Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction

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

Abstract

The main work in bilingual lexicon extraction from comparable corpora is based on the implicit hypothesis that corpora are balanced. However, the historical context-based projection method dedicated to this task is relatively insensitive to the sizes of each part of the comparable corpus. Within this context, we have carried out a study on the influence of unbalanced specialized comparable corpora on the quality of bilingual terminology extraction through different experiments. Moreover, we have introduced a regression model that boosts the observations of word co-occurrences used in the context-based projection method. Our results show that the use of unbalanced specialized comparable corpora induces a significant gain in the quality of extracted lexicons.

1 Introduction

The bilingual lexicon extraction task from bilingual corpora was initially addressed by using parallel corpora (i.e. a corpus that contains source texts and their translation). However, despite good results in the compilation of bilingual lexicons, parallel corpora are scarce resources, especially for technical domains and for language pairs not involving English. For these reasons, research in bilingual lexicon extraction has focused on another kind of bilingual corpora comprised of texts sharing common features such as domain, genre, sampling period, etc. without having a source text/target text relationship (McEnery and Xiao, 2007). These corpora, well known now as *comparable corpora*, have also initially been introduced as *non-parallel corpora* (Fung, 1995; Rapp, 1995), and *non-aligned corpora* (Tanaka and Iwasaki, 1996). According to Fung and Cheung (2004), who range bilingual corpora from parallel corpora to quasi-comparable corpora going through comparable corpora, there is a continuum from parallel to comparable corpora (i.e. a kind of filiation).

The bilingual lexicon extraction task from comparable corpora inherits this filiation. For instance, the historical context-based projection method (Fung, 1995; Rapp, 1995), known as the *standard approach*, dedicated to this task seems implicitly to lead to work with balanced comparable corpora in the same way as for parallel corpora (i.e. each part of the corpus is composed of the same amount of data).

In this paper we want to show that the assumption that comparable corpora should be balanced for bilingual lexicon extraction task is unfounded. Moreover, this assumption is prejudicial for specialized comparable corpora, especially when involving the English language for which many documents are available due the prevailing position of this language as a standard for international scientific publications. Within this context, our main contribution consists in a re-reading of the standard approach putting emphasis on the unfounded assumption of the balance of the specialized comparable corpora. In specialized domains, the comparable corpora are traditionally of small size (around 1 million words) in comparison with comparable corpus-based general language (up to 100 million words). Consequently, the observations of word co-occurrences which is the basis of the standard approach are unreliable. To make them more reliable, our second contribution is to contrast different regression models in order to boost the observations of word co-occurrences. This strategy allows to improve the quality of extracted bilingual lexicons from comparable corpora.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1284–1293,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

2 Bilingual Lexicon Extraction

In this section, we first describe the standard approach that deals with the task of bilingual lexicon extraction from comparable corpora. We then present an extension of this approach based on regression models. Finally, we discuss works related to this study.

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, 1994, 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 (e.g. a seven-word window approximates syntactic dependencies). In order to emphasize significant words in the context vector and to reduce word-frequency effects, the context vectors are normalized according to an association measure. Then, the translation is obtained by comparing the source context vector to each translation candidate vector after having translated each element of the source vector with a general dictionary.

The implementation of the standard approach can be carried out by applying the following three steps (Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Morin et al., 2007; Laroche and Langlais, 2010, among others):

Computing context vectors We collect all the words in the context of each word i and count their occurrence frequency in a window of n words around i. For each word i of the source and the target languages, we obtain a context vector v_i which gathers the set of co-occurrence words j associated with the number of times that j and i occur together $cooc(i, j)$. In order to identify specific words in the lexical context and to reduce word-frequency effects, we normalize context vectors using an association score such as Mutual Information, Log-likelihood, or the discounted log-odds (LO) (Evert, 2005) (see equation 1 and Table 1 where $N = a + b + c + d$).

Transferring context vectors Using a bilingual dictionary, we translate the elements of the source context vector. If the bilingual dictionary provides several translations for an element, we consider all of them but weight the different translations according to their frequency in the target language.

