Bilingual Lexicon Extraction from Comparable Corpora as Metasearch

Emmanuel Morin, Amir Hazem, Sebastián Peña Saldarriaga

Bilingual Lexicon Extraction from Comparable Corpora as Metasearch

Amir Hazem and Emmanuel Morin
Université de Nantes,
LINA - UMR CNRS 6241
2 rue de la Houssinière,
BP 92208 44322 Nantes Cedex 03
amir.hazem@univ-nantes.fr
emmanuel.morin@univ-nantes.fr

Sebastian Peña Saldarriaga
1100 rue Notre-Dame Ouest,
Montréal, Québec,
Canada H3C 1K3
spena@synchromedia.ca

Abstract

In this article we present a novel way of looking at the problem of automatic acquisition of pairs of translationally equivalent words from comparable corpora. We first present the standard and extended approaches traditionally dedicated to this task. We then reinterpret the extended method, and motivate a novel model to reformulate this approach inspired by the metasearch engines in information retrieval. The empirical results show that performances of our model are always better than the baseline obtained with the extended approach and also competitive with the standard approach.

1 Introduction

Bilingual lexicon extraction from comparable corpora has received considerable attention since the 1990s (Rapp, 1995; Fung, 1998; Fung and Lo, 1998; Peters and Picchi, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002a; Déjean et al., 2002; Gaussier et al., 2004; Morin et al., 2007; Laroche and Langlais, 2010, among others). This attention has been motivated by the scarcity of parallel corpora, especially for countries with only one official language and for language pairs not involving English. Furthermore, as a parallel corpus is comprised of a pair of texts (a source text and a translated text), the vocabulary appearing in the translated text is highly influenced by the source text, especially in technical domains. Consequently, comparable corpora are considered by human translators to be more trustworthy than parallel corpora (Bowker and Pearson, 2002). Comparable corpora are clearly of use in the enrichment of bilingual dictionaries and thesauri (Chiao and Zweigenbaum, 2002b; Déjean et al., 2002), and in the improvement of cross-language information retrieval (Peters and Picchi, 1998).

According to (Fung, 1998), bilingual lexicon extraction from comparable corpora can be approached as a problem of information retrieval (IR). In this representation, the query would be the word to be translated, and the documents to be found would be the candidate translations of this word. In the same way that as documents found, the candidate translations are ranked according to their relevance (i.e. a document that best matches the query). More precisely, in the *standard approach* dedicated to bilingual lexicon extraction from comparable corpora, a word to be translated is represented by a vector context composed of the words that appear in its lexical context. The candidate translations for a word are obtained by comparing the translated source context vector with the target context vectors through a general bilingual dictionary. Using this approach, good results on single word terms (SWTs) can be obtained from large corpora of several million words, with an accuracy of about 80% for the top 10-20 proposed candidates (Fung and McKeown, 1997; Rapp, 1999). Cao and Li (2002) have achieved 91% accuracy for the top three candidates using the Web as a comparable corpus. Results drop to 60% for SWTs using specialized small size language corpora (Chiao and Zweigenbaum, 2002a; Déjean and Gaussier, 2002; Morin et al., 2007).

In order to avoid the insufficient coverage of the bilingual dictionary required for the translation of source context vectors, an *extended approach* has

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 35–43,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

been proposed (Déjean et al., 2002; Daille and Morin, 2005). This approach can be seen as a query reformulation process in IR for which similar words are substituted for the word to be translated. These similar words share the same lexical environments as the word to be translated without appearing with it. With the extended approach, (Déjean et al., 2002) obtained for single French-English words 43% and 51% precision out of the ten and twenty first candidates applied to a medical corpus of 100 000 words (respectively 44% and 57% with the standard approach) and 79% and 84% precision on the ten and twenty first candidates applied to a social science corpus of 8 million words (respectively 35% and 42% with the standard approach). Within this context, we want to show how metasearch engines can be used for bilingual lexicon extraction from specialized comparable corpora. In particular, we will focus on the use of different strategies to take full advantage of similar words.

The remainder of this paper is organized as follows. Section 2 presents the standard and extended approaches based on lexical context vectors dedicated to word alignment from comparable corpora. Section 3 describes our metasearch approach that can be viewed as the combination of different search engines. Section 4 describes the different linguistic resources used in our experiments and evaluates the contribution of the metasearch approach on the quality of bilingual terminology extraction through different experiments. Finally, Section 5 presents our conclusions.

