Towards a Generic Approach 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 problem associated to *polysemous words* found in the seed bilingual lexicon when translating source context vectors. To improve the adequacy of context vectors, the use of a WordNet-based Word Sense Disambiguation process is tested. Experimental results on four specialized French-English comparable corpora show that our method outperforms two state-of-the-art approaches.

1 Introduction

Bilingual lexicons play an important role in many natural language processing applications such as machine translation or cross-language information retrieval (Shi, 2009). Research on lexical extraction from multilingual corpora have largely focused on parallel corpora. The scarcity of such corpora in particular for specialized domains and for language pairs not involving English pushed researchers to investigate the use of comparable corpora (Fung, 1998; Rapp, 1995; Chiao and Zweigenbaum, 2003), in which texts are not exact translation of each other but share common features.

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 lexical environments represented by *context vectors* of source and target words. In order to enable the comparision 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*. Identifying which translations provided by a given bilingual dictionary are most relevant impacts the quality of the extracted bilingual lexicons. 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. For instance, in the financial domain, translating *action* into *deed* or *lawsuit* would introduce noise in context vectors.

In this paper, we present a novel approach which addresses the word polysemy problem neglected in the standard approach. We exploit 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 relatedness measures and use a *data fusion* method that merges the results obtained by each measure. We test our approach on four specialized French-English comparable corpora \\\\\\\(*financial, medical, wind energy and mobile technology)* and report improved results compared to two state-of-the-art approaches.

The remainder of this paper is organized as follows: the next section describes the standard ap-

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 143–150.

proach and previous works addressing the task of bilingual lexicon extraction from comparable corpora. In Section 3, we present our context vector disambiguation process. Before concluding in section 5, we describe the experimental protocol we followed and discuss the obtained results in section 4.

2 Bilingual Lexicon extraction

2.1 Standard Approach

Most previous works addressing the task of bilingual lexicon extraction from comparable corpora are based on the standard approach (Fung, 1998; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010). Formally, this approach is composed of the following three steps:

1. **Building context vectors**: Vectors are first extracted by identifying the words that appear around the term to be translated S in a window of N words. Generally, an association measure like the mutual information (Morin and Daille, 2006), the log-likelihood (Morin and Prochasson, 2011) or the Discounted Odds-Ratio (Laroche and Langlais, 2010) are employed to shape the context vectors.

2. **Translation of context vectors**: To enable the comparison of source and target vectors, source terms vectors are translated in the target language by using a seed bilingual dictionary. Whenever it provides several translations for an element, all proposed translations are considered. Words not included in the bilingual dictionary are simply ignored.

3. **Comparison of source and target vectors**: Translated vectors are compared to target ones using a similarity measure. The most widely used is the cosine similarity, but many authors have studied alternative metrics such as the Weighted Jaccard index (Prochasson et al., 2009) or the City-Block distance (Rapp, 1999). According to similarity values, a ranked list of translations for S is obtained.

2.2 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. 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 dictionary to translate context vectors.

Few works have however focused on the ambiguity problem revealed by the seed bilingual dictionary. Hazem and Morin (2012) filtered the entries of the bilingual dictionary on the basis of part-of-speech tags and of domain relevance criteria but no improvement 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. However, only small improvements are reported. 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. Recently, Morin and Prochasson (2011) modified the standard approach by weighting the different translations according to their frequency in the target corpus.

3 Context Vector Disambiguation

Our approach includes the three steps of the standard approach. As we mentioned in Section 1, when 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 WSD process that aims at improving the adequacy of context vectors and therefore improve the results of the standard approach. The overall architecture of the lexical extraction process is shown in Figure 1. In this section, we first describe the semantic resource on which our approach is based. Then, we present in details the method we propose to

3.1 Semantic resource

A large number of WSD techniques were previously proposed in the literature. The most widely used ones are those that compute semantic relatedness[1] with WordNet. This thesaurus 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

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

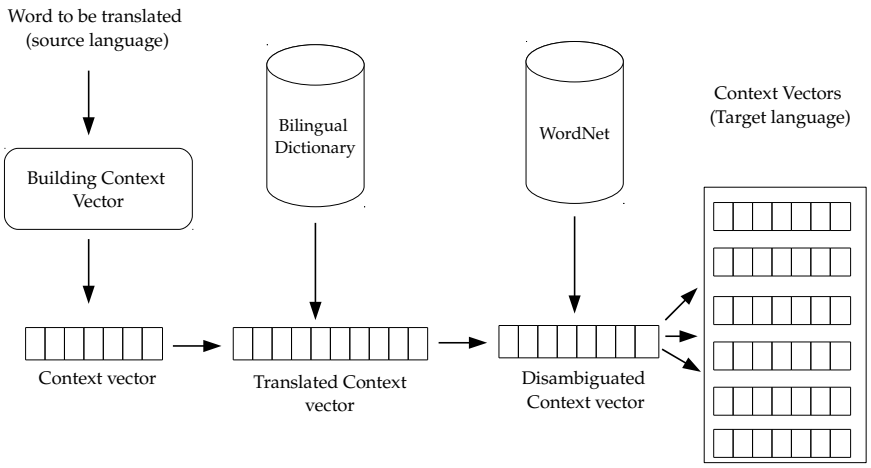

Figure 1: Overview of the lexical extraction approach

systems. In this work, we use WordNet to derive a semantic relatedness among 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 relatedness measures using WordNet, we distinguish: (1) measures based on path length which simply count 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 relatedness measures, compare their performances and then combine them. 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 (i.e. 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 therefore uninfluenced by sparse data.

