Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora

Audrey Laroche
OLST – Dép. de linguistique et de traduction
Université de Montréal
audrey.laroche@umontreal.ca

Philippe Langlais
RALI – DIRO
Université de Montréal
felipe@iro.umontreal.ca

Abstract

Context-based projection methods for identifying the translation of terms in comparable corpora has attracted a lot of attention in the community, *e.g.* (Fung, 1998; Rapp, 1999). Surprisingly, none of those works have systematically investigated the impact of the many parameters controlling their approach. The present study aims at doing just this. As a test-case, we address the task of translating terms of the medical domain by exploiting pages mined from Wikipedia. One interesting outcome of this study is that significant gains can be obtained by using an association measure that is rarely used in practice.

1 Introduction

Identifying translations of terms in comparable corpora is a challenge that has attracted many researchers. A popular idea that emerged for solving this problem is based on the assumption that the context of a term and its translation share similarities that can be used to rank translation candidates (Fung, 1998; Rapp, 1999). Many variants of this idea have been implemented.

While a few studies have investigated pattern matching approaches to compare source and target contexts (Fung, 1995; Diab and Finch, 2000; Yu and Tsujii, 2009), most variants make use of a bilingual lexicon in order to translate the words of the context of a term (often called *seed words*). Déjean et al. (2005) instead use a bilingual thesaurus for translating these.

Another distinction between approaches lies in the way the context is defined. The most common practice, the so-called window-based approach, defines the context words as those cooccuring significantly with the source term within windows centered around the term.[1] Some studies have reported gains by considering syntactically motivated co-occurrences. Yu and Tsujii (2009) propose a resource-intensive strategy which requires both source and target dependency parsers, while Otero (2007) investigates a lighter approach where a few hand coded regular expressions based on POS tags simulate source parsing. The latter approach only requires a POS tagger of the source and the target languages as well as a small parallel corpus in order to project the source regular expressions.

Naturally, studies differ in the way each co-occurrence (either window or syntax-based) is weighted, and a plethora of association scores have been investigated and compared, the likelihood score (Dunning, 1993) being among the most popular. Also, different similarity measures have been proposed for ranking target context vectors, among which the popular cosine measure.

The goal of the different authors who investigate context-projection approaches also varies. Some studies are tackling the problem of identifying the translation of general words (Rapp, 1999; Otero, 2007; Yu and Tsujii, 2009) while others are addressing the translation of domain specific terms. Among the latter, many are translating single-word terms (Chiao and Zweigenbaum, 2002; Déjean et al., 2005; Prochasson et

[1] A stoplist is typically used in order to prevent function words from populating the context vectors.

al., 2009), while others are tackling the translation of multi-word terms (Daille and Morin, 2005). The type of discourse might as well be of concern in some of the studies dedicated to bilingual terminology mining. For instance, Morin et al. (2007) distinguish popular science versus scientific terms, while Saralegi et al. (2008) target popular science terms only.

The present discussion only focuses on a few number of representative studies. Still, it is already striking that a direct comparison of them is difficult, if not impossible. Differences in resources being used (in quantities, in domains, etc.), in technical choices made (similarity measures, context vector computation, etc.) and in objectives (general versus terminological dictionary extraction) prevent one from establishing a clear landscape of the various approaches.

Indeed, many studies provide some figures that help to appreciate the influence of some parameters in a given experimental setting. Notably, Otero (2008) studies no less than 7 similarity measures for ranking context vectors while comparing window and syntax-based methods. Morin et al. (2007) consider both the log-likelihood and the mutual information association scores as well as the Jaccard and the cosine similarity measures.

Ideally, a benchmark on which researchers could run their translation finder would ease the comparison of the different approaches. However, designing such a benchmark that would satisfy the evaluation purposes of all the researchers is far too ambitious a goal for this contribution. Instead, we investigate the impact of some major factors influencing projection-based approaches on a task of translating 5,000 terms of the medical domain (the most studied domain), making use of French and English Wikipedia pages extracted monolingually thanks to an information retrieval engine. While the present work does not investigate all the parameters that could potentially impact results, we believe it constitutes the most complete and systematic comparison made so far with variants of the context-based projection approach.

