Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora

Bo Li, Eric Gaussier
Laboratoire d'Informatique de Grenoble (LIG)
Université de Grenoble
firstname.lastname@imag.fr

Abstract

Previous work on bilingual lexicon extraction from comparable corpora aimed at finding a good representation for the usage patterns of source and target words and at comparing these patterns efficiently. In this paper, we try to work it out in another way: improving the quality of the comparable corpus from which the bilingual lexicon has to be extracted. To do so, we propose a measure of comparability and a strategy to improve the quality of a given corpus through an iterative construction process. Our approach, being general, can be used with any existing bilingual lexicon extraction method. We show here that it leads to a significant improvement over standard bilingual lexicon extraction methods.

1 Introduction

Bilingual dictionaries are an essential resource in many multilingual natural language processing (NLP) tasks such as machine translation (Och and Ney, 2003) and cross-language information retrieval (CLIR) (Ballesteros and Croft, 1997). Hand-coded dictionaries are of high quality, but expensive to build and researchers have tried, since the end of the 1980s, to automatically extract bilingual lexicons from parallel corpora (see (Chen, 1993; Kay and Röscheisen, 1993; Melamed, 1997a; Melamed, 1997b) for early work). Parallel corpora are however difficult to get at in several domains, and the majority of bilingual collections are comparable and not parallel. Due to their low cost of acquisition, several researchers have tried to exploit such comparable corpora for bilingual lexicon extraction (Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Déjean et al., 2002; Gaussier et al., 2004; Robitaille et al., 2006; Morin et al., 2007; Yu and Tsujii, 2009). The notion of comparability is however a loose one, and comparable corpora range from lowly comparable ones to highly comparable ones and parallel ones. For data-driven NLP techniques, using better corpora often leads to better results, a fact which should be true for the task of bilingual lexicon extraction. This point has largely been ignored in previous work on the subject. In this paper, we develop a well-founded strategy to improve the quality of a comparable corpus, so as to improve in turn the quality of the bilingual lexicon extracted. To do so, we first propose a measure of comparability which we then use in a method to improve the quality of the existing corpus.

The remainder of the paper is organized as follows: Section 2 introduces the experimental materials used for the different evaluations; comparability measures are then presented and evaluated in Section 3; in Section 4, we detail and evaluate a strategy to improve the quality of a given corpus while preserving its vocabulary; the method used for bilingual lexicon extraction is then described and evaluated in Section 5. Section 6 is then devoted to a discussion, prior to the conclusion given in Section 7.

2 Experimental Materials

For the experiments reported here, several corpora were used: the parallel English-French *Europarl* corpus (Koehn, 2005), the TREC

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 644–652,
Beijing, August 2010

(http://trec.nist.gov/) *Associated Press* corpus (*AP*, English) and the corpora used in the multilingual track of CLEF (http://www.clef-campaign.org) which includes the *Los Angeles Times* (*LAT94*, English), *Glasgow Herald* (*GH95*, English), *Le Monde* (*MON94*, French), *SDA French 94* (*SDA94*, French) and *SDA French 95* (*SDA95*, French). In addition to these existing corpora, two monolingual corpora from the Wikipedia dump[1] were built. For English, all the articles below the root category *Society* with a depth less than 4[2] were retained. For French, all the articles with a depth less than 7 below the category *Société* are extracted. As a result, the English corpus *Wiki-En* consists of 367,918 documents and the French one *Wiki-Fr* consists of 378,297 documents.

The bilingual dictionary used in our experiments is constructed from an online dictionary. It consists of 33,372 distinct English words and 27,733 distinct French words, which constitutes 75,845 translation pairs. Standard preprocessing steps: tokenization, POS-tagging and lemmatization are performed on all the linguistic resources. We will directly work on lemmatized forms of content words (nouns, verbs, adjectives, adverbs).

3 Measuring Comparability

As far as we can tell, there are no practical measures with which we can judge the degree of comparability of a bilingual corpus. In this paper, we propose a comparability measure based on the expectation of finding the translation for each word in the corpus. The measure is light-weighted and does not depend on complex resources like the machine translation system. For convenience, the following discussions will be made in the context of the English-French comparable corpus.

