Ranking Translation Candidates Acquired from Comparable Corpora

Rima Harastani and Béatrice Daille and Emmanuel Morin
LINA UMR CNRS 6241 - University of Nantes
2 rue de la Houssinière, BP 92208
44322 Nantes, France
{rima.harastani,beatrice.daille,emmanuel.morin}@univ-nantes.fr

Abstract

Domain-specific bilingual lexicons extracted from domain-specific comparable corpora provide for one term a list of ranked translation candidates. This study proposes to re-rank these translation candidates. We suggest that a term and its translation appear in comparable sentences that can be extracted from domain-specific comparable corpora. For a source term and a list of translation candidates, we propose a method to identify and align the best source and target sentences that contain the term and its translation candidates. We report results with two language pairs (French-English and French-German) using domain-specific comparable corpora. Our method significantly improves the top 1, top 5 and top 10 precisions of a domain-specific bilingual lexicon, and thus, provides a better user-oriented results.

1 Introduction

Comparable corpora have been the subject of interest for extracting bilingual lexicons by several researchers (Rapp, 1995; Fung and Mckeown, 1997; Rapp, 1999; Koehn and Knight, 2002; Morin et al., 2008; Bouamor et al., 2013, among others). Rapp (1995) was the first to suggest that if a word A co-occurs frequently with another word B in one language, then the translation of A and the translation of B should co-occur frequently in another language. Approaches emerging from (Rapp, 1995) make different assumptions to extract bilingual lexicon from comparable corpora. However, they are all based on the assumption that a translation pair shares some similar context in comparable corpora. We refer to such approaches that depend on co-occurrences of words to extract a bilingual lexicon by *distributional approaches*. Results obtained from distributional approaches vary according to many parameters. For example, one of the parameters that impacts the performance of distributional approaches is the way the context of a word is defined. Various approaches defined contexts differently: windows (Rapp, 1999), sentences or paragraphs (Fung and Mckeown, 1997), or by taking into consideration syntax dependencies based on POS tags (Gamallo, 2007). However, the most common way the context of a word is defined is by choosing words within windows centered around the word (Laroche and Langlais, 2010), usually of small sizes (e.g. a window of size 3 is used by Rapp (1999)).

Domain-specific comparable corpora have been used for bilingual terminology extraction. These corpora are of modest sizes since large domain-specific corpora are not available for many domains (Morin et al., 2008). As a matter of fact, distributional approaches perform best with large comparable corpora, and thus they often give lower precisions when applied to domain-specific comparable corpora (Chiao and Zweigenbaum, 2002).

The goal of our work is to find translations of terms in domain-specific comparable corpora. Taking a list of ranked translation candidates (provided by a distributional method) for a term, we aim to improve the ranking of the correct translations that are not ranked first in the list. Obviously, the more translation candidates for a term are considered, the more correct translations are found. For example, Rapp (1999) obtains a precision of 72% when only the first translation candidate is considered correct. However, he reports an 89% precision when the first 10 translation candidates are provided as translations for a word.

This study proposes to take the best translation candidates provided by a distributional approach,

International Joint Conference on Natural Language Processing, pages 401–409,
Nagoya, Japan, 14-18 October 2013.

and tries to re-rank them in order to improve the top 1, top 5 and top 10 precisions. We suggest that a source term and its correct translation appear in comparable sentences. Comparable sentences are sentences that share parallel data (e.g. word overlap, long matched sequences, bilingual compound nouns). We proceed by first extracting sentences for a source term, as well as sentences for each of its provided translation candidates. For each translation pair (i.e. source term and a translation candidate), each extracted source sentence is aligned with at most one of the extracted sentences for the translation candidate. The aligned sentences are used to re-rank the translation candidates of the source term.

Besides being used by our approach to re-rank translations, comparable sentences that contain a term and its translation in corpora are promising, as they may be useful examples to a user or a human translator that needs to verify a translation pair.

