Clustering Comparable Corpora For Bilingual Lexicon Extraction

Bo Li, Eric Gaussier
UJF-Grenoble 1 / CNRS, France
LIG UMR 5217
firstname.lastname@imag.fr

Akiko Aizawa
National Institute of Informatics
Tokyo, Japan
aizawa@nii.ac.jp

Abstract

We study in this paper the problem of enhancing the comparability of bilingual corpora in order to improve the quality of bilingual lexicons extracted from comparable corpora. We introduce a clustering-based approach for enhancing corpus comparability which exploits the homogeneity feature of the corpus, and finally preserves most of the vocabulary of the original corpus. Our experiments illustrate the well-foundedness of this method and show that the bilingual lexicons obtained from the homogeneous corpus are of better quality than the lexicons obtained with previous approaches.

1 Introduction

Bilingual lexicons are an important resource in multilingual natural language processing tasks such as statistical machine translation (Och and Ney, 2003) and cross-language information retrieval (Ballesteros and Croft, 1997). Because it is expensive to manually build bilingual lexicons adapted to different domains, researchers have tried to automatically extract bilingual lexicons from various corpora. Compared with parallel corpora, it is much easier to build high-volume comparable corpora, i.e. corpora consisting of documents in different languages covering overlapping information. Several studies have focused on the extraction of bilingual lexicons from comparable corpora (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; Garera et al., 2009; Yu and Tsujii, 2009; Shezaf and Rappoport, 2010). The basic assumption behind most studies on lexicon extraction from comparable corpora is a distributional hypothesis, stating that words which are translation of each other are likely to appear in similar context across languages. On top of this hypothesis, researchers have investigated the use of better representations for word contexts, as well as the use of different methods for matching words across languages. These approaches seem to have reached a plateau in terms of performance. More recently, and departing from such traditional approaches, we have proposed in (Li and Gaussier, 2010) an approach based on improving the comparability of the corpus under consideration, prior to extracting bilingual lexicons. This approach is interesting since there is no point in trying to extract lexicons from a corpus with a low degree of comparability, as the probability of finding translations of any given word is low in such cases. We follow here the same general idea and aim, in a first step, at improving the comparability of a given corpus while preserving most of its vocabulary. However, unlike the previous work, we show here that it is possible to guarantee a certain degree of *homogeneity* for the improved corpus, and that this homogeneity translates into a significant improvement of both the quality of the resulting corpora and the bilingual lexicons extracted.

2 Enhancing Comparable Corpora: A Clustering Approach

We first introduce in this section the comparability measure proposed in former work, prior to describing the clustering-based algorithm to improve the

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 473–478,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

quality of a given comparable corpus. For convenience, the following discussion will be made in the context of the English-French comparable corpus.

2.1 The Comparability Measure

In order to measure the degree of comparability of bilingual corpora, we make use of the measure M developed in (Li and Gaussier, 2010): Given a comparable corpus $\mathcal{P}$ consisting of an English part $\mathcal{P}_e$ and a French part $\mathcal{P}_f$, the degree of comparability of $\mathcal{P}$ is defined as the expectation of finding the translation of any given source/target word in the target/source corpus vocabulary. Let σ be a function indicating whether a translation from the translation set $\mathcal{T}_w$ of the word w is found in the vocabulary $\mathcal{P}^v$ of a corpus $\mathcal{P}$, i.e.:

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

and let $\mathcal{D}$ be a bilingual dictionary with $\mathcal{D}_e^v$ denoting its English vocabulary and $\mathcal{D}_f^v$ its French vocabulary. The comparability measure M can be written as:

$$M(\mathcal{P}_e, \mathcal{P}_f) \quad (1)$$
$$= \frac{\sum_{w \in \mathcal{P}_e \cap \mathcal{D}_e^v} \sigma(w, \mathcal{P}_f) + \sum_{w \in \mathcal{P}_f \cap \mathcal{D}_f^v} \sigma(w, \mathcal{P}_e)}{\#_w(\mathcal{P}_e \cap \mathcal{D}_e^v) + \#_w(\mathcal{P}_f \cap \mathcal{D}_f^v)}$$

where $\#_w(\mathcal{P})$ denotes the number of different words present in $\mathcal{P}$. One can find from equation 1 that M directly measures the proportion of source/target words translated in the target/source vocabulary of $\mathcal{P}$.