Finding candidate translations For a word to be translated, we compute the similarity between the translated context vector and all target vectors through vector distance measures such as Jaccard or Cosine (see equation 2 where $assoc_j^i$ stands for "association score", v_k is the transferred context vector of the word k to translate, and v_l is the context vector of the word l in the target language). Finally, the candidate translations of a word are the target words ranked following the similarity score.

	j	$\neg j$
i	$a = cooc(i, j)$	$b = cooc(i, \neg j)$
$\neg i$	$c = cooc(\neg i, j)$	$d = cooc(\neg i, \neg j)$

Table 1: Contingency table

$$LO(i, j) = \log \frac{(a + \frac{1}{2}) \times (d + \frac{1}{2})}{(b + \frac{1}{2}) \times (c + \frac{1}{2})} \qquad (1)$$

$$Cosine_{v_l}^{v_k} = \frac{\sum_t assoc_t^l \, assoc_t^k}{\sqrt{\sum_t {assoc_t^l}^2} \sqrt{\sum_t {assoc_t^k}^2}} \qquad (2)$$

This approach is sensitive to the choice of parameters such as the size of the context, the choice of the association and similarity measures. The most complete study about the influence of these parameters on the quality of word alignment has been carried out by Laroche and Langlais (2010).

The standard approach is used by most researchers so far (Rapp, 1995; Fung, 1998; Peters and Picchi, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Gaussier et al., 2004; Morin et al., 2007; Laroche and Langlais, 2010; Prochasson and Fung, 2011;

References	Domain	Languages	Source/Target Sizes
Tanaka and Iwasaki (1996)	Newspaper	EN/JP	30/33 million words
Fung and McKeown (1997)	Newspaper	EN/JP	49/60 million bytes of data
Rapp (1999)	Newspaper	GE/EN	135/163 million words
Chiao and Zweigenbaum (2002)	Medical	FR/EN	602,484/608,320 words
Déjean et al. (2002)	Medical	GE/EN	100,000/100,000 words
Morin et al. (2007)	Medical	FR/JP	693,666/807,287 words
Otero (2007)	European Parliament	SP/EN	14/17 million words
Ismail and Manandhar (2010)	European Parliament	EN/SP	500,000/500,000 sentences
Bouamor et al. (2013)	Financial	FR/EN	402,486/756,840 words
-	Medical	FR/EN	396,524/524,805 words

Table 2: Characteristics of the comparable corpora used for bilingual lexicon extraction

Bouamor et al., 2013, among others) with the implicit hypothesis that comparable corpora are balanced. As McEnery and Xiao (2007, p. 21) observe, a specialized comparable corpus is built as balanced by analogy with a parallel corpus: "*Therefore, in relation to parallel corpora, it is more likely for comparable corpora to be designed as general balanced corpora.*". For instance, Table 2 describes the comparable corpora used in the main work dedicated to bilingual lexicon extraction for which the ratio between the size of the source and the target texts is comprised between 1 and 1.8.

In fact, the assumption that words which have the same meaning in different languages should have the same lexical context distributions does not involve working with balanced comparable corpora. To our knowledge, no attention[1] has been paid to the problem of using unbalanced comparable corpora for bilingual lexicon extraction. Since the context vectors are computed from each part of the comparable corpus rather than through the parts of the comparable corpora, the standard approach is relatively insensitive to differences in corpus sizes. The only precaution for using the standard approach with unbalanced corpora is to normalize the association measure (for instance, this can be done by dividing each entry of a given context vector by the sum of its association scores).

2.2 Prediction Model

Since comparable corpora are usually small in specialized domains (see Table 2), the discriminative power of context vectors (i.e. the observations of word co-occurrences) is reduced. One way to deal with this problem is to re-estimate co-occurrence counts by a prediction function (Hazem and Morin, 2013). This consists in assigning to each observed co-occurrence count of a small comparable corpora, a new value learned beforehand from a large training corpus.

