Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora

Dhouha Bouamor
CEA, LIST, Vision and
Content Engineering Laboratory,
91191 Gif-sur-Yvette CEDEX
France
dhouha.bouamor@cea.fr

Nasredine Semmar
CEA, LIST, Vision and Content
Engineering Laboratory,
91191 Gif-sur-Yvette
CEDEX France
nasredine.semmar@cea.fr

Pierre Zweigenbaum
LIMSI-CNRS,
F-91403 Orsay CEDEX
France
pz@limsi.fr

Abstract

This paper presents an approach that extends the standard approach used for bilingual lexicon extraction from comparable corpora. We focus on the unresolved problem of *polysemous words* revealed by the bilingual dictionary and introduce a use of a Word Sense Disambiguation process that aims at improving the adequacy of context vectors. On two specialized French-English comparable corpora, empirical experimental results show that our method improves the results obtained by two state-of-the-art approaches.

1 Introduction

Over the years, bilingual lexicon extraction from comparable corpora has attracted a wealth of research works (Fung, 1998; Rapp, 1995; Chiao and Zweigenbaum, 2003). The basic assumption behind most studies is a *distributional* hypothesis (Harris, 1954), which states that words with a similar meaning are likely to appear in similar contexts across languages. The so-called ***standard approach*** to bilingual lexicon extraction from comparable corpora is based on the characterization and comparison of *context vectors* of source and target words. Each element in the context vector of a source or target word represents its association with a word which occurs within a window of N words. To enable the comparison of source and target vectors, words in the source vectors are translated into the target language using an existing bilingual dictionary.

The core of the standard approach is the bilingual dictionary. Its use is problematic when a word has several translations, whether they are synonymous or polysemous. For instance, the French word *action* can be translated into English as *share, stock, lawsuit* or *deed*. In such cases, it is difficult to identify in flat resources like bilingual dictionaries which translations are most relevant. The standard approach considers all available translations and gives them the same importance in the resulting translated context vectors independently of the domain of interest and word ambiguity. Thus, in the financial domain, translating *action* into *deed* or *lawsuit* would introduce noise in context vectors.

In this paper, we present a novel approach that addresses the word polysemy problem neglected in the standard approach. We introduce a Word Sense Disambiguation (WSD) process that identifies the translations of polysemous words that are more likely to give the best representation of context vectors in the target language. For this purpose, we employ five WordNet-based semantic *similarity* and *relatedness* measures and use a *data fusion* method that merges the results obtained by each measure. We test our approach on two specialized French-English comparable corpora *(financial and medical)* and report improved results compared to two state-of-the-art approaches.

2 Related Work

Most previous works addressing the task of bilingual lexicon extraction from comparable corpora are based on the standard approach. In order to improve the results of this approach, recent researches based on the assumption that more the context vectors are representative, better is the bilingual lexicon extraction were conducted. In these works, additional linguistic resources such as specialized dictionaries (Chiao and Zweigenbaum, 2002) or transliterated words (Prochasson et al., 2009) were combined with the bilingual dic-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 759–764,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

tionary to translate context vectors. Few works have however focused on the ambiguity problem revealed by the seed bilingual dictionary. (Hazem and Morin, 2012) propose a method that filters the entries of the bilingual dictionary on the base of a POS-Tagging and a domain relevance measure criteria but no improvements have been demonstrated. Gaussier et al. (2004) attempted to solve the problem of word ambiguities in the source and target languages. They investigated a number of techniques including canonical correlation analysis and multilingual probabilistic latent semantic analysis. The best results, with an improvement of the F-Measure ($+0.02$ at Top20) were reported for a mixed method. Recently, (Morin and Prochasson, 2011) proceed as the standard approach but weigh the different translations according to their frequency in the target corpus. Here, we propose a method that differs from Gaussier et al. (2004) in this way: If they focus on words ambiguities on source and target languages, we thought that it would be sufficient to disambiguate only translated source context vectors.

3 Context Vector Disambiguation

3.1 Semantic similarity measures

A large number of WSD techniques were proposed in the literature. The most widely used ones are those that compute semantic similarity[1] with the help of WordNet. WordNet has been used in many tasks relying on word-based similarity, including document (Hwang et al., 2011) and image (Cho et al., 2007; Choi et al., 2012) retrieval systems. In this work, we use it to derive a semantic similarity between lexical units within the same context vector. To the best of our knowledge, this is the first application of WordNet to bilingual lexicon extraction from comparable corpora.