3.1 The Comparability Measure

For the comparable corpus $\mathcal{C}$, if we consider the translation process from the English part $\mathcal{C}_e$ to the French part $\mathcal{C}_f$, a comparability measure M_{ef} can be defined on the basis of the expectation of finding, for each English word w_e in the vocabulary $\mathcal{C}_e^v$ of $\mathcal{C}_e$, its translation in the vocabulary $\mathcal{C}_f^v$ of $\mathcal{C}_f$. Let σ be a function indicating whether a translation from the translation set $\mathcal{T}_w$ of w is found in the vocabulary $\mathcal{C}^v$ of a corpus $\mathcal{C}$, i.e.:

$$\sigma(w, \mathcal{C}^v) = \begin{cases} 1 & \text{iff } \mathcal{T}_w \cap \mathcal{C}^v \neq \emptyset \\ 0 & \text{else} \end{cases}$$

M_{ef} is then defined as:

$$\begin{aligned} M_{ef}(\mathcal{C}_e, \mathcal{C}_f) &= \mathbb{E}(\sigma(w, \mathcal{C}_f^v) | w \in \mathcal{C}_e^v) \\ &= \sum_{w \in \mathcal{C}_e^v} \underbrace{\sigma(w, \mathcal{C}_f^v) \cdot Pr(w \in \mathcal{C}_e^v)}_{A_w} \\ &= \frac{|\mathcal{C}_e^v|}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v|} \sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} A_w \end{aligned}$$

where $\mathcal{D}_e^v$ is the English part of a given, independent bilingual dictionary $\mathcal{D}$, and where the last equality is based on the fact that, the comparable corpus and the bilingual dictionary being independent of one another, the probability of finding the translation in $\mathcal{C}_f^v$ of a word w is the same for w is in $\mathcal{C}_e^v \cap \mathcal{D}_e^v$ and in $\mathcal{C}_e^v \backslash \mathcal{D}_e^v$[3]. Furthermore, the presence of common words suggests that one should rely on a presence/absence criterion rather than on the number of occurrences to avoid a bias towards common words. Given the natural language text, our evaluation will show that the simple presence/absence criterion can perform very well. This leads to $Pr(w \in \mathcal{C}_e^v) = 1/|\mathcal{C}_e^v|$, and finally to:

$$M_{ef}(\mathcal{C}_e, \mathcal{C}_f) = \frac{1}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v|} \sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v)$$

This formula shows that M_{ef} is actually the proportion of English words translated in the French part of the comparable corpus. Similarly, the counterpart of M_{ef}, M_{fe}, is defined as:

$$M_{fe}(\mathcal{C}_e, \mathcal{C}_f) = \frac{1}{|\mathcal{C}_f^v \cap \mathcal{D}_f^v|} \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v)$$

[1] The Wikipedia dump files can be downloaded at http://download.wikimedia.org. In this paper, we use the English dump file on July 13, 2009 and the French dump file on July 7, 2009.

[2] There are several cycles in the category tree of Wikipedia. It is thus necessary to define a threshold on the depth to make the iterative process feasible.

[3] The fact can be reliable only when a substantial part of the corpus vocabulary is covered by the dictionary. Fortunately, the constraint is satisfied in most applications where the common but not the specialized corpora like the medical corpora are involved.

and measures the proportion of French words in $\mathcal{C}_f^v$ translated in the English part of the comparable corpus. A symmetric version of these measures is obtained by considering the proportion of the words (both English and French) for which a translation can be found in the corpus:

$$M(\mathcal{C}_e, \mathcal{C}_f) = \frac{\sum_{w \in \mathcal{C}_e^v \cap \mathcal{D}_e^v} \sigma(w, \mathcal{C}_f^v) + \sum_{w \in \mathcal{C}_f^v \cap \mathcal{D}_f^v} \sigma(w, \mathcal{C}_e^v)}{|\mathcal{C}_e^v \cap \mathcal{D}_e^v| + |\mathcal{C}_f^v \cap \mathcal{D}_f^v|}$$

We now present an evaluation of these measures on artificial test corpora.

