Using WordNet and Semantic Similarity for Bilingual Terminology Mining 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 extension of the standard approach used for bilingual lexicon extraction from comparable corpora. We study of the ambiguity problem revealed by the seed bilingual dictionary used to translate context vectors. For this purpose, we augment the standard approach by a Word Sense Disambiguation process relying on a WordNet-based semantic similarity measure. The aim of this process is to identify the translations that are more likely to give the best representation of words in the target language. On two specialized French-English comparable corpora, empirical experimental results show that the proposed method consistently outperforms the standard approach.

1 Introduction

Bilingual lexicons play a vital role in many Natural Language Processing applications such as Machine Translation (Och and Ney, 2003) 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; Chiao and Zweigenbaum, 2003). These corpora are comprised of texts which are not exact translation of each other but share common features such as domain, genre, sampling period, etc.

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 specialized French-English comparable cor-

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 16–23,
Sofia, Bulgaria, August 8, 2013.

pora *(financial and medical)* and report improved results, especially when many of the words in the corpus are ambiguous.

The remainder of the paper is organized as follows: Section 2 presents the standard approach and recalls in some details previous work addressing the task of bilingual lexicon extraction from comparable corpora. In section 3 we present our context disambiguation process. Before concluding and presenting directions for future work, we describe in section 4 the experimental protocol we followed and discuss the obtained results.

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. Prochasson et al. (2009) 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.

A large number of Word Sense Disambiguation 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 resource groups English words into sets of synonyms called *synsets*, provides short, general definitions and records various semantic relations (hypernymy, meronymy, etc.) between these synonym sets. 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.

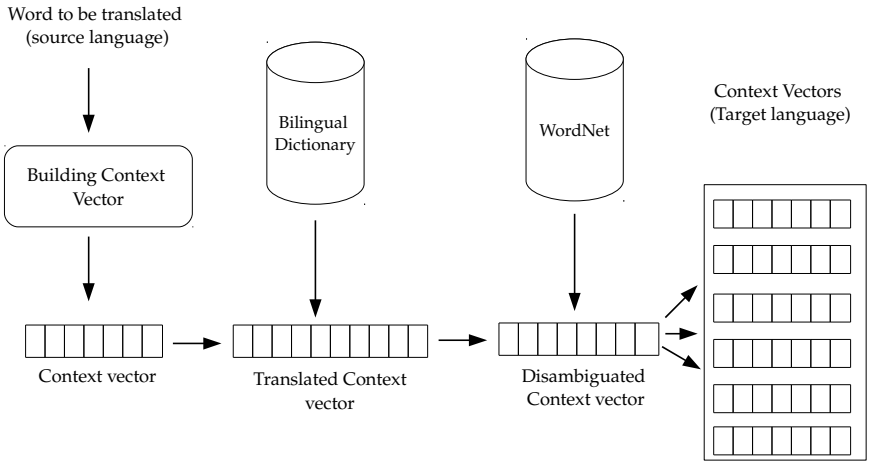

Figure 1: Overall architecture of the lexical extraction approach

3 Context Vector Disambiguation

The approach we propose includes the three steps of the standard approach. As it was 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 WordNet-based WSD process that aims at improving the adequacy of context vectors and therefore improve the results of the standard approach. Figure 1 shows the overall architecture of the lexical extraction 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.

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.

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

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:

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

Context Vector	Translations	Comparison	Ave_Sim
liquidité	liquidity	–	–
action	act	Sem_{Sim}(act,liquidity), Sem_{Sim}(act,dividend)	0.2139
	action	Sem_{Sim}(action,liquidity), Sem_{Sim}(action,dividend)	0.4256
	stock	Sem_{Sim}(stock,liquidity), Sem_{Sim}(stock,dividend)	***0.5236***
	deed	Sem_{Sim}(deed,liquidity), Sem_{Sim}(deed,dividend)	0.1594
	lawsuit	Sem_{Sim}(lawsuit,liquidity), Sem_{Sim}(lawsuit,dividend)	0.1212
	fact	Sem_{Sim}(fact,liquidity), Sem_{Sim}(fact,dividend)	0.1934
	operation	Sem_{Sim}(operation,liquidity), Sem_{Sim}(operation,dividend)	0.2045
	share	Sem_{Sim}(share,liquidity), Sem_{Sim}(share,dividend)	***0.5236***
	plot	Sem_{Sim}(plot,liquidity), Sem_{Sim}(plot,dividend)	0.2011
dividende	dividend	–	–

Table 1: Disambiguation of the context vector of the French term *bénéfice [income]* in the *corporate finance* domain. *liquidité* and *dividende* are monosemic and are used to infer the most similar translations of the term *action*.