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

3.2 Disambiguation process

We augment the standard approach by proposing a disambiguation process after translating context vectors. 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 relatedness measures, a relatedness value Rel_{Value}

Context Vector of *bénéfice*	Translation	Similarity
liquidité	liquidity	–
action	act	0.2139
	action	0.4256
	stock	***0.5236***
	deed	0.1594
	lawsuit	0.1212
	fact	0.1934
	operation	0.2045
	share	***0.5236***
	plot	0.2011
dividende	dividend	–

Table 1: Context vector disambiguation of the French term *bénéfice [income]* in the *corporate finance* domain. Similarity here is the Ave_Rel given by WUP. *liquidité* and *dividende* are monosemic and are used to infer the most similar translations of the term *action*.

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 relatedness between two words w_1 and w_2 is defined as the *maximum* of Rel_{Value} between the synset or the synsets that include the $synsets(w_1)$ and $synsets(w_2)$ according to the following equation:

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

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:

$$Ave\_Rel(w_p^j) = \frac{\sum_{i=1}^{N} Sem_{Rel}(w_i, w_p^j)}{N} \quad (2)$$

where N is the total number of monosemic words and Sem_{Rel} is the similarity value of w_p^j and the i^{th} monosemic word. Hence, according to average relatedness values $Ave\_Rel(w_p^j)$, we obtain for each polysemous word w_p an ordered list of translations $w_p^1 \dots w_p^n$. Table 1 displays the results of the disambiguation process for the context

[2]a group of a synonymous words in WordNet

Corpus	French	English	$P_R(\%)$
Corporate finance	402,486	756,840	41
Breast cancer	396,524	524,805	47
Wind energy	145,019	345,607	51
Mobile technology	197,689	144,168	37

Table 2: Comparable corpora's sizes in term of words and their polysemy rates (P_R)

.

vector of the French term *bénéfice* in the financial domain. This vector contains the words *action*, *dividende*, *liquidité* and others. The bilingual dictionary provides the following translations {*act, stock, action, deed, lawsuit, fact, operation, plot, share*} for the French polysemous word *action*. We use the monosemic words *dividende* and *liquidité* to disambiguate the word *action*. From observing relatedness values, we notice that the words *share* and *stock* are on the top of the list and therefore are most likely to represent the source word *action* in this context.

4 Experiments and Results

4.1 Resources and Experimental Setup

We conducted our experiments on four French-English comparable corpora specialized on the *corporate finance*, *breast cancer*, *wind energy* and the *mobile technology* sub-domains. The two first corpora were extracted from Wikipedia[3]. We considered 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., subcategories in Wikipedia) to construct a domain-specific *category tree*. Then we collected all articles belonging to one of these categories and used inter-language links to build the comparable corpus. Concerning the corpora related to the *wind energy* and *mobile technology* domains, we used the corpora used in the TTC project[4]. The four corpora were normalized through the following linguistic preprocessing steps: tokenisation, part-of-speech tagging, lemmatisation, and function word removal. The resulting corpora[5] sizes as well as their polysemy rate P_R are given in Table 2. P_R indicates the percentage of words that are associated to more than one translation in the seed

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

[4]http://www.ttc-project.eu/index.php/releases-publications

[5]Comparable corpora will be shared publicly

a) Corporate Finance

	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)				0.172			
	MP11				0.336			
Single measure	WUP	0.241	0.284	*0.301*	0.275	0.258	0.215	0.224
Single measure	PATH	0.250	0.284	*0.301*	*0.284*	0.258	0.215	0.215
Single measure	LEACOCK	0.250	*0.293*	*0.301*	0.275	*0.275*	0.241	*0.232*
Single measure	LESK	*0.272*	*0.293*	0.293	0.275	0.258	*0.250*	0.215
Single measure	VECTOR	0.267	0.310	0.284	*0.284*	0.232	0.232	*0.232*
	CONDORCET$_{Merge}$	**0.362**	**<u>0.379</u>**	**0.353**	**0.362**	**0.336**	**0.275**	**0.267**

b) Breast Cancer

	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)				0.493			
	MP11				0.553			
Single measure	WUP	0.481	0.566	*0.566*	0.542	0.554	0.542	*0.554*
Single measure	PATH	0.542	0.542	0.554	0.566	*0.578*	0.554	*0.554*
Single measure	LEACOCK	0.506	*0.578*	0.554	0.566	0.542	0.554	0.542
Single measure	LESK	0.469	0.542	0.542	***0.590***	0.554	0.554	0.542
Single measure	VECTOR	*0.518*	0.566	0.530	0.566	0.542	*0.566*	*0.554*
	CONDORCET$_{Merge}$	**0.566**	**<u>0.614</u>**	**0.600**	**0.590**	**0.600**	**0.578**	**0.578**