3.2 Validation

In order to test the comparability measures, we developed gold-standard comparability scores from the *Europarl* and *AP* corpora. We start from the parallel corpus, *Europarl*, of which we degrade the comparability by gradually importing some documents from either *Europarl* or *AP*. Three groups (G_a, G_b, G_c) of comparable corpora are built in this fashion. Each group consists of test corpora with a gold-standard comparability ranging, arbitrarily, from 0 to 1 and corresponding to the proportion of documents in "parallel" translation. The first group G_a is built from *Europarl* only. First, the *Europarl* corpus is split into 10 equal parts, leading to 10 parallel corpora (P_1, P_2, ..., P_{10}) with a gold-standard comparability arbitrarily set to 1. Then for each parallel corpus, e.g. P_i, we replace a certain proportion p of the English part with documents of the same size from another parallel corpus $P_j (j \neq i)$, producing the new corpus P_i' with less comparability which is the gold-standard comparability $1 - p$. For each P_i, as p increases, we obtain several comparable corpora with a decreasing gold-standard comparability score. All the P_i and their descendant corpora constitute the group G_a. The only difference between G_b and G_a is that, in G_b, the replacement in P_i is done with documents from the *AP* corpus and not from *Europarl*. In G_c, we start with 10 final, comparable corpora P_i' from G_a. These corpora have a gold-standard comparability of 0 in G_a, and of 1 in G_c. Then each P_i' is further degraded by replacing certain portions with documents from the *AP* corpus.

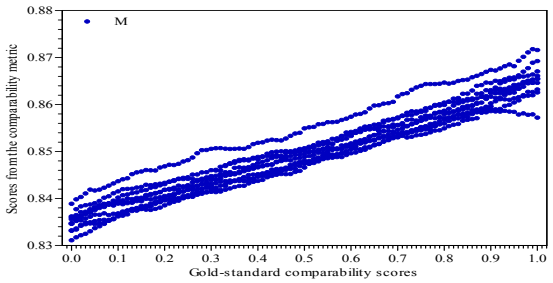

Figure 1: Evolution of M wrt gold-standard on the corpus group G_c (x-axis: gold-standard scores, y-axis: *M* scores)

We then computed, for each comparable corpus in each group, its comparability according to one of the comparability measures. Figure 1 plots the measure M for ten comparable corpora and their descendants from G_c, according to their gold-standard comparability scores. As one can note, the measure M is able to capture almost all the differences in comparability and is strongly correlated with the gold-standard. The correlation between the different measures and the gold-standard is finally computed with Pearson correlation coefficient. The results obtained are listed in Table 1. As one can note, M_{fe} performs worst among the three measures, the reason being that the process to construct G_b and G_c yields unbalanced bilingual corpora, the English section being larger than the French one. Translations of French words are still likely to be found in the English corpus, even though the corpora are not comparable. On all the 3 groups, M performs best and correlates very well with the gold standard, meaning that M was able to capture all the differences in comparability artificially introduced in the degradation process we have considered. This is the measure we will retain in the following parts.

	M_{ef}	M_{fe}	M
G_a	0.897	0.770	0.936
G_b	0.955	0.190	0.979
G_c	0.940	-0.595	0.960

Table 1: Correlation scores for the different comparability measures on the 3 groups of corpora

Having established a measure for the degree of comparability of bilingual corpora, we now turn to the problem of improving the quality of comparable corpora.

4 Improving Corpus Quality

We here try to improve the quality of a given corpus $\mathcal{C}$, which we will refer to as the *base corpus*, by extracting the highly comparable subpart $\mathcal{C}_H$ which is above a certain degree of comparability η (Step 1), and by enriching the lowly comparable part $\mathcal{C}_L$ with texts from other sources (Step 2). As we are interested in extracting information related to the vocabulary of the base corpus, we want the newly built corpus to contain a substantial part of the base corpus. This can be achieved by preserving in Step 1 as many documents from the base corpus as possible (e.g. by considering low values of η), and by using in step 2 sources close to the base corpus.

4.1 Step 1: Extracting $\mathcal{C}_H$

The strategy consisting of building all the possible sub-corpora of a given size from a given comparable corpora is not realistic as soon as the number of documents making up the corpora is larger than a few thousands. In such cases, better ways for extracting subparts have to be designed. The strategy we have adopted here aims at efficiently extracting a subpart of $\mathcal{C}$ above a certain degree of comparability and is based on the following property.