2 Related Work

In this section, we first describe the standard approach dedicated to word alignment from comparable corpora. We then present an extension of this approach.

2.1 Standard Approach

The main work in bilingual lexicon extraction from comparable corpora is based on lexical context analysis and relies on the simple observation that a word and its translation tend to appear in the same lexical contexts. The basis of this observation consists in the identification of first-order affinities for each source and target language: *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word* (Grefenstette, 1994a, p. 279). These affinities can be represented by context vectors, and each vector element represents a word which occurs within the window of the word to be translated (for instance a seven-word window approximates syntactical dependencies).

The implementation of this approach can be carried out by applying the following four steps (Rapp, 1995; Fung and McKeown, 1997):

Context characterization

All the lexical units in the context of each lexical unit i are collected, and their frequency in a window of n words around i extracted. For each lexical unit i of the source and the target languages, we obtain a context vector $\mathbf{i}$ where each entry, $\mathbf{i}_j$, of the vector is given by a function of the co-occurrences of units j and i. Usually, association measures such as the mutual information (Fano, 1961) or the log-likelihood (Dunning, 1993) are used to define vector entries.

Vector transfer

The lexical units of the context vector $\mathbf{i}$ are translated using a bilingual dictionary. Whenever the bilingual dictionary provides several translations for a lexical unit, all the entries are considered but weighted according to their frequency in the target language. Lexical units with no entry in the dictionary are discarded.

Target language vector matching

A similarity measure, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, is used to score each lexical unit, t, in the target language with respect to the translated context vector, $\bar{\mathbf{i}}$. Usual measures of vector similarity include the cosine similarity (Salton and Lesk, 1968) or the weighted jaccard index (WJ) (Grefenstette, 1994b) for instance.

Candidate translation

The candidate translations of a lexical unit are the target lexical units ranked following the similarity score.

2.2 Extended Approach

The main shortcoming of the standard approach is that its performance greatly relies on the coverage of the bilingual dictionary. When the context vectors

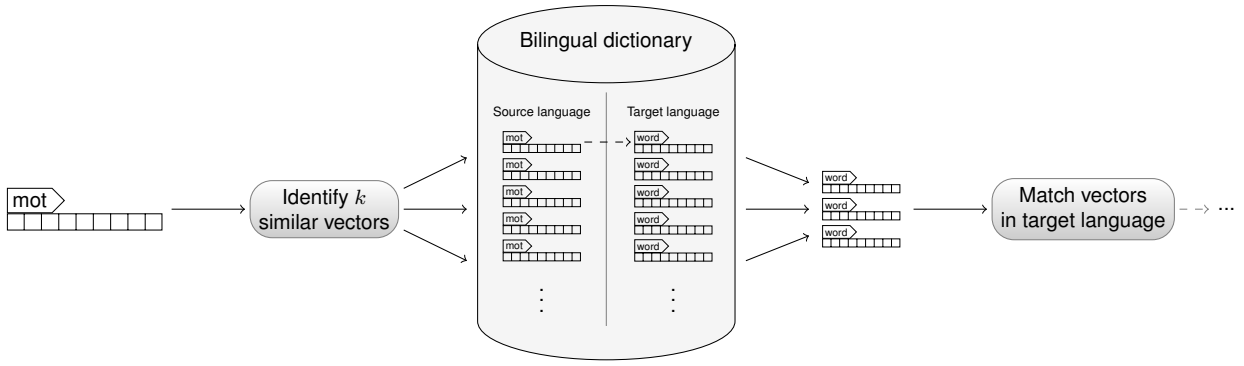

Figure 1: Illustration of the extended approach.

are well translated, the translation retrieval rate in the target language improves.

Although, the coverage of the bilingual dictionary can be extended by using specialized dictionaries or multilingual thesauri (Chiao and Zweigenbaum, 2003; Déjean et al., 2002), translation of context vectors remains the core of the approach.

In order to be less dependent on the coverage of the bilingual dictionary, Déjean and Gaussier (2002) have proposed an extension to the standard approach. The basic intuition of this approach is that words sharing the same meaning will share the same environments. The approach is based on the identification of second-order affinities in the source language: *Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar* (Grefenstette, 1994a, p. 280).