In Section 2, we present our approach and assumptions. In Section 3, we describe our method to extract sentences that best represent a term in corpora. In Section 4, we explain a method to score a sentence containing a term with a sentence containing its translation candidate. We evaluate our approach in Section 5 on two domain-specific corpora for the French-English and French-German language pairs, and report improvements in the top 1, top 5, and top 10 precisions. We conclude in Section 6.

2 Assumptions and Approach

A term may appear in several contexts, but some can be more interesting and more informative than others. In Table 1, an example of two sentences in which the term "tumor" appears is given. These sentences were extracted from an English corpus related to the domain of "Breast Cancer". Sentence (A) is considered to be more informative and more representative of the context of "tumor" than sentence (B). It also contains terms that are highly related to the "Breast Cancer" subject (e.g. chemotherapy, histological).

Our assumption is that the best context (represented by sentences) can be extracted for a term as well as for its translation candidates, and that these extracted sentences can be aligned in order to re-rank the translation candidates of the term.

After obtaining some candidate translations for a term by applying a distributional method, we score a source term (t_s) with its target translation candidate (t_t) as follows: we first extract the n best sentences that contain t_s in the source corpus as well as the n best sentences that contain its translation candidate in the target corpus. Then, we align each of the best sentences of t_s with at most a sentence of t_t using a method that depends on lexical similarity. Finally, the translation pair (t_s,t_t) is scored according to the scores of the aligned sentences between t_s and t_t. The scoring method is illustrated in Figure 1. We combine the resulting score with its initial score that is provided by a distributional method. Combined scores are then used to re-rank translation candidates of the specific term.

(A)	**Chemotherapy** was also administered to patients with smaller primary tumors with **histological** grade 2 or 3 or with negative hormone receptors.
(B)	The size of any captured image corresponding to the tumor was estimated.

Table 1: Sentence (A) and (B) containing the term "tumor"

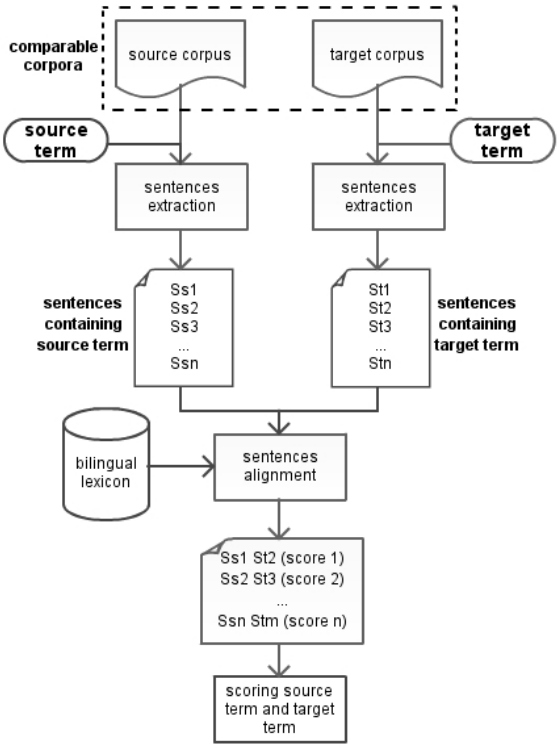

Figure 1: Method to score a translation pair (source term and target term)

Parallel sentence (or fragment) extraction from comparable corpora has received the attention of a number of researchers (Fung and Cheung, 2004; Munteanu and Marcu, 2005; Munteanu and Marcu, 2006; Smith et al., 2010; Hunsicker et al., 2012, among others), to enrich parallel text used by statistical machine translation (SMT) systems. They conducted experiments with large corpora (mainly news stories) which were noisy parallel, comparable (contain topic alignments or articles published in similar circumstances), or very non-parallel (Fung and Cheung, 2004). Usually, these approaches perform document-level alignments before extracting parallel sentences. The domain-specific corpora we use contain few documents (ranging from 38 to 262 documents for each corpus) and no parallel sentences. Furthermore, they are of modest size (about 0.3 M to 0.5 M words), so even if there were some parallel fragments, this phenomenon would be rare. Nevertheless, we assume that some features used in state-of-the-art parallel sentence extraction methods can be used to identify comparable sentences that contain a translation pair.

