (2014) proposed that discourse relations can be used as indicators for paradigmatic relations, including antonymy.

4 *APAnt*: an Average-Precision-based measure

Antonyms are often similar in every dimension of meaning except one (e.g. *giant* and *dwarf* are very similar and they differ only in respect to the size).

This peculiarity of antonymy – called by Cruse (1986) the *paradox of simultaneous similarity and difference* – has an important distributional correlate. Antonyms occur in similar contexts exactly as much as synonyms do, making the DSMs models unable to discriminate them. However, according to Cruse's definition, we can expect there to be a dimension of meaning in which antonyms have a different distributional behaviour. We can also hypothesize that this dimension of meaning is a salient one and that it can be used to discriminate antonyms from synonyms. For example, *size* is the salient dimension of meaning for the words *giant* and *dwarf* and we can expect that while *giant* occurs

more often with words such as *big*, *huge*, etc., *dwarf* is more likely to occur in contexts such as *small*, *hide*, and so on.

To verify this hypothesis, we select the *N* most salient contexts of the two target words (*N*=100[1]). We define the salience of a context for a specific target word by ranking the contexts through *Local Mutual Information* (LMI, Evert, 2005) and collecting the first *N*, as already done by Santus et al. (2014a). Once the *N* most salient contexts for the two target words have been identified, we verify the extent and the salience of the contexts shared by both the words. We predict that synonyms share a number of salient contexts that is significantly higher than the one shared by antonyms. To estimate the extent and the salience of the shared contexts, we adapt the Average Precision measure (AP; Voorhees and Harman, 1999), a common Information Retrieval (IR) evaluation metric already used by Kotlerman et al. (2010) to identify lexical entailment. In IR systems, this measure is used to evaluate the ranked documents returned for a specific query. It assigns high values to the rankings in which most or all the relevant documents are on the top (recall), while irrelevant documents are either removed or in the bottom (precision). For our purposes, we modify this measure in order to increase the scores as a function of (1) the size of the intersection and (2) the salience of the common features for the target words. To do so, we consider the common contexts as relevant documents and the maximum salience among the two target words as their rank. In this way, the score will be promoted when the context is highly salient for at least one of the two target words in the pair. For instance, in the pair *dog* – *cat*, if *home* is a common context, and it has salience=1 for *dog* and salience=*N*-1 for *cat*, we will consider *home* as a relevant document with rank=1. Formula (1) below provides the formal definition of *APAnt* measure:

$$APAnt = 1/\sum_{f \epsilon F_1 \cap F_2} \frac{1}{min(rank_1(f_1), rank_2(f_2))} \quad (1)$$

where *Fx* is the set of the *N* most salient features of a term *x* and $rank_x(f_x)$ is the position of the feature f_x in the salience ranked feature list for the term *x*. It is important to note that *APAnt* is defined as a reciprocal measure, so that the higher scores are assigned to antonyms.

5 Experiments and Evaluation

The evaluation includes two parts. The first part is a box-plot visualization to summarize the distributions of scores per relation. In the second part, the Average Precision (AP; Kotlerman et al., 2010) is used to compute the ability of our proposed measure to discriminate antonyms from synonyms. For comparison, we take as the baseline a model using the co-occurrence frequency of the target pairs.

5.1 The DSM and the Dataset

For the evaluation, we use a standard window-based DSM recording co-occurrences with context window of the nearest 2 content words both to the left and right of each target word. Co-occurrences are extracted from a combination of the freely available ukWaC and WaCkypedia corpora (with 1.915 billion and 820 million words, respectively) and weighted with LMI.

To assess *APAnt*, we rely on a subset of English word-pairs collected by Lenci and Benotto in 2012/13 using Amazon Mechanical Turk, following the method described by Schulte im Walde and Köper (2013). Among the criteria used for the collection, Lenci and Benotto balanced target items across word categories and took in consideration the frequency, the degree of ambiguity and the semantic classes.

Our subset contains 2.232 word-pairs[2], including 1.070 antonymy-related pairs and 1.162 synonymy-related pairs. Among the antonymy-related pairs, we have 434 noun-pairs (e.g. *parody-reality*), 262 adjective-pairs (e.g. *unknown-famous*) and 374 verb-pairs (e.g. *try-procrastinate*); among the synonymy-related pairs, we have 409 noun-pairs (e.g. *completeness-entirety*), 364 adjective-pairs (e.g. *determined-focused*) and 389 verb-pairs (e.g. *picture-illustrate*).

[1] *N=100* is the result of an optimization of the model against the dataset. Also the following suboptimal values have been tried: 50 and 150. In all the cases, the model outperformed the baseline.

[2] The sub-set include all the pairs for which both the target words exist in the DSM.

5.2 Results

5.2.1 *APAnt* Values Distribution

Figures 1 and 2 show the box-plots summarizing respectively the logarithmic distributions of *APAnt* and baseline scores for antonyms and synonyms. The logarithmic distribution is used to normalize the range of data, which would be otherwise too large and sparse for the box-plot representation.

Box-plots display the median of a distribution as a horizontal line within a box extending from the first to the third quartile, with whiskers covering 1.5 of the interquartile range in each direction from the box, and outliers plotted as circles.

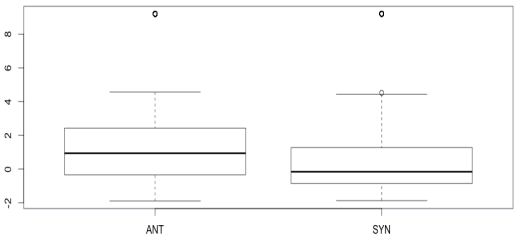


Figure 1: Logarithmic distribution of *APAnt* scores (*N=100*)

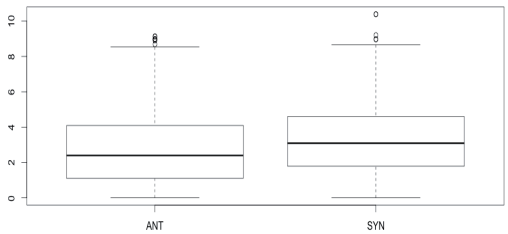


Figure 2: Logarithmic distribution of the baseline scores[3].

In Figure 2, we can observe that the baseline promotes synonyms over antonyms and also that there is a large range of overlap among synonyms and antonyms distributions, showing the weakness of the co-occurrence hypothesis on our data. On the other hand, in Figure 1 we can observe that, on average, *APAnt* scores are much higher for antonymy-related pairs and that the overlap is much smaller. In terms of distribution of values, in fact, synonyms have much lower values for *APAnt*.

[3] 410 pairs with co-occurrence equal to zero on a total of 2.232 have been removed to make the box-plot readable (i.e. *log(0)=-inf*).

5.2.2 Average Precision

Table 1 shows the second performance measure we used in our evaluation, the Average Precision (Lenci and Benotto, 2012; Kotlerman et al., 2010) per relation for both *APAnt* and baseline scores. As already mentioned above, AP is a method used in Information Retrieval to combine precision, relevance ranking and overall recall. The best possible score would be 1 for antonymy and 0 for synonymy.

	ANT	SYN
APAnt	**0.73**	0.55
Baseline	0.56	0.74

Table 1: Average Precision (AP).

Table 1 shows that *APAnt* is a much more effective measure for antonymy identification as it achieves +0.17 compared to the baseline. This value results in a 30% improvement for antonymy identification. This improvement comes together with a higher ability in discriminating antonyms from synonyms. The results confirm the trend shown in the box-plots of Figure 1 and Figure 2. *APAnt* clearly outperforms the baseline, confirming the robustness of our hypothesis.

6 Conclusions and Ongoing Work

This paper introduces *APAnt*, a new distributional measure for the identification of antonymy (an extended version of this paper will appear in Santus et al., 2014b).

APAnt is evaluated in a discrimination task in which both antonymy- and synonymy-related pairs are present. In the task, *APAnt* has outperformed the baseline implementing the *co-occurrence hypothesis* (Fellbaum, 1995; Justeson and Katz, 1991; Charles and Miller, 1989) by 17%. *APAnt* performance supports our hypothesis, according to which synonyms share a number of salient contexts that is significantly higher than the one shared by antonyms.

Ongoing research includes the application of *APAnt* to discriminate antonymy also from other semantic relations and to automatically extract antonymy-related pairs for the population of ontologies and lexical resources. Further work can be conducted to apply *APAnt* to other languages.

Acknowledgments

This work is partially supported by HK PhD Fellowship Scheme under PF12-13656.

References

Baroni, Marco and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Charles, Walter G. and George A. Miller. 1989. Contexts of antonymous adjectives. *Applied Psychology*, 10:357–375.

Cruse, David A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Fellbaum, Christiane. 1995. Co-occurrence and antonymy. *International Journal of Lexicography*, 8:281–303.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(23):146–162.

Hearst, Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–546, Nantes.

Justeson, John S. and Slava M. Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17:1–19.

Kempson, Ruth M. 1977. *Semantic Theory*. Cambridge University Press, Cambridge.

Kotlerman, Lili, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Lenci, Alessandro and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *SEM 2012 – The First Joint Conference on Lexical and Computational Semantics*, 2:75–79, Montréal, Canada.

Lin, Dekang, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1,492–1,493, Acapulco.

Lobanova, Anna, Tom van der Kleij, and Jennifer Spenader. 2010. Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography*, 23(1):19–53.

Mohammad, Saif, Bonnie Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Mohammad, Saif, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 982–991, Waikiki, HI.

Padó, Sebastian and Lapata, Mirella. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113-120, Sydney, Australia.

Roth, Michael and Sabine Schulte im Walde. 2014. Combining word patterns and discourse markers for paradigmatic relation classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2:524–530, Baltimore, Maryland, USA.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. dissertation, Department of Linguistics, Stockholm University.

Santus, Enrico, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014a. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2:38–42, Gothenburg, Sweden.

Santus, Enrico, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014b. Taking Antonymy Mask off in Vector Space. To Appear in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC),* Phuket, Thailand.

Schulte im Walde, Sabine and Maximilian Köper. 2013. Pattern-based distinction of paradigmatic relations for German nouns, verbs, adjectives. In *Language Processing and Knowledge in the Web*, 184-198. Springer.

Turney, Peter D. and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of

Semantics. *Journal of Articial Intelligence Research*, 37:141–188.

Turney, Peter D. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 905–912, Manchester.

Voorhees, Ellen M. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, 63–71, Columbus, OH.

10.12871/CLICIT2014164

Methods of textual archive preservation

Eva Sassolini
Istituto di Linguistica Computazionale
"Antonio Zampolli"
eva.sassolini@ilc.cnr.it

Sebastiana Cucurullo
Istituto di Linguistica Computazionale
"Antonio Zampolli"
nella.cucurullo@ilc.cnr.it

Manuela Sassi
Istituto di Linguistica Computazionale
"Antonio Zampolli"
manuela.sassi@ilc.cnr.it

Abstract

English. Over its fifty-years of history the Institute for Computational Linguistics "Antonio Zampolli" (ILC) has stored a great many texts and corpora in various formats and record layouts. The consolidated experience in the acquisition, management and analysis of texts has allowed us to formulate a plan of recovery and long-term digital preservation of such texts. In this paper, we describe our approach and a specific case study in which we show the results of a real attempt of text recovery. The most important effort for us has been the study and comprehension of more or less complex specific data formats, almost always tied to an obsolete technology.

Italiano. *L'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) nella sua storia cinquantennale ha accumulato una grande quantità di testi e corpora che sono stati conservati in vari formati e tracciati record. L'esperienza storica nell'acquisizione, gestione e analisi del testo ci ha permesso di formulare un piano di recupero e conservazione digitale a lungo termine di materiali testuali. In questo articolo, descriviamo il nostro approccio e un caso di studio specifico in cui sono riportati i risultati di una reale operazione di recupero. Il maggiore impegno è stato dedicato alla comprensione di particolari specifiche di formato più o meno complesse, ma quasi sempre legate ad obsolescenza tecnologica.*

1 Introduction

The international scientific communities consider electronic resources as a central part of cultural and intellectual heritage. Many institutions are involved in international initiatives[1] directed to the preservation of digital materials. The Digital Preservation Europe project (DPE) is an example of "best practice" realization. The principal issues concern the techniques and processes of digital memory management, but also of concerted action at both the national and international levels. *Digital preservation* is certainly a very challenging task for individual institutions.

The means of *digital preservation* can be explained by the following definition: "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time." (ALA 2007:2)

In our specific case we are engaged in looking for systems and techniques necessary for the long-term management of digital textual materials, stored at ILC over many years of work. At the beginning we did not know the age of all the materials. However, at a later stage, we found a number of digital materials, some of which dated as far back as the 70's. For this reason, the work of recovery is extremely complex and demanding. Firstly, the format of file/text encoding was often obsolete and not associated with exhaustive documentation. Secondly, many texts contained disused linguistic annotation schemas. Conse-

[1] Digital Preservation Europe (DPE) was a collaborative European digital preservation project that ran from 2006 to 2009, aimed at creating collaborations and synergies among many existing national initiatives across the European Research Area.

quently, we have adopted different strategies for "textual resource" retrieval, before focusing our attention on the application of standardized measures for conservation of the texts.

2 Text analysis

A *format specification*, when available, provides the details necessary to build a file from a text, and it establishes the admitted encodings and software applications able to decode the file and make its contents accessible. These documents can be of extremely variable size depending on the complexity of the format. However, the file *format specification* has not always evolved with the related software. Obsolete software and file formats, as well as storage medium, are today open issues.

A file format may become obsolete for several reasons:

- the latest software versions do not support the previous files;
- the format itself is superseded by a new one, or becomes more complex;
- the format is not so widely adopted, or the scientific community does not support the creation of compatible software;
- the format is no longer compatible with the current computers;
- the software supporting the format has declined.

Digital formats are a challenge for text conservation. In the early decades of computing, only few people were aware of the threat posed by the obsolescence of file formats for long-term digital preservation. A systematic effort for collecting software documentation or all the specifications necessary for the conservation of textual files was missing. With no proper documentation, the task of interpreting the contents of an old file is very demanding. It is only recently that we have started to catalogue, document, and understand these contents, together with their relationships and variations.

While most of the software is regularly updated, the relevant files become sometimes obsolete and therefore unable to meet the new format requirements, thus making even the latest versions of the software unreadable. Moreover, if the older versions are no longer available, or do not run on a recent computer or in the current version of the operating system, the data is lost. Owing to the complexity and nature of some file formats, it can be extremely complex to know whether a converted file in another format has retained all its features.

2.1 Conservation measures

Preserving the information should be the main goal. It is the information content of a document (tokens, linguistic annotations, critical apparatus, figures, etc.) that should be maintained in compliance with international standards. The standards usually need to respond to a large community of users, not linked to individual economic interests[2].

However, compatibility with the standards available is not generally priority for data producers, because either it is costly, or because there are commercial pressures that render the older formats quickly obsolete.

On the other hand, standard formats are not necessarily the best choice for all situations, but they offer great advantages for long-term preservation and storage. Finally, to reduce the risk of obsolescence a standardization process is required, which should primarily concern the formats at greatest risk, like the ones created by obsolete or outdated software versions.

3 First texts in electronic format

Electronic processing has always been articulated into three basic steps: input (input or acquisition of data within the procedure), processing (analysis of the material and specific processing depending on the intended purpose) and output (output, on suitable media, of the results of the previous stages). The output of any processing phase may be considered as a final result in its own right, even if - in a specific project - it can be an intermediate analysis subject to subsequent processing phases.

A fundamental parameter for the whole process is the type of storage medium used to preserve the material at the different processing stages. In the past, the only one choice available was the magnetic tape, which required sequential access to the data: in order to read (or write) a piece of data recorded on that medium it was necessary to read (or rewrite) all the preceding data in sequential order. This technology entailed objective

[2] The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its principal deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. (`www.tei-c.org`)

limitations concerning the possibility of implementing particularly sophisticated and efficient algorithms for data access.

3.1 Description of textual materials

During the analyses of textual materials we identified various levels of recovery in compliance with international recommendations, and at the same time to archive a standardized and universally recognized format that would allow the exchange and storage of materials. For the data to be better understood, we must explain the procedures which have produced them. For many years the "input" phases have only been possible through the preparation of punched cards that seem to belong to the pre-history of computer science. Units of data entry have come later, able to record on magnetic media or to operate in direct connection with the electronic computer.
The "output" phases normally consisted in the creation of two types of results:

- storing the results both as recovery of the same results, and as intermediate input to further processing;
- printing the results on fanfold paper to obtain the final output of the process and drafts used to check the correctness of the work performed, or to dispose of a working medium that can be enriched with new data or used to classify the previously stored data.

4 Main text problems

The first problem of management of a text concerns character encoding, which included sets of diacritics, or languages with non-Latin alphabets, or texts with multiple levels of annotation (e.g. comments, notes in margin, various types of footnotes, structured or dramatic text, etc.).
In the past the "ANSI format" (characters encoding belonging to the ISO 8859 family) generally represented the standard, with all the problems related to the sharing of sets of positions between the tables of the ISO 8859 family. Today the development of new encoding standards at the international level imposes shared models of representation of the data: XML TEI and UNICODE encoding.

4.1 Text acquisition strategy

If we retrace the steps of text acquisition for which ILC was among the pioneers in the industry, we see that there is no single conversion mapping, but that it is necessary to assess different types of material and their specific recovery paths.

At present, it is possible to make only an estimate about textual heritage. However, this is sufficient to set up a common procedure and useful to evaluate the costs of the entire operation. Depending on the types of material (from texts on magnetic tapes to machine readable and editable digital texts) we have hypothesized different phases of recovery. Therefore it is impossible to define a series of procedures valid for all types of contents or data.

We cannot forget the software DBT[3] that has often been used for the treatment of ILC texts.

For example at least three phases are required in order to convert a text file with obsolete character encodings: a first mapping involves the conversion into an intermediate format, typically an ANSI encoding; a second format is produced by the recognition, management and remapping of all the annotations inserted in the text; finally, the last phase involves the construction of a parser that can read these annotations and convert them into appropriate TEI-XML tags.

Source text	Perc.	Transition phases (TP) required	Meta data
Text on magnetic tape	10%	Many TP type	study and research in the ILC archives
Text divided into separate resources	5%	TP>3	recovered from paper-based data
Text in obsolete file	10%	TP>2	recovered from paper-based data
Digital text with obsolete character encoding	10%	2<TP<3	recovered from paper-based data / digital format
Digital text	65%	One TP	recovered from the digital format

Table 1: acquisition strategy

A more complex case is represented by lemmatized texts, where the annotations are at the level of words and then become more extensive. Even for the annotation of lemmatized texts there has been a wide use of the DBT software. In the acquisition protocol for this type of text, this level of analysis is added to the others together with the evaluation of the type of software tool that was used at the time.

[3] DBT (Data Base Testuale, Textual Data Base) is a specific software for the management and analysis of textual and lexical material.

Source text	Transition phases (TP) required	specific annotations type encoding	Meta data
Texts on magnetic tape	Many type TP	?	long and difficult work
Text divided into separate resources	TP>3	DBT type encoding	recovered from paper-based data
Text in obsolete file	TP>2	Obsolete type encoding	recovered from paper-based data
Digital text but obsolete character encoding	2<TP<3	Specific type encoding	recovered from paper-based data / digital format
Digital text	One TP	ILC text encoding	recovered from digital format

Table 2: annotated text acquisition strategy

5 Results

A concrete example of application of the procedure for recovery of texts belonging to the heritage of texts of the Institute (briefly "ILC Archives") is related to the work resulting from a scientific agreement between ILC and the "*Accademia della Crusca"* of Florence. The researchers of the *Accademia* were especially interested in recovering the lemmatized corpus of "*Periodici Milanesi"*[4].

The archive dates back to the early 80's and originates as post-elaboration of the lemmatization procedure implemented by ILC researchers and technicians in the 70's. The output format consists in files made up of fixed fields, each containing several types of information. The first challenge consisted in interpreting and decoding both the file format and the complex annotation scheme.

The most complex part of the decoding of the "starting-point" files (in TCL format/ASCII) concerned the retrieval of text and related annotations: lemma, POS (part of speech) and any other semantic type of information. For a correct interpretation of the data records contained in the lemmatized texts, a preliminary study phase was made. An example is shown in the figure below.

```
T1 162468  0404□4   per
L1per                          per
T3 162469  1212□4   inserirvelo
L1inserire                     inserire
L2egli                         egli
L3vi                           vi
T01162470  0101 180'
```

[4] Digital materials extracted from "La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze", edited by Giardini (Pisa, 1983), authors: Stefania De Stefanis Ciccone, Ilaria Bonomi, Andrea Masini. Management of the text required the advice of Eugenio Picchi and Remo Bindi.

The fragment of original text encoding shows the complex representation of the format used. As a matter of fact, the information is expressed by a complex annotation scheme, whose interpretation and decoding represented the first phase of work.

The complexity of the original format required a conversion in two steps:

- a first step in which the texts were converted from the original format to a DBT-like format to favor a simple check on the correct decoding of the source format with no loss of information;
- a second step required the representation of the text in XMT TEI with Unicode encoding.

The archive of "*Periodici Milanesi*" contains a collection of 58 newspapers (1800-1847), organized in seven main categories: Political Information, Literary Magazines, Magazines varieties, Technical journals, Magazines theater, Almanacs, Strennas.

Corpus analysis and results:

- 879,129 tokens;
- 59,639 different forms;
- a TEI P5 XML file for each article of the corpus (2277 files), where all lemmas are appropriately coded.

Extraction of the main linguistic features:

- index of 975 spelling variants of the words;
- index of 312 different multi-words in the corpus;
- list of 710 Latin and French forms that have been coded as "foreign words".

6 Conclusion

The preservation of that data produced with outdated technologies should be handled especially by public institutions, as this is part of the historical heritage. Therefore, it is necessary for us to complete this work, so that the resources can be reused. This will be possible only through a joint effort of the institutions involved at the regional, national and international levels. ILC is currently establishing a number of co-operation agreements like the one with the "*Accademia della Crusca*", in an attempt to gather data resources for maintenance, preservation and re-use by third parties.

References

Alessandra Cinini, Sebastiana Cucurullo, Paolo Picchi, Manuela Sassi, Eva Sassolini, Stefano Sbrulli. 2013. *I testi antichi: un patrimonio culturale da conservare e riutilizzare*. In: 27a DIDAMATICA 2013, Tecnologie e Metodi per la Didattica del Futuro, Pisa. (Pisa, 7-8-9 may 2013). Proceedings, AICA.867-870.

Eugenio Picchi, Maria L. Ceccotti, Sebastiana Cucurullo, Manuela Sassi, Eva Sassolini. 2004. *Linguistic Miner*. *An Italian Linguistic Knowledge System* In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 may 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.).1811–1814.

Eugenio Picchi. 2003. *Pisystem: sistemi integrati per l'analisi testuale*. In Linguistica Computazionale, Vol. XVIII-IX, I.L.C. and Computational Linguistics, special issue, A. Zampolli, N. Calzolari, L. Cignoni, (Eds.), I.E.P.I., Pisa-Roma. 2003.597-627

Eugenio Picchi. 2003. *Esperienze nel settore dell'analisi di corpora testuali: software e strumenti linguistici, Informatica e Scienze Umane*. In LEO (Lessico Intellettuale Europeo), a cura di Marco Veneziani, S. Olschki editore Maggio 2003. 129-155

ALA (American Library Association). 2007. *Definitions of digital preservation.* Chicago: American Library Association. Available at: `http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf`

Ingeborg Verheul. 2006. *Networking for Digital Preservation: Current Practice in 15 National Libraries.* Munchen: K.G. Saur 2006.

10.12871/CLICIT2014165

Combining unsupervised syntactic and semantic models of thematic fit

Asad Sayeed and Vera Demberg
Computational Linguistics and Phonetics / MMCI
Saarland University
D-66123 Saarbrücken
{asayeed,vera}@coli.uni-saarland.de

Abstract

English. We explore the use of the SENNA semantic role-labeller to define a distributional space to build a fully unsupervised model of event-entity thematic fit judgements. Existing models use syntactic dependencies for this. Our Distributional Memory model outperforms a syntax-based model by a wide margin, matches an augmented model that uses hand-crafted rules, and provides results that can be easily combined with the augmented model, improving matching over multiple thematic fit judgement tasks.

Italiano. *I giudizi di Thematic Fit tra eventi ed entità sono stati modellati in passato facendo ricorso a dipendenze sintattiche. Il nostro modello utilizza invece uno spazio distribuzionale costruito in maniera non supervisionata con un Semantic Role Labeler (SENNA). Il nostro modello ottiene risultati nettamente migliori rispetto a un modello basato su dipendenze sintattiche e comparabili a quelli di un modello potenziato, che sfrutta regole sviluppate manualmente in aggiunta alle dipendenze. Combinando il nostro modello e il modello potenziato si ottiene un ulteriore miglioramento dei risultati su diversi compiti di giudizio di Thematic Fit.*

1 Introduction

It is perfectly conceivable that automated tasks in natural language semantics can be accomplished entirely through models that do not require the contribution of semantic features to work at high accuracy. Unsupervised semantic role labellers such as that of Titov and Klementiev (2011) and Lang and Lapata (2011) do exactly this: predict semantic roles strictly from syntactic realizations. In other words, for practical purposes, the relevant and frequent semantic cases might be completely covered by learned syntactic information. For example, given a sentence *The newspaper was put on the table*, such SRL systems would identify that *the table* should receive a "location" role purely from the syntactic dependencies centered around the preposition *on*.

We could extend this thinking to a slightly different task: thematic fit modelling. It could well be the case that the *the table* could be judged a more appropriate filler of a location role for *put* than, e.g., *the perceptiveness*, entirely due to information about the frequency of word collocations and syntactic dependencies collected through corpus data, handmade grammars, and so on. In fact, today's distributional models used for modelling of selectional preference or thematic fit generally base their estimates on syntactic or string co-occurrence models (Baroni and Lenci, 2010; Ritter et al., 2010; Séaghdha, 2010). The Distributional Memory (DM) model by Baroni and Lenci (2010) is one example of an unsupervised model based on syntactic dependencies, which has been successfully applied to many different distributional similarity tasks, and also has been used in compositional models (Lenci, 2011).

While earlier work has shown that syntactic relations and thematic roles are related concepts (Levin, 1993), there are also a large number of cases where thematic roles assigned by a role labeller and their best-matching syntactic relations do not correspond (Palmer et al., 2005). However, it is possible that this non-correspondence is not a problem for estimating typical agents and patients from large amounts of data: agents will most of the time coincide with subjects, and patients will most of the time coincide with syntactic objects. On the other hand, the best resource

for estimating thematic fit should be based on labels that most closely correspond to the target task, i.e. semantic role labelling, instead of syntactic parsing. In this paper, we want to test how far a DM trained directly on a role labeller which produces PropBank style semantic annotations can complement the syntax-based DM model on thematic fit tasks, given a similar corpus of training data. We maintain the unsupervised nature of both models by combining their ratings by averaging without any weight estimation (we "guess" 50%) and show that we get an improvement in matching human judgements collected from previous experiments on agent/patient roles, location, and manner roles. We demonstrate that a fully unsupervised model based on a the SENNA role-labeller (Collobert et al., 2011) outperforms a corresponding model based on MaltParser dependencies (DepDM) by a wide margin. Furthermore, we show that the SENNA-based model can almost match B&L's better performing TypeDM model, which involves hand-crafted rules, and demonstrate that the SENNA-based model makes a contribution over and above the syntactic model in a range of thematic role labelling tasks.