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:

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

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.

In Table 1, we present the results of the disambiguation process for the context vector of the French term *bénéfice* in the *corporate finance* corpus. 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 average similariy values (Ave_Sim), 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.

Corpus	French	English
Corporate finance	402,486	756,840
Breast cancer	396,524	524,805

Table 2: Comparable corpora sizes in term of words.

4 Experiments and Results

4.1 Resources

4.1.1 Comparable corpora

We conducted our experiments on two French-English comparable corpora specialized on the *corporate finance* and the *breast cancer* domains. Both corpora were extracted from Wikipedia[1]. We consider the topic in the source language (for instance *finance des entreprises* [corporate finance]) as a query to Wikipedia and extract all its sub-topics (i.e., sub-categories in Wikipedia) to construct a domain-specific *category tree*. A sample of the *corporate finance* sub-domain's category tree is shown in Figure 2. 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 2.

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

[2] Comparable corpora will be shared publicly

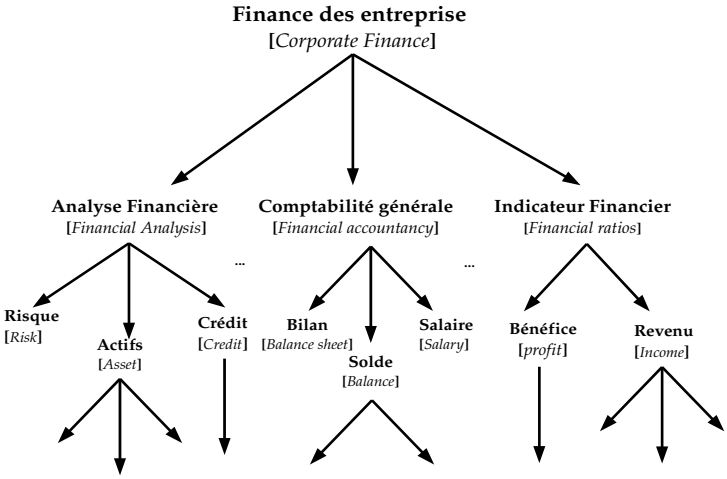

Figure 2: Wikipedia categories tree of the *corporate finance* sub-domain.

4.1.2 Bilingual dictionary

The 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 on both corpora has on average, 7 translations in the bilingual dictionary.

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 two reference lists[3] for the *corporate finance* and the *breast cancer* domains. The first list is composed of 125 single terms extracted from the glossary of bilingual micro-finance terms[4]. The second list contains 96 terms extracted from the French-English MESH and the UMLS thesauri[5]. 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, we use the results of the standard approach (SA) described in section 2.1 as a reference. We evaluate the performance of both the SA and ours with respect to Top$N$ precision (P$_N$), recall (R_N) and Mean Reciprocal Rank (MRR) (Voorhees, 1999). 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

[3]Reference lists will be shared publicly
[4]http://www.microfinance.lu/en/
[5]http://www.nlm.nih.gov/

a) Corporate Finance

Method	P1	P10	P20	R1	R10	R20	MRR
Standard Approach (SA)	0.046	0.140	0.186	0.040	0.120	0.160	0.064
WN-T_1	0.065	0.196	0.261	0.056	0.168	0.224	0.089
WN-T_2	0.102	**0.252**	0.308	0.080	**0.216**	0.264	0.122
WN-T_3	0.102	0.242	**0.327**	0.088	0.208	**0.280**	0.122
WN-T_4	**0.112**	0.224	0.299	**0.090**	0.190	0.250	**0.124**
WN-T_5	0.093	0.205	0.280	0.080	0.176	0.240	0.110
WN-T_6	0.084	0.205	0.233	0.072	0.176	0.200	0.094
WN-T_7	0.074	0.177	0.242	0.064	0.152	0.208	0.090

b) Breast Cancer

Method	P1	P10	P20	R1	R10	R20	MRR
Standard Approach (SA)	0.342	0.542	0.585	0.250	0.395	0.427	0.314
WN-T_1	0.257	0.500	0.571	0.187	0.364	0.416	0.257
WN-T_2	0.314	0.614	**0.671**	0.229	0.447	**0.489**	0.313
WN-T_3	0.342	**0.628**	**0.671**	0.250	**0.458**	**0.489**	0.342
WN-T_4	0.342	0.571	0.642	0.250	0.416	0.468	0.332
WN-T_5	**0.357**	0.571	0.657	**0.260**	0.416	0.479	**0.348**
WN-T_6	0.357	0.571	0.652	0.260	0.416	0.468	0.347
WN-T_7	0.357	0.585	0.657	0.260	0.427	0.479	0.339