Generally speaking, a bilingual dictionary is a bridge between two languages established by its entries. The extended approach is based on this observation and avoids explicit translation of vectors as shown in Figure 1. The implementation of this extended approach can be carried out in four steps where the first and last steps are identical to the standard approach (Déjean and Gaussier, 2002; Daille and Morin, 2005):

Reformulation in the target language

For a lexical unit i to be translated, we identify the k-nearest lexical units (k nlu), among the dictionary entries corresponding to words in the source language, according to $\mathrm{sim}(\mathbf{i}, \mathbf{s})$. Each nlu is translated via the bilingual dictionary, and the vector in the target language, $\overline{\mathbf{s}}$, corresponding to the translation is selected. If the bilingual dictionary provides several translations for a given unit, $\overline{\mathbf{s}}$ is given by the union of the vectors corresponding to the translations. It is worth noting that the context vectors are not translated directly, thus reducing the influence of the dictionary.

Vector matching against reformulations

The similarity measure, $\mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t})$, is used to score each lexical unit, t, in the target language with respect to the k nlu. The final score assigned to each unit, t, in the target language is given by:

$$\mathrm{sim}(\mathbf{i}, \mathbf{t}) = \sum_{s \in k\mathrm{NLU}} \mathrm{sim}(\mathbf{i}, \mathbf{s}) \times \mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t}) \quad (1)$$

An alternate scoring function has been proposed by Daille and Morin (2005). The authors computed the centroid vector of the k nlu, then scored target units with respect to the centroid.

3 The Metasearch Approach

3.1 Motivations

The approach proposed by Déjean and Gaussier (2002) implicitly introduces the problem of selecting a good k. Generally, the best choice of k depends on the data. Although several heuristic techniques, like cross-validation, can be used to select a good value of k, it is usually defined empirically.

The application of the extended approach (EA) to our data showed that the method is unstable with respect to k. In fact, for values of k over 20, the precision drops significantly. Furthermore, we cannot ensure result stability within particular ranges of

values. Therefore, the value of k should be carefully tuned.

Starting from the intuition that each nearest lexical unit (nlu) contributes to the characterization of a lexical unit to be translated, our proposition aims at providing an algorithm that gives a better precision while ensuring higher stability with respect to the number of nlu. Pushing the analogy of IR style approaches (Fung and Lo, 1998) a step further, we propose a novel way of looking at the problem of word translation from comparable corpora that is conceptually simple: a metasearch problem.

In information retrieval, metasearch is the problem of combining different ranked lists, returned by multiple search engines in response to a given query, in such a way as to optimize the performance of the combined ranking (Aslam and Montague, 2001). Since the k nlu result in k distinct rankings, metasearch provides an appropriate framework for exploiting information conveyed by the rankings.

In our model, we consider each list of a given nlu as a response of a search engine independently from the others. After collecting all the lists of the selected nlu's, we combine them to obtain the final similarity score. It is worth noting that all the lists are normalized to maximize in such a way the contribution of each nlu. A good candidate is the one that obtains the highest similarity score which is calculated with respect to the selected k. If a given candidate has a high frequency in the corpus, it may be similar not only to the selected nearest lexical units (k), but also to other lexical units of the dictionary. If the candidate is close to the selected nlu's and also close to other lexical units, we consider it as a potential noise (the more neighbours a candidate has, the more it's likely to be considered as noise). We thus weight the similarity score of a candidate by taking into account this information. We compare the distribution of the candidate with the k nlu and also with all its neighbours. This leads us to suppose that a good candidate should be closer to the selected nlu's than the rest of its neighbours, if it's not the case there is more chances for this candidate to be a wrong translation.

3.2 Proposed Approach

In the following we will describe our extension to the method proposed by Déjean and Gaussier (2002). The notational conventions adopted are reviewed in Table 1. Elaborations of definitions will be given when the notation is introduced. In all our experiments both terms and lexical units are single words.

Symbol	Definition
l	a list of a given lexical unit.
k	the number of selected nearest lexical units (lists).
$freq(w,k)$	the number of lists (k) in which a term appears.
n	all the neighbours of a given term.
u	all the lexical units of the dictionary.
w_l	a term of a given list l.
$\mathrm{s}(w_l)$	the score of the term w in the list l.
$\max_l$	the maximum score of a given list l.
$\max_{All}$	the maximum score of all the lists.
$\mathrm{s}_{norm}(w_l)$	the normalized score of term w in the list l.
$\mathrm{s}(w)$	the final score of a term w.
θ_w	the regulation parameter of the term w.