In order to make co-occurrence counts more discriminant and in the same way as Hazem and Morin (2013), one strategy consists in addressing this problem through regression: given training corpora of small and large size (abundant in the general domain), we predict word co-occurrence counts in order to make them more reliable. We then apply the resulting regression function to each word co-occurrence count as a pre-processing step of the standard approach. Our work differs from Hazem and Morin (2013) in two ways. First, while they experienced the linear regression model, we propose to contrast different regression models. Second, we apply regression to unbalanced comparable corpora and study the impact of prediction when applied to the source texts, the target texts and both source and target texts of the used comparable corpora.

We use regression analysis to describe the relationship between word co-occurrence counts in a large corpus (the response variable) and word co-occurrence counts in a small corpus (the predictor variable). As most regression models have already been described in great detail (Christensen, 1997; Agresti, 2007), the derivation of most models is only briefly introduced in this work.

As we can not claim that the prediction of word co-occurrence counts is a linear problem, we consider in addition to the simple linear regression

[1] We only found mention of this aspect in Diab and Finch (2000, p. 1501) "*In principle, we do not have to have the same size corpora in order for the approach to work*".

model (Lin), a generalized linear model which is the logistic regression model ($Logit$) and non linear regression models such as polynomial regression model ($Poly^n$) of order n. Given an input vector $x \in \mathbb{R}^m$, where x_1,...,x_m represent features, we find a prediction $\hat{y} \in \mathbb{R}^m$ for the co-occurrence count of a couple of words $y \in \mathbb{R}$ using one of the regression models presented below:

$$\hat{y}_{Lin} = \beta_0 + \beta_1 x \tag{3}$$

$$\hat{y}_{Logit} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))} \tag{4}$$

$$\hat{y}_{Poly^n} = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n \tag{5}$$

where β_i are the parameters to estimate.

Let us denote by f the regression function and by $cooc(w_i, w_j)$ the co-occurrence count of the words w_i and w_j. The resulting predicted value of $cooc(w_i, w_j)$, noted $c\hat{o}oc(w_i, w_j)$ is given by the following equation:

$$c\hat{o}oc(w_i, w_j) = f(cooc(w_i, w_j)) \tag{6}$$

2.3 Related Work

In the past few years, several contributions have been proposed to improve each step of the standard approach.

Prochasson et al. (2009) enhance the representativeness of the context vector by strengthening the context words that happen to be transliterated words and scientific compound words in the target language. Ismail and Manandhar (2010) also suggest that context vectors should be based on the most important contextually relevant words (in-domain terms), and thus propose a method for filtering the noise of the context vectors. In another way, Rubino and Linarès (2011) improve the context words based on the hypothesis that a word and its candidate translations share thematic similarities. Yu and Tsujii (2009) and Otero (2007) propose, for their part, to replace the window-based method by a syntax-based method in order to improve the representation of the lexical context.

To improve the transfer context vectors step, and increase the number of elements of translated context vectors, Chiao and Zweigenbaum (2003) and Morin and Prochasson (2011) combine a standard general language dictionary with a specialized dictionary, whereas Déjean et al. (2002) use the hierarchical properties of a specialized thesaurus. Koehn and Knight (2002) automatically induce the initial seed bilingual dictionary by using identical spelling features such as cognates and similar contexts. As regards the problem of words ambiguities, Bouamor et al. (2013) carried out word sense disambiguation process only in the target language whereas Gaussier et al. (2004) solve the problem through the source and target languages by using approaches based on CCA (Canonical Correlation Analysis) and multilingual PLSA (Probabilistic Latent Semantic Analysis).

The rank of candidate translations can be improved by integrating different heuristics. For instance, Chiao and Zweigenbaum (2002) introduce a heuristic based on word distribution symmetry. From the ranked list of candidate translations, the standard approach is applied in the reverse direction to find the source counterparts of the first target candidate translations. And then only the target candidate translations that had the initial source word among the first reverse candidate translations are kept. Laroche and Langlais (2010) suggest a heuristic based on the graphic similarity between source and target terms. Here, candidate translations which are cognates of the word to be translated are ranked first among the list of translation candidates.