1.1 Thematic role typicality

Thematic roles describe the relations that entities take in an event or relation. Thematic role fit correlates with human plausibility judgments (Padó et al., 2009; Vandekerckhove et al., 2009), which can be used to evaluate whether a distributional semantic model can be effectively encoded in the distributional space.

A suitable dataset is the plausibility judgment data set by Padó (2007), which includes 18 verbs with up to twelve nominal arguments, totalling 414 verb-noun-role triples. The words were chosen based on their frequency in the Penn Treebank and FrameNet. Human subjects were asked to how common the nominal arguments were as agents or as patients for the verbs. We also evaluate the DM models on a data set by McRae et al. (2005), which contains thematic role plausibility judgments for 1444 verb-role-noun triples calculated over the course of several experiments.

While the first two data sets only contain plausibility judgments for verbs and their agents and patients, we additionally use two data sets containing judgments for locations (274 verb-location pairs) and instruments (248 verb-instrument pairs) (McRae et al., 2005), to see how well these models apply to roles other than agent and patient. All ratings were on a scale of 1 to 7.

1.2 Semantic role labelling

Semantic role labelling (SRL) is the task of assigning semantic roles such as agent, patient, location, etc. to entities related to a verb or predicate. Structured lexica such as FrameNet, VerbNet and PropBank have been developed as resources which describe the roles a word can have and annotate them in text corpora such as the PTB. Both supervised and unsupervised techniques for SRL have been developed. Some build on top of a syntactic parser, while others work directly on word sequences. In this paper, we use SENNA, whose advantage is being very fast and robust (not needing parsed text) and is able to label large, noisy corpora such as UKWAC.

2 Distributional Memory

Baroni and Lenci (2010) present a framework for recording distributional information about linguistic co-occurrences in a manner explicitly designed to be multifunctional rather than being tightly designed to reflect a particular task. Distributional Memory (DM) takes the form of an order-3 tensor, where two of the tensor axes represent words or lemmas and the third axis represents the syntactic link between them.

B&L construct their tensor from a combination of corpora: the UKWAC corpus, consisting of crawled UK-based web pages, the British National Corpus (BNC), and a large amount of English Wikipedia. Their linking relation is based on the dependency-parser output of MaltParser (Nivre et al., 2007), where the links consist of lexicalized dependency paths and lexico-syntactic shallow patterns, selected by handcrafted rules.

The tensor is represented as a sparse array of triples of the form *(word, link, word)* with values as local mutual information (LMI), calculated as $O \log \frac{O}{E}$ where O is the observed occurrence count of the triple and E the count expected under independence. B&L propose different versions of representing the link between the words (encoding the link between the words in different degrees of detail) and ways of counting frequencies. Their DepDM model encodes the link as the (partially lexicalized) dependency path between words and counts occurrence frequencies of triples to cal-

model	*coverage* (%)	ρ
BagPack	100	60
ST-MeanDM	99	58
TypeDM	100	51
SENNA-DepDM	99	51
Padó	97	51
ParCos	98	48
DepDM	100	35

Table 1: Comparison on Padó data, results of other models from Baroni and Lenci (2010).

culate LMI. The more successful TypeDM model uses the same dependency path encoding as a link but bases the LMI estimates on type frequencies (counted over grammatical structures that link the words) rather than token frequencies.

The tensor also contains inverse links: if *(monster, sbj_tr eat)* appears in the tensor with a given LMI, another entry with the same LMI will appear as *(eat, sbj_tr*$^{-1}$*, monster)*.

B&L provide algorithms to perform computations relevant to various tasks in NLP and computational psycholinguistics. These operations are implemented by querying slices of the tensor. To assess the fit of a noun w_1 in a role r for a verb w_2, they construct a centroid from the 20 top fillers for r with w_2 selected by LMI, using subject and object link dependencies instead of thematic roles. To illustrate, in order to determine how well *table* fits as a location for *put*, they would construct a centroid of other locations for *put* that appear in the DM, e.g. *desk, shelf, account* ...

The cosine similarity between w_1's vector and the centroid represents the preference for the noun in that role for that verb. The centroid used to calculate the similarity represents the characteristics of the verb's typical role-fillers in all the other contexts in which they appear.

B&L test their procedure against the Padó et al. similarity judgements by using Spearman's ρ. They compare their model against the results of a series of other models, and find that they achieve full coverage of the data with a ρ of 0.51, higher than most of the other models except for the BagPack algorithm (Herdağdelen and Baroni, 2009), the only supervised system in the comparison, which achieved 0.60. Using the TypeDM tensor they freely provide, we replicated their result using our own tensor-processing implementation.

3 SENNA

SENNA (Collobert and Weston, 2007; Collobert et al., 2011) is a high performance role labeller well-suited for labelling a corpus the size of UKWAC and BNC due to its speed. It uses a multi-layer neural network architecture that learns in a sliding window over token sequences in a process similar to a conditional random field, working on raw text instead of syntactic parses. SENNA extracts features related to word identity, capitalization, and the last two characters of each word. From these features, the network derives features related to verb position, POS tags and chunking. It uses hidden layers to learn latent features from the texts which are relevant for the labelling task.

SENNA was trained on PropBank and large amounts of unlabelled data. It achieves a role labelling F score of 75.49%, which is slightly lower than state-of-the-art SRL systems which use parse trees as input (around 78% F score).

4 Implementation

We constructed a DM from the corpora used by B&L by running the sentences individually through SENNA and counting the *(assignee, role, assigner)* triples that emerged from the SENNA labelling. However, we omit the Wikipedia data included by Baroni and Lenci; results were better without them (ρ=48 on Padó), possibly an effect of genre.

SENNA assigns roles to entire phrases, but we only accepted head nouns and NN-composita. We used the part-of-speech tagging done by SENNA to identify head words and accepted only the first consecutive series of non-possessive noun-tagged words. If these are multiple words in this series (as in the case of composita), each of them is listed as a separate assignee. There is a very small amount of data loss due to parser errors and software crashes. Our implementation corresponds to B&L's DepDM model over MaltParser dependencies. The SENNA-based tensors are used to evaluate thematic fit data as in the method of B&L described above[1].

5 Experiments

We ran experiments with our tensor (henceforth SENNA-DepDM) on the following sources of thematic fit data: the Padó dataset, agents/patients from McRae, instrumental roles from McRae, and location roles from McRae. For each dataset, we calculated Spearman's ρ with respect to human plausibility judgments. We compared this against

[1]Our tensor will be provided via our web sites after this paper officially appears.

	TypeDM		SENNA-DepDM		ST-MeanDM		TDM/SENNA correl.
	cov. (%)	ρ	*cov.* (%)	ρ	*cov.* (%)	ρ	ρ
Padó	100	53	99	51	99	58	64
McRae agent/patient	95	32	96	24	95	32	59
McRae instrumental	93	36	94	19	92	38	23
McRae location	99	23	<100	19	<100	27	26

Table 2: Comparison of TypeDM to SENNA-DepDM and ST-MeanDM.

the performance of TypeDM given our implementation of B&L's thematic fit query system. We then took the average of the scores of SENNA-DepDM and TypeDM—we will call this ST-MeanDM—for each of these human judgement sources and likewise report ρ. We also report coverage for all these experiments.

During centroid construction, we used the ARG0 and ARG1 roles to find typical nouns for subject and object respectively. For the instrument role data, we mapped the verb-noun pairs to PropBank roles ARG2, ARG3 for verbs that have an INSTRUMENT in their frame, otherwise ARGM-MNR. We used "with" as the link for TypeDM-centroids; the same PropBank roles work with SENNA. For location roles, we used ARGM-LOC; TypeDM centroids are built with "in", "at", and "on" as locative prepositions.

6 Results and discussion

For all our results, we report coverage and Spearman's ρ. Spearman's ρ is calculated with missing items (due to absence in the tensor on which the result was based) removed from the calculation.

Our SENNA-based tensors are taken directly from SENNA output in a manner analogous to B&L's construction of DepDM from MaltParser dependency output. Both of them do much better than the reported results for DepDM (see Table 1) and one of them comes close to the performance of TypeDM on the Padó data. This suggests that improvements can be made to SENNA-DepDM by developing a procedure to determining lexicalized relation types mediated by PropBank roles, and calculating LMI values based on partially lexicalized types instead of tokens, similar to TypeDM.

Tables 2 shows that the MaltParser-based TypeDM and the SENNA-based DepDM models in combination achieve improved correlation with human judgments compared to TypeDM by itself[2]. The only exception was the McRae agent/patient data, which stayed the same. We also include the correlation between the TypeDM and SENNA-DepDM cosine similarities on each data set. These values suggest that even when their correlations with human judgements are similar, they only partly model the same aspects of thematic fit.

We calculated ρ on a per-verb basis for the Padó data on TypeDM and the SRL-augmented combined results and examined the differences. Augmentation by averaging with the SENNA-DepDM output improves ρ most strongly on verbs like "increase" and "ask". For example, SENNA-DepDM produces much sharper differences in judgements about whether "amount" can be the agent or patient of "increase", closer to human performance. Averaging with SENNA-DepDM also reduces the cosine similarities for both agent and patient roles of "state" with "ask", more in line with lower human judgements in both cases relative to the other nouns tested with "ask".

7 Conclusions

We have constructed a distributional memory based on SENNA-annotated thematic roles and shown an improved correlation with human data when combining it with the high-performing syntax-based TypeDM. We found that, even when built on similar corpora, SRL brings something to the table over and above syntactic parsing. In addition, our SENNA-based DM model was constructed in a manner roughly equivalent to B&L's simpler DepDM model, and yet it performs at a level far higher than DepDM on the Padó data set, on its own approaching the performance of TypeDM. It is likely that an SRL-based equivalent to TypeDM would further improve performance, and is thus a possible path for future work.

Our work also contributes the first evaluation of structured distributional models of semantics for thematic role plausibility for roles other than agent and patient.

[2]Baroni and Lenci used a version of the Pado data that erroneously swapped the judgments for some ARG0 vs. ARG1. We here evaluate on the original Pado data, with ARG2 for communicative verbs (*tell, ask, caution*) set to ARG1, as this is how SENNA labels the recipient of the utterances. This caused a small upward shift in the TypeDM results (from ρ=51 to 53), but should not cause DepDM (not made publicly available) to catch up.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Ronan Collobert and Jason Weston. 2007. Fast semantic extraction using a novel neural network architecture. In *Annual meeting-association for computational linguistics*, volume 45, page 560.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Amaç Herdağdelen and Marco Baroni. 2009. BagPack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40, Athens, Greece, March. Association for Computational Linguistics.

Joel Lang and Mirella Lapata. 2011. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1117–1126, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2Nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '11, pages 58–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7):1174–1184.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Ulrike Padó. 2007. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 424–434, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diarmuid Ó. Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan Titov and Alexandre Klementiev. 2011. A bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1445–1455, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bram Vandekerckhove, Dominiek Sandra, and Walter Daelemans. 2009. A robust and extensible exemplar-based model of thematic fit. In *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 826–834.

10.12871/CLICIT2014166

Deep neural network adaptation for children's and adults' speech recognition

Romain Serizel and Diego Giuliani
HLT research unit
Fondazione Bruno Kessler (FBK)
Trento, Italy
(serizel,giuliani)@fbk.eu

Abstract

English. This paper introduces a novel application of the hybrid deep neural network (DNN) - hidden Markov model (HMM) approach for automatic speech recognition (ASR) to target groups of speakers of a specific age/gender. We target three speaker groups consisting of children, adult males and adult females, respectively. The group-specific training of DNN is investigated and shown to be not always effective when the amount of training data is limited. To overcome this problem, the recent approach that consists in adapting a general DNN to domain/language specific data is extended to target age/gender groups in the context of hybrid DNN-HMM systems, reducing consistently the phone error rate by 15-20% relative for the three different speaker groups.

Italiano. *Questo articolo propone l'applicazione del modello ibrido "rete neurale artificiale multistrato - modelli di Markov nascosti" al riconoscimento automatico del parlato per gruppi di parlanti di una specifica fascia di età o genere che in questo caso sono costituiti da: bambini, maschi adulti e femmine adulte. L'addestramente della rete neurale multistrato si è dimostrato non sempre efficace quando i dati di addestramento erano disponibili solo in piccola quantità per uno specifico gruppo di parlanti. Per migliorare le prestazioni, un recente approccio proposto per adattare una rete neurale multistrato pre-addestrata ad un nuovo domino, o ad una nuova lingua, è stato esteso al caso di gruppi di parlanti di diverse età e genere. L'adozione di una rete multistrato adattata per ciascun gruppo di parlanti ha consentito di ottenere una riduzione dell'errore nel riconoscimento di fonemi del 15-20% relativo per ciascuno dei tre gruppi di parlanti considerati.*

1 Introduction

Speaker-related acoustic variability is one of the major source of errors in automatic speech recognition. In this paper we cope with age group differences, by considering the relevant case of children versus adults, as well as with male/female differences. Here DNN is used to deal with the acoustic variability induced by age and gender differences.

When an ASR system trained on adults' speech is employed to recognise children's speech, performance decreases drastically, especially for younger children (Wilpon and Jacobsen, 1996; Das et al., 1998; Claes et al., 1998; Potamianos and Narayanan, 2003; Giuliani and Gerosa, 2003; Gerosa et al., 2007). A number of attempts have been reported in literature to contrast this effect. Most of them try to compensate for spectral differences caused by differences in vocal tract length and shape by warping the frequency axis of the speech power spectrum of each test speaker or transforming acoustic models (Potamianos and Narayanan, 2003; Das et al., 1998; Claes et al., 1998). However, to ensure good recognition performance, age-specific acoustic models trained on speech collected from children of the target age, or group of ages, is usually employed (Wilpon and Jacobsen, 1996; Hagen et al., 2003; Nisimura et al., 2004; Gerosa et al., 2007). Typically, much less training data are available for children than for adults. The use of adults' speech for reinforcing the training data in the case of a lack of children's speech was investigated in the past (Wilpon and Jacobsen, 1996; Steidl et al., 2003). However,

This work was partially funded by the European project EU-BRIDGE, under the contract FP7-287658.

in order to achieve a recognition performance improvement when training with a mixture of children's and adults' speech, speaker normalisation and speaker adaptive training techniques are usually needed (Gerosa et al., 2009).

During the past years, DNN has proven to be an effective alternative to HMM - Gaussian mixture modelisation (GMM) based ASR (HMM-GMM) (Bourlard and Morgan, 1994; Hinton et al., 2012) obtaining good performance with context dependent hybrid DNN-HMM (Mohamed et al., 2012; Dahl et al., 2012).

Capitalising on their good classification and generalisation skills the DNN have been used widely in multi-domain and multi-languages tasks (Sivadas and Hermansky, 2004; Stolcke et al., 2006). The main idea is usually to first exploit a task independent (multi-lingual/multi-domain) corpus and then to use a task specific corpus. One approach consists in using the different corpora at different stages of the DNN training. The task independent corpus is used only for the pre-training (Swietojanski et al., 2012) or for a general first training (Le et al., 2010; Thomas et al., 2013) and the task specific corpus is used for the final training/adaptation of the DNN.

This paper introduces the use of the DNN-HMM approach for phone recognition in age and gender dependent groups, extending the idea introduced in (Yochai and Morgan, 1992) to the DNN context. Three target groups of speakers are considered here, that is children, adult males and adult females. There is only a limited amount of labeled data for such groups. To overcome this problem, a DNN trained on speech data from all the three groups of speakers is adapted to the age/gender group specific corpora. First it is shown that training a DNN only from a group specific corpus is not effective when only limited labeled data is available. Then the method proposed in (Thomas et al., 2013) is adapted to the age/gender specific problem.

The rest of this paper is organized as follows. Section 2 introduces the general training and adaptation methods. Experimental setup is described in Section 3 and results are presented in Section 4. Finally, conclusions are provided in Section 5.

2 DNN training and adaptation

In ASR what is called DNN is a feedforward network with at least one hidden layer (generally more than three). When applied in a hybrid context, the DNN is used to classify the acoustic features into HMM states. The output of the DNN is then used to estimate the HMM's state emission likelihoods. Recent experiments exhibit that DNN-HMM provides better performance for ASR than shallow networks (Dahl et al., 2012).

2.1 Age/gender independent training

The general training procedure described above can be applied, by using all training data available, in an attempt to achieve a system with strong generalisation capabilities. Estimating the DNN parameters on speech from all groups of speakers, that is children, adult males and adult females, may however, have some limitation due to the inhomogeneity of the speech data that may negatively impact on the classification accuracy compared to group-specific DNN.

2.2 Age/gender adaptation

ASR systems provide their best recognition performances when the operating (or testing) conditions are consistent with the training conditions. To be effective, the general training procedure described above requires that a sufficient amount of labeled data is available. Therefore, when considering training for under-resourced population groups (such as children or males/females in particular domains of applications) it might be more effective to train first a DNN on a large amount of data (including the target group specific corpora) and then to adapt this DNN to the group specific corpora. A similar approach has been proposed in (Thomas et al., 2013) for the case of multilingual training. Here the language does not change and the targets of the DNN remain the same when going from age/gender independent training to group specific adaptation. The DNN trained on speech data from all groups of speakers can then be used directly as initialisation to the adaptation procedure where the DNN is trained to convergence with back-propagation only on group specific corpora.

3 Experimental setup

3.1 Speech corpora

For this study we relied on two Italian speech copora: the ChildIt corpus consisting of children speech and the APASCI corpus consisting of adults' speech. Both corpora were used for evaluation purposes, while the ChildIt and the APASCI

provided similar amount of training data for children and adults, respectively.

3.1.1 ChildIt

The ChildIt corpus (Giuliani and Gerosa, 2003; Gerosa et al., 2007) is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. The overall duration of audio recordings in the corpus is 10h:48m hours. Speech was collected from 171 children. The corpus was partitioned into: a training set consisting of data from 115 speakers for a total duration of 7h:15m; a development set consisting of data from 14 speakers, for a total durations of 0h:49m; a test set consisting of data from 42 speakers balanced with respect to age and gender for a total duration of of 2h:20m.

3.1.2 APASCI

The APASCI speech corpus (Angelini et al., 1994) is a task-independent, high quality, acoustic-phonetic Italian corpus. APASCI was developed at ITC-irst and consists of speech data collected from 194 adult speakers for a total durations of 7h:05m. The corpus was partitioned into: a training set consisting of data from 134 speakers for a total duration of 5h:19m; a development set consisting of data from 30 speakers balanced per gender, for a total durations of 0h:39m; a test set consisting of data from 30 speakers balanced per gender, for a total duration of 0h:40m.

3.2 ASR systems

3.2.1 DNN-HMM

The DNN use 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. The context spans on a 31 frames window on which Hamming windowing is applied. This 403 dimensional feature vector is then projected on a 208 dimensional feature vector by applying Discrete Cosine Transform (DCT) and normalised before being used as input to the DNN. The targets of the DNN are the 3039 tied-states obtained from a HMM-GMM system trained on the mixture of adults' and children's speech (ChildIt + APASCI). The DNN have 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can be summarised as follows: 208 x 1500 x 1500 x 1500 x 1500 x 3039.

The DNN are trained with the TNet software package (Veselỳ et al., 2010). The DNN weights are initialised randomly and pre-trained with Restricted Boltzmann Machines (RBM) (Hinton et al., 2006; Erhan et al., 2010) with mini-batch size of 250. For the back propagation training the starting learning rate is 0.02 and the mini-batch size is 512. In both pre-training and training, a first-order momentum of 0.5 is applied.

The DNN are trained either on all speech data available (ChildIt + APASCI) or on group specific corpora (ChildIt, adult female speech in APASCI, adult male speech in APASCI).

3.2.2 Age/gender adapted DNN for DNN-HMM

One option is to adapt an already trained general DNN to group specific corpora. The data architecture is the same as described above. The initial DNN weights are the weights obtained with a pre-training/training procedure applied on all training data available (ChildIt+APASCI). The DNN is then trained with back propagation on a group specific corpus (ChildIt, adult female speech in APASCI and adult male speech in APASCI). The learning rate follows the same rule as above.

4 Experiment results

The experiments presented here are designed to verify the validity of the following statements:

- The age/gender group specific training of the DNN does not necessarily lead to improved performance, specially when a small amount of data is available
- The age/gender group adaptation of a general DNN can help to design group specific systems, even when only a small amount of data is available.

During the experiments the language model weight is tuned on the development set and used to decode the test set. Results were obtained with a phone loop language model and the PER was computed based on 28 phone labels. Variations in recognition performance were validated using the matched-pair sentence test (Gillick and Cox, 1989) to ascertain whether the observed results were inconsistent with the null hypothesis that the output of two systems were statistically identical. Considered significance levels were .05, .01 and .001.

4.1 Age/gender specific training for DNN-HMM

In this experiment, DNN are trained on group specific corpora (children's speech in ChildIt,

Training Set	Evaluation Set					
	ChildIt		APASCI (f)		APASCI (m)	
	Dev	Test	Dev	Test	Dev	Test
Mixture	13.98%	15.56%	**10.12%**	**10.91%**	**10.70%**	**8.62%**
ChildIt	**12.08%**	**12.76%**	24.46%	29.59%	50.93%	46.16%
APASCI (f)	32.23%	34.23%	10.92%	12.75%	36.01%	31.21%
APASCI (m)	53.85%	56.11%	29.73 %	30.81%	11.36%	9.83%

Table 1: Phone error rate achieved with the DNN-HMM trained age/gender groups specific data.

Adaptation Set	Evaluation Set					
	ChildIt		APASCI (f)		APASCI (m)	
	Dev	Test	Dev	Test	Dev	Test
No adaptation	13.98%	15.56%	10.12%	10.91%	10.70%	8.62%
ChildIt	**11.68%**	**12.43%**	13.82 %	16.93%	28.89 %	24.96%
APASCI (f)	19.77%	21.91%	**8.30%**	**9.65%**	20.40%	17.01%
APASCI (m)	30.04 %	32.33%	16.78 %	16.99%	**9.33%**	**7.61%**

Table 2: Phone error rate achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to speficic age/gender groups.

adult female speech in APASCI and adult male speech in APASCI) and performance are compared with the DNN-HMM baseline introduced above. Recognition results are reported in Table 1, which includes results achieved with the DNN-HMM baseline in the row *Mixture*. In ChildIt there is about 7h of training data which is apparently sufficient to train an effective DNN and we can observe an improvement of 2.8% PER ($p < .001$), from 15.56% to 12.76%. However, in adult data there is only about 2h:40m of data for each gender. This is apparently not sufficient to train a DNN. In fact, the DNN-HMM system based on a DNN that is trained on gender specific data consistently degrades the PER. The degradation is 1.84% PER on female speakers in APASCI ($p < .001$) and 1.21% PER on male speakers in APASCI ($p < .001$).

4.2 Age/gender adapted DNN-HMM

In this experiment the DNN trained on all available corpora is adapted to each group specific corpus and recognition performance is compared with that obtained by the DNN-HMM baseline (where the DNN is trained on all available corpora). PER performance is presented in Table 2 which also reports the results achieved by the DNN-HMM baseline (in row *No adaptation*). The group adapted DNN-HMM consistently improve the PER compared to the DNN-HMM baseline. On children's speech the PER improvement is of 3.13% ($p < .001$), from 15.56% to 12.43%, for adult female speakers in APASCI the PER improvement is 1.26% ($p < .001$), from 10.91% to 9.65% and for adult male speakers in APASCI the PER improvement is of 1.01% ($p < .05$), from 8.62% to 7.61%.

5 Conclusions

In this paper we have investigated the use of the DNN-HMM approach in a phone recognition task targeting three groups of speakers, that is children, adult males and adult females. It has been shown that, in under-resourced condition, group specific training does not necessarily lead to PER improvements. To overcome this problem a recent approach, which consists in adapting a task independent DNN for tandem ASR to domain/language specific data, has been extended to age/gender specific DNN adaptation for DNN-HMM. The DNN-HMM adapted on a low amount of group specific data have been shown to improve the PER by 15-20% relative with respect to the DNN-HMM baseline system trained on speech data from all the three groups of speakers.

In this work we have proven the effectiveness of the hybrid DNN-HMM approach when training with limited amount of data and targeting speaker populations of different age/gender. Future work will be devoted to embed the results presented here in a large vocabulary speech recogniser especially targeting under-resourced groups of speakers such as children.

References

B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. 1994. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proc. of ICSLP*, pages 1391–1394, Yokohama, Japan, Sept.

Herve A Bourlard and Nelson Morgan. 1994. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer.

T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle. 1998. A Novel Feature Transformation for Vocal Tract Length Normalisation in Automat ic Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549–557, Nov.

G.E. Dahl, Dong Yu, Li Deng, and A. Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, Jan.

S. Das, D. Nix, and M. Picheny. 1998. Improvements in Children's Speech Recognition Performance. In *Proc. of IEEE ICASSP*, Seattle,WA, May.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660.

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2007. Acoustic variability and automatic recognition of childrens speech. *Speech Communication*, 49(1011):847 – 860.

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2009. Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499 – 509.

L. Gillick and S. Cox. 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. of IEEE ICASSP*, pages I–532–535, Glasgow, Scotland, May.

D. Giuliani and M. Gerosa. 2003. Investigating Recognition of Children Speech. In *Proc. of IEEE ICASSP*, volume 2, pages 137–140, Hong Kong, Apr.

A. Hagen, B. Pellom, and R. Cole. 2003. Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. of IEEE ASRU Workshop*, St. Thomas Irsee, US Virgin Islands, Dec.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov.

Viet-Bac Le, L. Lamel, and J. Gauvain. 2010. Multistyle ML features for BN transcription. In *Proc. of IEEE ICASSP*, pages 4866–4869, March.

A. Mohamed, G.E. Dahl, and G. Hinton. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, Jan.

R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. 2004. Public Speech-Oriented Guidance System with Adult and Child Discrimination Capabi lity. In *Proc. of IEEE ICASSP*, Montreal, Canada, May.

A. Potamianos and S. Narayanan. 2003. Robust Recognition of Children's Speech. *IEEE Trans. on Speech and Audio Processing*, 11(6):603–615, Nov.

S. Sivadas and H. Hermansky. 2004. On use of task independent training data in tandem feature extraction. In *Proc. of IEEE ICASSP*, volume 1, pages I–541–4, May.

S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann. 2003. Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In *Pattern Recognition, 25th DAGM Symposium*, pages 600–607, Sep.

A. Stolcke, F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *Proc. of IEEE ICASSP*, volume 1, pages 321–334, May.

P. Swietojanski, A. Ghoshal, and S. Renals. 2012. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. of IEEE SLT Workshop*, pages 246–251, Dec.

S. Thomas, M.L. Seltzer, K. Church, and H. Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. of IEEE ICASSP*, pages 6704–6708, May.

Karel Veselỳ, Lukáš Burget, and František Grézl. 2010. Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, pages 439–446. Springer.

J. G. Wilpon and C. N. Jacobsen. 1996. A Study of Speech Recognition for Children and Elderly. In *Proc. of IEEE ICASSP*, pages 349–352, Atlanta, GA, May.

Konig Yochai and Nelson Morgan. 1992. GDNN: a gender-dependent neural network for continuous speech recognition. In *Proc. of Iternational Joint Conference on Neural Networks*, volume 2, pages 332–337, Jun.