Table 1: Notational conventions.

The first step of our method is to collect each list of each nlu. The size of the list has its importance because it determines how many candidates are close to a given nlu. We noticed from our experiments that, if we choose lists with small sizes, we should lose information and if we choose lists with large sizes, we could keep more information than necessary and this should be a potential noise, so we consider that a good size of each list should be between 100 and 200 terms according to our experiments.

After collecting the lists, the second step is to normalize the scores. Let us consider the equation 2 :

$$\mathrm{s}_{norm}(w_l) = \mathrm{s}(w_l) \times \frac{\max_l}{\max_{All}} \qquad (2)$$

We justify this by a rationale derived from two observations. First, scores in different rankings are compatible since they are based on the same similarity measure (i.e., on the same scale). The second observations follows from the first: if $\max(l) \gg$

$\max(m)$, then the system is more confident about the scores of the list l than m.

Using scores as fusion criteria, we compute the similarity score of a candidate by summing its scores from each list of the selected nlu's :

$$\mathrm{s}(w) = \theta_w \times \frac{\sum_{l=1}^{k} \mathrm{s}_{norm}(w_l)}{\sum_{l=1}^{n} \mathrm{s}_{norm}(w_l)} \quad (3)$$

the weight θ is given by :

$$\theta_w = freq(w,k) \times \frac{(u - (k - freq(w,k)))}{(u - freq(w,n))} \quad (4)$$

The aim of this parameter is to give more confidence to a term that occurs more often with the selected nearest neighbours (k) than the rest of its neighbours. We can not affirm that the best candidate is the one that follows this idea, but we can nevertheless suppose that candidates that appear with a high number of lexical units are less confident and have higher chances to be wrong candidates (we can consider those candidates as noise). So, θ allows us to regulate the similarity score, it is used as a confident weight or a regulation parameter. We will refer to this model as the multiple source (MS) model. We also use our model without using θ and refer to it by (LC), this allows us to show the impact of θ in our results.

4 Experiments and Results

4.1 Linguistic Resources

We have selected the documents from the Elsevier website[1] in order to obtain a French-English specialized comparable corpus. The documents were taken from the medical domain within the subdomain of 'breast cancer'. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term 'cancer du sein' in French and 'breast cancer' in English. We thus collected 130 documents in French and 118 in English and about 530,000 words for each language. The documents comprising the French/English specialized comparable corpus have been normalized through the following linguistic pre-processing steps: tokenisation, part-of-speech tagging, and lemmatisation. Next, the function words were removed and the words occurring less than twice (i.e. hapax) in the French and the English parts were discarded. Finally, the comparable corpus comprised about 7,400 distinct words in French and 8,200 in English.

The French-English bilingual dictionary required for the translation phase was composed of dictionaries that are freely available on the Web. It contains, after linguistic pre-processing steps, 22,300 French single words belonging to the general language with an average of 1.6 translations per entry.

In bilingual terminology extraction from specialized comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs are often composed of 100 single-word terms (SWTs) (180 SWTs in (Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002a), and 100 SWTs in (Daille and Morin, 2005)). To build our reference list, we selected 400 French/English SWTs from the UMLS[2] meta-thesaurus and the *Grand dictionnaire terminologique*[3]. We kept only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 122 French/English SWTs were extracted.

4.2 Experimental Setup

Three major parameters need to be set to the extended approach, namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. Laroche and Langlais (2010) carried out a complete study about the influence of these parameters on the quality of bilingual alignment.

As similarity measure, we chose to use the weighted jaccard index:

$$\mathrm{sim}(\mathbf{i},\mathbf{j}) = \frac{\sum_t \min(\mathbf{i}_t, \mathbf{j}_t)}{\sum_t \max(\mathbf{i}_t, \mathbf{j}_t)} \quad (5)$$

The entries of the context vectors were determined by the log-likelihood (Dunning, 1993), and we used a seven-word window since it approximates syntactic dependencies. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

[1] www.elsevier.com

[2] http://www.nlm.nih.gov/research/umls

[3] http://www.granddictionnaire.com/

4.3 Results

To evaluate the performance of our method, we use as a baseline, the extended approach (EA) proposed by Déjean and Gaussier (2002). We compare this baseline to the two metasearch strategies defined in Section 3: the metasearch model without the regulation parameter θ (LC); and the one which is weighted by $theta$ (MS). We also provide results obtained with the standard approach (SA).