In the remainder of this paper, we describe the projection-based approach to translation spotting in Section 2 and detail the parameters that directly influence its performance. The experimental protocol we followed is described in Section 3 and we analyze our results in Section 4. We discuss the main results in the light of previous work and propose some future avenues in Section 5.

2 Projection-based variants

The approach we investigate for identifying term translations in comparable corpora is similar to (Rapp, 1999) and many others. We describe in the following the different steps it encompasses and the parameters we are considering in the light of typical choices made in the literature.

2.1 Approach

Step 1 A comparable corpus is constructed for each term to translate. In this study, the source and target corpora are sets of Wikipedia pages related to the source term (S) and its reference translation (T) respectively (see Section 3.1). The degree of corpus preprocessing varies greatly from one study to another. Complex linguistic tools such as terminological extractors (Daille and Morin, 2005), parsers (Yu and Tsujii, 2009) or lemmatizers (Rapp, 1999) are sometimes used.

In our case, the only preprocessing that takes place is the deletion of the Wikipedia symbols pertaining to its particular syntax (*e.g.* `[[ ]]`).[2] It is to be noted that, for the sake of simplicity and generality, our implementation does not exploit interlanguage links nor structural elements specific to Wikipedia documents, as opposed to (Yu and Tsujii, 2009).

Step 2 A context vector v_s for the source term S is built (see Figure 1 for a made-up example). This vector contains the words that are in the context of the occurrences of S and are strongly correlated to S. The definition of "context" is one of the parameters whose best value we want to find. Context length can be based on a number of units, for instance 3 sentences (Daille and Morin, 2005), windows of 3 (Rapp, 1999) or 25 words (Prochasson et al., 2009), etc. It is an important parameter of the projection-based approach. Should the context length be too small, we would miss words that would be relevant in finding the translation. On the other hand, if the context is too large, it

[2] We used a set of about 40 regular expressions to do this.

might contain too much noise. At this step, a stoplist made of function words is applied in order to filter out context words and reduce noise in the context vector.

Additionally, an association measure is used to score the strength of correlation between S and the words in its contexts; it serves to normalize corpus frequencies. Words that have a high association score with S are more prominent in the context vector. The association measure is the second important parameter we want to study. As already noted, most authors use the log-likelihood ratio to measure the association between collocates; some, like (Rapp, 1999), informally compare the performance of a small number of association measures, or combine the results obtained with different association measures (Daille and Morin, 2005).

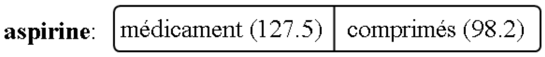

Figure 1: Step 2

Step 3 Words in v_s are projected into the target language with the help of the bilingual seed lexicon (Figure 2). Each word in v_s which is present in the bilingual lexicon is translated, and those translations define the projected context vector v_p. Words that are not found in the bilingual lexicon are simply ignored. The size of the seed lexicon and its content are therefore two important parameters of the approach. In previous studies, seed lexicons vary between 16,000 (Rapp, 1999) and 65,000 (Déjean et al., 2005) entries, a typical size being around 20,000 (Fung, 1998; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005).

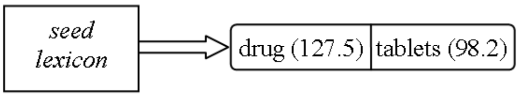

Figure 2: Step 3

Step 4 Context vectors v_t are computed for each candidate term in the target language corpus (Figure 3). The dimension of the target-vector space is defined to be the one induced by the projection mechanism described in Step 3. The context vector v_t of each candidate term is computed as in Step 2. Therefore, in Step 4, the parameters of context definition and association measure are important and take the same values as those in Step 2. Note that in this study, on top of all single terms, we also consider target bigrams as potential candidates (99.5 % of our reference target terms are composed of at most two words). As such, our method can handle complex terms (of up to two words), as opposed to most previous studies, without having to resort to a separate terminological extraction as in (Daille and Morin, 2005).