Property 1. *Let d_e^1 and d_e^2 (resp. d_f^1 and d_f^2) be two English (resp. French) documents from a bilingual corpus $\mathcal{C}$. We consider, as before, that the bilingual dictionary $\mathcal{D}$ is independent from $\mathcal{C}$. Let $(d_e^{1'}, d_f^{1'})$ be such that: $d_e^{1'} \subseteq d_e^1$, $d_f^{1'} \subseteq d_f^1$, which means $d_e^{1'}$ is a subpart of d_e^1 and $d_f^{1'}$ is a subpart of d_f^1.*
We assume:

$$\textit{(i)} \quad \frac{|d_e^1 \cup d_e^2|}{|d_e^2|} = \frac{|d_f^1 \cup d_f^2|}{|d_f^2|}$$

$$\textit{(ii)} \quad M_{ef}(d_e^{1'}, d_f^1) \geq M_{ef}(d_e^2, d_f^2)$$
$$M_{fe}(d_e^1, d_f^{1'}) \geq M_{fe}(d_e^2, d_f^2)$$

Then:

$$M(d_e^2, d_f^2) \leq M(d_e^1 \cup d_e^2, d_f^1 \cup d_f^2)$$

Proof [sketch]: Let $B = (d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v) \backslash (d_e^2 \cap \mathcal{D}_e^v)$. One can show, by exploiting condition (ii), that:

$$\sum_{w \in B} \sigma(w, d_f^1 \cup d_f^2) \geq |B| M_{ef}(d_e^2, d_f^2)$$

and similarly that:

$$\sum_{w \in d_e^2 \cap \mathcal{D}_e^v} \sigma(w, d_f^1 \cup d_f^2) \geq |d_e^2 \cap \mathcal{D}_e^v| M_{ef}(d_e^2, d_f^2)$$

Then exploiting condition (i), and the independence between the corpus and the dictionary, one arrives at:

$$\frac{\sum_{w \in (d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v} \sigma(w, d_f^1 \cup d_f^2)}{|(d_e^1 \cup d_e^2) \cap \mathcal{D}_e^v| + |(d_f^1 \cup d_f^2) \cap \mathcal{D}_f^v|} \geq \frac{|d_e^2 \cap \mathcal{D}_e^v| M_{ef}(d_e^2, d_f^2)}{|d_e^2 \cap \mathcal{D}_e^v| + |d_f^2 \cap \mathcal{D}_f^v|}$$

The same development on M_{fe} completes the proof. □

Property 1 shows that one can incrementally extract from a bilingual corpus a subpart with a guaranteed minimum degree of comparability η by iteratively adding new elements, provided (a) that the new elements have a degree of comparability of at least η and (b) that they are less comparable than the currently extracted subpart (conditions (ii)). This strategy is described in Algorithm 1. Since the degree of comparability is always above a certain threshold and since the new documents selected (d_e^2, d_f^2) are the most comparable among the remaining documents, condition (i) is likely to be satisfied, as this condition states that the increase in the vocabulary from the second documents to the union of the two is the same in both languages. Similarly, considering new elements by decreasing comparability scores is a necessary step for the satisfaction of condition (ii), which states that the current subpart should be uniformly more comparable than the element to be added. Hence, the conditions for property 1 to hold are met in Algorithm 1, which finally yields a corpus with a degree of comparability of at least η.

4.2 Step 2: Enriching $\mathcal{C}_L$

This step tries to absorb knowledge from other resources, which will be called *external corpus*,

Algorithm 1

Input:

English document set $\mathcal{C}_e^d$ of $\mathcal{C}$
French document set $\mathcal{C}_f^d$ of $\mathcal{C}$
Threshold η

Output:

$\mathcal{C}_H$, consisting of the English document set $\mathcal{S}_e$ and the French document set $\mathcal{S}_f$

Initialize $\mathcal{S}_e = \emptyset, \mathcal{S}_f = \emptyset$, temp $= 0$;
repeat
  $(d_e, d_f) = \underset{d_e \in \mathcal{C}_e^d, d_f \in \mathcal{C}_f^d}{\arg\max} M(d_e, d_f)$;
  temp $= \underset{d_e \in \mathcal{C}_e^d, d_f \in \mathcal{C}_f^d}{\max} M(d_e, d_f)$;
  if temp $\geq \eta$ **then**
    Add d_e into $\mathcal{S}_e$ and add d_f into $\mathcal{S}_f$;
    $\mathcal{C}_e^d = \mathcal{C}_e^d \backslash d_e, \mathcal{C}_f^d = \mathcal{C}_f^d \backslash d_f$;
  end if
until $\mathcal{C}_e^d = \emptyset$ or $\mathcal{C}_f^d = \emptyset$ or temp $< \eta$
return $\mathcal{C}_H$;