Our goal is not to extract parallel sentences, but rather we need to find, for a translation pair, bilingual sentences that are comparable. For example, consider that we need to score the correct translation pair (FR[1] clinique, EN[2] clinical), and that we have two sentences, the first contains "clinique" and the second contains "clinical" (see Figure 2). The two sentences are not parallel, however, they both contain the following information: a clinical examination detects the size of a tumor. Finding this kind of comparability in sentences would help in increasing the score of correct translation pairs.

3 Best Sentences Extraction for a Term

For a term (t), we aim to extract the *n* best sentences that represent its context in the corpus. We suggest that sentences that best represent t contain words that are: (a) strongly associated with t in the corpus, (b) highly specific to the domain of the corpus. A word in a sentence containing t is scored by means of two measures: association and domain specificity, that are presented in the following.

1. Association with t: word associations are computed according to log-likelihood scores that are based on the co-occurrences of words in a window of size (s=7) around t. The top (m=30) associated words and their scores with t are denoted by v_m (context vector of t of size m). The association between a word (w) and t is computed from occurrences that are resumed in the contingency table (see Table 2), where occ(t,w) is the number of occurrences of t and w, and $\neg w$ signifies all words except w.

	w	**¬w**
t	a=occ(t,w)	b=occ(t,¬w)
¬t	c=occ(¬t,w)	d=occ(¬t,¬w)

Table 2: Contingency table for t and w

The log-likelihood association measure is computed as follows:

$$\begin{aligned}\text{association}(t,w) &= a\log(a) + b\log(b)\\ &+ c\log(c) + d\log(d) + (N)\log(N)\\ &- (a+b)\log(a+b) - (a+c)\log(a+c)\\ &- (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \quad (1)$$

where $N = a+b+c+d$. The association between w and t is then divided by the biggest association score obtained with t to have a score $\in$[0,1].

2. Domain specificity: the specificity of a word is its relative frequency in the domain-specific corpus (dc=$\{w_1,w_2,..,w_n\}$) divided by its relative frequency in a general language corpus (gc=$\{w_1{}',w_2{}',..,w_m{}'\}$), it is defined in (Khurshid et al., 1994) as follows:

$$\text{ds}(w) = \frac{rvf_{dc}(w)}{rvf_{gc}(w)} \quad (2)$$

where rvf_{dc}=$\frac{freq_{dc}(w)}{\sum_{w_i \in dc} freq_{dc}(w_i)}$ is the relative frequency in the specific corpus, $rvf_{gc}(w)$=$\frac{freq_{gc}(w)}{\sum_{w_i' \in gc} freq_{gc}(w_i')}$ is the relative frequency in the general corpus, and $freq$ signifies frequency. The specificity of a term is normalized by being divided by the value of the biggest specificity in the corpus.

To extract the n best sentences for term t, we give a score to each sentence S that contains t and words

[1]FR signifies French
[2]EN signifies English

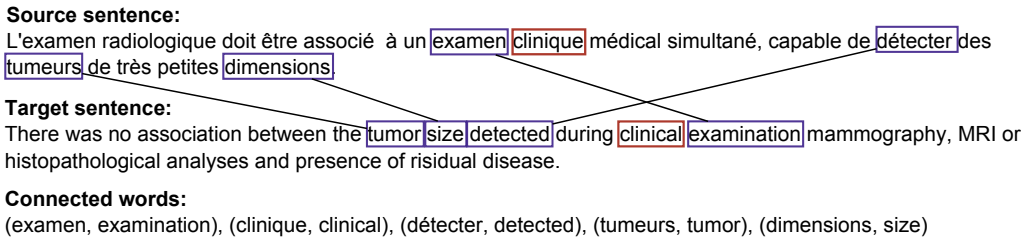

Figure 2: Example of source and target sentences that contain the translation pair (FR clinique and EN clinical)