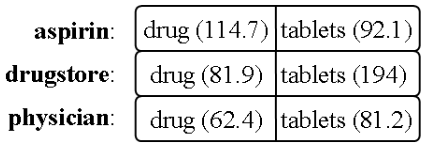

Figure 3: Step 4

Step 5 Context vectors v_t are ranked in decreasing order of their similarity with v_p (Figure 4). The similarity measure between context vectors varies among studies: city-block measure (Rapp, 1999), cosine (Fung, 1998; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Prochasson et al., 2009), Dice or Jaccard indexes (Chiao and Zweigenbaum, 2002; Daille and Morin, 2005), etc. It is among the parameters whose effect we experimentally evaluate.

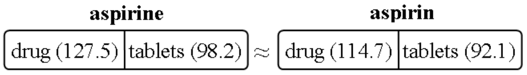

Figure 4: Step 5

2.2 Parameters studied

The five steps we described involve many parameters, the values of which can influence at varying degrees the performance of a translation spotter. In the current study, we considered the following parameter values.

Context We considered contexts defined as the current sentence or the current paragraph involving S. We also considered windows of 5 and 25 words on both sides of S.

Association measure Following the aforementioned studies, we implemented these popular measures: pointwise mutual information (PMI), log-likelihood ratio (LL) and chi-square (χ^2). We also implemented the discounted log-odds (LO) described by (Evert, 2005, p. 86) in his work on collocation mining. To our knowledge, this association measure has not been used yet in translation spotting. It is computed as:

$$\text{odds-ratio}_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})}$$

where O_{ij} are the cells of the 2×2 contingency matrix of a word token s cooccurring with the term S within a given window size.[3]

Similarity measure We implemented four measures: city-block, cosine, as well as Dice and Jaccard indexes (Jurafsky and Martin, 2008, p. 666). Our implementations of Dice and Jaccard are identical to the *DiceMin* and *JaccardMin* similarity measures reported in (Otero, 2008) and which outperformed the other five metrics he tested.

Seed lexicon We investigated the impact of both the size of the lexicon and its content. We started our study with a mixed lexicon of around 5,000 word entries: roughly 2,000 of them belong to the medical domain, while the other entries belong to the general language. We also considered mixed lexicons of 7,000, 9,000 and 11,000 entries (where 2,000 entries are related to the medical domain), as well as a 5,000-entry general language only lexicon.

2.3 Cognate heuristic

Many authors are embedding heuristics in order to improve their approach. For instance, Chiao and Zweigenbaum (2002) propose to integrate a reverse translation spotting strategy in order to improve precision. Prochasson et al. (2009) boost the strength of context words that happen to be transliterated in the other language. A somehow generalized version of this heuristic has been described in (Shao and Ng, 2004).

In this work, we examine the performance of the best configuration of parameters we found, combined with a simple heuristic based on graphic similarity between source and target terms, similar to the orthographic features in (Haghighi et al., 2008)'s generative model. This is very specific to our task where medical terms often (but not always) share Latin or Greek roots, such as *microvillosités* in French and *microvilli* in English.

In this heuristic, translation candidates which are cognates of the source term are ranked first among the list of translation candidates. In our implementation, two words are cognates if their first four characters are identical (Simard et al., 1992). One interesting note concerns the word-order mismatch typically observed in French and English complex terms, such as in *ADN mitochondrial* (French) and *mitochondrial DNA* (English). We do treat this case adequately.

3 Experimental protocol

In order to pinpoint the best configuration of values for the parameters identified in Section 2.2, four series of experiments were carried out. In all of them, the task consists of spotting translation candidates for each source language term using the resources[4] described below. The quality of the results is evaluated with the help of the metrics described in Section 3.2.

3.1 Resources

Corpora The comparable corpora are made of the (at most) 50 French and English Wikipedia documents that are the most relevant to the source term and to its reference translation respectively. These documents are retrieved with the NLGbAse Information Retrieval tool.[5] The average token count of all the 50-document corpora as well as the average frequency of the source and target terms in these corpora for our four series of experiments are listed in Table 1.