to enrich the lowly comparable part $\mathcal{C}_L$ which is the left part in $\mathcal{C}$ during the creation of $\mathcal{C}_H$. One choice for obtaining the external corpus $\mathcal{C}_T$ is to fetch documents which are likely to be comparable from the Internet. In this case, we first extract representative words for each document in $\mathcal{C}_L$, translate them using the bilingual dictionary and retrieve associated documents via a search engine. An alternative approach is of course to use existing bilingual corpora. Once $\mathcal{C}_T$ has been constructed, the lowly comparable part $\mathcal{C}_L$ can be enriched in exactly the same way as in section 4.1: First, Algorithm 1 is used on the English part of $\mathcal{C}_L$ and the French part of $\mathcal{C}_T$ to get the high-quality document pairs. Then the French part of $\mathcal{C}_L$ is enriched with the English part of $\mathcal{C}_T$ by the same algorithm. All the high-quality document pairs are then added to $\mathcal{C}_H$ to constitute the final result.

4.3 Validation

We use here *GH95* and *SDA95* as the base corpus $\mathcal{C}^0$. In order to illustrate that the efficiency of the proposed algorithm is not confined to a specific external resource, we consider two external resources: (a) $\mathcal{C}_T^1$ made of *LAT94*, *MON94* and *SDA94*, and (b) $\mathcal{C}_T^2$ consisting of *Wiki-En* and *Wiki-Fr*. The number of documents in all the corpora after elimination of short documents (< 30 words) is listed in Table 2.

	$\mathcal{C}^0$	$\mathcal{C}_T^1$	$\mathcal{C}_T^2$
English	55,989	109,476	367,918
French	42,463	87,086	378,297

Table 2: The size of the corpora in the experiments

For the extraction of the highly comparable part $\mathcal{C}_H$ from the base corpus $\mathcal{C}^0$, we set η to 0.3 so as to extract a substantial subpart of $\mathcal{C}^0$. After this step, corresponding to Algorithm 1, we have 20,124 English-French document pairs in $\mathcal{C}_H$. The second step is to enrich the lowly comparable part $\mathcal{C}_L$ of the base corpus documents from the external resources $\mathcal{C}_T^1$ and $\mathcal{C}_T^2$. The final corpora we obtain consist of 46,996 document pairs for $\mathcal{C}^1$ (with $\mathcal{C}_T^1$) and of 54,402 document pairs for $\mathcal{C}^2$ (with $\mathcal{C}_T^2$), size similar to the one of $\mathcal{C}^0$. The proportion of documents (columns "D-e" and "D-f"), sentences (columns "S-e" and "S-f") and vocabulary (columns "V-e" and "V-f") of $\mathcal{C}^0$ found in $\mathcal{C}^1$ and $\mathcal{C}^2$ is given in Table 3. As one can note, the final corpora obtained through the method presented above preserve most of the information from the base corpus. Especially for the vocabulary, the final corpora cover nearly all the vocabulary of the base corpus. Considering the comparability scores, the comparability of $\mathcal{C}^1$ is 0.912 and the one of $\mathcal{C}^2$ is 0.916. Both of them are more comparable than the base corpus of which the comparability is 0.882.

From these results of the intrinsic evaluation, one can conclude that the strategy developed to improve the corpus quality while preserving most of its information is efficient: The corpora obtained here, $\mathcal{C}^1$ and $\mathcal{C}^2$, are more comparable than the base corpus $\mathcal{C}^0$ and preserve most of its information. We now turn to the problem of extracting bilingual lexicons from these corpora.

5 Bilingual Lexicon Extraction

Following standard practice in bilingual lexicon extraction from comparable corpora, we rely on the approach proposed by Fung and Yee (1998). In this approach, each word w is represented as a

	D-e	D-f	S-e	S-f	V-e	V-f
$\mathcal{C}^1$	0.669	0.698	0.821	0.805	0.937	0.981
$\mathcal{C}^2$	0.785	0.719	0.893	0.807	0.968	0.987

Table 3: Proportion of documents, sentences and vocabulary of $\mathcal{C}^0$ covered by the result corpora

context vector consisting of the weight $a(w^c)$ of each context word w^c, the context being extracted from a window running through the corpus. Once context vectors for English and French words have been constructed, a general bilingual dictionary $\mathcal{D}$ can be used to bridge them by accumulating the contributions from words that are translation of each other. Standard similarity measures, as the cosine or the Jaccard coefficient, can then be applied to compute the similarity between vectors. For example, the cosine leads to:

$$s_c(w_e, w_f) = \frac{\sum_{(w_e^c, w_f^c)\in\mathcal{D}} a(w_e^c)a(w_f^c)}{\|\overrightarrow{w_e}\| \cdot \|\overrightarrow{w_f}\|} \quad (1)$$