3 Linguistic Resources

In this section, we outline the different textual resources used for our experiments: the comparable corpora, the bilingual dictionary and the terminology reference lists.

3.1 Specialized Comparable Corpora

For our experiments, we used two specialized French/English comparable corpora:

Breast cancer corpus This comparable corpus is composed of documents collected from the Elsevier website[2]. The documents were taken from the medical domain within the sub-domain 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 collected 130 French documents (about 530,000 words) and 1,640 English documents (about

[2] `http://www.elsevier.com`

7.4 million words). We split the English documents into 14 parts each containing about 530,000 words.

Diabetes corpus The documents making up the French part of the comparable corpus have been craweled from the web using three keywords: *diabète* (diabetes), *alimentation* (food), and *obésité* (obesity). After a manual selection, we only kept the documents which were relative to the medical domain. As a result, 65 French documents were extracted (about 257,000 words). The English part has been extracted from the medical website PubMed[3] using the keywords: *diabetes*, *nutrition* and *feeding*. We only kept the free fulltext available documents. As a result, 2,339 English documents were extracted (about 3,5 million words). We also split the English documents into 14 parts each containing about 250,000 words.

The French and English documents were then normalised through the following linguistic preprocessing steps: tokenisation, part-of-speech tagging, and lemmatisation. These steps were carried out using the TTC TermSuite[4] that applies the same method to several languages including French and English. Finally, the function words were removed and the words occurring less than twice in the French part and in each English part were discarded. Table 3 shows the number of distinct words (# words) after these steps. It also indicates the comparability degree in percentage (comp.) between the French part and each English part of each comparable corpus. The comparability measure (Li and Gaussier, 2010) is based on the expectation of finding the translation for each word in the corpus and gives a good idea about how two corpora are comparable. We can notice that all the comparable corpora have a high degree of comparability with a better comparability of the breast cancer corpora as opposed to the diabetes corpora. In the remainder of this article, [breast cancer corpus i] for instance stands for the breast cancer comparable corpus composed of the unique French part and the English part i ($i \in [1, 14]$).

3.2 Bilingual Dictionary

The bilingual dictionary used in our experiments is the French/English dictionary ELRA-M0033 available from the ELRA catalogue[5]. This resource is a general language dictionary which contains only a few terms related to the medical domain.

	Breast cancer	Diabetes
	# words (comp.)	# words (comp.)
French		
Part 1	7,376	4,982
English		
Part 1	8,214 (79.2)	5,181 (75.2)
Part 2	7,788 (78.8)	5,446 (75.9)
Part 3	8,370 (78.8)	5,610 (76.6)
Part 4	7,992 (79.3)	5,426 (74.8)
Part 5	7,958 (78.7)	5,610 (75.0)
Part 6	8,230 (79.1)	5,719 (73.6)
Part 7	8,035 (78.3)	5,362 (75.6)
Part 8	8,008 (78.8)	5,432 (74.6)
Part 9	8,334 (79.6)	5,398 (74.2)
Part 10	7,978 (79.1)	5,059 (75.6)
Part 11	8,373 (79.4)	5,264 (74.9)
Part 12	8,065 (78.9)	4,644 (73.4)
Part 13	7,847 (80.0)	5,369 (74.8)
Part 14	8,457 (78.9)	5,669 (74.8)

Table 3: Number of distinct words (# words) and degree of comparability (comp.) for each comparable corpora

3.3 Terminology Reference Lists

To evaluate the quality of terminology extraction, we built a bilingual terminology reference list for each comparable corpus. We selected all French/English single words from the UMLS[6] meta-thesaurus. We kept only i) the French single words which occur more than four times in the French part and ii) the English single words which occur more than four times in each English part i[7]. As a result of filtering, 169 French/English single words were extracted for the breast cancer corpus and 244 French/English single words were extracted for the diabetes corpus. It should be noted that the evaluation of terminology extraction using specialized comparable corpora of-

[3] http://www.ncbi.nlm.nih.gov/pubmed/

[4] http://code.google.com/p/ttc-project

[5] http://www.elra.info/

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

[7] The threshold sets to four is required to build a bilingual terminology reference list composed of about a hundred words. This value is very low to obtain representative context vectors. For instance, Prochasson and Fung (2011) showed that the standard approach is not relevant for infrequent words (since the context vectors are very unrepresentative i.e. poor in information).