10.12871/CLICIT2014167

An Italian Corpus for Aspect Based Sentiment Analysis of Movie Reviews

Antonio Sorgente
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy
a.sorgente@cib.na.cnr.it

Giuseppe Vettigli
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy

Francesco Mele
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy

Abstract

English. In this paper we will present an Italian corpus focused on the domain of movie reviews, developed in order to support our ongoing research for the development of new models about Sentiment Analysis and Aspect Identification in Italian language. The corpus that we will present contains a set of sentences manually annotated according to the various aspects of the movie that have been discussed in the sentence and the polarity expressed towards that particular aspect. In this paper we will present the annotation guidelines applied, some statistics about the corpus and the preliminary results about the identification of the aspects.

Italiano. *In questo lavoro presenteremo una nuova risorsa linguistica sviluppata per la creazione di nuovi modelli per la Sentiment Analysis Ascpect Based in lingua Italiana. Di seguito saranno introdotte le linee guida adottate per l'annotazione del corpus ed alcuni risultati preliminari riguardanti l'identificazione di aspetti.*

1 Introduction

Nowadays, on the Web there is a huge amount of unstructured information about public opinion and it continues growing up rapidly. Analysing the opinions expressed by the users is an important step to evaluate the quality of a product. In this scenario, the tools provided by Sentiment Analysis and Opinion Mining are crucial to process this information. In the particular case of movie reviews, we have that the number of reviews that a movie receives on-line grows quickly. Some popular movies can receive hundreds of reviews and, furthermore, many reviews are long and sometimes they contain only few sentences expressing the actual opinions. This makes hard for a potential viewer to read them and make an informed decision about whether to watch a movie or not. In the case that one only reads a few reviews, the choice may be biased. The large number of reviews also makes it hard for movie producers to keep track of viewer's opinions. The recent advances in Sentiment Analysis have shown that coarse overall sentiment scores fails to adequately represent the multiple potential aspects on which an entity can be evaluated (Socher et al., 2013). For example, if we consider the following review from Amazon.com about the movie *Inception*:

> *"By far one of the best movies I've ever seen. Visually stunning and mentally challenging. I would recommend this movie to people who are very deep and can stick with a movie to get the true meaning of the story."*

One can see that, even if the review is short, it not only expresses an overall opinion but also contains opinions about other two aspects of the movie: the photography and the story. So, in order to obtain a more detailed sentiment, an analysis that considers different aspects is required.

In this work, we present an Italian corpus focused on the domain of movie reviews developed in order to support our ongoing effort for the development of new models about Sentiment Analysis and Aspect Identification in Italian language.

The paper is structured as follows. In the Section 2 we present the motivations that led us to the creation of a new corpus and a short survey about related resources that already exist. Section 3 describes the guideline used to annotate the corpora, while Section 4 presents some statistical information about it. In section 5 we present some preliminary experiments about the identification of the as-

pects. Finally, in section 6 some conclusions will be offered.

2 Motivations

During the last years many studies focused on how to combine Sentiment Analysis and Aspect Identification.

The first attempt to combine Sentiment Analysis and Aspect Identification was made in (Hu and Liu, 2004), where a system to summarize the reviews was proposed. The system extracts terms that are related to various aspects of the textual comments and they tested their approach using 500 reviews about five types of electronics products (digital cameras, DVD players, mp3 players and mobile phones). The reviews were taken from Amazon.com and Cnet.com.

In (Ganu et al., 2009) a corpus of about 3400 sentences, gathered from a set of reviews about restaurants, have been annotated according to specific aspects of the restaurant domain with the related sentiment. The authors used the corpus to develop and test a regression-based model for the Sentiment Analysis. The same data and a set of about 1000 reviews on various topics collected from Amazon.com were used in (Brody and Elhadad, 2010). In this work the authors presented an unsupervised model able to extract the aspects and determine the related sentiments.

In the SemEval-2014 challenge, a task with the aim to identify the aspects of given target entities and the sentiment expressed towards each aspect has been proposed. The task is focused on two domain specific datasets of over 3000 sentences, the first one contains restaurant reviews extracted from the same data used in (Ganu et al., 2009) and the other one contains laptop reviews. Regarding the movie reviews, in (Thet et al., 2010) a method to determine the sentiment orientation and the strength of the reviewers towards various aspects of a movie was proposed. The method is based on a linguistic approach which uses the grammatical dependencies and a sentiment lexicon. The authors validated their method on a corpus of 34 reviews from which 1000 sentences were selected.

From the works mentioned, it follows that the approaches based on sentence-level analysis are predominant for the detection of the aspects in the reviews.

There are few studies that use Italian language because Italian lacks resources for sentiment analysis of natural language, although, some interesting resources have been produced using Twitter as data source. For example, in (Basile and Nissim, 2013) a dataset that contains 100 million tweets was proposed. This dataset contains 2000 tweets annotated according to the sentiment they express, 1000 of them regard general topics, while the remaining 1000 regard politics. In the Senti-TUT project (Bosco et al., 2013), an Italian corpus which consists of a collection of texts taken from Twitter and annotated with respect to irony was created. EVALITA (Evaluation of NLP and Speech Tools for Italian) has provided many interesting activities and resources, until now it has not hosted any activity or task about Sentiment Analysis aspect based. However, to the best of our knowledge, there is only one Italian corpus, which has been used in (Croce et al., 2013), where both the aspects and their polarity are taken in account. In particular, the corpus is focused on review of Italian wine products and it has been used to build a model able to classify opinions about wines according to the aspect of the analyzed product, such the flavor or taste of a wine, and the polarity.

3 Annotation Guidelines

To build our corpus, the annotators were instructed through a manual with the annotation guidelines. The guidelines were designed to be as specific as possible about the use of the aspects and the sentiment labels to be assigned.

3.1 Aspects

The aspects of the movies that we have considered for the annotation were suggested by some movie experts. For each of them we have provided a short guideline that helps the annotator with the identification of the aspect in the sentences:

- **Screenplay**: the sentence describes one or more scenes of a movie, their temporal distribution during the story, the quality and, the type or the complexity of the plot of the story. For example: *"Invictus è certo un film edificante di buone volontà, ma anche un bel film di solida struttura narrativa"* (*"Invictus is certainly an enlightened film of goodwill, but also a good movie with a solid narrative structure."*).

- **Cast**: the sentence expresses the importance

of the actors and their popularity. For example: *"Sono andato a vedere il film perchè c'era il mitico Anthony Hopkins"* (*"I went to see the movie because there was the legendary Anthony Hopkins"*).

- **Acting**: the sentence expresses an opinion on the actors' performances. For example: *"Grande come sempre la Bullock!"* (*"Bullock is always great!"*).
- **Story**: the sentence references to the storyline of the film. For example: *"Il finale di questo film è particolarmente amaro."* (*"The ending of this movie is particularly bittersweet."*).
- **Photography**: the sentence is about the colors, close-ups, countershot and shot used in the movie. For example: *"La fotografia è magistrale."* (*"The photography is great."*).
- **Soundtrack**: the sentence refers to the soundtrack and music used in the movie. For example: *"Fantastiche le basi musicali usate durante il film e l'indimenticabile sigla di chusura"* (*"The backing tracks used during the film and the unforgettable theme song in the ending are fantastic"*).
- **Direction**: the sentence is on the work of the director. For example: *" Tim Burton è pazzo ma anche un genio!"* (*"Tim Barton is crazy, but he is also a genius!"*).
- **Overall**: this aspect is used when the sentence doesn't report the description of any particular aspect of the movie but a general opinion or description. For example: *"Un film veramente bello, da vedere!!!"* (*"A really nice movie, to watch!!!"*).

3.2 Polarity Labels

We used 5 sentiment labels to represent the polarity: *strongly negative*, *negative*, *neutral*, *positive* and *strongly positive*. The description of each label follows:

- **Strongly Negative**: there is an extremely negative opinion.
- **Negative**: there is a negative opinion.
- **Neutral**: there is no opinion or if it expresses an opinion boundary between positive and negative.
- **Positive**: there is a positive opinion.
- **Strongly Positive**: there is an extremely positive opinion.

Aspect	Count
Overall	1370
Screenplay	226
Cast	165
Acting	338
Story	647
Photography	55
Soundtrack	40
Direction	235

Table 1: Distribution of the aspects in the Corpus.

4 Corpus description

The corpus contains 2648 sentences. Each sentence has been manually annotated according to the various aspects of the movie that have been discussed in the sentence; then, each aspect found has been annotated with the polarity expressed towards that particular aspect. So, for each sentence annotated we have a set of aspect-polarity pairs. Also, the sentences that were extracted from the same review are linked by an index in order to enable the study of the contex.

The sentences of the corpus have been extracted from over 700 user reviews in the website FilmUp.it. The user reviews of this website have been also studied in (Casoto et al., 2008), where the authors focused on the classification of the reviews according to the sentiment without considering the specific aspects referred in the reviews and without providing a sentence level study.

The distribution of the aspects is reported in Table 1, while the distribution of the sentiment labels is reported in Table 2. We can see that 23% of the labels are Negative or Strongly Negative, the 21% are Neutral and that the 53% are Positive or Strongly Positive.

It is important to notice that, the size of the corpus is comparable to the size of the English corpora with the same purpose.

In order to evaluate the accuracy of the annotation, 800 sentences have been annotated by two different annotators and the agreement among the annotators has been evaluated using the Cohen's Kappa (κ) measure (Carletta, 1996). This metric

Polarity	Count
S. Negative	250
Negative	526
Neutral	706
Positive	1238
S. Positive	491

Table 2: Distribution of the polarity labels in the Corpus.

measures the agreement between two annotators taking into account the possibility of chance agreement. It is computed as

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

where $P(a)$ is the relative observed agreement between two annotators and $P(e)$ is the expected agreement between two annotators, using the observed data to calculate the probabilities of each observer randomly saying each category. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators other than what would be expected by chance, $\kappa = 0$.

The inter-annotator agreement computed on our data is substantial for the aspect categories (0.7) and very good for the sentiment categories (above 0.8).

In case of disagreement the final annotation has been selected by a third annotator.

5 Preliminary experiments

In this section we report the results of a preliminary experiment on Aspect Identification. To do this, we have used Linear Discriminant Analysis (LDA) (Hastie et al., 2001) in order to build a set of classifiers, one for each aspect, able to recognize if a sentence is about or not a given aspect. The features used to train the classifiers were computed using the *tf-idf* (term frequency–inverse document frequency) model (Baeza-Yates and Ribeiro-Neto, 1999). In this model, the features extracted by a sentence are given by a set of terms, for each term t the value of the feature is computed as

$$tf(t, s) \times idf(t)$$

where $tf(t, s)$ is the count of t in the sentence s and $idf(t)$ is defined as

$$idf(t) = \log \frac{|S|}{1 + |\{s : t \in s\}|}$$

where S is the collection of sentences. Each classifier was trained on a different set of features (before the features extraction, stop-words were removed), and the terms for the features extraction were selected according to the χ^2 measure respect to the given aspect. This statistic measures the dependence between the features and a target variable and it is often used to discover which features are more relevant for statistical classification (Yang and Pedersen, 1997). Then, we have performed 5-fold cross validation and used accuracy, precision and recall to evaluate the quality of the classification. The Table 3 shows the error estimated using cross validation. With this basic model, we had a high accuracy (89% on average) and a good precision (70% on average). The recall was moderate (50% on average). In that Table, the evaluations with respect to the aspects *Photography* and *Soundtrack* are not reported because the samples for these categories are not enough to train and test a classification model.

Aspect	Accuracy	Precision	Recall
Overall	72%	70%	93%
Screenplay	92%	73%	42%
Cast	94%	71%	28%
Acting	90%	78%	52%
Story	82%	81%	49%
Direction	93%	78%	50%

Table 3: Aspect identification results.

6 Conclusion

We introduced an Italian corpus of sentences extracted by movie reviews. The corpus has been specifically designed to support the development of new tools for the Sentiment Analysis in Italian. We believe that corpus can be used to train and test new models for sentence-level sentiment classification and aspect-level opinion extractions.

In the paper, various aspects of the corpus we created have been described. Also, the results of some preliminary experiments about the automatic identification of the aspects have been showed.

7 Availability and license

The proposed Corpus is made available under a Creative Commons License (CC BY 3.0) and can be requested contacting one of the authors of this paper.

References

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *roceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June.

Paolo Casoto, Antonina Dattolo, Paolo Omero, Nirmala Pudota, and Carlo Tasso. 2008. A new machine learning based approach for sentiment classification of italian documents. In Maristella Agosti, Floriana Esposito, and Costantino Thanos, editors, *IRCDL*, pages 77–82. DELOS: an Association for Digital Libraries.

Danilo Croce, Francesco Garzoli, Marco Montesi, Diego De Cao, and Roberto Basili. 2013. Enabling advanced business intelligence in divino. In *DART@AI*IA*, pages 61–72.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Information Science*, 36(6):823–848.

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, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

10.12871/CLICIT2014168

Il *Perugia Corpus*: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione

Stefania Spina
Dipartimento di Scienze Umane e Sociali, Università per Stranieri di Perugia
stefania.spina@unistrapg.it

Abstract

Italiano Il *Perugia Corpus* (PEC) è un corpus dell'italiano contemporaneo scritto e parlato, che comprende oltre 26 milioni di parole. L'obiettivo che ha guidato la sua costituzione è quello di ovviare alla mancanza di un corpus di riferimento dell'italiano. In questo articolo vengono descritti i criteri alla base della sua composizione, la sua strutturazione in 10 sezioni e sottosezioni e la sua annotazione multilivello, con la relativa valutazione.

English *The Perugia Corpus (PEC) is a corpus of contemporary written and spoken Italian of more than 26 million words. Its aim is to fill the gap of the lack of an Italian reference corpus. This paper describes its composition and organization in 10 sections and sub-sections, and its multilevel annotation and evaluation.*

1 Introduzione

Il Perugia Corpus (PEC) è un corpus di riferimento dell'italiano contemporaneo, scritto e parlato[1]; è composto da oltre 26 milioni di parole, distribuite in 10 differenti sezioni, corrispondenti ad altrettanti generi testuali, e dotato di una annotazione multilivello. Il PEC intende ovviare alla mancanza di un corpus di riferimento (scritto e parlato), di cui hanno finora sofferto gli studi sull'italiano. Per la sua natura di risorsa di riferimento (EAGLES, 1996), il PEC è progettato per fornire informazioni linguistiche il più possibile generali sull'italiano e le sue principali varietà scritte e parlate.

La filosofia che ha guidato la composizione del PEC è dunque radicalmente diversa da quella che è alla base di alcuni web corpora (Baroni e Bernardini, 2006; Kilgarriff e Grefenstette, 2003) dell'italiano di ultima generazione come *Paisà* (Lyding et al., 2014), *itWac* (Baroni e Kilgarriff, 2006) o *itTenTen* (Jakubíček et al., 2013), ma anche da quella di corpora meno recenti come *Repubblica* (Baroni et al., 2004) e CORIS/CODIS (Rossini Favretti et al., 2002): la scelta è stata infatti quella di privilegiare la differenziazione dei generi testuali, includendo anche il parlato, a scapito delle dimensioni del corpus. Inoltre, si è puntato sulla riutilizzazione di risorse già esistenti e disponibili (Zampolli, 1991), ma a volte disperse e di difficile consultazione; ad esse sono stati aggiunti dati nuovi, raccolti col duplice scopo di riempire vuoti in cui non erano disponibili dati per l'italiano, ed aggiornare risorse esistenti, ma ormai datate. Il PEC può dunque essere considerato un corpus di riferimento "low cost", di dimensioni contenute ma con una buona rappresentatività delle diverse varietà scritte e parlate dell'italiano. Le dimensioni contenute del PEC presentano inoltre due vantaggi: permettono di gestire, in fase di interrogazione, quantità di risultati più maneggevoli (Hundt e Leech, 2012), e consentono di ottenere una buona accuratezza nell'annotazione (vedi par. 3.2).

2 Composizione del corpus

Il PEC è suddiviso in 10 sezioni, a loro volta articolate in sottosezioni; complessivamente, i testi inseriti nel corpus sono 41.401, con una lunghezza media di 12.500 tokens per testo. In linea con quanto avviene per corpora di riferimento di altre lingue, anche di dimensioni maggiori, come il *British National Corpus* (Burnard, 2007), lo scritto copre l'85% del totale del PEC, ed il parlato il restante 15%. La tab.1 presenta un quadro riassuntivo del corpus, con i dati relativi alle 10 sezioni; nei paragrafi che seguono, sarà invece descritta, per ogni sezione, la sua composizione interna.

[1] Il PEC è stato realizzato all'Università per Stranieri di Perugia tra il 2011 e il 2012.

sezione	n. testi	tokens	media tokens	% totale	types	TTR	frasi	tokens x frase
SCRITTO								
letteratura	60	3.545.459	59.091	13,38	103.141	54,78	229.361	15,46
saggi	79	2.354.996	29.810	8,89	97.795	63,73	102.130	23,06
stampa	8.232	5.772.040	701	21,78	147.707	61,48	225.827	25,56
accademico	240	1.113.590	4.640	4,20	54.658	51,80	32.736	34,02
scuola	4.054	1.257.842	310	4,75	46.981	41,89	51.208	24,56
amministrazione	119	1.160.334	9.751	4,38	28.562	26,52	31.950	36,32
web	27.383	7.359.460	269	27,78	225.190	83,01	295.041	29,94
TOT. SCRITTO	*40.167*	*22.563.721*		*85,16*	*704.034*		*969.059*	
PARLATO								
tv	127	1.147.151	9.033	4,33	50.643	47,28	73.950	15,51
film	66	626.487	9.492	2,36	31.967	40,39	99.858	6,27
parlato	1.041	2.158.522	2.074	8,15	67.987	46,28	80.354	26,86
TOT. PARLATO	*1.234*	*3.932.160*		*14,84*	*150.597*		*254.162*	
TOTALE	**41.401**	**26.495.881**	**12.517**		**854.631**		**1.223.221**	

Tabella 1 - La composizione delle 10 sezioni del PEC; la type-token ratio (TTR) è calcolata usando l'indice di Guiraud (Guiraud, 1954), per ovviare alla non omogeneità nel numero dei tokens.

2.1 Letteratura

La sezione dedicata alla letteratura comprende campioni estratti da 60 romanzi contemporanei, pubblicati tra il 1990 e il 2012 da 45 autori italiani diversi.

2.2 Saggistica

La saggistica comprende campioni estratti da 79 saggi di argomento diverso, ma riconducibili a quattro aree tematiche (attualità, biografia, politica e tempo libero). Tutti i saggi sono stati pubblicati da autori italiani dal 1990 al 2010.

2.3 Stampa

Gli 8.232 testi della sezione della stampa sono suddivisi tra articoli di quotidiani (79%) e di settimanali (21%): sono infatti tratti dal *Corriere della Sera* e da *Il Sole 24 ore* del 2012, e da *L'Espresso* del 2011 e del 2012. La tab. 2 riporta l'ulteriore suddivisione degli articoli dei quotidiani in 9 sottocategorie, con il rispettivo numero di tokens.

argomento	tokens	% totale
editoriale	436.570	7,6
politica	1.023.021	17,7
economia	565.641	9,8
cronaca	1.555.654	27,0
esteri	681.769	11,8
cultura	667.181	11,6
sport	424.416	7,4
lettere	120.178	2,1
spettacolo	297.610	5,2

Tabella 2 - Tipologie di articoli di quotidiani

2.4 Scritto accademico

In questa sezione è stato incorporato e riutilizzato integralmente, con alcune integrazioni, il *Corpus di Italiano Accademico* (Spina, 2010). In essa sono incluse quattro sottosezioni (tesi di laurea, dispense, manuali e articoli scientifici), a loro volta ripartite fra tre macroaree tematiche (umanistica, giuridico-economica e scientifica). La tab. 3 riporta i dati delle varie sottosezioni.

	tesi	dispense	manuali	articoli	TOT.
umanistica	55.311	170.501	36.077	114.581	*376.470*
giur-eco	58.087	176.206	64.020	75.814	*374.127*
scientifica	54.460	203.197	34.464	70.872	*362.993*
TOT.	*167.858*	*549.904*	*134.561*	*261.267*	*1.113590*

Tabella 3 - Sottosezioni dello scritto accademico

2.5 Scritto scolastico

La sezione è costituita da 4.054 temi svolti da studenti delle scuole medie e superiori tra il 2010 e il 2011, su 21 argomenti diversi; i temi sono stati estratti in modo automatico dal sito di *Repubblica Scuola*. Le due sottosezioni in cui la sezione è articolata sono i 2.431 temi della scuola media (652.749 tokens) e i 1.623 della scuola superiore (605.093 tokens).

2.6 Scritto amministrativo

La sezione amministrativa è composta per il 75% da testi di leggi (europee, statali, regionali), e per il restante 25% da regolamenti e documenti amministrativi più brevi.

2.7 Web

I testi scritti estratti dal web rappresentano la sezione più ampia del PEC, a sua volta suddivisa in testi di interazione e testi di riferimento, come

descritto nella tab. 4. Per quanto riguarda i blog, sono stati estratti i soli testi dei post, senza i commenti, da una cinquantina di blog di genere personale, giornalistico o aziendale. I testi di Wikipedia sono stati prelevati nel gennaio 2012 dalla versione italiana integrale, e selezionati in modo casuale. La sottosezione dei social network comprende 24.424 post tratti da profili di Facebook e di Twitter delle tre tipologie personale, politico e aziendale.

	tokens	% totale
INTERAZIONE		
blog	2.812.439	38,22
forum	171.111	2,33
chat	119.279	1,62
social network	603.630	8,20
TOT. INTERATTIVI	*3.706.459*	*50,36*
RIFERIMENTO		
Wikipedia	3.653.001	49,64
TOT. RIFERIMENTO	*3.653.001*	*49,64*

Tabella 4 - Sottosezioni della sezione web

2.8 Parlato

Per la sezione del parlato si è fatto ampio ricorso a corpora già esistenti e disponibili per uso accademico: il PEC contiene infatti i seguenti corpora o materiali testuali già trascritti, pari circa a 450.000 tokens:

- i testi del *LIP* (De Mauro et al., 1993), nella versione resa disponibile dal sito *Badip* (http://badip.uni-graz.at/it/);
- la sezione italiana del corpus *Saccodeyl*, un progetto *Minerva* sulla lingua dei giovani europei (Pérez-Paredes e Alcaraz-Calero, 2007);
- alcune trascrizioni del corpus CLIPS (Albano Leoni, 2007), tratte dalle sezioni elicitate attraverso map task e test delle differenze.

Questi dati già esistenti sono stati (ri)annotati secondo i criteri previsti dal PEC ed aggiunti al resto dei testi, raccolti ex novo.

La bipartizione principale della sezione del parlato è quella tra parlato dialogico (1.020.264 tokens) e parlato monologico (1.138.258 tokens); ad un livello successivo, il parlato dialogico, sulla base di una distinzione fondamentale derivata dall'analisi della conversazione (Drew e Heritage, 1992), è stato suddiviso in dialogo tra pari (faccia a faccia o telefonico) e dialogo istituzionale (in vari contesti, come quello scolastico-accademico, processuale, medico ecc.). Il parlato monologico, invece, è suddiviso nelle 7 sottosezioni descritte nella tab. 5.

	tokens	% tot.
DIALOGICO		
a. tra pari	*471.097*	*21,82*
- faccia a faccia	187.454	8,68
- telefonico	283.643	13,14
b. istituzionale (lezioni, processi...)	*549.167*	*25,44*
TOT. DIALOGICO	*1.020.264*	*47,27*
MONOLOGICO		
conferenze	168.051	7,79
lezioni	156.128	7,23
processi	174.728	8,09
istituzioni	158.967	7,36
politica	158.346	7,34
religione	168.917	7,83
testi di canzoni[2]	153.121	7,09
TOT. MONOLOGICO	*1.138.258*	*52,73*

Tabella 5 - Sottosezioni del parlato

2.9 Televisione

I dati televisivi inclusi nel PEC derivano dal *Corpus di Italiano Televisivo* (Spina, 2005), in una versione riorganizzata e ampliata. Le 127 trasmissioni comprese nel PEC appartengono ai due macrogeneri "informazione" e "intrattenimento", e sono suddivise nelle 6 sottosezioni descritte nella tab. 6. Nella categoria "approfondimento" rientrano i programmi come *Annozero*, *Report*, *In mezz'ora*, *Ballarò*, che costituiscono appunto un approfondimento delle notizie principali. "Talk show" sono invece le trasmissioni di argomento più conviviale come *Parla con me* o *Le invasioni barbariche*.

	tokens	% totale
INFORMAZIONE		
telegiornali	229.324	19,99
approfondimento	345.929	30,16
TOT. INFORMAZIONE	*575.253*	*50,15*
INTRATTENIMENTO		
talk show	221.008	19,27
fiction	127.026	11,07
sport	113.181	9,87
spettacolo	110.683	9,65
TOT. INTRATTENIMENTO	*571.898*	*49,86*

Tabella 6 - Sottosezioni della sezione tv

2.10 Film

La sezione comprende la trascrizione integrale dei dialoghi di 66 film italiani prodotti tra il 1995 e il 2011. Le trascrizioni sono state ottenute at-

[2] La ridotta estensione dei dati raccolti per i testi di canzoni (Werner, 2012) ha motivato il loro inserimento nella sezione del parlato monologico anziché una sezione autonoma del PEC.

traverso alcuni siti di condivisione di sottotitoli di film (come *opensubtitles.org*), e successivamente controllate e corrette manualmente.

3 Annotazione

Il PEC è dotato di un'annotazione multilivello, che comprende due fasi distinte: l'annotazione della struttura dei testi e l'annotazione linguistica.

3.1 Annotazione della struttura dei testi

I testi che compongono il PEC sono stati in primo luogo annotati in linguaggio XML, per distinguerli ed etichettarli a livello di genere testuale. Ad un livello ulteriore di dettaglio, ciascun testo è stato etichettato in base alle sue caratteristiche più specifiche: nel parlato, ad esempio, sono annotati i singoli turni di parola e alcune caratteristiche sociolinguistiche dei parlanti (ove possibile, sesso, età e provenienza geografica). E' stato utilizzato un set di tag analogo a quello dello standard della *Text Encoding Initiative* (Burnard e Bauman, 2014); la scelta è stata quella di adottare un tipo di annotazione minimalista, basato su alcune raccomandazioni essenziali (Hardie, 2014).

L'esempio che segue mostra l'annotazione XML di un testo parlato dialogico, prodotto nel corso di un'interazione faccia a faccia da un parlante di 25 anni, di sesso maschile, proveniente dalla Calabria:

```
<text id="427" type="par"
sub="dialogo">
<div type="pari" sub="faf">
<u who="L" sex="m" age="25"
prov="Calabria">
```

3.2 Annotazione linguistica

Il PEC è stato annotato per categoria grammaticale; il pos-tagging (Tamburini, 2007; Attardi e Simi, 2009) è stato effettuato usando *TreeTagger* (Schmid, 1994), con un tagset creato ad hoc[3]; il lessico, pur derivato da quello della distribuzione originale, è stato sensibilmente ampliato, fino a quasi 550.000 entrate. *TreeTagger* è stato addestrato con testi annotati manualmente, appartenenti a tutte le sezioni del corpus: il training set conteneva infatti, in misura uguale, campioni casuali estratti da ciascuna delle 10 sezioni (10.000 parole per sezione, per 100.000 parole totali), per fare in modo che ciascuno dei dieci generi testuali, con le sue peculiarità linguistiche, contribuisse in misura uguale al training del tagger (Jurafsky e Martin 2000; Giesbrecht e Evert, 2009).