5.1 Using Algorithm 1 pseudo-Alignments

The process we have defined in the previous section to improve the quality of a given corpus while preserving its vocabulary makes use of highly comparable document pairs, and thus provides some loose alignments between the two corpora. One can thus try to leverage the above approach to bilingual lexicon extraction by re-weighting $s_c(w_e, w_f)$ by a quantity which is large if w_e and w_f appear in many document pairs with a high comparability score, and small otherwise. In this section, we can not use the alignments in algorithm 1 directly because the alignments in the comparable corpus should not be 1 to 1 and we did not try to find the precise 1 to 1 alignments in algorithm 1.

Let η be the threshold used in algorithm 1 to construct the improved corpus and let $\phi(d_e, d_f)$ be defined as:

$$\phi(d_e, d_f) = \begin{cases} 1 & \text{iff } M(d_e, d_f) \geq \eta \\ 0 & \text{else} \end{cases}$$

Let $\mathcal{H}_e$ (resp. $\mathcal{H}_f$) be the set of documents containing word w_e (resp. w_f). We define the joint probability of w_e and w_f as being proportional to the number of comparable document pairs they belong to, where two documents are comparable if their comparability score is above η, that is:

$$p(w_e, w_f) \propto \sum_{d_e\in\mathcal{H}_e, d_f\in\mathcal{H}_f} \phi(d_e, d_f)$$

The marginal probability $p(w_e)$ can then be written as:

$$\begin{aligned} p(w_e) &\propto \sum_{w_f\in\mathcal{C}_f^v} p(w_e, w_f) \\ &\propto \sum_{d_e\in\mathcal{H}_e} \sum_{d_f\in\mathcal{C}_f^d} |d_f| \cdot \phi(d_e, d_f) \end{aligned}$$

Assuming that all d_f in $\mathcal{C}_f^d$ have roughly the same vocabulary size and all d_e have the same number of comparable counterparts in $\mathcal{C}_f^d$, then the marginal probability can be simplified as: $p(w_e) \propto |\mathcal{H}_e|$. By resorting to the exponential of the point-wise mutual information, one finally obtains the following weight:

$$\begin{aligned} \pi(w_e, w_f) &= \frac{p(w_e, w_f)}{p(w_e) \cdot p(w_f)} \\ &\propto \frac{1}{|\mathcal{H}_e| \cdot |\mathcal{H}_f|} \sum_{d_e\in\mathcal{H}_e, d_f\in\mathcal{H}_f} \phi(d_e, d_f) \end{aligned}$$

which has the desired property: It is large if the two words appear in comparable document pairs more often than chance would predict, and small otherwise. We thus obtain the revised similarity score for w_e and w_f:

$$s_{cr}(w_e, w_f) = s_c(w_e, w_f) \cdot \pi(w_e, w_f) \quad (2)$$

5.2 Validation

In order to measure the performance of the bilingual lexicon extraction method presented above, we divided the original dictionary into 2 parts: 10% of the English words (3,338 words) together with their translations are randomly chosen and used as the evaluation set, the remaining words (30,034 words) being used to compute context vectors and similarity between them. In this study, the weight $a(w^c)$ used in the context vectors (see above) are taken to be the tf-idf score of w^c: $a(w^c) = \text{tf-idf}(w^c)$. English words not

present in $\mathcal{C}_e^v$ or with no translation in $\mathcal{C}_f^v$ are excluded from the evaluation set. For each English word in the evaluation set, all the French words in $\mathcal{C}_f^v$ are then ranked according to their similarity with the English word (using either equation 1 or 2). To evaluate the quality of the lexicons extracted, we first retain for each English word its N first translations, and then measure the precision of the lists obtained, which amounts in this case to the proportion of lists containing the correct translation (in case of multiple translations, a list is deemed to contain the correct translation as soon as one of the possible translations is present). This evaluation procedure has been used in previous work (e.g. (Gaussier et al., 2004)) and is now standard for the evaluation of lexicons extracted from comparable corpora. In this study, N is set to 20. Furthermore, several studies have shown that it is easier to find the correct translations for frequent words than for infrequent ones (Pekar et al., 2006). To take this fact into account, we distinguished different frequency ranges to assess the validity of our approach for all frequency ranges. Words with frequency less than 100 are defined as low-frequency words (W_L), whereas words with frequency larger than 400 are high-frequency words (W_H), and words with frequency in between are medium-frequency words (W_M).