	Breast cancer corpus													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Balanced	26.1	26.2	21.0	27.0	22.8	27.1	26.3	25.8	29.2	23.3	21.7	**29.6**	29.1	26.1
Unbalanced	26.1	31.9	34.7	36.0	37.7	36.4	36.6	37.2	39.8	40.5	40.6	**42.3**	40.9	41.6

	Diabetes corpus													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Balanced	13.6	13.5	11.9	14.6	14.6	11.0	**16.5**	10.5	12.9	13.3	15.2	11.8	13.0	14.3
Unbalanced	13.6	17.5	18.9	21.2	23.4	23.8	24.8	24.7	24.7	24.4	24.8	25.2	**26.0**	24.9

Table 4: Results (MAP %) of the standard approach using the balanced and unbalanced comparable corpora

ten relies on lists of a small size: 95 single words in Chiao and Zweigenbaum (2002), 100 in Morin et al. (2007), 125 and 79 in Bouamor et al. (2013).

4 Experiments and Results

In this section, we present experiments to evaluate the influence of comparable corpus size and prediction models on the quality of bilingual terminology extraction.

We present the results obtained for the terms belonging to the reference list for English to French direction measured in terms of the Mean Average Precision (MAP) (Manning et al., 2008) as follows:

$$MAP(Ref) = \frac{1}{|Ref|} \sum_{i=1}^{|Ref|} \frac{1}{r_i} \qquad (7)$$

where $|Ref|$ is the number of terms of the reference list and r_i the rank of the correct candidate translation i.

4.1 Standard Approach Evaluation

In order to evaluate the influence of corpus size on the bilingual terminology extraction task, two experiments have been carried out using the standard approach. We first performed an experiment using each comparable corpus independently of the others (we refer to these corpora as balanced corpora). We then conducted a second experiment where we varied the size of the English part of the comparable corpus, from 530,000 to 7.4 million words for the breast cancer corpus in 530,000 words steps, and from 250,000 to 3.5 million words for the diabetes corpus in 250,000 words steps (we refer to these corpora as unbalanced corpora). In the experiments reported here, the size of the context window w was set to 3 (i.e. a seven-word window that approximates syntactic dependencies), the retained association and similarity measures were the discounted log-odds and the Cosine (see Section 2.1). The results shown were those that give the best performance for the comparable corpora used individually.

Table 4 shows the results of the standard approach on the balanced and the unbalanced breast cancer and diabetes comparable corpora. Each column corresponds to the English part i ($i \in [1, 14]$) of a given comparable corpus. The first line presents the results for each individual comparable corpus and the second line presents the results for the cumulative comparable corpus. For instance, the column 3 indicates the MAP obtained by using a comparable corpus that is composed i) only of [breast cancer corpus 3] (MAP of 21.0%), and ii) of [breast cancer corpus 1, 2 and 3] (MAP of 34.7%).

As a preliminary remark, we can notice that the results differ noticeably according to the comparable corpus used individually (MAP variation between 21.0% and 29.6% for the breast cancer corpora and between 10.5% and 16.5% for the diabetes corpora). We can also note that the MAP of all the unbalanced comparable corpora is always higher than any individual comparable corpus. Overall, starting with a MAP of 26.1% as provided by the balanced [breast cancer corpus 1], we are able to increase it to 42.3% with the unbalanced [breast cancer corpus 12] (the variation observed for some unbalanced corpora such as [diabetes corpus 12, 13 and 14] can be explained by the fact that adding more data in the source language increases the error rate of the translation phase of the standard approach, which leads to the introduction of additional noise in the translated context vectors).