$w1, w2, ..., w_n$ as follows:

$$score(S) = \sum_{i=1}^{n} \Big(\text{ds}(w_i) + \text{association}_{(\text{if } w_i \in v_m)}(w_i, t) \Big) \quad (3)$$

We discard any sentence with a length of less than 5 words (after removing the stop words). All sentences containing t are then ranked according to their scores. For a translation pair (t_s,t_t), the n best sentences for t_s as well as for t_t are extracted following the method explained above.

The next step consists of aligning the n best sentences of a source term t_s with n best sentences of each of its proposed translations.

4 Sentences Alignment for Translation Pairs

We suggest that if a source term (t_s) is translated by a target term (t_t), then they must share some comparable sentences. The more a translation pair shares sentences with high comparability, the higher its score should be.

The ratio between the lengths of two comparable sentences should be less than 2, following (Munteanu and Marcu, 2005). We also suppose that the overlap between two comparable sentences should be greater than 3 (including the translation pair). Like previous works on extracting parallel sentences from comparable corpora, our approach depends mostly on lexical information between sentences by using a bilingual lexicon.

Suppose that we have a source sentence S_s=$\{w_1$,w_2,t_s,...,$w_n\}$[3] and a target sentence S_t=$\{w_1^{'}$,$w_2^{'}$,t_t...,$w_n^{'}\}$[4] (after removing the stop words), with a set of possible connected words M=$\{(w_1,w_1^{'}),(w_2,w_2^{'}),..,(w_n,w_n^{'})\}$ obtained using a bilingual dictionary. An optimal alignment A (each word in the sentence S_s is connected to at most one word in the sentence S_t) is estimated according to a linear function.

Taking the optimal alignment A, feature functions (where each $\in$ [0,1]) are utilized to compute a score between the two sentences.

1. The cosine similarity between the two sentences (Fung and Cheung, 2004) penalized by the number of unconnected words: each word in S_s (respectively S_t) is weighted by its score in the context vector v_m (respectively $v_m^{'}$) of t_s (respectively t_t). If a word is missing from the context vector, it would be associated a fixed minimal weight. The first feature function is defined as follows:

$$f_1(S_{t_s}, S_{t_t}) = \frac{\text{cosine}(S_{t_s}, S_{t_t})}{|\text{UnConnectedWords}|} \quad (4)$$

where |UnConnectedWords| is the number of unconnected words between the two sentences.

2. Positions of connected words in the source sentence (target sentence respectively) in comparison to the position of source term (target term respectively): the nearer the connected words are from the term t in the sentence, the greater the score of this feature function will be. Besides, we suppose that for two connected words (w_i,$w_i^{'}$), the distance between w_i and t_s should be close to the distance between $w_i^{'}$ and t_t. The positions distance is defined as follows:

$$pos_{\text{distance}}(S_{t_s}, S_{t_t}) = \sum_{w_i, w_i^{'} \in A} \frac{(pos_s + pos_t + |pos_s - pos_t|)}{|S_{t_s}| + |S_{t_t}| + |S_{t_s} - S_{t_t}|} \quad (5)$$

[3] t_s could be at any position in S_s
[4] t_t could be at any position in S_t

where $pos_s = |pos(w_i) - pos(t_s)|$ and $pos_t = |pos(w_i^{'}) - pos(t_t)|$.