Among semantic similarity measures using WordNet, we distinguish: (1) measures based on path length which simply counts the distance between two words in the WordNet taxonomy, (2) measures relying on information content in which a semantically annotated corpus is needed to compute frequencies of words to be compared and (3) the ones using gloss overlap which are designed to compute semantic relatedness. In this work, we use five similarity measures and compare their performances. These measures include three path-based semantic similarity measures denoted PATH,WUP (Wu and Palmer, 1994) and LEACOCK (Leacock and Chodorow, 1998). PATH is a baseline that is equal to the inverse of the shortest path between two words. WUP finds the depth of the least common subsumer of the words, and scales that by the sum of the depths of individual words. The depth of a word is its distance to the root node. LEACOCK finds the shortest path between two words, and scales that by the maximum path length found in the is–a hierarchy in which they occur. Path length measures have the advantage of being independent of corpus statistics, and therefor uninfluenced by sparse data.

Since semantic relatedness is considered to be more general than semantic similarity, we also use two relatedness measures: LESK (Banerjee and Pedersen, 2002) and VECTOR (Patwardhan, 2003). LESK finds overlaps between the glosses of word pairs, as well as words' hyponyms. VECTOR creates a co-occurrence matrix for each gloss token. Each gloss is then represented as a vector that averages token co-occurrences.

3.2 Disambiguation process

Once translated into the target language, the context vectors disambiguation process intervenes. This process operates *locally* on each context vector and aims at finding the most prominent translations of polysemous words. For this purpose, we use monosemic words as a seed set of disambiguated words to infer the polysemous word's translations senses. We hypothesize that a word is monosemic if it is associated to only one entry in the bilingual dictionary. We checked this assumption by probing monosemic entries of the bilingual dictionary against WordNet and found that 95% of the entries are monosemic in both resources. According to the above-described semantic similarity measures, a similarity value Sim_{Value} is derived between all the translations provided for each polysemous word by the bilingual dictionary and all monosemic words appearing within the same context vector. In practice, since a word can belong to more than one synset[2] in WordNet, the semantic similarity between two words w_1 and w_2 is defined as the *maximum* of Sim_{Value} between the synset or the synsets that include the $synsets(w_1)$ and

[1] For consiseness, we often use "semantic similarity" to refer collectively to both similarity and relatedness.

[2] a group of a synonymous words in WordNet

$synsets(w_2)$ according to the following equation:

$$Sem_{Sim}(w_1, w_2) = \max\{Sim_{Value}(s_1, s_2); \\ (s_1, s_2) \in synsets(w_1) \times synsets(w_2)\} \quad (1)$$

Then, to identify the most prominent translations of each polysemous unit w_p, an *average similarity* is computed for each translation w_p^j of w_p:

$$Ave\_Sim(w_p^j) = \frac{1}{N}\sum_{i=1}^{N} Sem_{Sim}(w_i, w_p^j) \qquad (2)$$

where N is the total number of monosemic words in each context vector and Sem_{Sim} is the similarity value of w_p^j and the i^{th} monosemic word. Hence, according to average similarity values $Ave\_Sim(w_p^j)$, we obtain for each polysemous word w_p an ordered list of translations $w_p^1 \ldots w_p^n$.

4 Experiments and Results

4.1 Resources and Experimental Setup

We conducted our experiments on two French-English comparable corpora specialized on the *corporate finance* and the *breast cancer* sub-domains. Both corpora were extracted from Wikipedia[3]. We consider the domain topic in the source language (for instance *cancer du sein* [breast cancer]) as a query to Wikipedia and extract all its sub-topics (i.e., sub-categories in Wikipedia) to construct a domain-specific *categories tree*. Then we collected all articles belonging to one of these categories and used inter-language links to build the comparable corpus. Both corpora have been normalized through the following linguistic preprocessing steps: tokenisation, part-of-speech tagging, lemmatisation and function words removal. The resulting corpora[4] sizes as well as their polysemy rate P_R are given in Table 1. The polysemy rate indicates how much words in the comparable corpora are associated to more than one translation in the seed bilingual dictionary. The dictionary consists of an in-house bilingual dictionary which contains about 120,000 entries belonging to the general language with an average of 7 translations per entry.

In bilingual terminology extraction from comparable corpora, a reference list is required to evaluate the performance of the alignment. Such lists are often composed of about 100 single terms (Hazem and Morin, 2012; Chiao and Zweigenbaum, 2002). Here, we created two reference lists[5] for the *corporate finance* and the *breast cancer* sub-domains. The first list is composed of 125 single terms extracted from the glossary of bilingual micro-finance terms[6]. The second list contains 79 terms extracted from the French-English MESH and the UMLS thesauri[7]. Note that reference terms pairs appear more than five times in each part of both comparable corpora.