La valutazione dell'accuratezza del pos-tagging, effettuata su un test set di oltre 22.000 parole, ottenute in modo bilanciato dalle 10 sezioni del corpus, ha evidenziato un valore del 97,3% (range = 96.6%-97.7%)[4]: la fig. 1 mostra una certa uniformità nei valori delle varie sezioni, con accuratezza leggermente più bassa nei testi parlati.

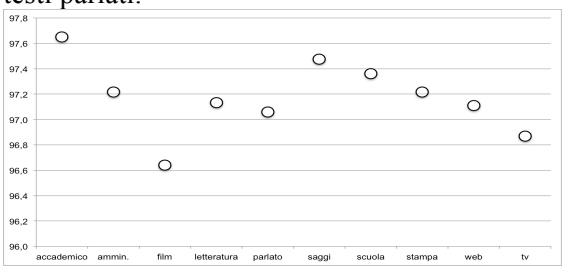


Figura 1 - Accuratezza del pos-tagging nelle 10 sezioni del PEC.

In una fase di "post-tagging", successiva all'annotazione, una serie di errori ricorrenti è stata corretta in modo automatico con l'aiuto di un guesser, basato su espressioni regolari, conformemente a quanto suggerito da Schmid et al. (2007). In tal modo, il pos-tagging ha superato il 98% di accuratezza.

4 Conclusioni

Il PEC rappresenta il primo corpus di riferimento dell'italiano contemporaneo scritto e parlato; nella sua composizione è stata privilegiata la differenziazione dei generi testuali, anche parlati, rispetto all'ampiezza delle dimensioni. Realizzato con risorse limitate e in tempi ristretti, attingendo, ove possibile, a risorse linguistiche già esistenti, il PEC costituisce un compromesso low cost tra creazione di risorse nuove e riuso di risorse esistenti.

L'interrogazione del PEC avviene attraverso l'interfaccia CWB e il *Corpus Query Processor* (Evert e Hardie, 2011), che consente di ricercare parole, sequenze di parole e annotazioni; è prevista la realizzazione di un'interfaccia di rete via *CQPweb* (Hardie, 2012), accessibile al pubblico[5].

[3] Il tagset (http://perugiacorpus.unistrapg.it/tagset.htm) comprende 53 etichette e trae spunto da quello descritto in Baroni et al. (2004).

[4] Sono stati conteggiati sia gli errori di categoria grammaticale che quelli di lemma.

[5] Tale interfaccia consentirà di interrogare il corpus online; non è invece prevista, per motivi di copyright, la disponibilità dei testi che compongono il corpus.

References

Federico Albano Leoni. 2007. Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS. *Bollettino d'Italianistica*, IV, (2), pp. 122-130.

Giuseppe Attardi e Maria Simi. 2009. Overview of the EVALITA 2009 Part-of-Speech Tagging Task. *EVALITA 2009*.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston e Marco Mazzoleni. 2004. Introducing the *"la Repubblica"* corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*.

Marco Baroni e Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Marco Baroni e Adam Kilgarriff. 2006. Large Linguistically-Processed Web Corpora for Multiple Languages. *EACL 2006 Proceedings*, 87-90.

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services, Oxford.

Lou Burnard e Syd Bauman. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, Charlottesville.

Tullio De Mauro, Federico Mancini, Massimo Vedovelli e Miriam Voghera. 1993. *Lessico di frequenza dell'italiano parlato*. EtasLibri, Milano.

Paul Drew e John Heritage (Eds.). 1992. *Talk at Work*. Cambridge University Press, Cambridge.

EAGLES. 1996. *Preliminary recommendations on Corpus Typology* EAG--TCWG--CTYP/P. Version of May, 1996.

Stefan Evert e Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Eugenie Giesbrecht e Stefan Evert. 2009. Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In I. Alegria, I. Leturia, and S. Sharoff (Eds.). *Proceedings of the 5th Web as Corpus Workshop (WAC5),* San Sebastian, Spain.

Paul Guiraud. 1954. *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.

Andrew Hardie. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380-409.

Andrew Hardie. 2014. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73-103

Marianne Hundt e Geoffrey Leech. 2012. Small is Beautiful – On the Value of Standard Reference Corpora for Observing Recent Grammatical Change. In T. Nevalainen & E. Traugott (Eds). *The Oxford Handbook of the History of English.* Oxford University Press, Oxford, pp. 175-188.

Daniel Jurafsky e James H. Martin. 2000. *Speech and language processing: An introduction to Natural Language Processng, Computational Linguistics and Speech Recognition*, Prentice-Hall, Upper Saddle River, NJ, USA.

Adam Kilgarriff e Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci e Vito Pirrelli. 2014. The PAISÀ Corpus of Italian Web Texts. *Proceedings of the 9th Web as Corpus Workshop (WaC-9),* Association for Computational Linguistics, Gothenburg, Sweden, April 2014. pp. 36-43

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář V, Pavel Rychlý e Vit Suchomel. 2013. The TenTen Corpus Family. *7th International Corpus Linguistics Conference*, Lancaster.

Pascual Pérez-Paredes e Jose M. Alcaraz-Calero. 2007. Developing annotation solutions for online data-driven learning. *EUROCALL 2007* - University of Ulster, 5 - 8 September**.**

Rema Rossini Favretti, Fabio Tamburini e C. De Santis. 2002. A corpus of written Italian: a defined and a dynamic model, in A. Wilson, P. Rayson, T. McEnery (eds.). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Lincom-Europa, Munich.

Helmut Schmid. 1994. *Probabilistic part-of-speech tagging using decision trees*. In Proceedings of the International Conference on New Methods in Language Processing.

Helmut Schmid, Marco Baroni, Eros Zanchetta e Achim Stein. 2007. Il sistema "TreeTagger arricchito". EVALITA 2007. *Intelligenza artificiale*, IV, 2007, 2, pp.22-23.

Stefania Spina. 2005. Il Corpus di Italiano Televisivo (CiT): struttura e annotazione, in Burr, E. (ed.), *Traditione & Innovazione. Il parlato: teoria – corpora – linguistica dei corpora*, Atti del VI Convegno SILFI (28 Giugno - 2 Luglio 2000, Gerhard-Mercator-Universität Duisburg, Germania). Franco Cesati, Firenze, pp. 413-426.

Stefania Spina. 2010. AIWL: una lista di frequenza dell'italiano accademico, in Bolasco S., Chiari I., Giuliano L., *Statistical Analysis of Textual Data*, Proceedings of the 10th JADT Conference (Rome, 9-11 june 2010), Editrice universitaria LED, pp. 1317-1325.

Fabio Tamburini. 2007. Evalita 2007: The Part-of-Speech Tagging Task. *Intelligenza artificiale,* IV, N° 2, pp. 4-7.

Valentin Werner. 2012. Love is all around: a corpus-based study of pop lyrics. *Corpora*, Vol. 7 (1), pp. 19-50

Antonio Zampolli. 1991. Towards reusable linguistic resources. *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*.

10.12871/CLICIT2014169

Are Quantum Classifiers Promising?

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. This paper presents work in progress on the development of a new general purpose classifier based on Quantum Probability Theory. We will propose a kernel-based formulation of this classifier that is able to compete with a state-of-the-art machine learning methods when classifying instances from two hard artificial problems and two real tasks taken from the speech processing domain.

Italiano. *Questo contributo presenta i primi risultati di un progetto per lo sviluppo di un classificatore basato sulla teoria della probabilità quantistica. Presenteremo un modello basato su kernel in grado di competere con i migliori metodi di machine learning considerando i due problemi artificiali complessi e i due casi reali sui quali è stato valutato.*

1 Introduction

Quantum Mechanics Theory (QMT) is one of the most successful theory in modern science. Despite its ability to properly describe most natural phenomena in the physics realm, the attempts to prove its effectiveness in other domains remain quite limited. Only in recent years some scholars tried to embody principles derived from QMT into their specific fields. This connection has been actively studied, for example, by the Information Retrieval community (Zuccon *et al.*, 2009; Melucci, van Rijsbergen, 2011; Gonzàlez, Caicedo, 2011) and in the domain of cognitive sciences and decision making (Busemeyer, Bruza, 2012). Also the NLP community started to look at QMT with interest and some studies using it have already been presented (Blacoe *et al.*, 2013; Liu *et al.*, 2013).

This paper presents work in progress on the development of a new classifier based on Quantum Probability Theory. Starting from the work presented in (Liu *et al.*, 2013) we will show all the limits of this simple quantum classifier and propose a new kernel-based formulation able to solve most of its problems and able to compete with a state-of-the-art classifier, namely Support Vector Machines, when classifying instances from two hard artificial problems and two real tasks taken from speech processing domain.

2 Quantum Probability Theory

A *quantum state* denotes an unobservable distribution which gives rise to various observable physical quantities (Yeang, 2010). Mathematically it is a vector in a complex Hilbert space. It can be written in Dirac notation as $|\psi\rangle = \sum_1^n \lambda_j |e_j\rangle$where λ_j are complex numbers and the $|e_j\rangle$ are the basis of the Hilbert space ($|.\rangle$ is a column vector, or a *ket*, while $\langle .|$ is a row vector, or a *bra*). Using this notation the inner product between two state vectors can be expressed as $\langle\psi|\phi\rangle$ and the outer product as $|\psi\rangle\langle\phi|$.

$|\psi\rangle$ is not directly observable but can be probed through measurements. The probability of observing the elementary event $|e_j\rangle$ is $|\langle e_j|\psi\rangle|^2 = |\lambda_j|^2$ and the probability of $|\psi\rangle$ collapsing on $|e_j\rangle$ is $P(e_j) = |\lambda_j|^2 / \sum_1^n |\lambda_i|^2$ (note that $\sum_1^n |\lambda_i|^2 = \||\psi\rangle\|^2$ where $\|\cdot\|$ is the vector norm). General events are subspaces of the Hilbert space.

A matrix can be defined as a *unitary operator* if and only if $UU^\dagger = I = U^\dagger U$, where $\dagger$ indicates the Hermitian conjugate. In quantum probability theory unitary operators can be used to evolve a quantum system or to change the state/space basis: $|\psi'\rangle = U|\psi\rangle$.

Quantum probability theory (see (Vedral, 2007) for a complete introduction) extends standard kolmogorovian probability theory and it is in principle adaptable to any discipline.

3 Quantum Classifiers

3.1 The Classifier by (Liu *et al.*, 2013)

In their paper Liu *et al.* presented a quantum classifier based on the early work of (Chen, 2002). Given an Hilbert space of dimension $n = n_i + n_o$, where n_i is the number of input features and n_o is the number of output classes, they use a unitary operator U to project the input state contained in the subspace spanned by the first n_i basis vectors into an output state contained in the subspace spanned by the last n_o basis vectors: $|\psi^o\rangle = U|\psi^i\rangle$. Input, $|\psi^i\rangle$, and output, $|\psi^o\rangle$, states are real vectors, the former having only the first n_i components different from 0 (assigned to the problem input features of every instance) and the latter only the last n_o components. From $|\psi^o\rangle$ they compute the probability of each class as $P(c_j) = |\psi^o_{ni+j}|^2 / \sum_1^{no} |\psi^o_{ni+i}|^2$ for $j = 1..n_o$.

The unitary operator U for performing instances classification can be obtained by minimising the loss function

$$err(T) = 1/\sum_{j=1}^{|T|} \langle \psi^o_j | \psi^t_j \rangle ,$$

where T is the training set and $|\psi^t\rangle$ is the target vector for output probabilities (all zeros except 1 for the target class) for every instance k, using standard optimisation techniques such as Conjugate Gradient (Hestenes, Stiefel, 1952), L-BFGS (Liu, Nocedal, 1989) or ASA (Ingber, 1989).

This classifier exhibits interesting properties. Let us examine its behaviour by using a standard non-linear problem: the XOR problem. The four instances of this problem are:

$$\begin{array}{ll} |\psi^i_1\rangle = (-1,-1,0,0) & |\psi^t_1\rangle = (0,0,1,0) \\ |\psi^i_2\rangle = (-1,1,0,0) & |\psi^t_2\rangle = (0,0,0,1) \\ |\psi^i_3\rangle = (1,-1,0,0) & |\psi^t_3\rangle = (0,0,0,1) \\ |\psi^i_4\rangle = (1,1,0,0) & |\psi^t_4\rangle = (0,0,1,0) \end{array}$$

Figure 1 depicts the probability functions for both classes as well as the decision boundaries where $P(c_1) > P(c_2)$ after a training session. Despite the relative simplicity of this classifier the two probability functions are non-linear, but the decision boundaries are linear. Nevertheless it is able to correctly classify the instances of the XOR problem.

The simplicity and the low power of this classifier emerge quite clearly when we test it with more difficult, though linearly separable, classification problems. Figure 2 shows the results of the (Liu *et al.*, 2013) classifier when applied to two simple problems. In both cases the classifier is not able

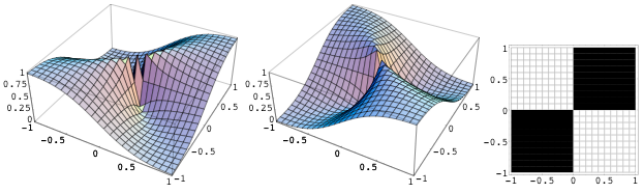


Figure 1: The probability functions for c_1 (left) and c_2 (center) for the XOR problem. At right, the decision boundaries between the two classes, where $P(c_1) > P(c_2)$ is marked in black.

to properly divide the input space into different regions corresponding to the required classes. Moreover, all the decision boundaries have to cross the origin of the feature space, a very limiting constraint for general classification problems, and problems that require strict non-linear decision boundaries cannot be successfully handled by this classifier. Nevertheless the ability of managing a classical non-linear problem, the XOR problem, is very promising and extending this method could lead, in our opinion, to interesting results.

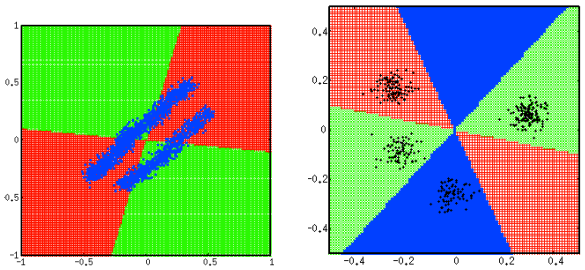


Figure 2: A two-class problem (left) and a four-class problem that cannot be successfully handled by the classifier proposed by (Liu *et al.*, 2013).

3.2 Kernel Quantum Classifier (KQC)

The goal of this paper is to extend the examined classifier in various direction in order to obtain a classification tool with higher performances.

A widely used technique to transform a linear classifier into a non-linear one involves the use of the "kernel trick". A non-linearly separable problem in the input space can be mapped to a higher-dimensional space where the decision borders between classes might be linear. We can do that through the mapping function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, with $m > n$, that maps an input state vector $|\psi^i\rangle$ to a new space. The interesting thing is that in the new space, for some particular mappings, the inner product can be calculated by using *kernel*

functions $k(x, y) = \langle \phi(x), \phi(y) \rangle$ without explicitly computing the mapping ϕ of the two original vectors.

We can express the unitary operator performing the classification process as a combination of the training input vectors in the new features space

$$\begin{aligned} |\psi^o\rangle &= U \ |\phi(\psi^i)\rangle \\ |\psi^o\rangle &= \sum_{j=1}^{|T|} |\alpha_j\rangle \ \langle \phi(\psi_j^i)| \ \ |\phi(\psi^i)\rangle \\ |\psi^o\rangle &= \sum_{j=1}^{|T|} |\alpha_j\rangle \ \langle \phi(\psi_j^i) | \phi(\psi^i)\rangle \end{aligned}$$

that can be rewritten using the kernel as

$$|\psi^o\rangle = \sum_{j=1}^{|T|} |\alpha_j\rangle \, k(\psi_j^i, \psi^i). \qquad (1)$$

Adding a bias term $|\alpha_0\rangle$ to the equation (1) lead to the final model governing this new classifier:

$$|\psi^o\rangle = |\alpha_0\rangle + \sum_{j=1}^{|T|} |\alpha_j\rangle \, k(\psi_j^i, \psi^i) \qquad (2)$$

In this new formulation we have to obtain all the $|\alpha_j\rangle$ vectors, $j = 0, ..., |T|$, through an optimisation process similar to the one of the previous case, minimising a standard euclidean loss function

$$err(T) = \sum_{j=1}^{|T|} \sum_{k=1}^{no} \left(P_j(c_k) - \psi_{j(ni+k)}^t \right)^2 + \gamma \sum_{j=0}^{|T|} \| |\alpha_j\rangle \|.$$

using a numerical optimisation algorithm, L-BFGS in our experiments, where $P(c)$ is the class probability defined in section 3.1 and $\gamma \sum \| |\alpha_j\rangle \|$ is an L_2 regularisation term on model parameters (the real and imaginary parts of $|\alpha_j\rangle$ components).

Once learned a good model from the training set T, represented by the $|\alpha_j\rangle$ vectors, we can use equation (2) and the definition of class probability for classifying new instance vectors.

It is worth noting that the KQC proposed here involves a large number of variables during the optimisation process (namely, $2 * no * (|T| + 1)$) that depends linearly on the number of instances in the training set T. In order to build a classifier applicable to real problems, we have to introduce special techniques to efficiently compute the gradient needed by optimisation methods. We relied on Automatic Differentiation (Griewank, Walther, 2008), avoiding any gradient approximation using finite differences that would require a very large number of error function evaluations. Using such techniques the training times of KQC are comparable to those of other machine learning methods.

Figure 3a and 3b show the classification results of KQC, using the linear kernel ($k(x, y) = \langle x, y \rangle$), when applied to the same problems analysed before to describe the behaviour of the (Liu *et al.*, 2013) classifier. KQC is able to discriminate efficiently between linearly separable binary or multiclass problems adapting the decision boundaries in the correct way. Moreover, using for example the RBF kernel $k(x, y) = exp(-\|x - y\|^2 / 2\sigma^2)$, is able to manage complex non-linear problems as in Figure 3c.

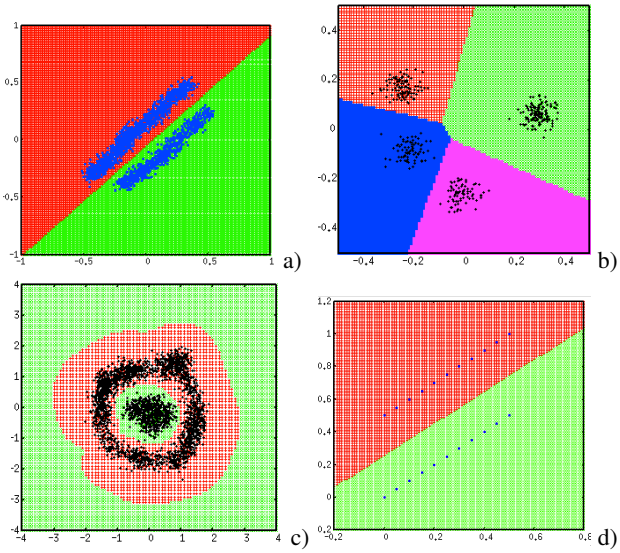


Figure 3: Some artificial problems used to verify KQC behaviour.

4 Experiments and Evaluation

In order to test quantitatively the effectiveness of the proposed quantum classifier – KQC – we set up a number of experiments, both using artificial benchmarks and real problems, and compared the KQC performances with one of the machine learning methods that usually achieve state-of-the-art performances on a large number of classification problems, that is Support Vector Machines. We relied on the SVM implementations in the SVM-light package (Joachims, 1999) and in the SVM-Multiclass package (Joachims *et al.*, 2009).

4.1 Artificial datasets

We used two artificial datasets: 2-SPIRALS and DECSIN as defined in (Segata, Blanzieri, 2009), without adding any noise to the data (see Figure 4).

They are both problems that involve a non-linear decision boundary and they are widely used for testing machine learning systems. The first dataset is composed by 628 instances and the second by 6280 instances. For both datasets $ni = no = 2$.

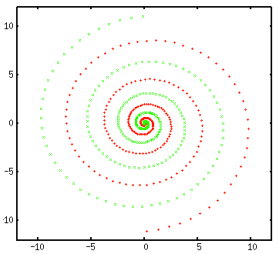
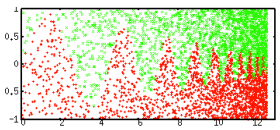

a)

b)

Figure 4: Artificial problems used for the evaluation. a) 2-SPIRALS, b) DECSIN.

4.2 Real problems

The two real problems used for the KQC evaluation are taken from the speech processing domain.

The first problem is a prominence identification task in connected speech (Tamburini, 2009). A subset of 382 utterances of the TIMIT Speech corpus (Garofolo *et al.*, 1990) has been manually annotated with binary prominence levels as described in (Tamburini, 2006). Extracting for each syllable the five acoustic features described in (Tamburini *et al.*, 2014), we formed a 35-feature input vector inserting the data from 3 syllables before and after the syllable. In total this dataset is composed of 4780 instance vectors.

The second problem is derived from an emotion recognition task. The E-Carini corpus (Tesser *et al.*, 2005) contains 322 utterances annotated with 7 fundamental emotions. From each utterance we extracted 1582 features using the OpenSMILE package (Eyben *et al.*, 2013) and the configuration file contained in the package for extracting the InterSpeech 2010 challenge feature set.

4.3 Results

Given the four dataset described above, we performed a number of experiments for comparing KQC with a SVM classifier. The reference metrics were precision/recall/F1 for the three binary-classified problems and the macro-averaged precision/recall/F1 for the Emotion multiclass dataset. All the experiments were performed executing a k-fold validation and optimising the classifiers parameters on a validation set. Table 1 outlines the different performances of the two classifiers when tested on the various evaluation datasets. KQC outperforms SVM in the experiments using artificial datasets and exhibit more or less the same performances of SVM on the real problems.

	KQC	**SVM**
2SPIRALS	RBF, σ=0.045	RBF, σ=0.02
5-fold valid.	γ=0.5	C=6e5
	P=1.0000	P=0.9532
	R=0.9969	R=0.9776
	F1=0.9984	F1=0.9650
DECSIN	RBF, σ=0.3	RBF, σ=5e-5
5-fold valid.	γ=0.5	C=1e3
	P=0.9851	P=0.9827
	R=0.9870	R=0.9805
	F1=0.9860	F1=0.9816
	KQC	**SVM**
Prominence	RBF, σ=18.0	LIN,
Detection	γ=0.5	C=30
8-fold valid.	P=0.8287	P=0.8200
	R=0.8153	R=0.8200
	F1=0.8216	F1=0.8200
Emotion	RBF, σ=75.0	LIN,
Recognition	γ=0.5	C=30
10-fold valid.	P=0.9479	P=0.9793
	R=0.9568	R=0.9728
	F1=0.9523	**F1=0.9760**

Table 1: KQC and SVM results (and optimal parameter sets) for the four evaluation problems.

5 Discussion and Conclusions

This paper presented a first attempt to produce a general purpose classifier based on Quantum Probability Theory. Considering the early experiments from (Liu *et al.*, 2013), KQC is more powerful and gains better performance. The results obtained on our experiments are quite encouraging and we are tempted to answer 'yes' to the question presented in the paper title.

This is a work in progress and the KQC is not free from problems. Despite its potential to outperform SVM using linear kernels, it is very complex to determine a tradeoff between the definition of decision boundaries with maximum margins and to maximise the classifier generalisation abilities. A long optimisation process on the training set maximise the margins between classes but could potentially lead to poor generalisations on new data. Making more experiments and evaluations in that directions is one of our future plans.

References

Blacoe W., Kashefi E. and Lapata M. 2013. A Quantum-Theoretic Approach to Distributional Semantics. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 847–857.

Busemeyer J.R. and Bruza P.D. 2012. *Quantum Models of Cognition and Decision*. Cambridge University Press.

Chen J.C.H. 2002. *Quantum Computation and Natural language Processing*. PhD thesis, University of Hamburg.

Eyben F., Weninger F., Gross F. and Schuller B. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. *ACM Multimedia (MM)*, Barcelona, 835–838.

Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett, D. and Dahlgren, N. 1990. *DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. National Institute of Standards and Technology.

Gonzàlez F.A. and Caicedo J.C. 2011. Quantum Latent Semantic Analysis. In A. Giambattista, F.Crestani (eds.),*Advances in Information Retrieval Theory*, LNCS, 6931, 52–63.

Griewank A. and Walther A. 2008. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Other Titles in Applied Mathematics 105 (2nd ed.), SIAM.

Hestenes M.R. and Stiefel E. 1952. Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49 (6), 409–436.

Ingber L. 1989. Very fast simulated re-annealing. *Mathl. Comput. Modelling*, 12 (8): 967–973.

Joachims T. 1999. Making large-Scale SVM Learning Practical. In B. Schlkopf, C. Burges, A. Smola (eds.),*Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 169–184.

Joachims T., Finley T. and Yu C-N. 2009. Cutting-Plane Training of Structural SVMs. *Machine Learning Journal*, 77 (1): 27–59.

Liu D.C. and Nocedal J. 1989. On the Limited Memory Method for Large Scale Optimization. *Mathematical Programming B*, 45 (3): 503–528.

Liu D., Yang X and Jiang M. 2013. A Novel Text Classifier Based on Quantum Computation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, 484–488.

Melucci M. and van Rijsbergen K. 2011. Quantum Mechanics and Information Retrieval. In M. Melucci and K. van Rijsbergen K (eds.),*Advanced Topics in Information Retrieval*, Springer, 33, 125–155.

Segata N. and Blanzieri E.. 2009. Empirical Assessment of Classification Accuracy of Local SVM. *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, Tilburg, 47–55.

Tamburini F. 2006. Reliable Prominence Identification in English Spontaneous Speech. *Proceedings of Speech Prosody 2006*, Dresden, PS1-9-19.

Tamburini F. 2009. Prominenza frasale e tipologia prosodica: un approccio acustico. *Linguistica e modelli tecnologici di ricerca, XL congresso internazionale di studi*, Societ di Linguistica Italiana, Vercelli, 437–455.

Tamburini F., Bertini, C., Bertinetto, P.M. 2014. Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. *Proceedings of Speech Prosody 2014*, Dublin, 285–289.

Tesser F., Cosi P., Drioli C. and Tisato G. 2005. Emotional FESTIVAL-MBROLA TTS synthesis. *Proceedings of the 9th European Conference on Speech Communication and Technology - InterSpeech2005*, Lisbon, 505–508.

Vedral V. 2007. *Introduction to Quantum Information Science*. Oxford University Press, USA.

Yeang C.H. 2010. A probabilistic graphical model of quantum systems. *Proceedings, the 9th International Conference on Machine Learning and Applications (ICMLA)*, Washington DC, 155–162.

Zuccon G., Azzopardi L.A. and van Rijsbergen K. 2009. The Quantum Probability Ranking Principle for Information Retrieval. In L.Azzopardi *et al.* (eds.),*Advances in Information Retrieval Theory*, LNCS, 5766, 232–240.