The pos_{distance} is then divided by $|A|$ to be normalized. The positions similarity is computed as follows:

$$f_2(S_{t_s}, S_{t_t}) = 1 - \frac{pos_{\text{distance}}}{|A|} \quad (6)$$

3. Longest contiguous span: it is defined by (Munteanu and Marcu, 2005) as being the longest "pair of substrings in which the words in one substring are connected only to words in the other substring". We assume that the length of a span must be greater than 2. The longest span is divided by the length of the smaller sentence, then:

$$f_3(S_{t_s}, S_{t_t}) = \frac{\text{span}(S_{t_s}, S_{t_t})}{\min(|S_{t_s}|, |S_{t_t}|)} \quad (7)$$

4. Number of connected bi-grams: this feature function is defined as the number of found connected bi-grams divided by the number of connected words in A, then:

$$f_4(S_{t_s}, S_{t_t}) = \frac{\text{bi-grams}(S_{t_s}, S_{t_t})}{|A|} \quad (8)$$

The optimal alignment A is the alignment that minimizes the squared Euclidean distance between the two sentence vectors and the pos_{distance}. Indeed, we choose this minimization function for a matter of optimization.

We follow (Hunsicker et al., 2012) in considering the final score between a sentence pair as the weighted sum of all feature functions, such as the following:

$$score(S_s, S_t) = \sum_{i=1}^{4} \left(w_i * f_i(S_{t_s}, S_{t_t})\right) \quad (9)$$

where $\sum_{i=1}^{4}(w_i) = 1$.

Contrary to previous works that use parallel corpora to train their models and define the weights of feature functions, we define the weights by guesswork. This is because we do not have an annotated parallel corpora. Nevertheless, this should not have a significant impact on our results since our goal is not to extract parallel sentences.

4.1 Reranking translation pairs

For a translation pair (t_s,t_t), each sentence of the n best representing sentences of t_s is aligned with at most one of the n best representing sentences of t_t. A target sentence can be aligned to multiple source sentences. The score between the translation pair is the average of the scores of the sentence alignments. We refer to this procedure as the sentence alignment method.

The re-ranking is done by combining the score obtained by the sentence alignment method for a translation pair with its initial score that is obtained by a distributional method. The scores are combined by the weighted geometric mean.

5 Evaluation

We first need to extract translations for a list of domain-specific terms in comparable corpora. In order to do this, we pre-process corpora and align terms with the free tool TermSuite[5] (Rocheteau and Daille, 2011). The distributional method that is implemented in TermSuite is the one described in (Rapp, 1999). TermSuite provides a chosen number of translations for a term. Translations are ranked according to the scores provided by the distributional method. We try to enhance the top candidate translations of each reference source term by applying our re-ranking method.

5.1 Data

To carry out the distributional approach with TermSuite, we need comparable corpora, bilingual dictionaries, and a list of source reference terms to translate. We need the same resources to perform experiments with our method as well as general language monolingual corpora.

- Comparable corpora: we carry out experiments with comparable corpora in two different domains and two language pairs French-English and French-German. The first are medical corpora in the sub-domain of *breast cancer*, these contain approximately 0.37 M to 0.5 M words for each language. The second corpora belong to the renewable energy domain, more specifically, to the sub-domain of *wind energy*, and contain about 0.3 M to 0.35 M words for each language. Breast Cancer corpora were collected from an online medical portal, while Wind Energy corpora have been crawled using Babouk crawler (Groc, 2011). Both corpora have been collected using some seed terms and contain no

[5]This tool is available on http://code.google.com/p/ttc-project/

parallel sentences. Table 3 resumes the sizes of monolingual parts of corpora.

Language	Breast Cancer	Wind Energy
French	531,240	313,943
English	528,428	314,549
German	378,474	358,602

Table 3: Sizes in number of words of corpora for each language and for each domain

- Bilingual dictionaries: general language bilingual dictionaries[6] for the French-English and French-German language pairs were obtained. The French-English dictionary contains 145,542 single-word entries and the French-German dictionary contains 118,776 single-word entries.