	Balanced breast cancer corpus													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$No\,prediction$	26.1	26.2	21.0	27.0	22.8	27.1	26.3	25.8	29.2	23.3	21.7	**29.6**	29.1	26.1
$Source_{pred}$	26.5	26.0	23.0	30.0	25.4	30.1	28.3	29.4	**32.1**	24.9	24.4	30.5	30.1	29.0
$Target_{pred}$	19.5	20.0	17.2	23.4	19.9	23.1	21.4	21.6	24.1	19.3	18.1	**26.6**	24.3	22.6
$Source_{pred} + Target_{pred}$	23.9	21.9	20.5	25.8	23.5	25.3	24.1	26.1	27.4	22.5	21.0	25.6	**28.5**	24.6
	Balanced diabetes corpus													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$No\,prediction$	13.6	13.5	11.9	14.6	14.6	11.0	**16.5**	10.5	12.9	13.3	15.2	11.8	13.0	14.3
$Source_{pred}$	13.9	14.3	12.6	15.5	14.9	10.9	**17.6**	11.1	14.0	14.2	16.4	13.3	13.5	15.7
$Target_{pred}$	09.8	09.0	08.3	11.9	10.1	08.0	**15.9**	07.3	10.8	10.0	10.1	08.8	10.8	10.2
$Source_{pred} + Target_{pred}$	10.9	11.0	09.0	13.6	11.8	08.6	**15.4**	07.7	12.8	11.5	11.9	10.5	11.7	11.8

Table 5: Results (MAP %) of the standard approach using the Lin regression model on the balanced breast cancer and diabetes corpora (comparison of predicting the source side, the target side and both sides of the comparable corpora)

4.2 Prediction Evaluation

The aim of this experiment is two-fold: first, we want to evaluate the usefulness of predicting word co-occurrence counts and second, we want to find out whether it is more appropriate to apply prediction to the source side, the target side or both sides of the bilingual comparable corpora.

	Breast cancer	Diabetes
$No\,prediction$	29.6	16.5
Lin	30.5	**17.6**
$Poly^2$	**30.6**	17.5
$Poly^3$	30.4	**17.6**
$Logit$	22.3	13.6

Table 6: Results (MAP %) of the standard approach using different regression models on the balanced breast cancer and diabetes corpora

4.2.1 Regression Models Comparison

We contrast the prediction models presented in Section 2.2 to findout which is the most appropriate model to use as a pre-processing step of the standard approach. We chose the balanced corpora where the standard approach has shown the best results in the previous experiment, namely [breast cancer corpus 12] and [diabetes corpus 7].

Table 6 shows a comparison between the standard approach without prediction noted No $prediction$ and the standard approach with prediction models. We contrast the simple linear regression model (Lin) with the second and the third order polynomial regressions ($Poly^2$ and $Poly^3$) and the logistic regression model ($Logit$). We can notice that except for the $Logit$ model, all the regression models outperform the baseline (No $prediction$). Also, as we can see, the results obtained with the linear and polynomial regressions are very close. This suggests that both linear and polynomial regressions are suitable as a preprocessing step of the standard approach, while the logistic regression seems to be inappropriate according to the results shown in Table 6.

That said, the gain of regression models is not significant. This may be due to the regression parameters that have been learned from a training corpus of the general domain. Another reason that could explain these results is the prediction process. We applied the same regression function to all co-occurrence counts while learning models for low and high frequencies should have been more appropriate. In the light of the above results, we believe that prediction can be beneficial to our task.