10.12871/CLICIT2014170

Geometric and statistical analysis of emotions and topics in corpora

Francesco Tarasconi
CELI S.R.L. / Turin, Italy
tarasconi@celi.it

Vittorio Di Tomaso
CELI S.R.L. / Turin, Italy
ditomaso@celi.it

Abstract

English. NLP techniques can enrich unstructured textual data, detecting topics of interest and emotions. The task of understanding emotional similarities between different topics is crucial, for example, in analyzing the Social TV landscape. A measure of how much two audiences share the same feelings is required, but also a sound and compact representation of these similarities. After evaluating different multivariate approaches, we achieved these goals by adapting Multiple Correspondence Analysis (MCA) techniques to our data. In this paper we provide background information and methodological reasons to our choice. We also provide an example of Social TV analysis, performed on Twitter data collected between October 2013 and February 2014.

Italiano. *Tecniche di NLP possono arricchire dati testuali non strutturati, individuando topic di interesse ed emozioni. Comprendere le somiglianze emotive fra diversi topic è un'attività cruciale, per esempio, nell'analisi della Social TV. E' richiesta una misura di quanto due tipi di pubblico condividano le stesse sensazioni, ma anche una rappresentazione compatta e coerente di queste somiglianze. Dopo aver valutato diversi approcci multivariati, abbiamo raggiunto questi obiettivi adattando tecniche di Multiple Correspondence Analysis (MCA) ai nostri dati. In questo articolo presentiamo background e ragioni metodologiche dietro tale scelta. Forniamo un esempio di analisi di Social TV, effettuata su dati Twitter raccolti fra ottobre 2013 e febbraio 2014.*

1 Introduction

Classification of documents based on *topics* of interest is a popular NLP research area; see, for example, Hamamoto et al. (2005). Another important subject, especially in the context of Web 2.0 and social media, is the sentiment analysis, mainly meant to detect polarities of expressions and opinions (Liu, 2012). A sentiment analysis task which has seen less contributions, but of growing popularity, is the study of *emotions* (Wiebe et al., 2005), which requires introducing and analyzing multiple variables (appropriate "emotional dimensions") potentially correlated. This is especially important in the study of the so-called Social TV (Cosenza, 2012): people can share their TV experience with other viewers on social media using smartphones and tablets. We define the empirical distribution of different emotions among viewers of a specific TV show as its *emotional profile*. Comparing at the same time the emotional profiles of several formats requires appropriate descriptive statistical techniques. During the research we conducted, we evaluated and selected geometrical methods that satisfy these requirements and provide an easy to understand and coherent representation of the results. The methods we used can be applied to any dataset of documents classified based on topics and emotions; they also represent a potential tool for the quantitative analysis of any NLP annotated data.
We used the Blogmeter platform[1] to download and process textual contents from social networks (Bolioli et al., 2013). Topics correspond to TV programs discussed on Twitter. Nine emotions are detected: the basic six according to Ekman (Ekman, 1972) (*anger, disgust, fear, joy, sadness, surprise*), *love* (a primary one in Parrot's classification) and *like/dislike* expressions, quite common on Twitter.

[1] www.blogmeter.it

2 Vector space model and dimension reduction

Let $\mathcal{D}$ be the initial data, a collection of m_D documents. Let $\mathcal{T}$ be the set of n_T distinct topics and $\mathcal{E}$ the set of n_E distinct emotions that the documents have been annotated with. Let $n = n_T + n_E$. A document $d_i \in \mathcal{D}$ can be represented as a vector of 1s and 0s of length n, where entry j indicates whether annotation j is assigned to the document or not. The *document-annotation* matrix $\mathbf{D}$ is defined as the $m_D \times n$ matrix of 1s and 0s, where row i corresponds to document vector d_i, $i = 1, \ldots, m_D$. For the rest of our analysis, we suppose all documents to be annotated with at least one topic and one emotion. $\mathbf{D}$ can be seen as a block matrix:

$$\mathbf{D}_{m_D \times n} = \begin{pmatrix} \mathbf{T}_{m_D \times n_T} & \mathbf{E}_{m_D \times n_E} \end{pmatrix},$$

where blocks $\mathbf{T}$ and $\mathbf{E}$ correspond to topic and emotion annotations.

The *topic-emotion* frequency matrix $\mathbf{T}_E$ is obtained by multiplication of $\mathbf{T}$ with $\mathbf{E}$:

$$\mathbf{T}_E = \mathbf{T}^T \mathbf{E},$$

thus $(\mathbf{T}_E)_{ij}$ is the number of co-occurrences of topic i and emotion j in the same document. In the Social TV context, rows of $\mathbf{T}_E$ represent emotional profiles of TV programs on Twitter. From documents we can obtain *emotional impressions* which are (*topic, emotion*) pairs. For example, a document annotated with {*topic = X Factor, emotion = fear, emotion = love*} generates distinct emotional impressions (*X Factor, fear*) and (*X Factor, love*). Let $\mathcal{J}$ be the set of all m_J emotional impressions obtained from $\mathcal{D}$. Then we can define, in a manner similar to $\mathbf{D}$, the corresponding *impression-annotation* matrix $\mathbf{J}$, a $m_J \times n$ matrix of 0s and 1s. $\mathbf{J}$ can be seen as a block matrix as well:

$$\mathbf{J} = \begin{pmatrix} \mathbf{T}_J & \mathbf{E}_J \end{pmatrix},$$

where blocks $\mathbf{T}_J$ and $\mathbf{E}_J$ correspond to topics and emotions of the impressions.

We can therefore represent documents or emotional impressions in a vector space of dimension n and represent topics in a vector space of dimension n_E. Our first idea was to study topics in the space determined by emotional dimensions, thus to obtain emotional similarities from matrix representation $\mathbf{T}_E$. These similarities can be defined using a distance between topic vectors or, in a manner similar to information retrieval and Latent Semantic Indexing (LSI) (Manning et al., 2008), the corresponding cosine. Our first experiments highlighted the following requirements:

1. To reduce the importance of (potentially very different) topic absolute frequencies (e.g. using cosine between topic vectors).
2. To reduce the importance of emotion absolute frequencies, giving each variable the same weight.
3. To graphically represent, together with computing, emotional similarities, as already mentioned.
4. To highlight why two topics are similar, in other words which emotions are shared.

In multivariate statistics, the problem of graphically representing an *observation-variable* matrix can be solved through dimension reduction techniques, which identify convenient projections (2-3 dimensions) of the observations. Principal Component Analysis (PCA) is probably the most popular of these techniques. See Abdi and Williams (2010) for an introduction. It is possible to obtain from $\mathbf{T}_E$ a reduced representation of topics where the new dimensions better explain the original variance. PCA and its variants can thus define and visualize reasonable emotional distances between topics. After several experiments, we selected Multiple Correspondence Analysis (MCA) as our tool, a technique aimed at analyzing categorical and discrete data. It provides a framework where requirements 1-4 are fully met, as we will show in section 3. An explanation of the relation between MCA and PCA can be found, for example, in Gower (2006).

3 Multiple Correspondence Analysis

(Simple) Correspondence Analysis (CA) is a technique that can be used to analyze two categorical variables, usually described through their *contingency table* $\mathbf{C}$ (Greenacre, 1983), a matrix that displays the frequency distribution of the variables. CA is performed through a Singular Value Decomposition (SVD) (Meyer, 2000) of the matrix of *standardized residuals* obtained from $\mathbf{C}$. SVD of a matrix finds its best low-dimensional approximation in quadratic distance. CA procedure yields new axes for rows and columns of $\mathbf{C}$ (variable categories), and new coordinates, called *principal coordinates*. Categories can be repre-

sented in the same space in principal coordinates (symmetric map). The reduced representation (the one that considers the first k principal coordinates) is the best k-dimensional approximation of row and column vectors in *chi-square* distance (Blasius and Greenacre, 2006). Chi-square distance between column (or row) vectors is an Euclidean-type distance where each squared distance is divided by the corresponding row (or column) average value. Chi-square distance can be read as Euclidean distance in the symmetric map and allow us to account for different volumes (frequencies) of categories. It is therefore desirable in the current application, but it is defined only between row vectors and between column vectors. CA measures the information contained in $\mathbf{C}$ through the *inertia* I, which corresponds to variance in the space defined by the chi-square distance, and aims to explain the largest part of I using the first few new axes. Matrix $\mathbf{T}_E$ can be seen as a contingency table for emotional impressions, and a representation of topics and emotions in the same plane can be obtained by performing CA. Superimposing topics and emotions in the symmetric map apparently helps in its interpretation, but the topic-emotion distance doesn't have a meaning in the CA framework. We have therefore searched for a representation where analysis of topic-emotion distances was fully justified.

MCA extends CA to more than two categorical variables and it is originally meant to treat problems such as the analysis of surveys with an arbitrary number of closed questions (Blasius and Greenacre, 2006). But MCA has also been applied with success to positive matrices (each entry greater or equal to zero) of different nature and has been recast (rigorously) as a geometric method (Le Roux and Rouanet, 2004). MCA is performed as the CA of the *indicator matrix* of a group of respondents to a set of questions or as the CA of the corresponding *Burt matrix* (Greenacre, 2006). The Burt matrix is the symmetric matrix of all two-way crosstabulations between the categorical variables. Matrix $\mathbf{J}$ can be seen as the indicator matrix for emotional impressions, where the questions are which topic and which emotion are contained in each impression. The corresponding Burt matrix $\mathbf{J}_B$ can be obtained by multiplication of $\mathbf{J}$ with itself:

$$\mathbf{J}_B = \mathbf{J}^T\mathbf{J} = \begin{pmatrix} \mathbf{T}_J^T\mathbf{T}_J & \mathbf{T}_J^T\mathbf{E}_J \\ \mathbf{E}_J^T\mathbf{T}_J & \mathbf{E}_J^T\mathbf{E}_J \end{pmatrix}.$$

Diagonal blocks $\mathbf{T}_J^T\mathbf{T}_J$ e $\mathbf{E}_J^T\mathbf{E}_J$ are diagonal matrices and all the information about correspondences between variables is contained in the off-diagonal blocks. From the CA of the indicator matrix we can obtain new coordinates in the same space both for respondents (impressions) and for variables (topics, emotions). From the CA of the Burt matrix it is only possible to obtain principal coordinates for the variables. MCAs performed on $\mathbf{J}$ and $\mathbf{J}_B$ yield similar principal coordinates. but with different scales (different singular values). Furthermore, chi-square distances between the columns/rows of matrix $\mathbf{J}_B$ include the contributions of diagonal blocks. For the same reason, the inertia of $\mathbf{J}_B$ can be extremely inflated.

Greenacre (2006) solves these problems by proposing an adjustment of inertia that accounts for the structure of diagonal blocks. Inertia explained in the first few principal coordinates is thus estimated more reasonably. MCA of the Burt matrix with adjustment of inertia also yields the same principal coordinates as the MCA of the indicator matrix. Finally, in the case of two variables, CA of the contingency table and MCA yield the same results. Thus the three approaches (CA, MCA in its two variants) are unified.

MCA offers possibilities common to other multivariate techniques. In particular, a measure on how well single topics and emotions are represented in the retained axes is provided (*quality* of representation).

Symmetric treatment of topics and emotions facilitates the interpretation of axes. Distances between emotions and topics can now be interpreted and, thanks to them, it is possible to establish why two topics are close in the reduced representation. An additional (and interesting) interpretation of distances between categories in terms of *sub-clouds of individuals* (impressions) is provided by Le Roux and Rouanet (2004).

4 Comparison between MasterChef and X Factor

Among the studies we conducted, we present a comparison between two popular Italian formats: X Factor (music talent show, seventh edition) and MasterChef (competitive cooking show, third edition). Each episode is considered as a different topic. Results are shown in figure 1. 82% of total inertia (after adjustment) is preserved in two dimensions, making the representation accurate.

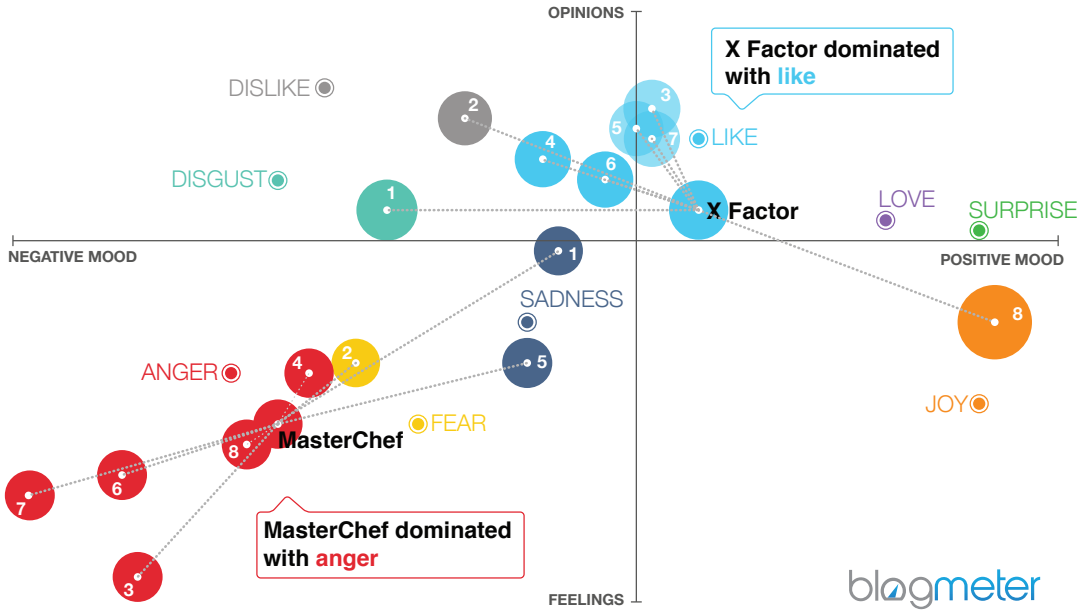


Figure 1: Comparison via MCA between X Factor and MasterChef formats, 2013-2014 editions.

Origin of axes in a MCA map acts as the barycenter, or weighted mean based on the number of emotional impressions, of all topics. In a similar way, we can consider the barycenters of X Factor/MasterChef episodes, highlighted in figure, as representants of the whole shows. Episodes are numbered progressively within each show: data were collected on a weekly basis, between 24 October and 12 December 2013 for X Factor, between 19 December 2013 and 6 February 2014 for MasterChef. X Factor obtained on average 47k emotional impressions for each episode; MasterChef an average of 8k impressions/episode. This difference in volume is reflected in the distances from the origin, which can be considered as the average profile, and therefore closer to X Factor.

By looking at the position of emotions, the first axis can be interpreted as the contrast between *moods* (positive and negative) of the public, and this is therefore highlighted as the most important structure in our dataset. X Factor was generally perceived in a more positive way than MasterChef. The advantage of incorporating emotions in our sentiment analysis is more manifest when we look at the second retained axis. We can say the audience of X Factor lives in a world of opinion dominated by *like/dislike* expressions, while the public of MasterChef is characterized by true and active feelings concerning the show and its protagonists. This is coherent with the fact that viewers of X Factor could directly evaluate the performances of contestants. This was not possible for the viewers of MasterChef, who focused instead on the most outstanding and emotional moments of the show. Reaching these conclusions would not have been possible by looking at simple polarity of impressions.

5 Conclusions and further researches

By applying carefully chosen multivariate statistical techniques, we have shown how to represent and highlight important emotional relations between topics. Further results in the MCA field can be experimented on datasets similar to the ones we used. For example, additional information about opinion polarity and document authors (such as Twitter users) could be incorporated in the analysis. The geometric approach to MCA (Le Roux and Rouanet, 2004) could be interesting to study in greater detail the *clouds* of impressions and documents (**J** and **D** matrices); authors could also be considered as mean points of well-defined sub-clouds.

Ancknowledgements

We would like to thank: V. Cosenza and S. Monotti Graziadei for stimulating these researches; the ISI-CRT foundation and CELI S.R.L. for the support provided through the Lagrange Project; A. Bolioli for the supervision and the essential help in the preparation of this paper.

References

Hervé Abdi and Lynne J. Williams. 2010. *Principal Component Analysis*, Wiley Interdisciplinary Reviews: Computational Statistics, Volume 2, Issue 4, pp. 433-459.

Jörg Blasius and Michael Greenacre. 2006. *Correspondence Analysis and Related Methods in Practice*, Multiple Correspondence Analysis and Related Methods, Chapter 1. CRC Press.

Andrea Bolioli, Federica Salamino and Veronica Porzionato. 2013. *Social Media Monitoring in Real Life with Blogmeter Platform*, ESSEM@AI*IA 2013, Volume 1096 of CEUR Workshop Proceedings, pp. 156-163. CEUR-WS.org.

Vincenzo Cosenza. 2012. *Social Media ROI*. Apogeo.

Paul Ekman, Wallace V. Friesen and Phoebe Ellsworth. 1972. *Emotion in the Human Face*. Pergamon Press.

Dario Galati. 2002. *Prospettive sulle emozioni e teorie del soggetto*. Bollati Boringhieri.

John C. Gower. 2006. *Divided by a Common Language: Analyzing and Visualizing Two-Way Arrays*, Multiple Correspondence Analysis and Related Methods, Chapter 3. CRC Press.

Michael Greenacre. 1983. *Theory and Applications of Correspondence Analysis*. Academic Press.

Michael Greenacre. 2006. *From Simple to Multiple Correspondence Analysis*, Multiple Correspondence Analysis and Related Methods, Chapter 2. CRC Press.

Masafumi Hamamoto, Hiroyuki Kitagawa, Jia-Yu Pan and Christos Faloutsos. 2005. *A Comparative Study of Feature Vector-Based Topic Detection Schemes for Text Streams*, Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Springer.

Brigitte Le Roux and Henry Rouanet. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data*. Kluwer.

Bing Liu. 2012. *Sentiment Analysis e Opinion Mining*. Morgan & Claypool Publishers.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Carl D. Meyer. 2000. *Matrix Analysis and Applied Linear Algebra*. Siam.

Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. *Annotating expressions of opinions and emotions in language*, Language Resources and Evaluation, Volume 39, Issue 2-3, pp. 165-210.

10.12871/CLICIT2014171

Corpus ICoN: una raccolta di elaborati di italiano L2 prodotti in ambito universitario

Mirko Tavosanis
Università di Pisa
Dipartimento di Filologia,
Letteratura e Linguistica
Via Santa Maria 36, 56126 Pisa
PI

tavosanis@ital.unipi.it

Abstract

Italiano. Il contributo presenta le caratteristiche essenziali del Corpus ICoN. Il corpus raccoglie elaborati realizzati nell'arco di 13 anni da studenti universitari; gli elaborati sono ripartiti in due sottocorpora equivalenti dedicati rispettivamente agli studenti che conoscono l'italiano come L1 e a quelli che lo conoscono come L2/LS.

English. *The paper describes the essential features of the Corpus ICoN. The corpus includes essays created over 13 years by university students; the essays are divided into two comparable subcorpora dedicated respectively to students who speak Italian as L1 and those who know the language as L2/FL.*

1 Introduzione

I corpora di testi realizzati come L2 sono da tempo uno strumento essenziale per lo studio dell'apprendimento delle lingue. Nel caso dell'italiano, tuttavia, anche se esistono prodotti importanti e ottimamente realizzati, il numero di corpora è ancora ritenuto insufficiente per molti tipi di ricerche (per una panoramica: Andorno e Rastelli 2009).

Il corpus in allestimento descritto qui di seguito mira a fornire in contributo in questo senso. Il lavoro si colloca all'interno delle attività del progetto PRIN "Scritture brevi" ed è previsto che il prodotto finale venga usato in primo luogo dall'Istituto di Linguistica Computazionale del CNR di Pisa per la messa a punto di strumenti di valutazione automatica dell'elaborato di apprendenti.

Il lavoro è attualmente ancora in corso. La conclusione delle attività è prevista per la fine del 2015, ma le caratteristiche complessive del corpus sono già ben definite e rendono quindi possibile una presentazione articolata.

2 Composizione del corpus

Il corpus è composto da circa 8000 elaborati complessivi. L'elaborazione e l'eliminazione delle irregolarità sono ancora in corso, ma le dimensioni del corpus finale sono al momento stimate in circa due milioni di token.

Il corpus si divide in due sottocorpora equivalenti tra loro come dimensione (circa un milione di token l'uno). Il primo è composto da elaborati realizzati da studenti che hanno l'italiano come L1; il secondo è composto da elaborati di studenti che conoscono l'italiano come LS/L2 e a un livello almeno pari al B2.

I due sottocorpora sono formati da testi realizzati dai relativi gruppi di studenti in circostanze identiche tra di loro. Ciò rende evidente la possibilità di un confronto tra i due corpora sul modello consolidato VALICO / VINCA.

3 Il Corso di Laurea ICoN

ICoN - Italian Culture on the Net – è un consorzio di università italiane (19, al momento della stesura di questo testo) che opera in collaborazione con il Ministero per gli Affari Esteri. Il Consorzio è stato fondato nel 1999 con il patro-

nato della Camera dei Deputati e con il supporto della Presidenza del Consiglio e del Ministero per l'Università e la Ricerca. Nella pratica, ICoN opera attraverso il proprio sito web, all'indirizzo: www.italicon.it (Tavosanis 2004).

Scopo del Consorzio è "promuovere e diffondere la lingua e la cultura italiana nel mondo" attraverso Internet e iniziative educative specifiche. Le attività mirate a questo scopo sono diverse, e includono per esempio la realizzazione di corsi di lingua e l'erogazione di Master universitari e corsi di aggiornamento. Il servizio più antico del Consorzio è però l'erogazione di un Corso di Laurea triennale in Lingua e cultura italiana per stranieri. Attivo dal 2001, il Corso di Laurea è erogato completamente via Internet ed è rivolto a due precise fasce di studenti: cittadini stranieri e cittadini italiani residenti all'estero. Di fatto, in oltre dieci anni di attività il Corso di Laurea ha avuto tra i propri iscritti un numero grosso modo equivalente di stranieri e di italiani (v. sezione 5). Dalle produzioni didattiche realizzate per il Corso è quindi possibile ricavare due corpora approssimativamente simili come estensione e del tutto comparabili come origine.

3.1 Criteri d'ammissione al Corso

I criteri di ammissione al Corso sono gli stessi di tutti i Corsi di Laurea delle università italiane. Per l'iscrizione è necessario possedere due requisiti: un titolo di studio che consenta l'accesso all'Università in Italia o nel paese di provenienza e una conoscenza della lingua italiana pari o superiore al livello B2.

3.2 Prove d'esame

Le prove scritte d'esame sono state svolte con modalità immutate fin dal primo anno accademico di operatività del Corso. Ogni corso all'interno del piano di studi si è quindi concluso con una prova scritta, che ogni studente ha dovuto realizzare al computer. Le prove si svolgono all'estero (e, in rari casi, in Italia, presso la sede del Consorzio) e sono composte da due parti. Lo studente deve infatti fornire le risposte a una batteria di trenta domande e scrivere un breve elaborato (descritto qui in dettaglio al punto 4). Per svolgere entrambi i compiti sono disponibili complessivamente 90 minuti, che ogni studente può dedicare all'una o all'altra parte nella proporzione che preferisce. Al termine del tempo stabilito il programma impedisce ulteriori modifiche; le prove vengono poi trasmesse in forma criptata alla sede centrale ICoN per la valutazione.

Durante le prove di esame gli studenti si trovano in ambienti controllati, in modo che non possano consultare libri o appunti, e il computer su cui operano è scollegato dalla rete fino al termine delle prove, in modo che sia impossibile fare riferimento a testi disponibili su Internet.

4 Gli elaborati

Gli elaborati del corso di laurea costituiscono il punto di partenza per la costituzione del corpus ICoN.

4.1 Caratteristiche dell'elaborato

La prova scritta è, in pratica, un piccolo tema. Il candidato può scegliere una traccia tra le tre alternative che gli sono proposte.

Esempi tipici di consegna sono:

Il restauro barocco di Maratta e il restauro ottocentesco di Cavalcaselle: metti a confronto due atteggiamenti diversi nei confronti della conservazione dell'opera d'arte.

Analizza il rapporto tra Petrarca e l'Umanesimo.

Illustra il concetto di equivalenza e il suo ruolo nella metodologia del confronto interlinguistico.

Il testo che segue è invece un esempio tipico di inizio di elaborato:

Raffaello Sanzio e' uno dei maggiori rappresentanti internazionali del Rinascimento italiano. Lui ha lavorato, come molti altri artisti famosi a quei tempi, in Roma - sede della Chiesa Cattolica e centro di grandi imprese artistiche com temi teologici. Uno dei posti din concentrazione dell'attivita' era il Vaticano, edificio accanto alla Bassilica di San Pietro. Giulio II era in ufficio nel momento in quale chiese a Raffaello di decorare le stanze private del papa, nel 1508. La prima stanza affrescata e' stata quella della Segantura. Qui le quattro pareti sono state divise usando il modello della divisione del sapere diritto, folosofia, poesia (al posto della medicina) e teologia. Cosi' si fa il percorso del bene, del vero e del bello.

Le caratteristiche testuali attese sono quelle degli elaborati prodotti in ambiente universitario. Nel sistema italiano, esami scritti di questo tipo

sembrano relativamente rari ma esistono anche nei corsi di laurea tradizionali.

4.2 Interfaccia

L'interfaccia di scrittura è formata da una finestra molto semplice. La finestra include un indicatore che mostra il tempo ancora disponibile per completare la prova e un contatore di caratteri che mostra la lunghezza dell'elaborato.

L'interfaccia non include invece strumenti avanzati di gestione del testo (cerca e sostituisci) e strumenti di formattazione (corsivi, grassetti e simili).

Inoltre, l'interfaccia non possiede funzioni di controllo ortografico. Questa caratteristica ha ovvie motivazioni didattiche ma si combina anche con un fattore esterno per produrre risultati linguisticamente interessanti. Poiché gli studenti sono residenti all'estero, le tastiere usate per scrivere le prove sono infatti solo raramente tastiere italiane. Ciò fa sé che spesso per gli studenti sia difficile inserire le lettere accentate. Esempi tipici di scrittura sono quindi:

La famiglia *e'* l'obiettivo principale, la vita professionale ancora non svolse per loro un ruolo di grande rilievo. Dagli anni settanta fino agli anni 90 *e'* diminuito il numero dei nuclei famigliari. Con il tempo si *noto'* sempre di *piu'* il processo della femminizzazione (iniziato *gia'* dopo I Guerra Mondiale), il quale ebbe un forte impatto sul nucleo famigliare.

Le commissioni d'esame sono a conoscenza della situazione e sono quindi invitate a ignorare deviazioni di questo genere dall'ortografia standard in tutti i casi in cui si può ragionevolmente ritenere che le ragioni a monte siano esclusivamente di questo tipo.

4.3 Archiviazione

Al termine delle prove gli elaborati vengono registrati all'interno di un file XML (la DTD di riferimento non è usata per validazioni in corso d'opera). Le prime righe dei file hanno di regola questo aspetto:

```
  <ESAMI idstudente="93969" nomestu-
dente="(eliminato)">
  <ESAME idnucleo="498" nome="Lingua
e letteratura latina/Letteratura la-
tina, medievale e umanistica 1 LET
A" status="F" datainizio="01/02/2011
08.09.28" datafine="01/02/2011
09.39.29" duratamax="90" ultimari-
sposta="00">
  <ESERCIZIO MODULEID="" unita=""
poolid="" testid="9440">
```

I file XML vengono poi criptati e inviati per posta elettronica alla sede operativa ICoN. A consegna avvenuta, i file vengono decriptati, controllati, caricati sul server di archiviazione e resi disponibili alle commissioni.