2.2 Clustering Documents for High Quality Comparable Corpora

If a corpus covers a limited set of topics, it is more likely to contain consistent information on the words used (Morin et al., 2007), leading to improved bilingual lexicons extracted with existing algorithms relying on the distributional hypothesis. The term *homogeneity* directly refers to this fact, and we will say, in an informal manner, that a corpus is homogeneous if it covers a limited set of topics. The rationale for the algorithm we introduce here to enhance corpus comparability is precisely based on the concept of homogeneity. In order to find document sets which are similar with each other (i.e. homogeneous), it is natural to resort to clustering techniques. Furthermore, since we need homogeneous corpora for bilingual lexicon extraction, it will be convenient to rely on techniques which allows one to easily prune less relevant clusters. To perform all this, we use in this work a standard hierarchical agglomerative clustering method.

2.2.1 Bilingual Clustering Algorithm

The overall process retained to build high quality, homogeneous comparable corpora relies on the following steps:

1. Using the bilingual similarity measure defined in Section 2.2.2, cluster English and French documents so as to get bilingual dendrograms from the original corpus $\mathcal{P}$ by grouping documents with related content;

2. Pick high quality sub-clusters by thresholding the obtained dendrograms according to the node depth, which retains nodes far from the roots of the clustering trees;

3. Combine all these sub-clusters to form a new comparable corpus $\mathcal{P}_H$, which thus contains homogeneous, high-quality subparts;

4. Use again steps (1), (2) and (3) to enrich the remaining subpart of $\mathcal{P}$ (denoted as $\mathcal{P}_L$, $\mathcal{P}_L = \mathcal{P} \setminus \mathcal{P}_H$) with external resources.

The first three steps aim at extracting the most comparable and homogeneous subpart of $\mathcal{P}$. Once this has been done, one needs to resort to new corpora if one wants to build an homogeneous corpus with a high degree of comparability from $\mathcal{P}_L$. To do so, we simply perform, in step (4), the clustering and thresholding process defined in (1), (2) and (3) on two comparable corpora: The first one consists of the English part of $\mathcal{P}_L$ and the French part of an external corpus $\mathcal{P}_T$; The second one consists of the French part of $\mathcal{P}_L$ and the English part of $\mathcal{P}_T$. The two high quality subparts obtained from these two new comparable corpora in step (4) are then combined with $\mathcal{P}_H$ to constitute the final comparable corpus of higher quality.

2.2.2 Similarity Measure

Let us assume that we have two document sets (i.e. clusters) $\mathcal{C}_1$ and $\mathcal{C}_2$. In the task of bilingual lexicon extraction, two document sets are similar to each other and should be clustered if the combination of the two can complement the content of each single set, which relates to the notion of homogeneity. In other words, both the English part $\mathcal{C}_1^e$ of $\mathcal{C}_1$ and the French part $\mathcal{C}_1^f$ of $\mathcal{C}_1$ should be comparable to their counterparts (respectively the same for the French part $\mathcal{C}_2^f$ of $\mathcal{C}_2$ and the English part $\mathcal{C}_2^e$ of $\mathcal{C}_2$). This leads to the following similarity measure for $\mathcal{C}_1$ and $\mathcal{C}_2$:

$$sim(\mathcal{C}_1, \mathcal{C}_2) = \beta \cdot M(\mathcal{C}_1^e, \mathcal{C}_2^f) + (1-\beta) \cdot M(\mathcal{C}_2^e, \mathcal{C}_1^f)$$

where $\beta\,(0 \leq \beta \leq 1)$ is a weight controlling the importance of the two subparts ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) and ($\mathcal{C}_2^e$, $\mathcal{C}_1^f$). Intuitively, the larger one, containing more information, of the two comparable corpora ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) and ($\mathcal{C}_2^e$, $\mathcal{C}_1^f$) should dominate the overall similarity $sim(\mathcal{C}_1, \mathcal{C}_2)$. Since the content relatedness in the comparable corpus is basically reflected by the relations between all the possible bilingual document pairs, we use here the number of document pairs to represent the scale of the comparable corpus. The weight β can thus be defined as the proportion of possible document pairs in the current comparable corpus ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) to all the possible document pairs, which is:

$$\beta = \frac{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f)}{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f) + \#_d(\mathcal{C}_2^e) \cdot \#_d(\mathcal{C}_1^f)}$$

where $\#_d(\mathcal{C})$ stands for the number of documents in $\mathcal{C}$. However, this measure does not integrate the relative length of the French and English parts, which actually impacts the performance of bilingual lexicon extraction. If a 1-to-1 constraint is too strong (i.e. assuming that all clusters should contain the same number of English and French documents), having completely unbalanced corpora is also not desirable. We thus introduce a penalty function ϕ aiming at penalizing unbalanced corpora:

$$\phi(\mathcal{C}) = \frac{1}{(1 + log(1 + \frac{|\#_d(\mathcal{C}^e) - \#_d(\mathcal{C}^f)|}{\min(\#_d(\mathcal{C}^e)), \#_d(\mathcal{C}^f))})} \quad (2)$$