4.2.2 Source versus Target Prediction

Table 5 shows a comparison between the standard approach without prediction noted $No\,prediction$ and the standard approach based on the prediction of the source side noted $Source_{pred}$, the target side noted $Target_{pred}$ and both sides noted $Source_{pred}$+$Target_{pred}$. If prediction can not replace a large amount of data, it aims at increasing co-occurrence counts as if large amounts of data were at our disposal. In this case, applying prediction to the source side may simulate a configuration of using unbalanced comparable corpora where the source side is n times bigger than the target side. Predicting the target side only, may

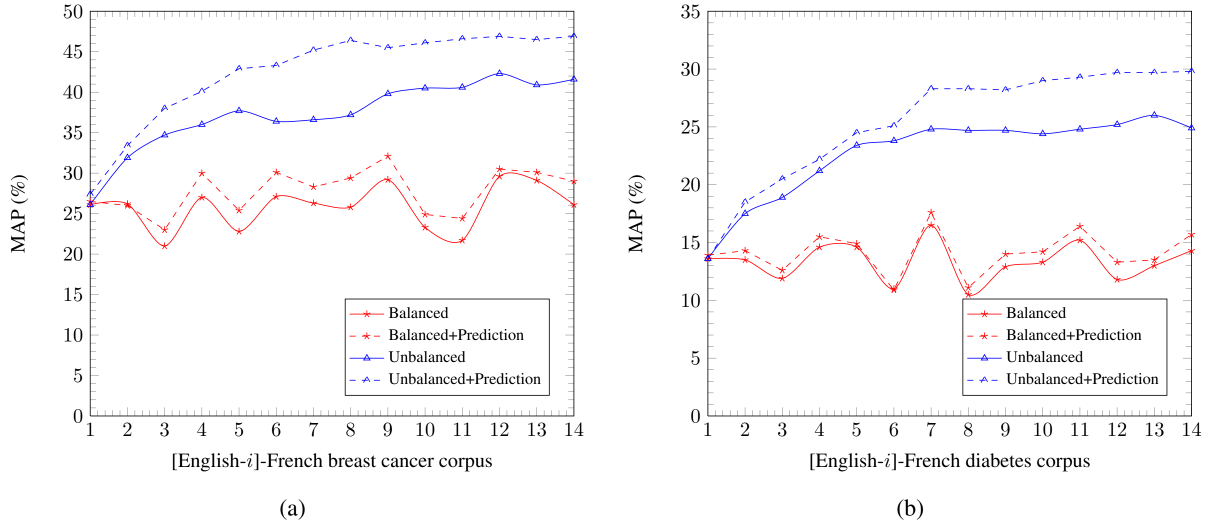

Figure 1: Results (MAP %) of the standard approach using the best configurations of the prediction models (Lin for $Balanced + Prediction$ and $Poly^2$ for $Unbalanced + Prediction$) on the breast cancer and the diabetes corpora

leads us to the opposite configuration where the target side is n times bigger than the source side. Finally, predicting both sides may simulate a large comparable corpora on both sides. In this experiment, we chose to use the linear regression model (Lin) for the prediction part. That said, the other regression models have shown the same behavior as Lin.

We can see that the best results are obtained by the $Source_{pred}$ approach for both comparable corpora. We can also notice that predicting the target side and both sides of the comparable corpora degrades the results. It is not surprising that predicting the target side only leads to lower results, since it is well known that a better characterization of a word to translate (given from the source side) leads to better results. We can deduce from Table 5 that source prediction is the most appropriate configuration to improve the quality of extracted lexicons. This configuration which simulates the use of unbalanced corpora leads us to think that using prediction with unbalanced comparable corpora should also increase the performance of the standard approach. This assumption is evaluated in the next Subsection.

4.3 Predicting Unbalanced Corpora

In this last experiment we contrast the standard approach applied to the balanced and unbalanced corpora noted $Balanced$ and $Unbalanced$ with the standard approach combined with the prediction model noted $Balanced + Prediction$ and $Unbalanced + Prediction$.

Figure 1(a) illustrates the results of the experiments conducted on the breast cancer corpus. We can see that the $Unbalanced$ approach significantly outperforms the baseline ($Balanced$). The big difference between the $Balanced$ and the $Unbalanced$ approaches would indicate that the latter is optimal. We can also notice that the prediction model applied to the balanced corpus ($Balanced + Prediction$) slightly outperforms the baseline while the $Unbalanced + Prediction$ approach significantly outperforms the three other approaches (moreover the variation observed with the $Unbalanced$ approach are lower than the $Unbalanced + Prediction$ approach). Overall, the prediction increases the performance of the standard approach especially for unbalanced corpora.