Table 3: 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.

bilingual dictionary. To translate context vectors, we used an in-house bilingual dictionary containing 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 usually composed of about 100 single terms (Hazem and Morin, 2012; Chiao and Zweigenbaum, 2002). Here, we created four reference lists[6] for the *corporate finance*, *breast cancer*, *wind energy* and the *mobile technology* subdomains. The first list is composed of 125 single terms extracted from the glossary of bilingual micro-finance terms[7]. The second list contains 96 terms extracted from the French-English MESH and the UMLS thesauri[8]. The thrid list is composed of 89 single terms extracted from a glossary of terms for the renewable domain. For the mobile technology domain, a list of 142 single terms is composed from. Note that reference terms pairs appear at least five times in each part of the comparable corpora.

Three other parameters need to be set up: (1) the window size, (2) the association measure and (3) 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.

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

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. The cosine measure is used to compute similarity.

4.2 Results of bilingual lexicon extraction

The performances of our approach are evaluated against the standard approach (SA) and the approach proposed by (Morin and Prochasson, 2011) (henceforth MP11). The experiments were performed using the five semantic relatedness 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

[6]Reference lists will be shared publicly

[7]http://www.microfinance.lu/en/

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

c) Wind Energy

	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)		0.080						
MP11		0.242						
Single measure	WUP	0.150	0.208	0.173	0.184	0.161	0.161	0.150
	PATH	0.231	0.231	0.196	0.161	0.127	0.127	0.104
	LEACOCK	0.184	0.173	0.197	0.197	0.150	0.150	0.138
	LESK	0.208	0.196	0.196	0.161	0.115	0.138	0.104
	VECTOR	*0.219*	0.208	0.219	0.161	0.138	0.115	0.127
CONDORCET$_{Merge}$		**<u>0.346</u>**	**0.300**	**0.289**	0.219	0.208	**0.184**	**0.173**

d) Mobile Technology

	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)		0.064						
MP11		0.057						
Single measure	WUP	0.151	*0.173*	0.137	0.101	0.101	0.086	0.086
	PATH	0.129	0.166	0.122	0.101	0.093	0.086	0.086
	LEACOCK	0.129	0.144	0.108	0.093	0.079	0.079	0.086
	LESK	0.137	0.129	0.108	0.086	0.101	0.086	0.086
	VECTOR	0.144	0.151	0.115	0.108	0.115	0.086	0.079
CONDORCET$_{Merge}$		**0.223**	**<u>0.245</u>**	**0.187**	**0.151**	**0.158**	**0.137**	**0.122**

Table 4: 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.

number of translations, mainly when a translation list contains synonyms. For this reason, we take into account in our experiments different numbers of translations, noted WN-T_i, ranging from the top translation ($i = 1$) to the seventh word in the translations 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.

As evaluation metric, we use the Top20 F-measure, which measures the harmonic mean of precision and recall. Precision is the total number of correct translations divided by the number of terms for which the system gave at least one answer. Recall is equal to the ratio of correct translation to the total number of terms.

The results obtained for the *corporate finance* corpus are presented in Table 3a. The first notable observation is that disambiguating context vectors using semantic relatedness measures outperforms 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. Similar results are obtained for the *wind energy* domain (Table 4c). Here, the best improvement of the SA is achieved by PATH when considering only the first translation for each polysemous word. Concerning the *breast cancer* corpus, Table 3b shows improvements in most cases over both 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. For the *mobile technology* domain, the obtained results, displayed in table 4d, show that the disambiguation of context vectors report high values of the F-measure compared to the SA and MP11 for all the configurations.

The five tested semantic relatedness measures provide complementary rankings of 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 chose a voting method from the Condorcet family, namely the *Condorcet data fusion method*. This method was widely used to combine docu-

ment 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 relatedness measures are the voters. The combination method referred to as CONDORCET$_{Merge}$ outperformed all the others (see Tables 3a, 3b, 4c and 4d): (1) individual measures, (2) SA, and (3) MP11. Even though the four corpora are fairly different (subject and polysemy rate), the optimal results are obtained for most domains, 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 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 disambiguation positively affects the overall results.

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. We also plan 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

Banerjee, Satanjeev 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.

Chiao, Yun-Chuang 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.

Chiao, Yun-Chuang 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.

Cho, Miyoung, 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.

Choi, Dongjin, 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).

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

Gaussier, Éric, 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.

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

Hazem, Amir 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.

Hwang, Myunggwon, 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.

Laroche, Audrey 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.

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

Montague, Mark 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.

Morin, Emmanuel and Béatrice Daille. 2006. Comparabilité de corpus et fouille terminologique multilingue. In *Traitement Automatique des Langues (TAL)*.

Morin, Emmanuel 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.

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

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

Prochasson, Emmanuel, 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.

Rapp, Reinhard. 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.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526. Association for Computational Linguistics.

Shi, Lei. 2009. Adaptive web mining of bilingual lexicons for cross language information retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1561–1564, New York, NY, USA. ACM.

Wu, Zhibiao 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.