Corpus	French	English	P_R
Corporate finance	402.486	756.840	41%
Breast cancer	396.524	524.805	47%

Table 1: Comparable corpora sizes in term of words and polysemy rates (P_R) associated to each corpus

Three other parameters need to be set up, namely the window size, the association measure and the similarity measure. We followed (Laroche and Langlais, 2010) to define these parameters. They carried out a complete study of the influence of these parameters on the bilingual alignment. The context vectors were defined by computing the Discounted Log-Odds Ratio (equation 3) between words occurring in the same context window of size 7.

$$\textit{Odds-Ratio}_{disc} = \log\frac{(O_{11}+\frac{1}{2})(O_{22}+\frac{1}{2})}{(O_{12}+\frac{1}{2})(O_{21}+\frac{1}{2})} \qquad (3)$$

where O_{ij} are the cells of the 2×2 contingency matrix of a token s co-occurring with the term S within a given window size. As similarity measure, we chose to use the cosine measure.

4.2 Results of bilingual lexicon extraction

To evaluate the performance of our approach, we used both the standard approach (SA) and the approach proposed by (Morin and Prochasson, 2011) (henceforth MP11) as baselines. The experiments were performed with respect to the five semantic similarity measures described in section 3.1. Each measure provides, for each polysemous word, a ranked list of translations. A question that arises here is whether we should introduce only the top-ranked translation into the context vector or consider a larger number of translations, mainly when a translation list contains synonyms. For this

[3] `http://dumps.wikimedia.org/`

[4] Comparable corpora will be shared publicly

[5] Reference lists will be shared publicly

[6] `http://www.microfinance.lu/en/`

[7] `http://www.nlm.nih.gov/`

		Method	WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
a) Corporate Finance	Standard Approach (SA)		0.172						
	MP11		0.336						
	Single measure	WuP	0.241	0.284	*0.301*	0.275	0.258	0.215	0.224
		Path	0.250	0.284	*0.301*	*0.284*	0.258	0.215	0.215
		Leacock	0.250	*0.293*	*0.301*	0.275	*0.275*	0.241	*0.232*
		Lesk	*0.272*	*0.293*	0.293	0.275	0.258	*0.250*	0.215
		Vector	0.267	0.310	0.284	*0.284*	0.232	0.232	*0.232*
	Condorcet$_{Merge}$		**0.362**	$\underline{\textbf{0.379}}$	**0.353**	**0.362**	**0.336**	0.275	0.267

		Method	WN-T$_1$	WN-T$_2$	WN-T$_3$	WN-T$_4$	WN-T$_5$	WN-T$_6$	WN-T$_7$
b) Breast Cancer	Standard Approach (SA)		0.493						
	MP11		0.553						
	Single measure	WuP	0.481	0.566	*0.566*	0.542	0.554	0.542	*0.554*
		Path	0.542	0.542	0.554	0.566	*0.578*	0.554	*0.554*
		Leacock	0.506	*0.578*	0.554	0.566	0.542	0.554	0.542
		Lesk	0.469	0.542	0.542	***0.590***	0.554	0.554	0.542
		Vector	*0.518*	0.566	0.530	0.566	0.542	*0.566*	*0.554*
	Condorcet$_{Merge}$		**0.566**	$\underline{\textbf{0.614}}$	**0.600**	**0.590**	**0.600**	**0.578**	**0.578**

Table 2: F-Measure at Top20 for the two domains; MP11 = (Morin and Prochasson, 2011). In each column, italics shows best single similarity measure, bold shows best result. Underline shows best result overall.

reason, we take into account in our experiments different numbers of translations, noted WN-T$_i$, ranging from the pivot translation ($i = 1$) to the seventh word in the translation list. This choice is motivated by the fact that words in both corpora have on average 7 translations in the bilingual dictionary. Both baseline systems use all translations associated to each entry in the bilingual dictionary. The only difference is that in MP11 translations are weighted according to their frequency in the target corpus.

The results of different works focusing on bilingual lexicon extraction from comparable corpora are evaluated on the number of correct candidates found in the first N first candidates output by the alignment process (the Top$N$). Here, we use the Top20 F-measure as evaluation metric. The results obtained for the *corporate finance* corpus are presented in Table 2a. The first notable observation is that disambiguating context vectors using semantic similarity measures outperforms the SA. The highest F-measure is reported by Vector. Using the top two words (WN-T$_2$) in context vectors increases the F-measure from 0.172 to 0.310. However, compared to MP11, no improvement is achieved. Concerning the *breast cancer* corpus, Table 2b shows improvements in most cases over both the SA and MP11. The maximum F-measure was obtained by Lesk when for each polysemous word up to four translations (WN-T$_4$) are considered in context vectors. This method achieves an improvement of respectively $+0.097$ and $+0.037\%$ over SA and MP11.