[3] For instance, O_{21} stands for the number of windows containing S but not s.

[4] Our resources are available at http://olst.ling.umontreal.ca/~audrey/coling2010/. They were acquired as described in (Rubino, 2009).

[5] http://nlgbase.org/

	Experiment			
	1	2	3	4
Tokens_s	89,431	73,809	42,762	90,328
Tokens_t	52,002	27,517	12,891	38,929
$\|S\|$	296	184	66	306
$\|T\|$	542	255	104	404

Table 1: 50-document corpora averages

The corpora are somewhat small (most corpora in previous studies are made of at least a million words). We believe this is more representative of a task where we try to translate domain specific terms. Some of the Wikipedia documents may contain a handful of parallel sentences (Smith et al., 2010), but this information is not used in our approach. The construction of the corpus involves a bias in that the reference translations are used to obtain the most relevant target language documents. However, since our objective is to compare the relative performance of different sets of parameters, this does not affect our results. In fact, as per (Déjean et al., 2005) (whose comparable corpora are English and German abstracts), the use of such an "ideal" corpus is common (as in (Chiao and Zweigenbaum, 2002), where the corpus is built from a specific query).

Seed lexicon The mixed seed lexicon we use is taken from the Heymans Institute of Pharmacology's *Multilingual glossary of technical and popular medical terms*.[6] Random general language entries from the FreeLang[7] project are also incorporated into the lexicon for some of our experiments.

Reference translations The test set is composed of 5,000 nominal single and multi-word pairs of French and English terms from the MeSH (*Medical Subject Heading*) thesaurus.[8]

3.2 Evaluation metrics

The performance of each set of parameters in the experiments is evaluated with Top N precision (P_N), recall (R_N) and F-measure (F_N), as well as Mean Average Precision (MAP). Precision is the number of correct translations (at most 1 per source term) divided by the number of terms for which our system gave at least one answer; recall is equal to the ratio of correct translations to the total number of terms. F-measure is the harmonic mean of precision and recall:

$$\text{F-measure} = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

The MAP represents in a single figure the quality of a system according to various recall levels (Manning et al., 2008, p. 147–148):

$$\text{MAP(Q)} = \frac{1}{|Q|} \sum_{|Q|}^{j=1} \frac{1}{m_j} \sum_{m_j}^{k=1} Precision(R_{jk})$$

where $|Q|$ is the number of terms to be translated, m_j is the number of reference translations for the j^{th} term (always 1 in our case), and $Precision(R_{jk})$ is 0 if the reference translation is not found for the j^{th} term or $1/r$ if it is (r is the rank of the reference translation in the translation candidates).

4 Experiments

In Experiment 1, 500 single and multi-word terms must be translated from French to English using each of the 64 possible configurations of these parameters: context definition, association measure and similarity measure. In Experiment 2, we submit to the 8 best variants 1,500 new terms to determine with greater confidence the best 2, which are again tested on the last 3,000 of the test terms (Experiment 3). In Experiment 4, using 1,350 frequent terms, we examine the effects of seed lexicon size and specificity and we apply a heuristic based on cognates.

4.1 Experiment 1

The results of the first series of experiments on 500 terms can be analysed from the point of view of each of the parameters whose values varied among 64 configurations (Section 2.2). The maximal MAP reached for each parametric value is given in Table 2.

The most notable result is that, of the four association measures studied, the log-odds ratio is

[6] http://users.ugent.be/~rvdstich/eugloss/welcome.html

[7] http://www.freelang.net/

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

	Param.	Value	Best MAP	In config.
association	LO	**0.536**	sentence_cosine
	LL	0.413	sentence_Dice
	PMI	0.299	sentence_city-block
	χ^2	0.179	sentence_Dice
similarity	cosine	**0.536**	sentence_LO
	Dice	0.520	sentence_LO
	Jaccard	0.520	sentence_LO
	city-block	0.415	sentence_LO
context	sentence	**0.536**	cosine_LO
	paragraph	0.460	cosine_LO
	25 words	0.454	cosine_LO
	5 words	0.361	Dice_LO

Table 2: Best MAP in Experiment 1

significantly superior to the others in every variant. There is as much as 34 % difference between LO and other measures for Top 1 recall. This is interesting since most previous works use the log-likelihood, and none use LO. Our best results for LO (with cosine_sentence) and LL (with Dice_sentence) are in Table 3. Note that the oracle recall is 93 % (7 % of the source and target terms were not in the corpus).