We then tested the standard method based on the cosine similarity (equation 1) on the corpora $\mathcal{C}^0$, $\mathcal{C}_H$, $\mathcal{C}'_H$, $\mathcal{C}^1$ and $\mathcal{C}^2$. The results obtained are displayed in Table 4, and correspond to columns 2-6. They show that the standard approach performs significantly better on the improved corpora $\mathcal{C}^1$/$\mathcal{C}^2$ than on the base corpus $\mathcal{C}^0$. The overall precision is increased by 5.3% on $\mathcal{C}^1$ (corresponding to a relative increase of 26%) and 9.5% on $\mathcal{C}^2$ (corresponding to a relative increase of 51%), even though the low-frequency words, which dominate the overall precision, account for a higher proportion in $\mathcal{C}^1$ (61.3%) and $\mathcal{C}^2$ (61.3%) than in $\mathcal{C}^0$ (56.2%). For the medium and high frequency words, the precision is increased by over 11% on $\mathcal{C}^1$ and 16% on $\mathcal{C}^2$. As pointed out in other studies, the performance for the low-frequency words is usually bad due to the lack of context information. This explains the relatively small improvement obtained here (only 2.2% on $\mathcal{C}^1$ and 6.7% on $\mathcal{C}^2$). It should also be noticed that the performance of the standard approach is better on $\mathcal{C}^2$ than on $\mathcal{C}^1$, which may be due to the fact that $\mathcal{C}^2$ is slightly larger than $\mathcal{C}^1$ and thus provides more information or to the actual content of these corpora. Lastly, if we consider the results on the corpus $\mathcal{C}_H$ which is produced by only choosing the highly comparable part from $\mathcal{C}^0$, the overall precision is increased by only 1.9%, which might come from the fact that the size of $\mathcal{C}_H$ is less than half the size of $\mathcal{C}^0$. We also notice the better results on $\mathcal{C}_H$ than on $\mathcal{C}'_H$ of the same size which consists of randomly choosing documents from $\mathcal{C}^0$.

The results obtained with the refined approach making use of the comparable document pairs found in the improved corpus (equation 2) are also displayed in Table 4 (columns "$\mathcal{C}^1$ new" and "$\mathcal{C}^2$ new"). From these results, one can see that the overall precision is further improved by 2.0% on $\mathcal{C}^1$ and 2.3% on $\mathcal{C}^2$, compared with the standard approach. For all the low, medium and high-frequency words, the precision has been improved, which demonstrates that the information obtained through the corpus enrichment process contributes to improve the quality of the extracted bilingual lexicons. Compared with the original base corpus $\mathcal{C}^0$, the overall improvement of the precision on both $\mathcal{C}^1$ and $\mathcal{C}^2$ with the refined approach is significant and important (respectively corresponding to a relative improvement of 35% and 62%), which also demonstrates that the efficiency of the refined approach is not confined to a specific external corpus.

6 Discussion

It is in a way useless to deploy bilingual lexicon extraction techniques if translation equivalents are not present in the corpus. This simple fact is at the basis of our approach which consists in constructing comparable corpora close to the original corpus and which are more likely to contain translation equivalents as they have a guaranteed degree of comparability. The pseudo-alignments identified in the construction process are then used to leverage state-of-the-art bilingual lexicon extraction methods. This approach to bilingual lexicon extraction from comparable corpora radically differs, to our knowledge, from previous approaches

	$\mathcal{C}^0$	$\mathcal{C}_H$	$\mathcal{C}'_H$	$\mathcal{C}^1$	$\mathcal{C}^2$	$\mathcal{C}^1$ new	$> \mathcal{C}^1, > \mathcal{C}^0$	$\mathcal{C}^2$ new	$> \mathcal{C}^2, > \mathcal{C}^0$
W_L	0.114	0.144	0.125	0.136	0.181	0.156	2.0%, 4.2%	0.205	2.4%, 9.1%
W_M	0.233	0.313	0.270	0.345	0.401	0.369	2.4%, 3.6%	0.433	3.2%, 20.0%
W_H	0.417	0.456	0.377	0.568	0.633	0.581	1.3%, 16.4%	0.643	1.0%, 22.6%
All	**0.205**	**0.224**	**0.189**	**0.258**	**0.310**	**0.278**	**2.0%, 7.3%**	**0.333**	**2.3%, 12.8%**