Table 3: Precision, Recall at Top$N$ (N=1,10,20) and MRR at Top20 for the two domains. In each column, bold show best results. Underline show best results overall.

the ratio of correct translation to the total number of terms. The MRR takes into account the rank of the first good translation found for each entry. Formally, it is defined as:

$$MRR = \frac{1}{Q} \sum_{|Q|}^{i=1} \frac{1}{rank_i} \quad (5)$$

where Q is the total number of terms to be translated and $rank_i$ is the position of the first correct translation in the translations candidates.

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 (*share* and *stock* in Table 1). 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 both corpora have on average 7 translations in the bilingual dictionary. The baseline (SA) uses all translations associated to each entry in the bilingual dictionary. Table 3a displays the results obtained for the *corporate finance* corpus. The first substantial observation is that our method which consists in disambiguating polysemous words within context vectors consistently outperforms the standard approach (SA) for all configurations. The best MRR is reported when for each polysemous word, we keep the most similar four translations (WN-T_4) in the context vector of the term to be translated. However, the highest Top20 precision and recall are obtained by WN-T_3. Using the top three word translations in the vector boosts the Top20 precision from 0.186 to 0.327 and the Top20 recall from 0.160 to 0.280. Concerning the Breast Cancer corpus, slightly different results were obtained. As Table 3b show, when the context vectors are totally disambiguated (i.e. each source unit is translated by at most one word in context vectors), all Top$N$ precision, recall and MRR decrease. However, we report improvements against the SA in most other cases. For WN-T_5, we obtain the maximum MRR score with an improvement of $+0.034$ over the SA. But, as for the *corporate finance* corpus, the best Top20 precision and recall are reached by the WN-T_3 method, with a gain of $+0.082$ in both Top10 and Top20 precision and of about $+0.06$ in Top10 and Top20 recall.

From observing result tables of both *corporate finance* and *breast cancer* domains, we notice that our approach performs better than the SA but with different degrees. The improvements achieved in

Corpus	Corpus P_R	Vectors P_R
Corporate finance	41%	$91,6\%$
Breast cancer	47%	$85,1\%$

Table 4: Comparable corpora's and context vector's Polysemy Rates P_R.

the *corporate finance* domain are higher than those reported in the *breast cancer* domain. The reason being that the vocabulary used in the *breast cancer* corpus is more specific and therefore less ambiguous than that used in *corporate finance* texts. The results given in table 4 validate this assumption. In this table, we give the polysemy rates of the comparable corpora (Corpus P_R) and that of context vectors (Vectors P_R). P_R indicates the percentage of words that are associated to more than one translation in the bilingual dictionary. The results show that *breast cancer* corpus is more polysemic than that of the *corporate finance*. Nevertheless, even if in both corpora, the candidates' context vectors are highly polysemous, *breast cancer*'s context vectors are less polysemous than those of the *corporate finance* texts. In this corpus, $91,6\%$ of the words used as entries to define context vectors are polysemous. This shows that the ambiguity present in specialized comparable corpora hampers bilingual lexicon extraction, and that disambiguation positively affects the overall results. Even though the two corpora are fairly different (subject and polysemy rate), the optimal Top20 precision and recall results are obtained when considering up to three most similar translations in context vectors. This behavior shows that the disambiguation method is relatively robust to domain change. We notice also that the addition of supplementary translations, which are probably noisy in the given domain, degrades the overall results but remains greater than the SA.

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 two highly polysemous specialized comparable corpora 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. In addition to the metric defined by (Wu and Palmer, 1994), we plan to apply other semantic similarity and relatedness measures and compare their performance. 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 Morin and Béatrice Daille. 2006. Comparabilité de corpus et fouille terminologique multilingue. In *Traitement Automatique des Langues (TAL)*.

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.

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

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.

Reinhard Rapp. 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.

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.

Lei Shi. 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.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *In Proceedings of TREC-8*, pages 77–82.

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.