Assoc.	R1	R20	P1	P20	F1	F20	MAP
LO	**39.4**	**84.8**	**42.3**	**91.0**	**40.8**	**87.8**	**0.536**
LL	29.0	75.2	31.3	81.0	30.1	78.0	0.413

Table 3: Best LO and LL configurations scores

Another relevant observation is that the parameters interact with each other. When the similarity measure is cosine, PMI results in higher Top 1 F-scores than LL, but the Top 20 F-scores are better with LL. PMI is better than LL when using city-block as a similarity measure, but LL is better than PMI when using Dice and Jaccard indexes. χ^2 gives off the worst MAP in all but 4 of the 64 parametric configurations.

As for similarity measures, the Dice and Jaccard indexes have identical performances, in accordance with the fact that they are equivalent (Otero, 2008).[9] Influences among parameters are also observable in the performance of similarity measures. When the association measure is LO, the cosine measure gives slightly better Top 1 F-scores, while the Dice index performs slightly better with regards to Top 20 F-scores. Dice is better when the association measure is LL, with a Top 1 F-score gain of about 15 % compared to the cosine.

Again, in the case of context definitions, relative performances depend on the other parameters and on the number of top translation candidates considered. With LO, sentence contexts have the highest Top 1 F-measures, while Top 20 F-measures are highest with paragraphs, and 5-word contexts are the worst.

4.2 Experiment 2

The best parametric values found in Experiment 1 were put to the test on 1,500 different test terms for scale-up verification. Along with LO, which was the best association measure in the previous experiment, we used LL to double-check its relative inefficiency. For all of the 8 configurations evaluated, LL's recall, precision and MAP remain worse than LO's. In particular, LO's MAP scores with the cosine measure are more than twice as high as LL's (respectively 0.33 and 0.124 for sentence contexts). As in Experiment 1, the Dice index is significantly better for LL compared to the cosine, but not for LO. In the case of LO, sentence contexts have better Top 1 performances than paragraphs, and vice versa for Top 20 performances (see Table 4; oracle recall is 93.5 %). Hence, paragraph contexts would be more useful in tasks consisting of proposing candidate translations to lexicographers, while sentences would be more appropriate for automatic bilingual lexicon construction.

Ctx	R1	R20	P1	P20	F1	F20	MAP
Sent.	**23.1**	63.9	**27.8**	76.6	**25.23**	69.68	**0.336**
Parag.	20.1	**70.0**	22.9	**79.7**	21.41	**74.54**	0.325

Table 4: LO_Dice configuration scores

The cosine and Dice similarity measures have similar performances when LO is used. Moreover, we observe the effect of source and target term frequencies in corpus. As seen in Table 1, these frequencies are on average about half smaller in Experiment 2 as they are in Experiment 1, which results in significantly lower performances for all

[9] For this reason, whenever "Dice" is mentioned from this point on, it also applies to the Jaccard index.

8 variants. As Figure 5 shows for the variant LO_cosine_sentence, terms that are more frequent have a greater chance of being correctly translated at better ranks.

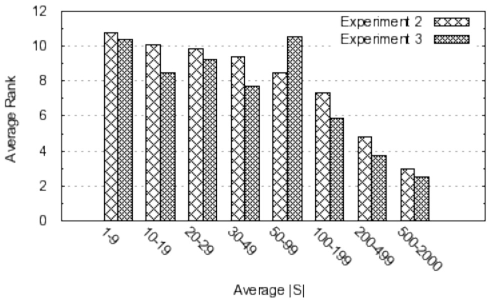

Figure 5: Average rank of correct translation according to average source term frequency

However, the relative performance of the different parametric configurations still holds.