The results of the experiments conducted on the diabetes corpus are shown in Figure 1(b). As for the previous experiment, we can see that the $Unbalanced$ approach significantly outperforms the $Balanced$ approach. This confirms the unbalanced hypothesis and would motivate the use of unbalanced corpora when they are available. We can also notice that the $Balanced + Prediction$ approach slightly outperforms the baseline while the $Unbalanced + Prediction$ approach gives the best results. Here also, the prediction increases the performance of the standard approach especially for unbalanced corpora. It is clear that in addition to the benefit of using unbalanced comparable corpora, prediction shows a positive impact on the performance of the standard approach.

5 Conclusion

In this paper, we have studied how an unbalanced specialized comparable corpus could influence the quality of the bilingual lexicon extraction. This aspect represents a significant interest when working with specialized comparable corpora for which the quantity of the data collected may differ depending on the languages involved, especially when involving the English language as many scientific documents are available. More precisely, our different experiments show that using an unbalanced specialized comparable corpus always improves the quality of word translations. Thus, the MAP goes up from 29.6% (best result on the balanced corpora) to 42.3% (best result on the unbalanced corpora) in the breast cancer domain, and from 16.5% to 26.0% in the diabetes domain. Additionally, these results can be improved by using a prediction model of the word co-occurrence counts. Here, the MAP goes up from 42.3% (best result on the unbalanced corpora) to 46.9% (best result on the unbalanced corpora with prediction) in the breast cancer domain, and from 26.0% to 29.8% in the diabetes domain. We hope that this study will pave the way for using specialized unbalanced comparable corpora for bilingual lexicon extraction.

Acknowledgments

This work is supported by the French National Research Agency under grant ANR-12-CORD-0020.

References

Alan Agresti. 2007. *An Introduction to Categorical Data Analysis (2nd ed.).* Wiley & Sons, Inc., Hoboken, New Jersey.

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 (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 (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In *The New Navigators: from Professionals to Patients, Actes Medical Informatics Europe*, pages 397–402.

Ronald Christensen. 1997. *Log-Linear Models and Logistic Regression.* Springer-Verlag, Berlin.

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.

Mona T. Diab and Steve Finch. 2000. A Statistical Word-Level Translation Model for Comparable Corpora. In *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (RIAO'00)*, pages 1500–1501, Paris, France.

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

Pascale Fung and Percy Cheung. 2004. Multilevel bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 1051–1057, Geneva, Switzerland.

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. 1995. Compiling Bilingual Lexicon Entries from a non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Cambridge, MA, USA.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora 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. 1994. 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.

Amir Hazem and Emmanuel Morin. 2013. Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP'13)*, pages 1392–1400, Nagoya, Japan.

Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 481–489, Beijing, China.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition (ULA'02)*, pages 9–16, Philadelphia, PA, 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.

Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 644–652, Beijing, China.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Anthony McEnery and Zhonghua Xiao. 2007. Parallel and comparable corpora: What are they up to? In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: Translation and the Linguist*, Multilingual Matters, chapter 2, pages 18–31. Clevedon, UK.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC'11)*, pages 27–34, Portland, OR, USA.

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.

Pablo Gamallo Otero. 2007. Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of the 11th Conference on Machine Translation Summit (MT Summit XI)*, pages 191–198, Copenhagen, Denmark.

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.

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 1327–1335, Portland, OR, USA.

Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings of the 12th Conference on Machine Translation Summit (MT Summit XII)*, pages 284–291, Ottawa, Canada.

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.

Raphaël Rubino and Georges Linarès. 2011. A multi-view approach for term translation spotting. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, pages 29–40, Tokyo, Japan.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 580–585, Copenhagen, Denmark.

Kun Yu and Junichi Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'09)*, pages 121–124, Boulder, CO, USA.