Table 4: Precision of the different approaches on different corpora

which are mainly variants of the standard method proposed in (Fung and Yee, 1998) and (Rapp, 1999). For example, the method developed in (Déjean et al., 2002) and (Chiao and Zweigenbaum, 2002) involves a representation of dictionary entries with context vectors onto which new words are mapped. Pekar et al. (2006) smooth the context vectors used in the standard approach in order to better deal with low frequency words. A nice geometric interpretation of these processes is proposed in (Gaussier et al., 2004), which furthermore introduces variants based on Fisher kernels, Canonical Correlation Analysis and a combination of them, leading to an improvement of the F1-score of 2% (from 0.14 to 0.16) when considering the top 20 candidates. In contrast, the approach we have developed yields an improvement of 7% (from 0.13 to 0.20) of the F-1 score on $\mathcal{C}^2$, again considering the top 20 candidates. More important, however, is the fact that the approach we have developed can be used in conjunction with any existing bilingual extraction method, as the strategies for improving the corpus quality and the re-weighting formula (equation 2) are general. We will assess in the future whether substantial gains are also attained with other methods.

Some studies have tried to extract subparts of comparable corpora to complement existing parallel corpora. Munteanu (2004) thus developed a maximum entropy classifier aiming at extracting those sentence pairs which can be deemed parallel. The step for choosing similar document pairs in this work resembles some of our steps. However their work focuses on high quality and specific documents pairs, as opposed to the entire corpus of guaranteed quality we want to build. In this latter case, the cross-interaction between documents impacts the overall comparability score, and new methods, as the one we have introduced, need to be proposed. Similarly, Munteanu and Marcu (2006) propose a method to extract subsentential fragments from non-parallel corpora. Again, the targeted elements are very specific (parallel sentences or sub-sentences) and limited, and the focus is put on a few sentences which can be considered parallel. As already mentioned, we rather focus here on building a new corpus which preserves most of the information in the original corpus. The construction process we have presented is theoretically justified and allows one to preserve ca. 95% of the original vocabulary.

7 Conclusion

We have first introduced in this paper a comparability measure based on the expectation of finding translation word pairs in the corpus. We have then designed a strategy to construct an improved comparable corpus by (a) extracting a subpart of the original corpus with a guaranteed comparability level, and (b) by completing the remaining subpart with external resources, in our case other existing bilingual corpora. We have then shown how the information obtained during the construction process could be used to improve state-of-the-art bilingual lexicon extraction methods. We have furthermore assessed the various steps of our approach and shown: (a) that the comparability measure we introduced captures variations in the degree of comparability between corpora, (b) that the construction process we introduced leads to an improved corpus preserving most of the original vocabulary, and (c) that the use of pseudo-alignments through simple re-weighting yields bilingual lexicons of higher quality.

Acknowledgements

This work was supported by the French National Research Agency grant ANR-08-CORD-009.

References

Ballesteros, Lisa and W. Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th ACM SIGIR*, pages 84–91, Philadelphia, Pennsylvania, USA.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, USA.

Chiao, Yun-Chuang and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Taipei, Taiwan.

Déjean, Hervé, Eric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Taipei, Taiwan.

Fung, Pascale 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, Hong Kong.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, pages 414–420, Montreal, Quebec, Canada.

Gaussier, E., J.-M. Renders, I. Matveeva, C. Goutte, and H. 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*, pages 526–533, Barcelona, Spain.

Kay, Martin and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*.

Melamed, I. Dan. 1997a. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312, Madrid, Spain.

Melamed, I. Dan. 1997b. A word-to-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Madrid, Spain.

Morin, Emmanuel, 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*, pages 664–671, Prague, Czech Republic.

Munteanu, Dragos Stefan 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.

Munteanu, Dragos Stefan, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the HLT-NAACL 2004*, pages 265–272, Boston, MA., USA.

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

Pekar, Viktor, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Rapp, Reinhard. 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*, pages 519–526, College Park, Maryland, USA.

Robitaille, Xavier, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. 2006. Compiling French-Japanese terminologies from the web. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, pages 225–232, Trento, Italy.

Yu, Kun and Junichi Tsujii. 2009. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of HLT-NAACL 2009*, pages 121–124, Boulder, Colorado, USA.