4.3 Experiment 3

In Experiment 3, we evaluate the two best configurations from Experiment 2 with 3,000 new terms in order to verify the relative performance of the cosine and Dice similarity measures. As Table 5 shows, cosine has slightly better Top 1 figures, while Dice is a little better when considering the Top 20 translation candidates. Therefore, as previously mentioned, the choice of similarity measure (cosine or Dice) should depend on the goal of translation spotting. Note that the scores in Experiment 3 are much lower than those of Experiments 1 and 2 because of low term frequencies in the corpus (see Table 1 and Figure 5). Also, oracle recall is only 71.1 %.

Sim.	R1	R20	P1	P20	F1	F20	MAP
Cosine	**9.8**	28.1	**20.7**	59.4	**13.3**	38.15	0.232
Dice	9.4	**28.9**	19.8	**61.2**	12.75	**39.26**	**0.286**

Table 5: LO_sentence configuration scores

4.4 Experiment 4

In the last series of experiments, we examine the influence of the bilingual seed lexicon specificity and size, using the 1,350 terms which have source and target frequencies ≥ 30 from the 1,500 and 3,000 sets used in Experiments 2 and 3 (oracle recall: 100 %). We tested the different lexicons (see Section 2.2) on the 4 parametric configurations made of sentence contexts, LO or LL association measures, and cosine or Dice similarity measures.

Yet again, LO is better than LL. MAP scores for LO in all variants are comprised in [0.466–0.489]; LL MAPs vary between 0.135 and 0.146 when the cosine is used and between 0.348 and 0.380 when the Dice index is used.

According to our results, translation spotting is more accurate when the seed lexicon contains (5,000) entries from both the medical domain and general language instead of general language words only, but only by a very small margin. Table 6 shows the results for the configuration LO_cosine_sentence. The fact that the difference is so small could be explained by our resources' properties. The reference translations from MeSH contain terms that are also used in other domains or in the general language, *e.g.* terms from the category "people" (Névéol and Ozdowska, 2006). Wikipedia documents retrieved by using those references may in turn not belong to the medical domain, in which case medical terms from the seed lexicon are not appropriate. Still, the relatively good performance of the general language-only lexicon supports (Déjean et al., 2005, p. 119)'s claim that general language words are useful when spotting translations of domain specific terms, since the latter can appear in generic contexts.

Lex.	R1	R20	P1	P20	F1	F20	MAP
Gen. + med.	**39.3**	87.0	**39.6**	87.6	**39.4**	87.3	**0.473**
Gen. only	38.8	**88.1**	39.0	**88.5**	38.9	**88.3**	0.471

Table 6: LO_cosine_sentence configuration scores

Lexicon sizes tested are 5,000 (the mixed lexicon used in previous experiments), 7,000, 9,000 and 11,000 entries. The performance (based on MAP) is better when 7,000- and 9,000-entry lexicons are used, because more source language context words can be taken into account. However, when the lexicon reaches 11,000, Top 1 MAP scores and F-measures are slightly lower than those obtained with the 7,000-entry one. This may happen because the lexicon is increased with general language words; 9,000 of the 11,000 entries are not from the medical domain, making it harder for the context words to be specific. It would be interesting to study the specificity of context vectors built from the source corpus. Still, the differences in scores are small, as Table 7 shows (see Table 6 for the results obtained with 5,000 entries). This is because, in our implementation, context vector size is limited to 20, as in (Daille and Morin, 2005), in order to reduce processing time. The influence of context vector sizes should be studied.

Lex. size	R1	R20	P1	P20	F1	F20	MAP
7,000	**41.5**	88.8	**41.6**	89.1	**41.5**	88.9	0.488
9,000	40.9	89.3	41.1	89.7	41.0	89.5	**0.489**
11,000	40.1	**89.8**	40.2	**90.1**	40.1	**89.9**	0.484

Table 7: LO_cosine_sentence configuration scores

The parameters related to the seed lexicon do not have as great an impact on the performance as the choice of association measure does: the biggest difference in F-measures for Experiment 4 is 2.9 %. At this point, linguistic-based heuristics such as graphic similarity should be used to significantly increase performance. We applied the cognate heuristic (Section 2.3) on the Top 20 translation candidates given by the variant LO_sentence_9,000-entry lexicon using cosine and Dice similarity measures. Without the heuristic, Top 1 performances are better with cosine, while Dice is better for Top 20. Applying the cognate heuristic makes the Top 1 precision go from 41.1 % to 55.2 % in the case of cosine, and from 39.6 % to 53.9 % in the case of Dice.

