Building Specialized Bilingual Lexicons Using Word Sense Disambiguation

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 extension of the standard approach used for bilingual lexicon extraction from comparable corpora. We study the ambiguity problem revealed by the seed bilingual dictionary used to translate context vectors and augment the standard approach by a Word Sense Disambiguation process. Our aim is to identify the translations of words that are more likely to give the best representation of words in the target language. On two specialized French-English and Romanian-English comparable corpora, empirical experimental results show that the proposed method consistently outperforms the standard approach.

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 main work in this research area could be seen as an extension of Harris's *distributional hypothesis* (Harris, 1954). It is based on the simple observation that a word and its translation are likely to appear in similar contexts across languages (Rapp, 1995). Based on this assumption, the alignment method, known as the *standard approach* builds and compares context vectors for each word of the source and target languages.

A particularity of this approach is that, to enable the comparison of context vectors, it requires the existence of a seed bilingual dictionary to translate source context vectors. The use of the bilingual dictionary 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, wherein entries are usually unweighted and unordered, 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 probably introduce noise in context vectors.

In this paper, we present a novel approach which addresses the word ambiguity problem neglected in the standard approach. We introduce a use of a WordNet-based semantic similarity measure permitting the disambiguation of translated context vectors. The basic intuition behind this method is that instead of taking all translations of each seed word to translate a context vector, we only use the translations that are more likely to give the best representation of the context vector in the target language. We test the method on two comparable corpora specialized on the Breast Cancer domain, for the French-English and Romanian-English pair of languages. This choice allows us to study the behavior of the disambiguation for a pair of languages that are richly represented and for a pair that includes Romanian, a language that has fewer associated resources than French and English.

2 Related Work

Recent improvements of the standard approach are based on the assumption that the more the context vectors are representative, the better the bilingual lexicon extraction is. Prochasson et al. (2009)

International Joint Conference on Natural Language Processing, pages 952–956,
Nagoya, Japan, 14-18 October 2013.

used transliterated words and scientific compound words as 'anchor points'. Giving these words higher priority when comparing target vectors improved bilingual lexicon extraction. In addition to transliteration, Rubino and Linarès (2011) combined the contextual representation within a thematic one. The basic intuition of their work is that a term and its translation share thematic similarities. Hazem and Morin (2012) recently proposed a method that filters the entries of the bilingual dictionary based upon POS-tagging and domain relevance criteria, but no improvements was demonstrated.

Gaussier et al. (2004) attempted to solve the problem of different 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 a very small improvement were reported for a mixed method. One important difference with Gaussier et al. (2004) is that they focus on words ambiguities on source and target languages, whereas we consider that it is sufficient to disambiguate only translated source context vectors.

3 Context Vector Disambiguation

The approach we propose augments the standard approach used for bilingual lexicons mining from comparable corpora. As it was mentioned in section 1, when the lexical extraction applies to a specific domain, not all translations in the bilingual dictionary are relevant for the target context vector representation. For this reason, we introduce a WordNet-based WSD process that aims at improving the adequacy of context vectors and therefore improve the results of the standard approach.

A large number of WSD techniques were previously proposed in the literature. The most popular ones are those that compute semantic similarity with the help of existing thesauri such as WordNet (Fellbaum, 1998). This thesaurus has been applied to 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 this resource 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 the task of bilingual lexicon extraction from comparable corpora.

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.

Formally, we derive a semantic similarity value between all the translations provided for each polysemous word by the bilingual dictionary and all monosemic words appearing whithin the same context vector. There is a relatively large number of word-to-word similarity metrics that were previously proposed in the literature, ranging from path-length measures computed on semantic networks, to metrics based on models of distributional similarity learned from large text collections. For simplicity, we use in this work, the Wu and Palmer (1994) (WUP) path-length-based semantic similarity measure. It was demonstrated by (Lin, 1998) that this metric achieves good performances among other measures. WUP computes a score (equation 1) denoting how similar two word senses are, based on the depth of the two synsets (s_1 and s_2) in the WordNet taxonomy and that of their Least Common Subsumer (LCS), i.e., the most specific word that they share as an ancestor.

$$Wup_{Sim}(s_1, s_2) = \frac{2 \times \textit{depth}(LCS)}{\textit{depth}(s_1) + \textit{depth}(s_2)} \quad (1)$$

In practice, since a word can belong to more than one synset in WordNet, we determine the semantic similarity between two words w_1 and w_2 as the maximum Wup_{Sim} between the synset or the synsets that include the $synsets(w_1)$ and $synsets(w_2)$ according to the following equation:

$$Sem_{Sim}(w_1, w_2) = \max\{Wup_{Sim}(s_1, s_2); \\ (s_1, s_2) \in synsets(w_1) \times synsets(w_2)\} \quad (2)$$

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:

$$Ave\_Sim(w_p^j) = \frac{\sum_{i=1}^{N} Sem_{Sim}(w_i, w_p^j)}{N} \quad (3)$$

Corpus	French	English
	396,524	524,805
Corpus	Romanian	English
	22,539	322,507

Table 1: Comparable corpora sizes in term of words.

where N is the total number of monosemic words and Sem_{Sim} is the similarity value of w_p^j and the i^{th} monosemic word. Hence, according to average relatedness values $Ave\_Sim(w_p^j)$, we obtain for each polysemous word w_p an ordered list of translations $w_p^1 \dots w_p^n$. This allows us to select translations of words which are more salient than the others to represent the word to be translated.

4 Experiments and Results

4.1 Resources

4.1.1 Comparable corpora

We conducted our experiments on two French-English and Romanian-English comparable corpora specialized on the *breast cancer* domain. Both corpora were extracted from Wikipedia[1]. We consider the 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 *category tree*. Then, based on the constructed tree, we collect all Wikipedia pages belonging to one of these categories and use *inter-language links* to build the comparable corpus. Both corpora were normalized through the following linguistic preprocessing steps: tokenisation, part-of-speech tagging, lemmatisation, and function word removal. The resulting corpora[2] sizes are given in Table 1.