We first investigate the stability of the metasearch strategies with respect to the number of nlu considered. Figure 2 show the precision at Top 20 as a function of k.

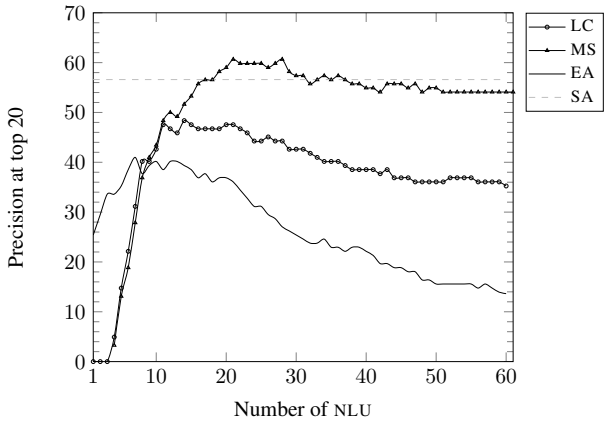

Figure 2: Precision at top 20 as a function of the number of nlu.

In order to evaluate the contribution of the parameter θ, we chose to evaluate the metasearch method starting from $k = 4$, this explains why the precision is extremely low for low values of k. We further considered that less than four occurrences of a term in the whole lexical units lists can be considered as noise. On the other side, we started from $k = 1$ for the extended approach since it makes no use of the parameter θ. Figure 2 shows that extended approach reaches its best performance at $k = 7$ with a precision of 40.98%. Then, after $k = 15$ the precision starts steadily decreasing as the value of k increases.

The metasearch strategy based only on similarity scores shows better results than the baseline. For every value of $k \geq 10$, the LC model outperform the extended approach. The best precision (48.36%) is obtained at $k = 14$, and the curve corresponding to the LC model remains above the baseline regardless of the increasing value of the parameter k. The curve corresponding to the MS model is always above the (EA) for every value of $k \geq 10$. The MS model consistently improves the precision, and achieves its best performance (60.65%) at $k = 21$.

We can notice from Figure 2 that the LC and MS models outperform the baseline (EA). More importantly, these models exhibit a better stability of the precision with respect to the k-nearest lexical units. Although the performance decrease as the value of k increases, it does not decrease as fast as in the baseline approach.

For the sake of comparability, we also provide results obtained with the standard approach (SA) (56.55%) represented by a straight line as it is not dependent on k. As we can see, the metasearch approach (MS) outperforms the standard approach for values of k bertween 20 and 30 and for greater values of k the precision remains more or less almost the same as the standard approach (SA). Thus, the metasearch model (MS) can be considered as a competitive approach regarding to its results as it is shown in the figure 2.

Finally, Figure 3 shows the contribution of each nlu taken independently from the others. This confirms our intuition that each nlu contribute to the characterization of a lexical unit to be translated, and supports our idea that their combination can improve the performances.

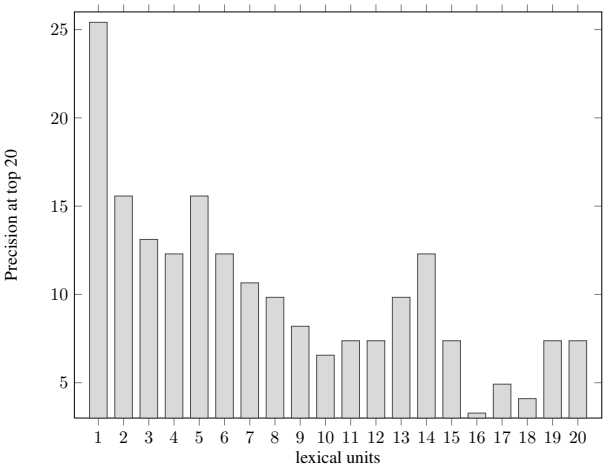

Figure 3: Precision at top 20 for each of the 20 nlu. The precision is computed by taking the each nlu independently from the others.