5 Discussion

Our results show that using the log-odds ratio as the association measure allows for significantly better translation spotting than the log-likelihood. A closer look at the translation candidates obtained when using LL, the most popular association measure in projection-based approaches, shows that they are often collocates of the reference translation. Therefore, LL may fare better in an indirect approach, like the one in (Daille and Morin, 2005).

Moreover, we have seen that the cosine similarity measure and sentence contexts give more correct top translation candidates, at least when LO is used. Indeed, the values of the different parameters influence one another in most cases. Parameters related to the seed lexicon (size, domain specificity) are not of great influence on the performance, but this may in part be due to our resources and the way they were built.

The highest Top 1 precision, 55.2 %, was reached with the following parameters: sentence contexts, LO, cosine and a 9,000-entry mixed lexicon, with the use of a cognate heuristic.

In future works, other parameters which influence the performance will be studied, among which the use of a terminological extractor to treat complex terms (Daille and Morin, 2005), more contextual window configurations, and the use of syntactic information in combination with lexical information (Yu and Tsujii, 2009). It would also be interesting to compare the projection-based approaches to (Haghighi et al., 2008)'s generative model for bilingual lexicon acquisition from monolingual corpora.

One latent outcome of this work is that Wikipedia is surprisingly suitable for mining medical terms. We plan to check its adequacy for other domains and verify that LO remains a better association measure for different corpora and domains.

Acknowledgments

We are deeply grateful to Raphaël Rubino who provided us with the data material we have been using in this study. We thank the anonymous reviewers for their suggestions.

References

Chiao, Yun-Chuang and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *19th International Conference on Computational Linguistics*, pages 1208–1212.

Daille, Béatrice and Emmanuel Morin. 2005. French-English terminology extraction from comparable corpora. In *2nd International Joint Conference on Natural Language Processing*, pages 707–718.

Déjean, Hervé, Éric Gaussier, Jean-Michel Renders, and Fatiha Sadat. 2005. Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2):111–124. Elsevier Science, New York.

Diab, Mona and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access*.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 236–243.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *3rd Conference of the Association for Machine Translation in the Americas*, pages 1–17.

Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Human Language Technology and Association for Computational Linguistics*, pages 771–779.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing*. Prentice-Hall.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Morin, Emmanuel, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining — using brain, not brawn comparable corpora. In *45th Annual Meeting of the Association for Computational Linguistics*, pages 664–671.

Névéol, Aurélie and Sylwia Ozdowska. 2006. Terminologie médicale bilingue anglais/français: usages cliniques et bilingues. *Glottopol*, 8.

Otero, Pablo Gamallo. 2007. Learning bilingual lexicons from comparable English and Spanish corpora. In *Machine Translation Summit 2007*, pages 191–198.

Otero, Pablo Gamallo. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *1st Workshop Building and Using Comparable Corpora*.

Prochasson, Emmanuel, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Machine Translation Summit XII*, pages 284–291.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 66–70.

Rubino, Raphaël. 2009. Exploring context variation and lexicon coverage in projection-based approach for term translation. In *Proceedings of the Student Research Workshop associated with RANLP–09*, pages 66–70.

Saralegi, X., I. San Vicente, and A. Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *1st Workshop Building and Using Comparable Corpora*.

Shao, Li and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *20th International Conference on Computational Linguistics*, pages 618–624.

Simard, Michel, George Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *4th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

Smith, Jason R., Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 403–411.

Yu, Kun and Junichi Tsujii. 2009. Bilingual dictionary extraction from Wikipedia. In *Machine Translation Summit XII*.