Le commissioni controllano eventuali anomalie nei test e valutano gli elaborati secondo griglie predefinite. La valutazione viene conservata su file separati.

5 Studenti

I dati anagrafici degli studenti rappresentano la prima fonte di informazione sociolinguistica per il corpus. La segreteria ICoN registra infatti le informazioni principali, incluse dichiarazioni sulla L1 e sulle L2 conosciute. Questi dati, opportunamente anonimizzati ai fini dell'analisi, sono poi raccordati ai singoli elaborati in modo da permettere la selezione dei testi attraverso diversi criteri.

Per la sostanza dei dati, va notato che la provenienza degli studenti è molto varia. I circa 300 laureati che hanno conseguito il titolo di studio entro l'estate del 2014 provengono infatti da 56 paesi diversi. Questa varietà segue una distribuzione spiccatamente da "coda lunga": i primi quattro paesi di provenienza dei laureati (Argentina, Germania, Brasile e Turchia) non solo corrispondono di regola a quattro diverse lingue madri ma forniscono complessivamente poco più di un quarto del totale dei laureati. Le L1 di origine sono quindi quasi altrettanto variate e tra gli studenti sono abbondantemente rappresentate lingue che vanno dal polacco al farsi.

Tuttavia, come accennato al punto 2, è possibile fare una distinzione molto forte tra due categorie di studenti: quelli che hanno l'italiano come L1 e quelli che invece lo conoscono come L2/LS.

5.1 Italiano come L1

In generale, si può dare per scontata la conoscenza dell'italiano a livello madrelingua da parte dei cittadini italiani che abbiano compiuto buona parte del proprio percorso formativo in Italia. Nelle scritture degli studenti residenti da molto tempo all'estero, però, sono presenti occasionalmente esempi di erosione dell'italiano o di interferenza da parte delle L2.

5.2 Italiano come L2/LS

Il livello degli studenti stranieri è molto variabile. Sebbene tutti siano accomunati da una conoscenza della lingua di livello almeno B2, la differenza tra studenti con conoscenza appena sufficiente e studenti con competenze assimilabili a L1 è molto vistosa.

Gestire questa diversità è sicuramente una delle sfide principali nell'elaborazione del corpus. In una prima fase, l'assegnazione del livello di competenze dovrà essere fatta interamente da valutatori umani. In una seconda fase, è possibile che l'operazione possa essere condotta in parte in modo automatico.

Nel corpus definitivo gli studenti che hanno avuto contatti con l'italiano come L2 saranno distinti da quelli per cui l'italiano è stato solo LS. In entrambi i casi inoltre, compatibilmente con la documentazione disponibile, saranno etichettati gli studenti che per vari motivi (origine familiare, ambiente, trasferimenti in Italia) hanno avuto un contatto con l'italiano diverso da quanto normalmente prevedibile per una L2/LS. La granularità di questa etichettatura non è ancora stata definita; soprattutto per gli studenti degli ultimi anni, che come parte della procedura di iscrizione forniscono spesso lettere di motivazione e descrizioni dei propri contatti con la lingua e la cultura italiana, è possibile che possa essere realizzata una descrizione molto dettagliata, probabilmente presentata sotto forma di testo articolato.

6 Inserimento degli elaborati all'interno del corpus

Il corpus prevede che gli elaborati vengano importati come testo semplice con codifica UTF-8. Un punto delicato è la gestione degli errori ortografici collegati a tastiere non italiane e descritti a 4.2. Tuttavia i primi esempi di analisi, condotti con il sistema READ-IT dell'Istituto di Linguistica Computazionale del CNR di Pisa (Dell'Orletta, Montemagni e Venturi 2011), mostrano che gli strumenti oggi esistenti possono ricondurre senza problemi gli errori di questo genere alle forme target, senza che sia necessaria neanche una fase di addestramento.

7 Distribuzione del corpus

Il prodotto finito sarà reso disponibile in forma mediata. Per ragioni collegate alla natura del corpus non sarà quindi possibile il libero scaricamento degli elaborati o il loro collegamento a consegne. Si prevede che la ricerca avvenga attraverso un'interfaccia web e che la dimensione dei contesti venga limitata, in alternativa, ai confini di frase o a un massimo di 300 caratteri.

8 Conclusioni

L'elaborazione del corpus è ancora in corso. Tuttavia, gli assaggi eseguiti fino a questo momento sono molto promettenti e rassicurano sull'utilità del progetto. Di particolare valore sembra la possibilità di confrontare i testi prodotti da studenti che hanno o meno l'italiano come L1 in circostanze in cui i fini comunicativi rispondono a una precisa realtà didattica.

Bibliografia

Cecilia Andorno - Stefano Rastelli, *Corpora di italiano L2. Tecnologie, metodi, spunti teorici*, Perugia: Guerra, 2009.

Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, *Read-it: Assessing Readability of Italian texts with a View to Text Simplification.* In: *Proceedings of the 2nd Workshop on Speech and Language processing for Assistive Technologies*, Edinburgh, 2011, pp. 73-83.

Mirko Tavosanis, *L'italiano del web*, Roma: Carocci, 2011.

Mirko Tavosanis, *Insegnamento di lingua e cultura italiana a stranieri: l'esperienza di ICoN*. In: *Italiano e italiani nel mondo. Italiani all'estero e stranieri in Italia: identità linguistiche e culturali*. Vol. 1, Roma: Bulzoni, pp. 1-13.

10.12871/CLICIT2014172

Coreference Resolution for Italian: Assessing the Impact of Linguistic Components

Olga Uryupina[1] and **Alessandro Moschitti**[2,1]
[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Qatar Computing Research Institute
uryupina@gmail.com, amoschitti@gmail.com

Abstract

English. This paper presents a systematic evaluation of two linguistic components required to build a coreference resolution system: mention detection and mention description. We compare gold standard annotations against the output of the modules based on the state-of-the-art NLP for Italian. Our experiments suggest the most promising direction for future work on coreference in Italian: we show that, while automatic mention description affects the performance only mildly, the mention detection module plays a crucial role for the end-to-end coreference performance. We also show that, while a considerable number of mentions in Italian are zero pronouns, their omission doesn't affect a general-purpose coreference resolver, suggesting that more specialized algorithms are needed for this subtask.

Italiano. *Questo articolo presenta una valutazione sistematica di due componenti linguistiche necessarie per costruire un sistema di risoluzione delle coreferenze: (i) selezione automatica delle menzioni ad entitá e (ii) la loro descrizione in termini di features. Per questo scopo si confrontano i risultati dei moduli che sono allo stato dell'arte per l'italiano basati sul NLP con le annotazioni gold standard. Questi esperimenti suggeriscono le direzioni di ricerca piú promettenti per futuri lavori sulla coreferenza per l'italiano: infatti, si dimostra che, mentre la descrizione automatica delle menzioni influisce sulle prestazioni solo leggermente, il modulo di selezione delle menzioni svolge un ruolo fondamentale per la prestazione del risolutore di coreferenze (end-to-end). Si dimostra anche che, mentre un numero considerevole di menzioni in italiano sono zero-pronouns, la loro omissione non pregiudica il risultato di coreferenza. Questo suggerisce che algoritmi piú specializzati sono necessari per questa sottoattivitá.*

1 Introduction

Coreference Resolution is an important prerequisite for a variety of Natural Language Processing tasks, in particular, for Information Extraction and Question Answering, Machine Translation or Single-document Summarization. It is, however, a challenging task, involving complex inference over heterogeneous linguistic cues. Several high-performance coreference resolvers have been proposed recently in the context of the CoNLL-2011 and CoNLL-2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012). These systems, however, are engineered to process English documents and cannot be directly applied to other languages: while the CoNLL-2012 shared task includes Arabic and Chinese datasets, most participants have not investigated any language-specific approaches and have relied on the same universal algorithm, retraining it for particular corpora.

To our knowledge, only very few systems have been proposed so far to provide end-to-end coreference resolution in Italian. In the context of the SemEval-2010 shared task (Recasens et al., 2010), four systems have attempted Italian coreference. Among these toolkits, only BART relied on any language-specific solutions at this stage (Broscheit et al., 2010; Poesio et al., 2010). The TANL system, however, was enhanced with language-specific information and integrated into the University of Pisa Italian pipeline later on (Attardi et al., 2012). At Evalita 2009 and 2011, different variants of coreference resolution were proposed as shared tasks (Lenzi and Sprugnoli, 2009; Uryupina and Poesio, 2012), in both cases, only one participant managed to submit the final run.

One of the bottlenecks in creating high-performance coreference resolvers lies in the complexity of their architecture. Coreference is a deep linguistic phenomenon and state-of-the-art

systems incorporate multiple modules for various related subtasks. Even creating a baseline end-to-end resolver is therefore a difficult engineering task. Going beyond the baseline is even more challenging, since it is generally unclear how different types of errors might affect the overall performance level. This paper focuses on systematic evaluation of different sub-modules of a coreference resolver to provide a better understanding of their impact on the system's performance and thus suggest more promising venues for future research. Starting with a gold pipeline, we gradually replace its components with automatic modules, assessing the impact. The ultimate goal of our study is to boost the performance level for Italian. We are focusing on improving the language-specific representation, leaving aside any comparison between coreference models (for example, mention-pair vs. mention-entity vs. graph-based).

2 Coreference Resolution Pipelines

End-to-end coreference resolvers operate on raw texts, requiring a full linguistic pipeline to preprocess the data. Below we describe the preprocessing pipeline used in our study and then proceed to the proper coreference pipeline.

2.1 Preprocessing pipeline

Our preprocessing pipeline for Italian is a part of the *Semantic Model Extractor* developed for the EU FP7 LiMoSINe project.[1] The LiMoSINe Semantic Model contains various levels of linguistic description, representing a document from different angles. It combines outputs of numerous linguistic preprocessors to provide a uniform and deep representation of document's semantics. Our Semantic Model is based on Apache UIMA—a framework for Unstructured Information Management,[2] successfully used for a number of NLP projects, e.g., for the IBM Watson system.

TextPro wrapper. To provide basic levels of linguistic processing, we rely on TextPro (Pianta et al., 2008).—a suite of Natural Language Processing tools for analysis of Italian (and English) texts. The suite has been designed to integrate various NLP components developed by researchers at Fondazione Bruno Kessler (FBK). The TextPro suite has shown exceptional performance for several NLP tasks at multiple EvalIta competitions. Moreover, the toolkit is being constantly updated and developed further by FBK. We can therefore be sure that TextPro provides state-of-the-art processing for Italian. TextPro combines rule-based and statistical methods. In addition, it allows for a straightforward integration of task-specific user-defined pre- and post-processing techniques. For example, one can customize TextPro to provide better segmentation for web data. We use Textpro to extract part-of-speech tags, named entities, lemmata and token-level morphology.

Parsing. A model has been trained for Italian on the Torino Treebank data[3] using the Berkeley parser by the Fondazione Bruno Kessler. The treebank being relatively small, a better performance can be achieved by enforcing TextPro part-of-speech tags when training and running the parser. Both the Torino Treebank itself and the parsing model use specific tagsets that do not correspond to the Penn TreeBank tags of the English parser. To facilitate cross-lingual processing and enable unlexicalized cross-lingual modeling for deep semantic tasks, we have mapped these tagsets to each other.

2.2 Coreference pipeline

Once the preprocessing pipeline has created a rich linguistics representation of the input documents, a statistical coreference resolver runs a sequence of sub-modules to provide appropriate information to its model, train/run its classifier and use the output to create coreference chains . This involves the following steps:

- **Mention extraction.** The goal of this step is to extract nominal *mentions* from the textual stream. The exact definition of what is to be considered a mention varies across different annotation schemes. Roughly speaking, nominal chunks, named entities and pronouns (including zeroes) are potential mentions. More fine-grained schemes distinguish between different type of mentions (e.g., referential vs. non-referential) and discard some of them from the scope of their annotation.
- **Mention description.** This component provides a meaningful representation of each mention, extracting its linguistic properties, for example: mention type, number, gender and semantic class.
- **Feature extraction.** This component relies on mention descriptions to create feature vectors for the classifier. The exact nature of

[1] http://www.limosine-project.eu

[2] http://uima.apache.org/

[3] http://www.di.unito.it/~tutreeb/

the feature vector depends on the selection of the underlying model. Thus, in the *mention-pair* model (Soon et al., 2001), used in our study, each vector corresponds to two mentions from the same document, the anaphor and the antecedent. The individual features, engineered manually, combine different bits of information from the corresponding descriptions. An example of such a feature is "the anaphor is a pronoun and it agrees in gender with the antecedent".

- **Modeling.** At the final step, the classifier is trained and tested on the feature vectors and its prediction is then passed to a clustering algorithm to create the resulting partition.

In this paper, we focus on the first two steps, since they require the largest language-specific engineering effort. We believe that the modeling part is relatively language-independent and that most high-performance state-of-the-art models can be applied to Italian if adequate feature representations can be extracted. In our study, we rely on the simple and fast mention-pair model (Soon et al., 2001). We have tested several machine learners (Decision Trees, SVMs and MaxEnt), observing that the highest performance is achieved with decision trees.

3 Experiments

For our experiment, we use a cleaned up version of the LiveMemories Wikipedia corpus (Rodríguez et al., 2010). The first version of the same dataset was adopted for the Anaphora Resolution track at Evalita-2011 (Uryupina and Poesio, 2012). We have invested considerable efforts in checking the consistency of the annotations and adjusting them when necessary. The second version of the corpus will be publicly available by the end of 2014. We refer the reader to the Evalita Anaphora Resolution track guidelines for a detailed description of the dataset, including the adopted methodology on defining mentions (for example, on the treatment of appositions) and mention boundaries (spans).

The LiveMemories Wikipedia corpus provides rich annotations of nominal mentions. In particular, each mention is characterized for its number, gender, semantic class and referentiality. We will not assess the impact of referentiality on the final performance in this paper, since no automatic referentiality detector has been proposed for Italian so far. However, the corpus does not contain any gold-standard annotations of the basic linguistic levels (for example, parse trees, part of speech tags): all the preprocessing was conducted using automatic modules.

In our experiments, we replace the LiveMemories basic levels with the LiMoSINe pipeline, since it relies on the more recent and robust technology. For coreference components, we start with the oracle pipeline: we extract mentions from the gold annotations and use gold attributes to provide mention descriptions. The performance level of the system with the oracle pipeline can be considered the upper bound for the selected feature extractors and model configurations. The first row of Table 1 summarize the performance level of such a system. We report F-scores for the three most commonly used metrics for coreference (MUC, B^3 and $CEAF_{\phi_3}$). We then gradually replace the oracle components with the automatic ones, measuring the drop in the system's performance.

3.1 Mention Description

In our first experiment, we take gold mention boundaries and try to describe mention properties automatically. To this end, we try to extract the head of each mention. We traverse the parse tree for mentions corresponding to parse nodes. For other mentions, we rely on simple heuristics for extracting head nouns. Once the head noun has been extracted, we consult the TexPro morphology to determine its number and gender. If the mention aligns with some named entity, as extracted by TextPro, we also assign it a semantic type.

This methodology may lead to incomplete or incorrect mention descriptions for various reasons. First, the head-finding rules, especially for mentions that do not correspond to any parsing nodes (this can happen, for example, if the parsing tree is erroneous itself), are not perfect. Second, the TextPro morphology may provide misleading cues. These two types of errors can be remedied in the future with the advancement of the NLP technology. The third group of errors are the cases when the LiveMemories annotators assign some attributes to a mention to agree with other members of its coreference chain. For example, pronouns often receive semantic type attributes that can not be inferred from the corresponding one-sentence contexts. For such cases, a joint model for mention description and coreference resolution might be beneficial. Denis and Baldridge (2009) propose an example of such a model for joint coreference resolution and NE classification.

Components	MUC	CEAF	B^3
Gold boundaries, gold descriptions	60.7	67.8	78.4
Gold boundaries, automatic descriptions	60.0	66.0	77.2
Gold boundaries with no zero pronouns, automatic descriptions	60.3	65.8	76.4
Automatic boundaries, automatic descriptions	46.8	50.3	52.2

Table 1: The system performance with automatic and oracle modules, F-scores.

The second row of Table 1 shows that while the system performance decreases with imperfect mention descriptions, the drop is not large. We believe that this can be explained by two factors:

- unlike many other datasets, the LiveMemories corpus provides two boundaries for each mention: the *minimal* and *maximal* span; since minimal spans are very short and typically contain 1-3 words, the head finding procedure is more robust;
- while the system is not always able to extract implicit properties for pronouns, the explicit morphology (number and gender) often provides enough information for the coreference resolver; this is in a sharp contrast with the same task for English, where the lack of explicit morphological marking on candidate antecedents makes it essential to extract implicit properties as well.

3.2 Mention Extraction

In our second experiment, we replace the mention extraction module with the automatic one. The automatic mention extractor is a rule-base system developed for English and adjusted for Italian (Poesio et al., 2010). Since the system cannot handle zero pronouns, we do the assessment in two steps. For the first run (row 3, Table 1), we take all gold mentions that are not zeroes and thus provide a more accurate upper bound for our approach. For the second run (row 4), we do the mention extraction fully automatically. Neither run uses any gold information about mention properties.

The most surprising results is the performance level of the system with no zero pronouns. When we remove them from the oracle, the performance doesn't decrease at all. This can be explained by the fact that zero pronouns are very different from other types of mentions and require special algorithms for their resolution. The general-purpose system cannot handle them correctly and produces too many errors. We believe that while zero pronouns pose a challenging problem, the more promising approach would treat them separately from other anaphors, capitalizing on various syntactic clues for their identification and resolution. An example of such an approach for Italian has been advocated by Iida and Poesio (2011).

Altogether, when gold mention boundaries are replaced with the automatic ones, the performance goes down considerably. This is a common problem for coreference and has been observed for many other languages. This finding suggests that the first step in boosting the performance level of a coreference resolver should focus on improving the mention extraction part.

4 Conclusion

In this paper, we have attempted an extensive evaluation of the impact of two language-specific components on the performance of a coreference resolver for Italian. We show that the mention extraction module plays a crucial role, while the contribution of the mention description model, while still important, is much less pronounced. This suggests that the mention extraction subtask should be in the primary focus at the beginning of the language-specific research on coreference. Our future work in this direction includes developing a robust statistical mention detector for Italian based on parse trees.

We also show that zero pronouns can not be handled by a general-purpose coreference resolver and should therefore be addressed by a separate system, combining their extraction and resolution.

Finally, our study has not addressed the last language-specific component of the coreference pipeline, the feature extraction module. Its performance cannot be assessed via a comparison with an oracle since there are no perfect gold features. In the future,we plan to evaluate the impact of this component by comparing different feature sets, engineered both manually and automatically.

Acknowledgments

The research described in this paper has been partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant #288024: LIMOSINE – Linguistically Motivated Semantic aggregation engiNes.

References

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2012. UNIPI participation in the Evalita 2011 Anaphora Resolution Task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Lecture Notes in Computer Science 7689*. Springer.

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, and Yannick Versley. 2010. BART: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. In *Procesamiento del Lenguaje Natural 42, Barcelona: SEPLN*.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 804–813.

Valentina Bartalesi Lenzi and Rachele Sprugnoli. 2009. EVALITA 2009: Description and results of the local entity detection and recognition (LEDR) task. In *Proceedings of Evalita-2009*.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of the Language Resources and Evaluation Conference*.

Massimo Poesio, Olga Uryupina, and Yannick Versley. 2010. Creating a coreference resolution system for Italian. In *Proceedings of the Linguistic Coreference Workshop at the International Conference on Language Resources and Evaluation (LREC'10)*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon, June.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Marta Recasens, Lluís Màrquez, Emili Sapena, M.Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of Wikipedia and blogs in the Live Memories corpus. In *Proceedings of the 7th International Conference on Language Ressources and Evaluation*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistic*, 27(4):521–544.

Olga Uryupina and Massimo Poesio. 2012. EvalIta 2011: Anaphora resolution task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Lecture Notes in Computer Science 7689*. Springer. (extended version).

10.12871/CLICIT2014173

A context based model for Sentiment Analysis in Twitter for the Italian Language

Andrea Vanzo(†), **Giuseppe Castellucci**(‡), **Danilo Croce**(†) **and Roberto Basili**(†)
(†)Department of Enterprise Engineering
(‡)Department of Electronic Engineering
University of Roma, Tor Vergata
Via del Politecnico 1, 00133 Roma, Italy
{vanzo,croce,basili}@info.uniroma2.it, castellucci@ing.uniroma2.it

Abstract

English. Recent works on Sentiment Analysis over Twitter leverage the idea that the sentiment depends on a single incoming tweet. However, tweets are plunged into streams of posts, thus making available a wider context. The contribution of this information has been recently investigated for the English language by modeling the polarity detection as a sequential classification task over streams of tweets (Vanzo et al., 2014). Here, we want to verify the applicability of this method even for a morphological richer language, i.e. Italian.

Italiano. *Studi recenti per la Sentiment Analysis in Twitter hanno tentato di creare modelli per caratterizzare la polaritá di un tweet osservando ciascun messaggio in isolamento. In realtà, i tweet fanno parte di conversazioni, la cui natura può essere sfruttata per migliorare la qualità dell'analisi da parte di sistemi automatici. In (Vanzo et al., 2014) è stato proposto un modello basato sulla classificazione di sequenze per la caratterizzazione della polarità dei tweet, che sfrutta il contesto in cui il messaggio è immerso. In questo lavoro, si vuole verificare l'applicabilità di tale metodologia anche per la lingua Italiana.*

1 Introduction

Web 2.0 and Social Networks allow users to write about their life and personal experiences. This huge amount of data is crucial in the study of the interactions and dynamics of subjectivity on the Web. Sentiment Analysis (SA) is the computational study and automatic recognition of opinions and sentiments. Twitter is a microblogging service that counts about a billion of active users. In Twitter, SA is traditionally treated as any other text classification task, as proved by most systems participating to the *Sentiment Analysis in Twitter* task in SemEval-2013 (Nakov et al., 2013). A Machine Learning (ML) setting allows to induce detection functions from real world labeled examples. However, the shortness of the message and the resulting semantic ambiguity represent a critical limitation, thus making the task very challenging. Let us consider the following message between two users:

`Benji:` *@Holly sono completamente d'accordo con te*

The tweet sounds like to be a reply to the previous one. Notice how no lexical or syntactic property allows to determine the polarity. Let's look now at the entire conversation:

`Benji:` *@Holly con un #RigoreAl90 vinci facile!!*
`Holly:` *@Benji Lui vince sempre però :) accanto a chiunque.. Nessuno regge il confronto!*
`Benji:` *@Holly sono completamente d'accordo con te*

The first is clearly a positive tweet, followed by a positive one that makes the third positive as well. Thus, through the conversation we can disambiguate even a very short message. We want to leverage on this to define a context-sensitive SA model for the Italian language, in line with (Vanzo et al., 2014). The polarity detection of a tweet is modeled as a sequential classification task through the SVM^{hmm} learning algorithm (Altun et al., 2003), as it allows to classify an instance (i.e. a tweet) within an entire sequence. First experimental evaluations confirm the effectiveness of the proposed sequential tagging approach combined with the adopted contextual information even in the Italian language.

A survey of the existing approaches is presented in Section 2. Then, Section 3 provides an account of the context-based model. The experimental evaluation is presented in Section 4.

2 Related Work

The spread of microblog services, where users post real-time opinions about "everything", poses different challenges in Sentiment Analysis. Classical approaches (Pang et al., 2002; Pang and Lee, 2008) are not directly applicable to tweets: they focus on relatively large texts, e.g. movie or product reviews, while tweets are short and informal and a finer analysis is required. Recent works tried to model the sentiment in tweets (Go et al., 2009; Davidov et al., 2010; Bifet and Frank, 2010; Zanzotto et al., 2011; Croce and Basili, 2012; Si et al., 2013). Specific approaches, e.g. probabilistic paradigms (Pak and Paroubek, 2010) or Kernel based (Barbosa and Feng, 2010; Agarwal et al., 2011; Castellucci et al., 2013), and features, e.g. *n*-grams, POS tags, polarity lexicons, have been adopted in the tweet polarity recognition task.

In (Mukherjee and Bhattacharyya, 2012) contextual information, in terms of discourse relations is adopted, e.g. the presence of conditionals and semantic operators like *modals* and *negations*. However, these features are derived by considering a tweet in isolation. The approach in (Vanzo et al., 2014) considers a tweet within its context, i.e. the stream of related posts. In order to exploit this information, a Markovian extension of a Kernel-based categorization approach is there proposed and it is briefly described in the next section.

3 A Context Based Model for SA

As discussed in (Vanzo et al., 2014), contextual information about one tweet stems from various aspects: an explicit conversation, the user attitude or the overall set of recent tweets about a topic (for example a hashtag like *#RigoreAl90*). In this work, we concentrate our analysis only on the explicit conversation a tweet belongs to. In line with (Vanzo et al., 2014), a conversation is a sequence of tweets, each represented as vectors of features characterizing different semantic properties. The Sentiment Analysis task is thus modeled as a sequential classification function that associates tweets, i.e. vectors, to polarity classes.

3.1 Representing Tweets

The proposed representation makes use of different representations that allow to model different aspects within a Kernel-based paradigm.

Bag of Word (BoWK). The simplest Kernel function describes the lexical overlap between tweets, thus represented as a vector, whose dimensions correspond to the presence or not of a word. Even if very simple, the BoW model is one of the most informative representation in Sentiment Analysis, as emphasized since (Pang et al., 2002).

Lexical Semantic Kernel (LSK). In order to generalize the BoW model, we provide a further representation. A vector for each word is obtained from a co-occurrence Word Space built according to the Distributional Analysis technique (Sahlgren, 2006). A word-by-context matrix M is built through large scale corpus analysis and then processed through *Latent Semantic Analysis* (Landauer and Dumais, 1997). Dimensionality reduction is applied to M through Singular Value Decomposition (Golub and Kahan, 1965): the original statistical information about M is captured by the new k-dimensional space, which preserves the global structure while removing low-variance dimensions, i.e. distribution noise. A word can be projected in the reduced Word Space: the distance between vectors surrogates the notion of paradigmatic similarity between represented words, e.g. the most similar words of *vincere* are *perdere* and *partecipare*. A vector for each tweet is represented through the linear combination of its word vectors.

Whenever the different representations are available, we can combine the contribution of both vector simply through a juxtaposition, in order to exploit both lexical and semantic properties.

3.2 SA as a Sequential Tagging Problem

Contextual information is embodied by the stream of tweets in which a message t_i is immersed. A stream gives rise to a sequence on which sequence labeling can be applied: the target tweet is here labeled within the entire sequence, where contextual constraints are provided by the preceding tweets. Let formally define a conversational context.

Conversational context. For every tweet $t_i \in \mathcal{T}$, let $r(t_i) : \mathcal{T} \rightarrow \mathcal{T}$ be a function that returns either the tweet to which t_i is a reply to, or $null$ if t_i is not a reply. Then, the *conversational context* $\Lambda_i^{C,l}$ of tweet t_i (i.e., the *target tweet*) is the sequence of tweet iteratively built by applying $r(\cdot)$, until l tweets have been selected or $r(\cdot) = null$.