Figure 3 shows the top 20 of each nlu. Notice

that the nlu are ordered from the most similar to the lexical unit to be translated to the less similar, and that each one of the nearest lexical units contains information that it is worth taking into account.

Although each nlu can only translate few terms, by using the metasearch idea we are able to improve the retrieval of translation equivalents. The main idea of the metasearch paradigm is to take into account the information conveyed by all the k nlu, using either similarity scores, their behaviour with all the neighbours, in order to improve the performance of the alignment process.

Although significant improvements can be obtained with the metasearch models (comparatively to the EA and SA approach), especially concerning precision stability with respect to the k nlu, we believe that we need to address the estimation of k beforehand. Rather than fixing the same k for all the units to be translated, there is the possibility to adapt an optimal value of k to each lexical unit, according to some criteria which have to be determined.

Approachs	Top 5	Top 10	Top 15	Top 20
SA	37.70	45.08	52.45	56.55
EA	21.31	31.14	36.88	40.98
MS	**40.98**	**54.91**	**56.55**	**60.65**

Table 2: Precision(%) at top 5, 10, 15, 20 for SA, EA and MS.

Finally, we present in table 2 a comparison between SA, EA and MS for the top 5, 10, 15 and 20. By choosing the best configuration of each method, we can note that our method outperforms the others in each top. In addition, for the top 10 our precision is very close to the precision of the standard approach (SA) at the top 20. we consider these results as encouraging for future work.

4.4 Discussion

Our experiments show that the parameter k remains the core of both EA and MS approaches. A good selection of the nearest lexical units of a term guarantee to find the good translation. It is important to say that EA and MS which are based on the k nlu's depends on the coverage of the terms to be translated. Indeed, these approaches face three cases : firstly, if the frequency of the word to be translated is high and the frequency of the good translation in the target language is low, this means that the nearest lexical units of the candidate word and its translation are unbalanced. This leads us to face a lot of noise because of the high frequency of the source word that is over-represented by its nlu's comparing to the target word which is under-represented. Secondly, we consider the inverse situation, which is: low frequency of the source word and high frequency of the target translation, here as well, we have both the source and the target words that are unbalanced regarding to the selected nearest lexical units. The third case, represents more or less the same distribution of the frequencies of source candidate and target good translation. This can be considered as the most appropriate case to find the good translation by applying the approaches based on the nlu's (EA or MS). Our experiments show that our method works well in all the cases by using the parameter θ which regulate the similarity score by taken into account the distribution of the candidate according to both : selected nlu's and all its neighbours. In resume, words to be translated as represented in case one and two give more difficulties to be translated because of their unbalanced distribution which leads to an unbalanced nlu's. Future works should confirm the possibility to adapt an optimal value of k to each candidate to be translated, according to its distribution with respect to its neighbours.

5 Conclusion

We have presented a novel way of looking at the problem of bilingual lexical extraction from comparable corpora based on the idea of metasearch engines. We believe that our model is simple and sound. Regarding the empirical results of our proposition, performances of the multiple source model on our dataset was better than the baseline proposed by Déjean and Gaussier (2002), and also outperforms the standard approach for a certain range of k. We believe that the most significant result is that a new approach to finding single word translations has been shown to be competitive. We hope that this new paradigm can lead to insights that would be unclear in other models. Preliminary tests in this perspective show that using an appropriate value of k for each word can improve the performance of the lexical extraction process. Dealing with this prob-

lem is an interesting line for future research.

6 Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-08-CORD-013.

References

Javed A. Aslam and Mark Montague. 2001. Models for Metasearch. In *SIGIR '01, proceedings of the 24th Annual SIGIR Conference*, pages 276–284.

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge, London/New York.

Yunbo Cao and Hang Li. 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 127–133, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002a. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002b. Looking for French-English Translations in Comparable Medical Corpora. *Journal of the American Society for Information Science*, 8:150–154.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Pascale Fung and Yuen Yee Lo. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, pages 414–420.

Pascale Fung and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From ParallelCorpora to Non-parallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Éric Gaussier, Jean-Michel Renders, Irena Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Gregory Grefenstette. 1994a. Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, pages 279–290, Amsterdam, The Netherlands.

Gregory Grefenstette. 1994b. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

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 (ACL'99)*, pages 519–526, College Park, MD, USA.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.