- General language corpora: for each language, a general language corpus is obtained and used in computing specificities of words to the domain-specific corpora. These contain 12003, 3903 and 44365 unique single words for French, English and German respectively.

- Reference lists: we have built a list of reference single-word terms (SWTs) for each corpora and for each language pair. Each source term in the list is domain-specific with a frequency greater than 5 in the source corpus and has been manually aligned with one golden translation that exists in the target corpus. For Breast Cancer corpora, for each language pair we built a list that that contains 122 translation pairs. As for Wind Energy corpora, for each language pair we built a list that includes 96 translation pairs.

5.2 Experimental Settings

For the sentence alignment method, we manually define the same parameters for Breast Cancer and Wind Energy corpora. For each term and each translation candidate, we extract the 70 best sentences, where sentences that have the same score are ranked at the same position. However, we take a maximum of 200 sentences for a term. If a term is less frequent that 70 in the corpus, we extract all the sentences that include this term. We do not extract a large number of sentences for a term because the alignment process will be computationally expensive, besides, our assumption is that if a translation pair is valid, then its best representative sentences are comparable. When extracting sentences for a term, we discard any sentence with a length of less than 5 words (after removing the stop words). A sentence is supposed to be simply delimited by punctuation marks ("?", "!", "."). We point out that the words, in a sentence containing a term t, that are used in computing the score of this sentence and as context for t are the words appearing at maximum in a window of size n=20 around t (10 words or less appearing before t in the sentence, and 10 words or less appearing after t in the sentence, after removing the stop words).

To score a translation pair by aligning its sentences (see equation 9), the biggest weight is set to 0.4 and is attributed to the first feature function (see equation 4). The remainder of weights are set equally to 0.2. When combining the scores of the distributional and the sentence alignment methods by the weighted geometric mean, the weight of the first is set to 0.3, and the weight of the second is set to 0.7.

5.3 Evaluation Measures

The precision of a bilingual lexicon is computed at different levels after taking several n best translations for each term (top 1, top 5, etc.). The precision is the number of the correct translations found divided by the number of source terms in the reference list.

The Mean Reciprocal Rank (MRR) is also used to evaluate the obtained results. The reciprocal rank for a given source term is the multiplicative inverse of the rank of the first correct target translation. The mean reciprocal rank is the average of the reciprocal ranks of the aligned source reference terms. MRR values are between 0 and 1, where higher values indicate a better performance of the system.

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{10}$$

where $|Q|$ is the number of source terms to be aligned. If a the correct translation of a term has not been found, then its corresponding "$\frac{1}{rank_i}$" is equal to 0.

[6]The dictionaries were obtained from http://catalog.elra.info/product_info.php?products_id=666 and http://catalog.elra.info/product_info.php?products_id=668

5.4 Experiments

The results of the distributional approach (baseline) with the language pairs and two corpora are given in Table 4 (P1 signifies the precision when 1 translation candidate is provided for a term). We notice that the results on Breast Cancer corpora are better than those obtained with Wind Energy. This may be justified by the fact that Wind Energy corpora are of smaller sizes and less technical.

The results are also significantly better with the French-English language pair than with the French-German language pair. In fact, domain-specific corpora contain many terms that are compound nouns. In the German language, many compound nouns may be written as single units (e.g. German term "Produktionsstandort" is translated into French by "site de production"). Therefore, the distributional approach may consider such German terms as one word when computing co-occurrences. One way to overcome this problem would be to perform splitting before applying the distributional approach (Macherey et al., 2011).

To analyze the results obtained by the distributional method in more depth, we measured the comparability of Wind Energy corpora for the different language pairs, using the comparability measure presented by Li et al. (2011). For the French-English corpora, we obtained a comparability value of 0.81. As for the French-German corpora, we obtained a comparability value of 0.70. This implies that our French-German corpora are less comparable than the French-English corpora, and partly justifies the reason behind obtaining worse results with the French-German pair using the distributional method.