A markovian approach. The sentiment prediction of a target tweet can be seen as a sequential classification task over a context, and the SVM^{hmm} algorithm can be applied. Given an input sequence $\mathbf{x} = (x_1 \dots x_l) \subseteq \mathcal{X}$, where $\mathbf{x}$ is a

tweet context, i.e. the *conversational context* previously defined, and x_i is a feature vector representing a tweet, the model predicts a tag sequence $\mathbf{y} = (y_1 \dots y_l) \in \mathcal{Y}^+$ after learning a linear discriminant function $F : \mathcal{P}(\mathcal{X}) \times \mathcal{Y}^+ \rightarrow \mathbb{R}$ over input/output pairs. The labeling $f(\mathbf{x})$ is defined as: $f(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathcal{Y}^+} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$. In these models, F is linear in some combined feature representation of inputs and outputs $\Phi(\mathbf{x}, \mathbf{y})$, i.e. $F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle$. As Φ extracts meaningful properties from an observation/label sequence pair $(\mathbf{x}, \mathbf{y})$, in *SVM*^{hmm} it is modeled through two types of features: interactions between attributes of the observation vectors x_i and a specific label y_i (i.e. **emissions** of a tweet w.r.t. a polarity class) as well as interactions between neighboring labels y_i along the chain (i.e. **transitions** of polarity labels in a conversation context.). Thus, through *SVM*^{hmm} the label for a target tweet is made dependent on its context history. The markovian setting acquires patterns across tweet sequences to recognize sentiment even for truly ambiguous tweets. Further details about the modeling and the *SVM*^{hmm} application to tweet labeling can be found in (Vanzo et al., 2014).

4 Experimental Evaluation

The aim of the experiments is to verify the applicability of the model proposed in (Vanzo et al., 2014) in a different language, i.e. Italian. In order to evaluate the models discussed above in an Italian setting, an appropriate dataset has been built by gathering[1] tweets from Twitter servers. By means of Twitter APIs[2], we retrieved the whole corpus by querying several Italian hot topics, i.e. `expo`, `mose`, `renzi`, `prandelli`, `mondiali`, `balotelli` and commonly used emoticons, i.e. `:)` and `:(` smiles. Each tweet t_i and its corresponding conversation $\Lambda_i^{C,l}$ have been included into the dataset if and only if the conversation itself was available (i.e. $|\Lambda_i^{C,l}| > 1$). Then, three annotators labeled each tweet with a sentiment polarity label among `positive`, `negative`, `neutral` and `conflicting`[3], obtaining a inter-annotator agreement of 0.829, measured as the mean accuracy computed between annotators pairs.

[1] The process has been run during June-July 2014

[2] `http://twitter4j.org/`

[3] A tweet is said to be conflicting when it expresses both a positive and negative polarity

As about 1,436 tweets, including conversations, were gathered from Twitter, a static split of 64%/16%/20% in *Training/Held-out/Test* respectively, has been carried out as reported in Table 1.

	train	**dev**	**test**
Positive	212	61	69
Negative	211	42	92
Neutral	387	72	87
Conflicting	129	26	48
	939	201	296

Table 1: Dataset composition

Tweets have been analyzed through the *Chaos* natural language parser (Basili et al., 1998). A normalization step is previously applied to each message: fully capitalized words are converted in lowercase; reply marks, hyperlinks and hashtags are replaced with the pseudo-tokens, and emoticons have been classified with respect to 13 different classes. LSK vectors are obtained from a Word Space derived from a corpus of about 3 million tweets, downloaded during July and September 2013. The methodology described in (Sahlgren, 2006) with the setting discussed in (Croce and Previtali, 2010) has been applied.

Performance scores are reported in terms of Precision, Recall and F-Measure. We also report both the F_1^{pnn} score as the arithmetic mean between the F_1s of *positive*, *negative* and *neutral* classes, and the F_1^{pnnc} considering even the *conflicting* class. It is worth noticing that a slightly different setting w.r.t. (Vanzo et al., 2014) has been used. In this work we manually labeled every tweet in each conversation and performance measures considers all the tweets. On the contrary in (Vanzo et al., 2014) only the last tweet of the conversation is manually labeled and considered in the evaluation.

4.1 Experimental Results

Experiments are meant to verify the ability of a context-based model in the Italian setting. As a baseline we considered a multi-class classifier within the *SVM*^{multiclass} framework (Tsochantaridis et al., 2004). Each tweet in a conversation is classified considering it in isolation, i.e. without using contextual information. In Table 2, performances of the Italian dataset are reported, while Table 3 shows the outcomes of experiments over the English dataset (Vanzo et al., 2014). Here, *w/o conv* results refer to a baseline computed with the *SVM*^{multiclass} algorithm, while *w/ conv* results refer to the application of the model described in the

	Precision				Recall				$\mathbf{F_1}$				$\mathbf{F_1^{pnn}}$	$\mathbf{F_1^{pnnc}}$
	pos	*neg*	*neu*	*conf*	*pos*	*neg*	*neu*	*conf*	*pos*	*neg*	*neu*	*conf*		
BoWK														
w/o conv	.705	.417	.462	.214	.449	.109	.690	.438	.549	.172	.553	.288	.425	.390
w conv	.603	.580	.379	.375	.507	.435	.701	.063	.551	.497	.492	.107	.513	.412
BoWK+LSK														
w/o conv	.507	.638	.416	.000	.493	.402	.793	.000	.500	.493	.545	.000	.513	.385
w conv	.593	.560	.432	.368	.464	.457	.736	.146	.520	.503	.545	.209	.523	.444

Table 2: Evaluation results of the Italian setting.

	Precision			Recall			$\mathbf{F_1}$			$\mathbf{F_1^{pnn}}$
	pos	*neg*	*neu*	*pos*	*neg*	*neu*	*pos*	*neg*	*neu*	
BoWK										
w/o conv	.713	.496	.680	.649	.401	.770	.679	.444	.723	.615
w/ conv	.723	.511	.722	.695	.472	.762	.709	.491	.741	.647
BoWK+LSK										
w/o conv	.754	.595	.704	.674	.486	.804	.712	.535	.751	.666
w/ conv	.774	.554	.717	.682	.542	.791	.725	.548	.752	.675

Table 3: Evaluation results on the English language from (Vanzo et al., 2014)

previous sections with the *SVM*^{hmm} algorithm. In the last setting, the whole *conversational context* of each tweet is considered.

Firstly, all *w/o conv* models beneficiate by the lexical generalization provided by the Word Space in the LSA model. In fact, the information derived from the Word Space seems beneficial in its relative improvement with respect to the simple BoW Kernel accuracy, up to an improvement of 20.71% of $\mathbf{F_1^{pnn}}$, from .425 to .513. However, it is not always true, in particular w.r.t. the conflicting class where the smoothing provided by the generalization negatively impact on the classifiers, that are not able to discriminate the contemporary presence of positive and negative polarity.

Most importantly, the contribution of conversations is confirmed in all context-driven models, i.e. *w/conv* improves w.r.t. their *w/o conv* counterpart. Every polarity category benefits from the introduction of contexts, although many tweets annotated with the conflicting (*conf*) class are not correctly recognized: contextual information unbalances the output of a borderline tweet with the polarity of the conversations. The impact of conversational information contribute to a statistically significant improvement of 20.71% in the BoWK setting, and of 1.95% in the BoWK+LSK setting.

In (Vanzo et al., 2014) a larger dataset (10,045 examples) has been used for the evaluation of contextual models in an English setting. The dataset is provided by *ACL SemEval-2013* (Nakov et al., 2013). Results are thus not directly comparable, as in this latter dataset, where even tweets without a conversational contexts are included, only the target tweet is manually labeled and the labels of remaining tweets have been automatically predicted in a semi supervised fashion, as discussed in (Vanzo et al., 2014). Additionally, the conflicting class, where a lexical overlap is observed with both positive and negative classes, is not considered. However, results in Table 3 show that the BoWK setting benefits by the introduction of the lexical generalization, given by the LSK, with a performance improvement of 8.29%. When the focus is held within the same Kernel setting, in both BoWK and BoWK+LSK, the conversational information seems to be beneficial as increases of 5.20% and 1.35%, respectively, are observed.

5 Conclusions

In this work, the role of contextual information in supervised Sentiment Analysis over Twitter is investigated for the Italian language. Experimental results confirm the empirical findings presented in (Vanzo et al., 2014) for the English language. Although the size of the involved dataset is still limited, i.e. about 1,400 tweets, the importance of contextual information is emphasized within the considered markovian approach: it is able to take advantage of the dependencies that exist between different tweets in a conversation. The approach is also largely applicable as all experiments have been carried out without the use of any manual coded resource, but mainly exploiting unannotated material within the distributional method. A larger experiment, eventually on an oversized dataset, such as *SentiTUT*[4], will be carried out.

[4] http://www.di.unito.it/~tutreeb/sentiTUT.html

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Y. Altun, I. Tsochantaridis, and T. Hofmann. 2003. Hidden Markov support vector machines. In *Proceedings of the International Conference on Machine Learning*.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING (Posters)*, pages 36–44. Chinese Information Processing Society of China.

Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 1998. Efficient parsing for information extraction. In *Proc. of the European Conference on Artificial Intelligence*, pages 135–139.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, DS'10, pages 1–15, Berlin, Heidelberg. Springer-Verlag.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2013. Unitor: Combining syntactic and semantic kernels for twitter sentiment analysis. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 369–374, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Danilo Croce and Roberto Basili. 2012. Grammatical feature engineering for fine-grained ir tasks. In Giambattista Amati, Claudio Carpineto, and Giovanni Semeraro, editors, *IIR*, volume 835 of *CEUR Workshop Proceedings*, pages 133–143. CEUR-WS.org.

Danilo Croce and Daniele Previtali. 2010. Manifold learning for the semi-supervised induction of framenet predicates: An empirical investigation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '10, pages 7–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING (Posters)*, pages 241–249. Chinese Information Processing Society of China.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

G. Golub and W. Kahan. 1965. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics*, 2(2).

T. Landauer and S. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment analysis in twitter with lightweight discourse analysis. In *Proceedings of COLING*, pages 1847–1864.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 104–, New York, NY, USA. ACM.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 2345–2354, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Fabio M. Zanzotto, Marco Pennaccchiotti, and Kostas Tsioutsiouliklis. 2011. Linguistic Redundancy in Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

10.12871/CLICIT2014174

Semantic role annotation of instrument subjects

Rossella Varvara
Università di Trento
rossella.varvara@unitn.it

Elisabetta Ježek
Università di Pavia
jezek@unipv.it

Abstract

English. Semantic role annotation has become widely used in NLP and lexical resource implementation. Even if attempts of standardization are being developed, discordance points are still present. In this paper we consider a problematic semantic role, the Instrument role, which presents differences in definition and causes problems of attribution. Particularly, it is not clear whether to assign this role to inanimate entities occurring as subjects or not. This problem is especially relevant 1- because of its treatment in practical annotation and semantic role labeling, 2- because it affects the whole definition of semantic roles. We propose arguments to sustain that inanimate nouns denoting instruments in subject positions are not instantiations of Instrument role, but are Cause, Agent or Theme. Ambiguities in the annotation of these cases are due to confusion between semantic roles and ontological types.

Italiano. *L'annotazione dei ruoli semantici è ormai molto utilizzata nell'ambito del NLP e della creazione di risorse lessicali. Sebbene si stia cercando uno standard condiviso, vi sono ancora punti di disaccordo. Nel presente articolo si considera un problematico ruolo semantico, il ruolo di Strumento, il quale causa ambiguità nell'annotazione e nella sua definizione. In particolare, tra i ricercatori non vi è ancora accordo nell'assegnare questo ruolo a casi di entità inanimate in posizione soggetto. Tale questione è certamente significativa 1- in ragione dell'annotazione pratica di questi casi 2- in quanto interessa la definizione generale di ruoli semantici. Sosteniamo che nomi di entità strumentali in posizione soggetto non sono casi del ruolo Strumento, ma dei ruoli di Causa, Agente o Tema. Ambiguità nella loro annotazione sono dovute alla confusione tra ruoli semantici e tipi ontologici.*

1 Background

Semantically annotated resources have become widely used and requested in the field of Natural Language Processing, growing as a productive research area. This trend can be confirmed by looking at the repeated attempts in the implementation of annotated resources (FrameNet, VerbNet, Propbank, SALSA, LIRICS, SensoComune) and in the task of automatic Semantic Role Labeling (Gildea and Jurafsky 2002, Surdeanu et al. 2007, Màrquez et al. 2008, Lang and Lapata 2010, Titov and Klementiev 2012 among others).

Since their first introduction by Fillmore (1967), semantic roles have been described and defined in many different ways, with different sets and different level of granularity - from macro-roles (Dowty 1991) to frame-specific ones (Fillmore et al. 2002). In order to reach a common standard of number and definition, the LIRICS (Linguistic Infrastructure for Interoperable ResourCes and Systems) project has recently evaluated several approaches for semantic role annotation and proposed an ISO (International Organization for Standardization) ratified standard that enables the exchange and reuse of (multilingual) language resources.

In this paper we examine some problematic issues in semantic role attribution. We will highlight a case, the Instrument role, whose definition and designation should be, in our opinion, reconsidered. The topic is particularly relevant since there is difference in its treatment in different lexical resources and since the theoretical debate is

still lively. Moreover, this matter highlights aspects of the nature of semantic roles, relevant both for their theoretical definition and for practical annotation, such as the difference between semantic roles and ontological types. The former refer to the role of participants in the particular event described by the linguistic utterance, the latter to the inherent properties of the entity. We argue that this is a main point in the annotation, because, even in the recent past, roles have been frequently tagged according to the internal properties of the entities involved, not, as it should be, because of their role in the particular event described.

This analysis arose from the first step of the implementation of the Senso Comune resource (Vetere et al. 2012). With the aim to provide it with semantic roles, a first annotation experiment was conducted to check the reliability of the set and the annotation procedure (Ježek et al. 2014). The dataset was composed of 66 examples without disambiguation, 3 for 22 target verbs, and it was annotated for semantic roles by 8 annotators. They were instructed with a guideline in which a set of 24 coarse-grained roles was defined, with examples and a taxonomy. During the evaluation process, the major cases of disagreement were highlighted. The present study is based on the evidence coming from these data: the Instrument role caused several misunderstandings (see also Varvara 2013). Nevertheless, our analysis will look primarily at examples from literature and other resources in order to rethink this role and to reach a standardization. We propose to consider what are called instrument subjects (Alexiadou and Schäfer 2006) as instances of three different roles (Cause, Agent and Theme) rather than as Instrument.

2 The case of instrument subjects

With "instrument subjects" we refer to examples in which a noun, denoting an inanimate entity frequently used as instrument by humans (and occurring in *with*-phrases), is the subject of the sentence, as in the examples below (Levin 1993:80, Schlesinger 1989:189): "**The hammer** broke the window", "**The stick** hit the horse". In the past, it has been frequently asserted that these subjects cover the role of Instrument (Fillmore 1967, Nilsen 1973, Dowty 1991), as much as the nouns preceded by the preposition *with*: "David broke the window **with a hammer**", "Marvin hit the horse **with a stick**". In Levin (1993)'s terms, these are called "Instrument-Subject alternation"[1]. On the other side, several authors have argued against this interpretation, suggesting other roles to these cases (Schlesinger 1989, DeLancey 1991, Van Valin and Wilkins 1996, Alexiadou and Schäfer 2006, Grimm 2013, among others). Although this interpretation is the most recent one and many scholars agree on that, in the implementation of lexical resources the trend is to consider instrument subjects as Instrument role. In Verbnet, instrument subjects are tagged with the role Instrument, as can be seen in the annotation of the verb *hit*: "**The stick** hit the fence"; "**The hammer** hit the window to pieces"; "**The stick** hit the door open". In the LIRICS guidelines (Schiffrin and Bunt 2007:38) the Instrument-Subject alternation is used as exemplification of the role definition: "He opened the door [with the key (Instrument)]"; "[The brick (Instrument)] hit the window and shattered it." The reason of the annotation of these last examples is not clear if we look at the role definition (as annotators usually do). It is said that the Instrument is the "participant in an event that is manipulated by an agent, and with which an intentional act is performed" (2007:38). Here, the agent and the intentionality of the act are explicitly mentioned, but while annotating the examples above a question arises: in order to tag a noun with the role Instrument, should the Agent be present in the context of the event in which the Instrument occurs, should it be inferable or could it be totally absent? We argue that, in order to assign the Instrument role, an Agent should be specified in the event representation and it should be linguistically expressed. From our data, it seems that, in presence of instrument subjects, there is not an Agent, neither expressed neither included inferentially in the scene. In the cases observed, it is clear that there are reasons for which speakers left the intentional Agent out of the scope of their utterance. Their intention could be to describe the instrument

[1]The traditional examples of "instrument subjects" cover also other Levin's alternations, such as Characteristic property alternation (1993:39) or Middle alternation (1993:26). Even the examples that will be a matter of discussion in the present study are ascribable to different alternations. We will then consider the term "instrument subject" in a broad way, taking into account every noun that can occurs both in a *with*-phrase, both in subject position. Even if this term may bring confusion with the real semantic role Instrument, we will adopt it because of a lack of other appropriate terms. To avoid difficulties, we will use the capital initial letter for semantic roles and the lower initial for the words in their common sense (e.g. Agent vs agent).

noun as an autonomous entity, as the only known source of causation, not as an Instrument manipulated by an Agent, and as such its role in the event should be considered. In the next section, we will list and group in classes the occurrences of instrument subjects that we have encountered so far, according to our proposal.

3 Why instrument subjects do not perform the Instrument role

Nowadays it is a shared opinion that semantic roles are relational notions that express the role of participants in reference to the event expressed by the verb. As pointed by Pethukova and Bunt (2008), semantic roles should be defined not as primitives "but rather as relational notions that link participants to an event, and describe the way the participant is involved in an event, rather than by internal properties"(2008:40). From this statement, we argue that semantic roles should be considered as semantic qualities attributed to a participant not only in a particular event, but in the specific linguistic representation of that event. The same event can be the object of two different sentences that represent the event from different perspectives. In the words of DeLancey (1991:350): "case roles, like any other semantic categories, encode construals of events rather than objective facts". This is the mistake that we make when we evaluate an instrument subject as Instrument role. Consider the examples "The janitor opened the lock with a key" and "The key opened the lock". "The underlying argument is that since "*the key*" in 19 (the first example) is an Instrument, and since 19 and 20 could refer to the same scenario, "*the key*" must be Instrument in 20 (the second example) as well" (DeLancey 1991:348). Actually, examples like the second one are often not realistic, invented by linguists. We believe that, looking at corpus data, it appears clearly that subjects like "the key" are not usually represented as an instrument used by an human, but as a Cause that substitutes an unknown Agent in the causal chain (as in the previous example) or as an entity whose a characteristic is described (e.g. the property of opening a lock in an example such as "This key opens the lock"). As referenced in the Introduction, our idea is that instrument subjects usually cover the role of Cause, Theme or, metaphorically, Agent.

3.1 Instrument subjects as Cause

Most frequently instrument subjects cover the role of Cause. It is usually the case when: 1- it is not possible to find another Agent or general causer other than the instrument inanimate subject; 2- it is possible to imagine an Agent that has "activated" the inanimate entity, but it is no longer present in the scene or it is not known. This could be a choice of the speaker that does not want to include or talk about the Agent or it could be the case with generic events with non specific agents. Consider the example "The clock was ticking so loudly that it woke the baby"(DeLancey 1991: 347): it is not possible to find another Agent other than the clock. The same can be seen in this sentence taken from the corpus ItTenTen (Jakubček et al. 2013): "Un masso caduto da una galleria ha messo fuori uso la metro. Il sasso ha rotto il pantografo, l'antenna che trasmette l'energia al treno, e ha interrotto la tensione per 600 metri di linea aerea" ("A stone falling down from a tunnel put out of order the metro. The stone has broken the pantograph, the spar that transmits the energy to the train, and it has interrupted the tension for 600 meters"). The stone is a Cause[2] because nobody has thrown it, but it has taken its own energy by its falling[3]. The same interpretation could be applicable to the sentence cited before from the LIRICS guidelines "The brick hit the window and shattered it": from this context we do not know if there is an agent that has thrown the brick; if we do not have evidence about that, we cannot consider "the brick" an Instrument in this sentence. There are cases in which our real-world knowledge enables us to understand that the instrument subject has been manipulated by somebody, but it has been focused in the sentence as the principal or the only known element of the causal chain[4]: "The poison killed its victim", "The camomile

[2] The definition of the role Cause in SensoComune is "participant (animate or inanimate) in an event that starts the event, but does not act intenctionally; it exists indipendently from the event".

[3] A reviewer pointed out that the real Cause is the event of falling, not the stone. Although this is a true inference, we argue that the stone is metonymically reinterpreted as the falling of the stone and for this reason the cause of the event. This interesting matter deserves a deeper analysis that will be subject of further work.

[4] Alexiadou and Schäfer note: "They are Causers by virtue of their being involved in an event without being (permanently) controlled by a human Agent. The fact that this involvement in an event might be the result of a human agent having introduced these Causers is a fact about the real world, not about the linguistic structure" (2006: 42-43).

cured the patient". There is a case of this sort in the dataset of the SensoComune's annotation experiment. The subject of the sentence "leggi che colpiscono il contrabbando" ("Laws that hit the smuggling") has been tagged by 2 annotators upon 8 as Instrument role instantiaton: it is possible that they have thought that there was an inferred Agent (the legislator) that was using the laws as an instrument. Putting instruments as subjects can be seen as a stylistic means adopted by the speaker to "defocus" the Agent: "ricorda teste sbattute contro il muro, saluto romano, ustioni con sigaretta e accendino. Un'altra le minacce mentre **le forbici** tagliavano ciocche di capelli" ('she remembers heads hit against the wall, cigarette and lighter burns. Another the threats while the scissors cut locks of hair"). Lastly, instrument subjects can be Cause if the sentence expresses a generic event with a non-specific agent: 'The piano addressed this by a mechanism where the way the key is struck changes the speed with which the **hammer** hits the string".

3.2 Instrument subjects as Agent

We argue that the cases in which an instrument subject covers the role of Agent are sporadic and involve metaphorical or metonymical interpretations (Jezek et al. 2014). It should be kept in mind that it is widely assumed that the Agent role implies animacy and intentionality; as such an inanimate entity like an instrument noun cannot be Agent. This view contrasts with what has been claimed by some linguists (Schlesinger 1989, Alexiadou and Schäfer 2006) that were arguing anyway against the Instrument role attribution to instrument subjects. The Agent role can be fulfilled by instrument subjects in case of personification or metaphorical extension of the meaning of the lexeme: "Un giorno una forbice gigante tagliò della carta a forma di burattino. Un altro giorno ha ritagliato due palle giganti che erano il sole giallo e la Terra" ("Once upon a time a giant scissor cut a paper into a puppet. Later, it cut two giant balls, the yellow sun and the Earth"); "Tante penne scrivono su Napoli, usano Napoli per vendere copie" ("A lot of pens (writers) write about Naples, they use Naples to sell"); "Tutto l'ufficio ha lavorato bene" ("All the office has worked well").

3.3 Instrument subjects as Theme

Analyzing the SensoComune dataset, a case has been found that has not been previously discussed in the literature on semantic roles. The examples to which we refer are: "**La penna** scrive nero" ("The pen writes black"), "**Forbici** che tagliano bene" ("Scissors that cut well"). These subjects have been tagged as Instrument by respectively 3/8 and 4/8 annotators. As previously claimed, the ambiguity is caused by the possibility of these nouns to occur as real Instrument with the preposition "*with*" (ex. "I have written the letter with this pen"). We suggest that in these cases the instrument subjects are neither Instrument, nor Cause, because they are not presented as causing an event or as being used by an Agent. The verb predicates a property of the subject and as such the Theme role is fulfilled. The Theme is defined in SensoComune as "participant in an event or state, which, if in an event, it is essential for it to take place, but it does not determine the way in which the event happens (it doesn't have control) and it is not structurally modified by it; if in a state, it is characterized by being in a certain condition or position throughout the state and it is essential to its occurring". In other resources, these examples could be referred to roles similar to our Theme, such as the role Pivot in LIRICS.

4 Conclusion

In this paper we have shown how theoretical and data analysis can be mutually improved by each other. Literature has offered critical discussion about the Instrument role and the case of instrument subjects, a discussion that can be useful for the definition and annotation of semantic roles in the implementation of lexical resources. Moreover, analysis of annotated data can reveal fallacies in the reliability of the set, coming back from application to theoretical topics. At last, our study highlights the importance of distinguishing between semantic roles - relational notions belonging to the level of linguistic representation - and ontological types, which refer to internal qualities of real-world entities. We believe that this topic, because of its importance, should be taken into consideration for a more complete treatment in future work.

Acknowledgments

Thanks to Guido Vetere, Laure Vieu, Fabio Zanzotto, Alessandro Oltramari and Aldo Gangemi for the ongoing collaborative work on semantic role annotation within the Senso Comune initiative, and to the participants of the 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation at LREC 2014 for feedback on early results of this work. We also acknowledge two anonymous reviewers for their very useful comments and suggestions.

References

Alexiadou, A. and F. Schäfer. 2006. Instrument subjects are agents or causers. *Proceedings of West Coast Conference on Formal Linguistics*, vol.25

Burchardt, A., E. Katrin, A. Frank, A. Kowalski, S.Padó, M. Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. *Proceedings of LREC 2006*

DeLancey, S. 1991. Event construal and case role assignment. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, vol.17

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language*, 126: 547-619

Fillmore, C.J. 1967. The case for case. *Universals in Linguistic Theory*, Bach and Harms (eds). New York, Holt, Rinehart and Winston edition.

Fillmore, C. J., C. F. Baker and H. Sato. 2002. The framenet database and software tools. *Proceedings of the Third International Conference on Language Resources and Evaluation*, vol.4

Gildea, D. and D. Jurafsky 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245288.

Grimm, S. 2013. The Bounds of Subjecthood: Evidence from Instruments. *Proceedings of the 33rd Meeting of the Berkeley Linguistic Society.*,Berkeley Linguistic Society.

Jakubiček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. 2013. The TenTen Corpus Family *Proceedings of the International Conference on Corpus Linguistics.*

Ježek, E., Vieu L., Zanzotto F.M., Vetere G., Oltramari A., Gangemi A., Varvara R. 2014. Extending 'Senso Comune' with Semantic Role Sets. *Proceedings 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, LREC 2014.

Lang, J., and Lapata, M. 2010. Unsupervised induction of semantic roles. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 939-947

Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press Chicago, IL.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. 2008. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2), 145-159.

Nilsen, Don L. F. 1973. *The instrumental case in english: syntactic and semantic considerations.*. The Hague; Paris: Mouton.

Petukhova, V. and Bunt, H.C. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2830.

Schiffrin, A. and Bunt, H.C. 2007. LIRICS Deliverable D4.3. Documented compilation of semantic data categories. *http://lirics.loria.fr.*.

Schlesinger, I.M. 1989. Instruments as agents: on the nature of semantic relations. *Journal of Linguistics*, 25(01). 189210.

Surdeanu, M., L. Màrquez, X. Carreras, and P. R. Comas. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research (JAIR)*, 29:105151.