4.1.2 Bilingual dictionary

The French-English bilingual dictionary used to translate context vectors consists of an in-house manually revised bilingual dictionary which contains about 120,000 entries belonging to the general domain. It is important to note that words has on average 7 translations in the bilingual dictionary. The Romanian-English dictionary consists of translation pairs extracted from Wikipedia. The resulting bilingual dictionary contains about 136,681 entries for Romanian-English with an average of 1 translation per word.

4.1.3 Evaluation list

In bilingual terminology extraction from comparable corpora, a reference list is required to evaluate the performance of the alignment. Such lists are usually composed of about 100 single terms (Hazem and Morin, 2012; Chiao and Zweigenbaum, 2002). Here, we created a reference list[3] for each pair of language. The French-English list contains 96 terms extracted from the French-English MESH and the UMLS thesauri[4]. The Romanian-English reference list was created by a native speaker and contains 38 pair of words. Note that reference terms pairs appear at least five times in each part of both comparable corpora.

4.2 Experimental setup

Three other parameters need to be set up: (1) the window size, (2) the association measure and the (3) similarity measure. To define context vectors, we use a seven-word window as it approximates syntactic dependencies. Concerning the rest of the parameters, we followed Laroche and Langlais (2010) for their definition. The authors carried out a complete study of the influence of these parameters on the bilingual alignment and showed that the most effective configuration is to combine the Discounted Log-Odds ratio (equation 4) with the cosine similarity. The Discounted Log-Odds ratio is defined as follows:

$$\textit{Odds-Ratio}_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \quad (4)$$

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.

4.3 Results and discussion

It is difficult to compare results between different studies published on bilingual lexicon extraction from comparable corpora, because of difference between (1) used corpora (in particular their construction constraints and volume), (2) target domains, and also (3) the coverage and relevance of linguistic resources used for translation. To the best of our knowledge, there is no common benchmark that can serve as a reference. For this reason,

[1] http://dumps.wikimedia.org/

[2] Comparable corpora will be shared publicly

[3] Reference lists will be shared publicly

[4] http://www.nlm.nih.gov/

b) FR-EN

Method		WN-T_1	WN-T_2	WN-T_3	WN-T_4	WN-T_5	WN-T_6	WN-T_7
Standard Approach(SA)		0.49						
Single measures	WUP	0.48	0.56	*0.56*	0.54	0.55	0.54	*0.55*
	PATH	0.54	0.54	0.55	0.56	*0.57*	0.55	*0.55*
	LEACOCK	0.50	*0.57*	0.55	0.56	0.54	0.55	0.54
	LESK	0.46	0.54	0.54	***0.59***	0.55	0.55	0.54
	VECTOR	*0.51*	0.56	0.53	0.56	0.54	*0.56*	*0.55*

b) RO-EN

Method		WN-T_1	WN-T_2	WN-T_3	WN-T_4	WN-T_5	WN-T_6	WN-T_7
Standard Approach(SA)		0.21						
Single measures	WUP	0.18	0.21	0.21	0.21	0.21	0.21	0.21
	PATH	0.18	0.21	0.21	0.21	0.21	0.21	0.21
	LEACOCK	0.15	0.18	0.18	0.18	0.18	0.18	0.18
	LESK	0.21	0.21	0.21	0.21	0.21	0.21	0.21
	VECTOR	0.18	0.21	0.21	0.21	0.21	0.21	0.21

Table 2: F-Measure at Top20 for the Breast Cancer domain for the two pairs of languages; In each column, italics shows best single similarity measure, bold shows best result. Underline shows best result overall.

we use the results of the standard approach (SA) as a reference. We evaluate the performance of both the SA and ours with respect to Top20 F-Measure which computes the harmonic mean between precision and recall.

Our method provides a ranked list of translations for each polysemous word. A question that arises here is whether we should introduce only the best ranked translation in the context vector or consider a larger number of words, especially when a translations list contain synonyms. For this reason, we take into account in our experiments different number of translations, noted WN-T_i, ranging from the pivot translation ($i = 1$) to the seventh word in the translations list. This choice is motivated by the fact that words in the French-English corpus have on average 7 translations in the bilingual dictionary. The baseline (SA) uses all translations associated to each entry in the bilingual dictionary. Table 2a displays the results obtained for the French-English comparable corpus. The first substantial observation is that our method which consists in disambiguating polysemous words within context vectors consistently outperforms the standard approach. 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 $+10\%$ and over the standard approach.

Concerning the Romanian-English pair of language, no improvements have been reported. The reason being that words in the bilingual dictionary are not heavily polysemous. Each word used to shape context vectors is associated to only one translation in the bilingual dictionary.

5 Conclusion

We presented in this paper a novel method that extends the standard approach used for bilingual lexicon extraction from comparable corpora. The proposed method disambiguates polysemous words in context vectors and selects only the translations that are most relevant to the general context of the corpus. Conducted experiments on a highly polysemous specialized comparable corpus show that integrating such process leads to a better performance than the standard approach. Although our initial experiments are positive, we believe that they could be improved in a number of ways. It would also be interesting to mine much more larger comparable corpora and focus on their quality as presented in (Li and Gaussier, 2010). We want also to test our method on bilingual lexicon extraction for a larger panel of specialized corpora, where disambiguation methods are needed to prune translations that are irrelevant to the domain.

References

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).

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel 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.

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.

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.

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.

Raphaël Rubino and Georges Linarès. 2011. A multi-view approach for term translation spotting. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 29–40.

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.