	Breast Cancer		**Wind Energy**	
	FR-EN	**FR-GR**	**FR-EN**	**FR-GR**
P1	26.22%	9.16%	16.66%	3.12%
P5	45.08%	18.85%	38.54%	9.37%
P10	53.27%	26.22%	45.83%	10.41%
P15	59.01%	29.50%	50.00%	12.50%
P20	60.65%	31.96%	57.29%	14.58%
P25	61.47%	32.78%	59.37%	14.58%

Table 4: Results obtained with distributional method (baseline). EN-FR signifies English-French, and FR-GR signifies French-German.

In order to improve these results, especially the top 1, top 5 and top 10 precisions, we try to re-rank the translation candidates for each source term by combining their initial scores with the scores obtained from aligning their sentences.

Let us suppose that for a source term t_s, we want to re-rank its top 5 translation candidates L_{top5}=$\{t_{t_1}$,t_{t_2},t_{t_3},t_{t_4},$t_{t_5}\}$ provided by the distributional method. Following the approach presented in Section 3, we extract the best ranked sentences for t_s. We do the same for each translation candidate in L_{top5}. Then, for each translation pair (e.g. t_s and t_{t_1}) we try to align each sentence that was extracted for t_s with one sentence that shares the highest score with it among the sentences extracted for t_{t_1}, using the approach described in Section 4. A source sentence can be aligned with at most one target sentence and is assigned a score (which is equal to 0 if the sentence is not aligned). The score between t_s and t_{t_1} is the average of the scores of the alignments.

Following the above explained procedure, we take the best n=20 translation candidates proposed by the distributional method for each term and re-rank the translation candidates. This evaluation strategy is denoted by RR1 in Tables 5 and 6 which resume the obtained results on our corpora with two language pairs. For example, using the French-English Breast Cancer list, we find that re-ranking the top 20 translation candidates provided for each source term improved the top 1 precision by approximately 5%. Moreover, before re-ranking, 43.24% of the correct translations found in the top 20 results were ranked at the 1^{st} position, after re-ranking, this percentage increases to 52.70%. Which means that the re-ranking has significantly improved the ranks of the correct translations. An improvement of approximately 6% in the top 1 precision is obtained when using 20 translation candidates to re-rank the results obtained with the French-English Wind Energy list. However, fewer improvements were obtained with the French-German language pair as there were not many correct translations in the first 20 translations provided for each term by the distributional method.

While performing experiments, we have noticed that re-ranking the first 5 translation candidates for each term may increase the top 1 precision more than if we, for example, re-ranked the first 20 translation candidates for each term. For that, we have decided to follow a different strategy (denoted by RR2) for re-ranking translations. To de-

	Breast Cancer			Wind Energy		
	Baseline	**RR1**	**RR2**	**Baseline**	**RR1**	**RR2**
P1	26.22%	31.96%	**35.24%**	16.66%	**23.95%**	22.91%
P5	45.08%	**52.45%**	**52.45%**	38.54%	**45.83%**	44.79%
P10	53.27%	**57.37%**	**57.37%**	45.83%	48.95%	**52.08%**
MRR	0.338	0.396	**0.419**	0.249	**0.324**	0.319

Table 5: Results obtained on both Breast Cancer and Wind Energy French-English Corpora

	Breast Cancer			Wind Energy		
	Baseline	**RR1**	**RR2**	**Baseline**	**RR1**	**RR2**
P1	9.16%	**11.47%**	**11.47%**	3.12%	**7.29%**	5.20%
P5	18.85%	**21.31%**	**21.31%**	9.37%	**10.41%**	**10.41%**
P10	26.22%	**27.04%**	**27.04%**	10.41%	**13.51%**	**13.51%**
MRR	0.139	0.160	**0.162**	0.051	**0.088**	0.075

Table 6: Results obtained on both Breast Cancer and Wind Energy French-German Corpora

termine which translation candidate will be ranked at the n (starting from 1) position for a term, we first re-rank the top m=$\big($round (2(n-1)+5) to the nearest multiple of 5$\big)$ translations proposed for each term. The translation candidate at position 1 will have the position n in the new ranked list and it will not be further re-ranked. Then, we determine the translation candidate that will be ranked at the position (n+1) in the new ranked list. We repeat this process until obtaining 10 translation candidates for each term in the new ranked list.