Titov, I., and Klementiev, A. 2012. A Bayesian approach to unsupervised semantic role induction. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp.12-22. Association for Computational Linguistics.

Van Valin, R. D. and D. P. Wilkins. 1996. The case for "effector": case roles, agents, and agency revisited. *Grammatical constructions: Their form and meaning.*, eds. Shibatani and Thompson, 289322. Oxford: Oxford University Press.

Varvara, R. 2013. *I ruoli tematici: Agente, Strumento e la nozione di causa.* Master thesis. University of Pavia

Vetere, G., A. Oltramari, I. Chiari, E. Jezek, L. Vieu, F.M. Zanzotto. 2012. 'Senso Comune': An Open Knowledge Base for Italian. *Revue TAL (Traitement Automatique des Langues), Journal Special Issue on Free Language Resources*, 52.3, 217-43.

10.12871/CLICIT2014175

The Italian Module for NooJ

Simonetta Vietri
Department of Political, Social and
Communication Sciences
University of Salerno, Italy
vietri@unisa.it

Abstract

English. This paper presents the Italian module for NooJ. First, we will show the basic linguistic resources: dictionaries, inflectional and derivational grammars, syntactic grammars. Secondly, we will show some results of the application of such linguistic resources: the annotation of date/time patterns, the processing of idioms, the extraction and the annotation of transfer predicates.

Italiano. *In questo articolo si presenta il modulo italiano per NooJ. In un primo momento si descrivono le risorse lessicali di base: i dizionari, le grammatiche flessive, derivazionali e sintattiche. Si presentano poi i risultati relativi all'applicazione di tali risorse: l'annotazione dei pattern temporali, il parsing delle frasi idiomatiche, l'estrazione e l'annotazione dei predicati di trasferimento.*

1 Introduction

NooJ is a development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to corpora, in real time. NooJ, whose author is Max Silberztein (Silberztein 2003-), is a knowledge-based system that makes use of huge linguistic resources.

Dictionaries, combined with morpho-syntactic grammars, are the basic linguistic resources without which it would be impossible to perform a text analysis. The system includes various modules for more than twenty languages, among them Italian (.nooj4nlp.net). Most of the Italian linguistic resources are completely new.

The goal of the NooJ project is twofold: to provide tools allowing linguists to implement exhaustive descriptions of languages, and to design a system which processes texts in natural language (see Silberztein 2014).

NooJ consists of higher and higher linguistics levels: tokenization, morphological analysis, disambiguation, named entity recognition, syntactic parsing[1].

Unlike other systems, for example TreeTagger, developed by Helmut Schmidt (1995)[2], NooJ is not a tagger, but the user can freely build disambiguation grammars and apply them to texts.

Section 2 describes the Italian dictionary and the inflectional/derivational grammars associated with it. Section 3 shows the extraction of date/time patterns, section 4 the parsing of idioms. Section 5 describes the XML annotation and extraction of transfer predicates.

2 The dictionaries and the inflectional grammars

The first Machine Italian dictionary was built at the Institute for Computational Linguistics, C.N.R, directed by Antonio Zampolli (see Bortolini et al (1971), Gruppo di Pisa (1979)). More than a decade later a group of researchers of the Linguistics Institute at the University of Salerno, directed by Annibale Elia, started to implement an electronic Italian dictionary on the principles of the Lexicon-Grammar framework (Gross 1968, 1979, Elia et al 1981)[3].

More recently Baroni and Zanchetta (2005) developed *Morph-it!*, that contains more than 505,000 entries and about 35,000 lemmas [4].

[1] See textpro.fbk.eu/docs.html for **TextPro**, an NLP system implemented at FBK. It is a suite of modules performing various tasks. **Unitex** is a system developed by Sébastien Paumier, see igm.univ-mlv.fr/~unitex/index.php?page=1.

[2] See cis.uni-muenchen.de/~schmid/tools/TreeTagger/ and elearning.unistrapg.it/TreeTaggerWeb/TreeTagger.html. See also the Venice Italian Treebank (**VIT**), the Turin University Treebank (**TUT**), the Italian Syntactic Semantic Treebank (**ISST**).

[3] For the literature on Lexicon-Grammar, see infolingu.univ-mlv.fr/english/. A very first version of the Italian dictionary was built for Intex. See De Bueriis and Montemagno (1995).

[4] See dev.sslmit.unibo.it/linguistics/morph-it.php. As concerns the corpus utilized, see Baroni et al. (2004).

The Italian dictionary of simple words (S_dic) for NooJ contains 129,000+ lemmas, whereas the dictionary of compounds includes 127,000+ nouns and 2,900+ adverbs Elia (1995). Furthermore, the Italian module consists of a number of satellite dictionaries including toponyms (1,000+), first and last names (2,000+), acronyms (200+). Some dictionaries are richer than others which are still under construction. The canonical forms of dictionary entries, either simple or compound, are of the following type:

```
americano,A+FLX=N88
il,DET+FLX=D301
su,PREP
surfista,N+FLX=N70
tavola a vela,N+FLX=C41
tavola,N+FLX=N41
volare,V+FLX=V3
```

Each entry is associated to an alphanumeric code that refers to an inflectional grammar, as the following example[5]:

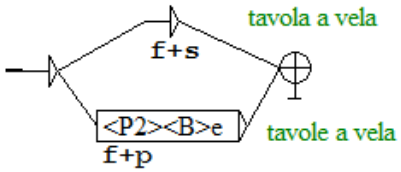


Fig. 1. Sample of an Inflectional Grammar

On the basis of the entries and the inflectional codes, NooJ generates the dictionaries of inflected forms (more than one million of simple forms, and 260.000+ of noun compounds) in a few seconds. By applying these resources, NooJ will annotate a sentence such as *Le surfiste volavano sulle tavole a vela* as follows:

```
le,il,DET+f+p
le,PRON
surfiste,surfista,N+f+p+Um
volavano,volare,V+IM+3+p+i+a+e
su,PREP
tavole,tavola,N+f+p
tavole a vela,tavola a vela,N+f+p
a,PREP
vela,N+f+s
vela,velare,V+PR+3+s+t
```

Each form is associated with morpho-syntactic information. Since NooJ is not a tagger, the annotations show the ambiguities (unless the user applies disambiguation grammars)[6]. For example, *vela* may be not only a feminine (+f) singular (+s) noun (N), but also the Present Indicative (+PR) form of the transitive (+t) verb *volare*, in the 3rd person (+3), singular (+s).

[5] For the FSA/FST grammars, see Silberztein (2003-).
[6] For reason of space, some annotations are not shown.

2.1 Proper Names and derivation

The dictionary of proper names is built according to the same criteria used for the main dictionary. Although proper names do not inflect, they are linked to derived forms. Such forms like *renzismo*, *antirenziano*, *renzista* are relatively new and are not included in the S_dic. The dictionary of proper names and a derivational grammar associated with it allow NooJ to annotate these very productive forms, as in the following:

```
renzismo,Matteo Renzi,N+Npr...
antirenziano,Matteo Renzi,A+Npr
```

2.2 The Annotation of Pronominal forms

Italian is particularly rich of agglutinated forms such as *vederti*, *mandandogliela*, *dimmi*, *compratala*, etc. which are constituted of a verb (infinitive, gerund, imperative, past participle) and one or more clitics. Although these forms are formally single words, they are analyzed by means of a morpho-syntactic grammar which separates the verb form from the pronoun. Therefore, the forms above will be annotated as follows:

```
vedere,V+t+a+INF
ti,PRON+Persona=2+s
mandando,mandare,V+G
gli,PRON+Persona=3+m+s
la,PRON+Persona=3+f+s
dì,dire,V+IMP+2+s+t+a
mi,PRON+Persona=1+s
comprata,comprare,V+PP+f+s
la,PRON+Persona=3+f+s
```

3 The extraction of date/time patterns

Among the syntactic resources, the Italian module includes a grammar for the extraction and annotation of date and time sequences. It's a complex net of local grammars which, applied to a text of 1MB (129,000+ word forms), extracts and annotates sequences like the following:

Nell'arco di tre mesi/<DURATA>
fino al giugno 2006/<DURATA>
intorno alle 23,20/<DATA>
Dal 1987 al 2004/<DURATA>
Per la fine di gennaio/<DATA>
Nel novembre del 2001/<DATA>
Il 18 e 19 dicembre scorsi/<DATA>
un mese dopo/<DATA>
in due giorni/<DURATA>
dieci anni fa/<DATA>
il prossimo 9 gennaio/<DATA>
dal 18 al 21 gennaio prossimi/<DURATA>
per 30 mesi/<DURATA>
nell'ottobre del 2004/<DATA>
fino al dicembre 2005/<DURATA>

4 The Annotation of Idioms

The formal representation and processing of idioms has always been a very debated issue (Abeillé 1995, Sag et al 2001, Fothergill et al 2012). In the NooJ dictionaries, Italian idioms (Vietri 2014a, 2014c) are represented as strings formed by a verb that requires one or more fixed elements as in the following (simplified) example:

```
alzare,V+C1+FLX=V3+DET=<il,DET+m+s>
+N=<gomito,N+m+s>
```

The verb *alzare* is associated with the determiner *il* and the fixed noun *gomito*. The idiom *alzare il gomito* ('lift one's elbow') belongs to class **C1** (+C1), the verb inflects (+FLX) according to the code **V3**, and the **DET**erminer has to be masculine singular (+m+s) because the noun *gomito* is obligatory masculine singular. NooJ is an "open" system, and the user can choose to assign a property like +Passive only to those idioms that accept this construction. In such a case, the property ±Passive can be recalled in the grammar which is associated with the dictionary of idioms.

The dictionary is associated with a grammar, since the fixed lexical elements have to be linked to each other. Figure 2 shows a simplified example of grammar where the variable (indicated by the rounded parentheses) containing the verb is directly linked to the determiner (**V$DET**) and to the noun (**V$N**). This formalism keeps the fixed elements linked together also in case of modifiers or adverbs insertion, or in case of discontinuous idioms such as *prendere qc. per la gola*.

The dictionary/grammar pair, whose formalism is explained in details in Silberztein (2012), allows NooJ to automatically annotate sequences like *alzare il gomito*. Since this construction is ambiguous, NooJ produces both the idiomatic annotation, signaled by the little curve, and the literal one, as shown in Figure 3.

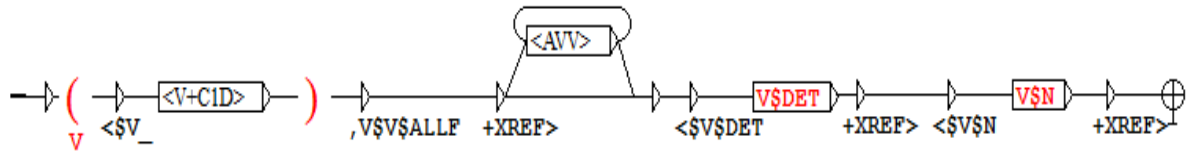


Fig. 2. The 'Active' Grammar

Maria alzò il gomito

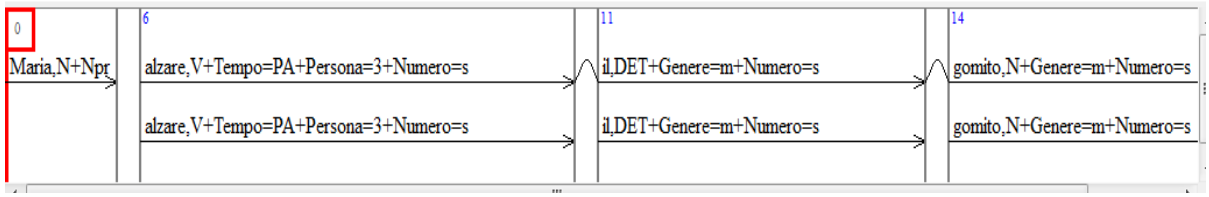


Fig. 3. Text Annotation

4.1 Parsing Idioms

Once NooJ has annotated idioms, it is possible to syntactically parse the sentence in question by applying an appropriate syntactic grammar. However, a sentence such as *Maria alzò il gomito* is ambiguous, therefore it has to be assigned a double representation. The representations in Figures 4 and 5 are flat trees which can be (re)designed according to the user's choice. Figure 4 represents the idiomatic construction: the blue boxes indicate that the lexical entries are linked.

The tree in Figure 5 represents instead the non-idiomatic construction, where the lexical entries are not linked.

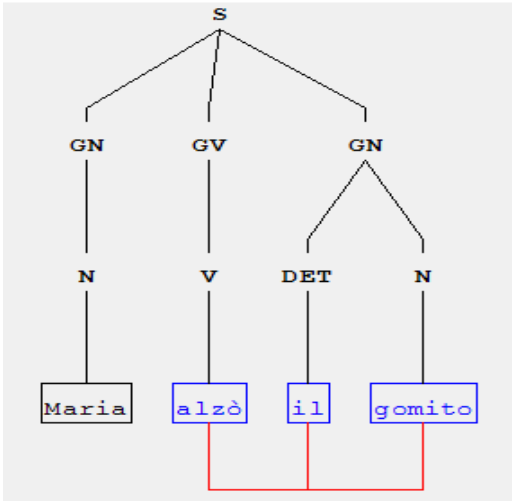


Fig. 4. Idiomatic Representation

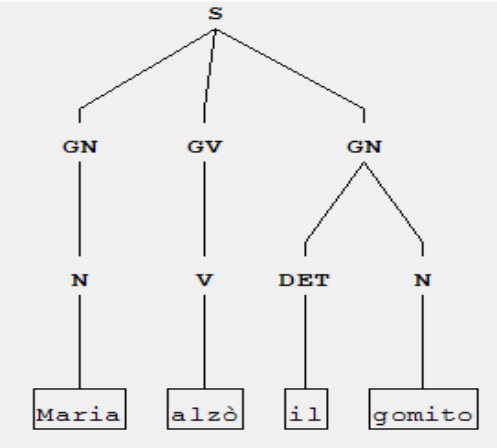


Fig. 5. Non-idiomatic Representation

Furthermore, the user can freely decide to assign only the idiomatic representation by means of the property +UNAMB.

5 Annotation of Transfer Predicates

The annotation of the Predicate-Argument structure of Transfer Predicates is described in details in Vietri (2014b). In the following examples, the transfer predicate is *consegnare* (= to deliver) in (1), *effettuare la consegna* (= make delivery) in (2), and *consegna* (= delivery) in (3):

(1) *Il fornitore consegna la merce al cliente*
The supplier delivers the goods to the customer

(2) *Il fornitore effettua la consegna della merce al cliente*
The supplier makes delivery of the goods to the customer

(3) *La consegna della merce al cliente dal fornitore*
The delivery of the goods to the customer by the supplier

They are all transfer predicates with three arguments: the Giver (*il fornitore* = the supplier), the Receiver (*il cliente* = the customer), and the Object (*la merce* = the goods) that is transferred from the Giver to the Receiver. Therefore, the Predicate-Argument structure is a function of the type **T (Giver, Object, Receiver)**. NooJ can build a concordance and annotate sequences such as (1)-(3), according to their Transfer Predicate-Argument Structure. This can be done by applying to a text/corpus a complex grammar that contains more than 70 sub-graphs. The annotated text can be exported as an XML document. Here is the XML text referring to the examples (1)-(3):

```
<G> Il fornitore </G> <T> consegna </T>
<O> la merce </O> al <R> cliente </R>,
ma prima di <T> effettuare la consegna
<\T> della <O> merce <\O> ...
<T> La consegna </T> della <O> merce
</O> al <R> cliente </R>.
```

The Transfer Grammar applied to the Italian Civil and Commercial Codes produce more than 2,600 occurrences. The most frequent Predicate-Argument structure is formed of the Transfer predicate **T** and the Object **O** (1,200 occurrences), immediately followed by the passive constructions where the Object **O** precedes the predicate **T** (387 occurrences).[7]

6 Conclusion

The application of the Italian module to a corpus of 100MB (La Stampa 1998) produced the following results: 33,866.028 tokens, 26,785.331 word forms. The unknown tokens are loan words, typos, acronyms, alterates[8].

The Italian module consists of exhaustive dictionaries/grammars formally coded and manually built on those distributional and morpho-syntactic principles as defined within the Lexicon-Grammar framework. Such a lingware (a) constitutes an invaluable linguistic resource because of the linguistic precision and complexity of dictionaries/grammars, (b) can be exploited by the symbolic as well as the hybrid approach to Natural Language Processing. The linguistic approach to NLP still constitutes a valid alternative to the statistical method that requires the (not always reliable) annotation of large corpora. If the annotated data contain errors, those systems based on them will produce inaccurate results. Moreover, corpora are never exhaustive descriptions of any language.

On the other hand, formalized dictionaries/grammars can be enriched, corrected and maintained very easily. Silberztein (2014) contains a detailed discussion on the limits, errors and naïveté of the statistical approach to NLP. The Italian module for NooJ constituted the basis of several research projects such as Elia et al. (2013), Monti et al. (2013), di Buono et al. (2014), Maisto et al. (2014). Therefore, it has been tested, verified and validated. The results constitute the basis for the updating of the module itself. Ultimately, the lexical resources of the Italian module can be easily exported into any format usable by other systems.

[7] In a different perspective, the *Lexit* project, directed by Alessandro Lenci, explores the distributional/semantic profiles of Italian nouns, verbs, and adjectives.
[8] The grammar that annotates alterates is under construction.

References

Anne Abeillé. 1995. The Flexibility of French Idioms: a Representation with Lexicalized Tree Adjoining Grammar. In M. Everaert, E-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: structural and psychological perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates: 15-41.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian, in *Proceedings of the Fourth Language Resources and Evaluation Conference*, (Lisbon: ELDA): 1771-1774.

Ugo Bortolini, Carlo Tagliavini, and Antonio Zampolli. 1971. *Lessico di Frequenza della Lingua Italiana*. Milano: Garzanti.

Giustino De Bueriis and Mario Monteleone. 1995. *Dizionario elettronico DELAS_I - DELAF_I ver. 1.0*, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno.

Maria Pia di Buono, Mario Monteleone, and Annibale Elia. 2014. How to populate ontology. Computational linguistics applied to the Cultural Heritage Domain. In E. Métais, M. Roche, and M. Teisseire (Eds.): *NLDB 2014 - 19th International Conference on Application of Natural Language to Information Systems*, 18-20 June 2014 - Montpellier, France: 55-58.

Annibale Elia, Daniela Guglielmo, Alessandro Maisto, and Serena Pelosi. 2013. A Linguistic-Based Method for Automatically Extracting Spatial Relations from Large Non-Structured Data. In *Algorithms and Architectures for Parallel Processing*. Springer International Publishing: 193-200.

Annibale Elia, Maurizio Martinelli, and Emilio D'Agostino. 1981. *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano*, Napoli: Liguori.

Annibale Elia. 1995. Chiaro e tondo, in *Tra sintassi e semantica. Descrizioni e metodi di elaborazione automatica della lingua d'uso*, E. D'Agostino (ed.), ESI: Salerno.

Richard Fothergill and Timothy Baldwin. 2012. Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*: 100-104.

Maurice Gross. 1968. *Syntaxe du verbe*. Paris: Larousse.

Maurice Gross. 1975. *Mèthodes en syntaxe*. Paris: Hermann.

Gruppo di Pisa. 1979. Il dizionario di macchina dell'italiano. In Daniele Gambarara, Franco Lo Piparo, Giulianella Ruggiero (eds), *Linguaggi e formalizzazioni*, Atti del Convegno internazionale di studi, Catania, 17-19 settembre 1976. Bulzoni, Roma: 683-707.

Alessandro Maisto and Serena Pelosi. 2014. A Lexicon-Based Approach to Sentiment Analysis. The Italian Module for Nooj. *Proceedings of the International Nooj 2014 Conference*, University of Sassari, Italy (forthcoming).

Johanna Monti, Mario Monteleone, Maria Pia di Buono, and Federica Marano. 2013. Natural Language Processing and Big Data. An Ontology-Based Approach for Cross-Lingual Information Retrieval**.** *Proceedings of the Social Computing (SocialCom)* - 2013 ASE/IEEE International Conference: 725-731.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing*. Berlin Heidelberg: Springer: 1-15.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland: 172-176.

Max Silberztein. 2003. *NooJ Manual*. Available for download at: www.nooj4nlp.net .

Max Silberztein. 2012. Variable Unification in NooJ v3. In K. Vučković, B. Bekavac, & M. Silberztein (Eds.), *Automatic Processing of Various Levels of Linguistic Phenomena*. Newcastle upon Tyne: Cambridge Scholars Publishing: 1-13.

Max Silberztein. 2014. *Formaliser les langues: l'approche de NooJ*. London: ISTE eds.(forthcoming).

Simonetta Vietri. 2014a. The Lexicon-Grammar of Italian Idioms. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, Coling 2014, Dublin: 137-146.

Simonetta Vietri. 2014b. The Construction of an Annotated Corpus for the Analysis of Italian Transfer Predicates, *Lingvisticae Investigationes*, 37-1, Amsterdam & Philadelphia*:* John Benjamins: 69-105.

Simonetta Vietri. 2014c. *Idiomatic Constructions in Italian. A Lexicon-Grammar Approch.* Linguisticae Investigationes Supplementa, 31. Amsterdam & Philadelphia*:* John Benjamins (forthcoming).

Eros Zanchetta and Marco Baroni. 2006. Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings of Corpus Linguistics 2005*, online at corpus.bham.ac.uk/PCLC/.

Author Index

Abad, Azad, 1
Alam, Firoj, 122
Albanesi, Davide, 53
Allisio, Leonardo, 322
Anastasopoulos, Antonios, 138
Angster, Marco, 191
Annesi, Paolo, 7
Assylbekov, Zhenisbek, 12
Attardi, Giuseppe, 17

Barbagli, Alessia, 23
Barbieri, Francesco, 28
Barlacchi, Gianni, 33
Bartolini, Roberto, 128
Basile, Pierpaolo, 38, 43
Basile, Valerio, 48
Basili, Roberto, 7, 379
Bell, Peter, 103
Bellandi, Andrea, 53, 58
Bellusci, Alessia, 53, 58
Benotto, Giulia, 58
Bentivogli, Luisa, 63
Bernardi, Raffaella, 208
Bernardini, Lorenzo, 67
Bertoldi, Nicola, 108
Bigi, Brigitte, 73
Bolioli, Andrea, 78
Boschetti, Federico, 83
Bosco, Cristina, 316, 322
Bozzi, Andrea, 53
Brunato, Dominique, 87

Cannella, Vincenzo, 296
Cappelli, Amedeo, 58
Caputo, Annalina, 38, 43
Caselli, Tommaso, 93, 103
Castagnoli, Sara, 234
Castellucci, Giuseppe, 379
Celli, Fabio, 98
Cervone, Alessandra, 103
Cettolo, Mauro, 108
Chiari, Isabella, 93, 113
Chiusaroli, Francesca, 117
Cornacchia, Michele, 311
Cozza, Vittoria, 17
Croce, Danilo, 7, 379
Cucurullo, Sebastiana, 334

Danieli, Morena, 122
De Felice, Irene, 128, 132
De Mauro, Tullio, 113
De Rosa, Aurelio, 213
de Souza, José G. C., 138
Del Gratta, Riccardo, 83
Dell'Orletta, Felice, 23, 87
Delmonte, Rodolfo, 144, 151, 158
Delogu, Cristina, 311
Demberg, Vera, 339
Di Tomaso, Vittorio, 365
Donati, Emanuele, 239
Donati, Margherita, 132

Ermolaeva, Marina, 164

Falcone, Mauro, 311
Federico, Marcello, 108
Feltracco, Anna, 306
Ferro, Marcello, 301
Ferrone, Lorenzo, 169
Frontini, Francesca, 175, 224

Gagliardi, Gloria, 180
Gangemi, Aldo, 93
Geraci, Carlo, 191
Ghosh, Arindam, 98
Giovannetti, Emiliano, 53, 58
Giuliani, Diego, 344
Gobbi, Andrea, 197
Généreux, Michel, 186

Haponchyk, Iryna, 203
Huang, Chu-Ren, 328
Hürlimann, Manuela, 208

Iacobini, Claudio, 213

Jezek, Elisabetta, 93, 219, 306, 384

Kermanidis, Katia Lida, 275
Khan, Fahad, 224

Lamé, Marion, 83
Lavelli, Alberto, 229
Lebani, Gianluca E., 234
Lenci, Alessandro, 234, 286, 328
Lisi, Eleonora, 239

Lu, Qin, 328
Lucisano, Pietro, 23
Lyding, Verena, 186

Magnini, Bernardo, 63, 244, 306
Magnolini, Simone, 244
Marchetti, Alessandro, 260
Marchi, Simone, 58
Marchioni, Eleonora, 78
Mariotti, Alice, 249
Marotta, Giovanna, 132
Marzi, Claudia, 301
Masini, Francesca, 234
Mazzei, Alessandro, 191, 254
Mele, Francesco, 349
Menini, Stefano, 270
Minard, Anne-Lyse, 260
Monachini, Monica, 128, 175
Montemagni, Simonetta, 23, 87
Monti, Johanna, 265
Moretti, Giovanni, 270
Moschitti, Alessandro, 1, 33, 203, 374
Mussa, Valeria, 322

Negri, Matteo, 138
Nicolas, Lionel, 186
Nicosia, Massimo, 33
Nikiforos, Stefanos, 275
Nissim, Malvina, 234, 249
Nurkas, Assulan, 12

Oltramari, Alessandro, 93

Palmerini, Maria, 281
Paoloni, Andrea, 311
Paperno, Denis, 208
Pareti, Silvia, 103
Passaro, Lucia, 286
Passarotti, Marco, 291
Patti, Viviana, 322
Petrone, Caterina, 73
Phuoc An Vo, Ngoc, 306
Pipitone, Arianna, 296
Pirrelli, Vito, 301
Pirrone, Roberto, 296
Popescu, Octavian, 306
Poroli, Fabio, 311
Prodanof, Irina, 67, 103

Quochi, Valeria, 128, 175

Riccardi, Giuseppe, 98, 122
Ronzano, Francesco, 28
Ruffo, Giancarlo, 322
Russo, Irene, 128

Saggion, Horacio, 28
Sanguinetti, Manuela, 316, 322
Santus, Enrico, 328
Sartiano, Daniele, 17
Sassi, Manuela, 334
Sassolini, Eva, 334
Savy, Renata, 281
Sayeed, Asad, 339
Schirato, Giovanna, 213
Semeraro, Giovanni, 38, 43
Serizel, Romain, 344
Sorgente, Antonio, 349
Speranza, Manuela, 260
Spina, Stefania, 197, 354
Sprugnoli, Rachele, 270
Stemle, Egon W., 186
Sulis, Emilio, 322

Tamburini, Fabio, 360
Tarasconi, Francesco, 365
Tavosanis, Mirko, 370
Todisco, Massimiliano, 311
Tonelli, Sara, 270
Turchi, Marco, 138

Uryupina, Olga, 374

Vanzo, Andrea, 379
Varvara, Rossella, 384
Ventaglio, Raffaella, 78
Venturi, Giulia, 23, 87
Vetere, Guido, 93
Vettigli, Giuseppe, 349
Vietri, Simonetta, 389
Vieu, Laure, 93, 219

Zanzotto, Fabio Massimo, 93, 169, 239