Each of the tested 5 semantic similarity measures provides a different view of how to rank the translations of a given test word. Combining the obtained ranked lists should reinforce the confidence in consensus translations, while decreasing the confidence in non-consensus translations. We have therefore tested their combination. For this, we used a voting method, and chose one in the Condorcet family the *Condorcet data fusion method*. This method was widely used to combine document retrieval results from information retrieval systems (Montague and Aslam, 2002; Nuray and Can, 2006). It is a single-winner election method that ranks the candidates in order of preference. It is a *pairwise voting*, i.e. it compares every possible pair of candidates to decide the preference of them. A matrix can be used to present the competition process. Every candidate appears in the matrix as a row and a column as well. If there are m candidates, then we need m^2 elements in the matrix in total. Initially 0 is written to all the elements. If d_i is preferred to d_j , then we add 1 to the element at row i and column j (a_{ij}). The process is repeated until all the ballots are processed. For every element a_{ij} , if $a_{ij} > m/2$, then d_i beats d_j; if $a_{ij} < m/2$, then d_j beats d_i; otherwise ($a_{ij} = m/2$), there is a draw between d_i and d_j. The total score of each candidate is quantified by summing the raw scores it obtains in all pairwise competitions. Finally the ranking is achievable based on the total scores calculated.

Here, we view the ranking of the extraction results from different similarity measures as a special instance of the voting problem where the Top20 extraction results correspond to candidates and different semantic similarity measures are the voters. The combination method referred to as CONDORCET$_{Merge}$ outperformed all the others (see Tables 2a and 2b): (1) individual measures, (2) SA, and (3) MP11. Even though the two corpora are fairly different (subject and polysemy rate), the optimal results are obtained when considering up to two most similar translations in context vectors. This behavior shows that the fusion method is robust to domain change. The addition of supplementary translations, which are probably noisy in the given domain, degrades the overall results. The F-measure gains with respect to SA are $+0.207$ for corporate finance and $+0.121$ for the breast cancer corpus. More interestingly, our approach outperforms MP11, showing that the role of disambiguation is more important than that of feature weighting.

5 Conclusion

We presented in this paper a novel method that extends the standard approach used for bilingual lexicon extraction. This method disambiguates polysemous words in context vectors by selecting only the most relevant translations. Five semantic similarity and relatedness measures were used for this purpose. Experiments conducted on two specialized comparable corpora indicate that the combination of similarity metrics leads to a better performance than two state-of-the-art approaches. This shows that the ambiguity present in specialized comparable corpora hampers bilingual lexicon extraction, and that methods such as the one introduced here are needed. The obtained results are very encouraging and can be improved in a number of ways. First, we plan to mine much larger specialized comparable corpora and focus on their quality (Li and Gaussier, 2010). We also plan to test our method on bilingual lexicon extraction from general-domain corpora, where ambiguity is generally higher and disambiguation methods should be all the more needed.

References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 136–145, London, UK, UK. Springer-Verlag.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5. Association for Computational Linguistics.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The effect of a general lexicon in corpus-based identification of french-english medical word translations. In *Proceedings Medical Informatics Europe, volume 95 of Studies in Health Technology and Informatics*, pages 397–402, Amsterdam.

Miyoung Cho, Chang Choi, Hanil Kim, Jungpil Shin, and PanKoo Kim. 2007. Efficient image retrieval using conceptualization of annotated images. Lecture Notes in Computer Science, pages 426–433. Springer.

Dongjin Choi, Jungin Kim, Hayoung Kim, Myunggwon Hwang, and Pankoo Kim. 2012. A method for enhancing image retrieval based on annotation using modified wup similarity in wordnet. In *Proceedings of the 11th WSEAS international conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, AIKED'12, pages 83–87, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In *Parallel Text Processing*, pages 1–17. Springer.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL*, pages 526–533.

Z.S. Harris. 1954. Distributional structure. *Word*.

Amir Hazem and Emmanuel Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings, 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May.

Myunggwon Hwang, Chang Choi, and Pankoo Kim. 2011. Automatic enrichment of semantic relation network and its application to word sense disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, 23:845–858.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China, Aug.

Claudia Leacock and Martin Chodorow, 1998. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press.

Bo Li and Ëric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, Aug.

Mark Montague and Javed A. Aslam. 2002. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 538–548, New York, NY, USA. ACM.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings, 4th Workshop on Building and Using Comparable Corpora (BUCC)*, page 27–34, Portland, Oregon, USA.

Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, May.

Siddharth Patwardhan. 2003. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness. Master's thesis, University of Minnesota, Duluth, August.

Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings, 12th Conference on Machine Translation Summit (MT Summit XII)*, page 284–291, Ottawa, Ontario, Canada.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322. Association for Computational Linguistics.

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. Association for Computational Linguistics.