For example, taking a list of translation candidates provided for a term: to determine which translation candidate will be ranked at the first position, we re-rank the list of top 5 (L_{top5}) translation candidates provided for the term, we put the translation now ranked in the first position in a list we name L_{taken}. To determine which translation candidate will be in the second position, we re-rank the list (L_{top5} - L_{taken}) and add the translation ranked in the first position to L_{taken}. Now to determine which translation will be ranked in the third position, we re-rank the (list of top 10 - L_{taken}), and put the translation ranked in the first position in L_{taken}, and so on. Results obtained using this strategy are presented in Tables 5 and 6 (under RR2).

RR2 strategy gave better top 1 precision and MRR than RR1 with French-English Breast Cancer corpora, and better top 10 precision with French-English Wind Energy corpora. RR1 strategy gave better MRR on Wind Energy corpora. In general, the results of the two strategies were comparable. This means that RR1 gave stable improvements when re-ranking a list of 20 candidates for each term. Both RR1 and RR2 significantly improved the baseline results for French-English and French-German language pairs.

6 Conclusion

In this paper, we proposed a method to re-rank the top translation candidates acquired by a distributional method from comparable corpora. We assumed that some sentences are more representative of a term than others, and that a term and its correct translation share comparable sentences that can be extracted from comparable corpora. We suggested aligning sentences that best represent a term with sentences that best represent its translation candidates to re-rank these translation candidates. Our experiments showed improvements in precision and MRR measures for two language pairs and two domains.

Our re-ranking method was tested with SWTs, and we aim to further evaluate it with multi-word terms (MWTs). Moreover, best aligned sentences for a term and its translation candidates can also be proposed for a user-oriented evaluation to see whether the aligned sentences can help in validating a translation pair.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable remarks. This work was supported by the French National Research Agency under grant ANR-12-CORD-0020.

References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2 of *ACL '13*, pages 759–764, Sofia, Bulgaria.

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 of *COLING '02*, pages 1–5.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 57–63, Barcelona, Spain.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

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

Clément De Groc. 2011. Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *The IEEE/WIC/ACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France.

Sabine Hunsicker, Radu Ion, and Dan Stefanescu. 2012. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, EAMT '12, Trento, Italy.

Ahmad Khurshid, Davies Andrea, Fulford Heather, and Rogers Margaret. 1994. What is a term? the semi-automatic extraction of terms from text. In *Translation Studies: An Interdiscipline, John Benjamins Publishing Company, Amsterdam*, pages 267–278.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

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.

Bo Li, Eric Gaussier, and Akiko Aizawa. 2011. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2 of *HLT '11*, pages 473–478, Portland, Oregon.

Klaus Macherey, Andrew M. Dai, David Talbot, Ashok C. Popat, and Franz Och. 2011. Language-independent compound splitting with morphological operations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT '11*, pages 1395–1404, Portland, Oregon.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2008. Brains, not brawn: The use of smart comparable corpora in bilingual terminology mining. *ACM Trans. Speech Lang. Process.*, 7(1):1–23.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia.

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, Cambridge, Massachusetts.

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, College Park, Maryland.

Jerome Rocheteau and Béatrice Daille. 2011. TTC TermSuite: A UIMA Application for Multilingual Terminology extraction from Comparable Corpora. In *the 5th International Joint Conference on Natural Language Processing*, IJCNLP '11, pages 9–12, Chiang Mai, Thailand.

Jason R. Smith, 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 Association for Computational Linguistics*, HLT '10, pages 403–411.