The above penalty function leads us to a new similarity measure sim_l which is the one finally used in the above algorithm:

$$sim_l(\mathcal{C}_1, \mathcal{C}_2) = sim(\mathcal{C}_1, \mathcal{C}_2) \cdot \phi(\mathcal{C}_1 \cup \mathcal{C}_2) \quad (3)$$

3 Experiments and Results

The experiments we have designed in this paper aim at assessing (a) whether the clustering-based algorithm we have introduced yields corpora of higher quality in terms of comparability scores, and (b) whether the bilingual lexicons extracted from such corpora are of higher quality. Several corpora were used in our experiments: the TREC[1] *Associated Press* corpus (*AP*, English) and the corpora used in the CLEF[2] campaign including the *Los Angeles Times* (*LAT94*, English), the *Glasgow Herald* (*GH95*, English), *Le Monde* (*MON94*, French), *SDA French 94* (*SDA94*, French) and *SDA French 95* (*SDA95*, French). In addition, two monolingual corpora *Wiki-En* and *Wiki-Fr* were built by respectively retrieving all the articles below the category *Society* and *Société* from the Wikipedia dump files[3]. The bilingual dictionary used in the experiments is constructed from an online dictionary. It consists of 33k distinct English words and 28k distinct French words, constituting 76k translation pairs. In our experiments, we use the method described in this paper, as well as the one in (Li and Gaussier, 2010) which is the only alternative method to enhance corpus comparability.

3.1 Improving Corpus Quality

In this subsection, the clustering algorithm described in Section 2.2.1 is employed to improve the quality of the comparable corpus. The corpora *GH95* and *SDA95* are used as the original corpus $\mathcal{P}^0$ (56k English documents and 42k French documents). We consider two external corpora: $\mathcal{P}_T^1$ (109k English documents and 87k French documents) consisting of the corpora *LAT94*, *MON94* and *SDA94*; $\mathcal{P}_T^2$ (368k English documents and 378k French documents) consisting of *Wiki-En* and *Wiki-Fr*.

[1] http://trec.nist.gov

[2] http://www.clef-campaign.org

[3] 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.

	$\mathcal{P}^0$	$\mathcal{P}^{1'}$	$\mathcal{P}^{2'}$	$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^1 > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^0$
Precision	0.226	0.277	0.325	0.295	**0.461**	0.069, 30.5%	0.235, 104.0%
Recall	0.103	0.122	0.145	0.133	**0.212**	0.030, 29.1%	0.109, 105.8%

Table 1: Performance of the bilingual lexicon extraction from different corpora (best results in bold)

After the clustering process, we obtain the resulting corpora $\mathcal{P}^1$ (with the external corpus $\mathcal{P}^1_T$) and $\mathcal{P}^2$ (with $\mathcal{P}^2_T$). As mentioned before, we also used the method described in (Li and Gaussier, 2010) on the same data, producing resulting corpora $\mathcal{P}^{1'}$ (with $\mathcal{P}^1_T$) and $\mathcal{P}^{2'}$ (with $\mathcal{P}^2_T$) from $\mathcal{P}^0$. In terms of lexical coverage, $\mathcal{P}^1$ (resp. $\mathcal{P}^2$) covers 97.9% (resp. 99.0%) of the vocabulary of $\mathcal{P}^0$. Hence, most of the vocabulary of the original corpus has been preserved. The comparability score of $\mathcal{P}^1$ reaches 0.924 and that of $\mathcal{P}^2$ is 0.939. Both corpora are more comparable than $\mathcal{P}^0$ of which the comparability is 0.881. Furthermore, both $\mathcal{P}^1$ and $\mathcal{P}^2$ are more comparable than $\mathcal{P}^{1'}$ (comparability 0.912) and $\mathcal{P}^{2'}$ (comparability 0.915), which shows homogeneity is crucial for comparability. The intrinsic evaluation shows the efficiency of our approach which can improve the quality of the given corpus while preserving most of its vocabulary.

3.2 Bilingual Lexicon Extraction Experiments

To extract bilingual lexicons from comparable corpora, we directly use here the method proposed by Fung and Yee (1998) which has been referred to as the *standard approach* in more recent studies (Déjean et al., 2002; Gaussier et al., 2004; Yu and Tsujii, 2009). In this approach, each word w is represented as a context vector consisting of the words co-occurring with w in a certain window in the corpus. The context vectors in different languages are then bridged with an existing bilingual dictionary. Finally, a similarity score is given to any word pair based on the cosine of their respective context vectors.

3.2.1 Experiment Settings

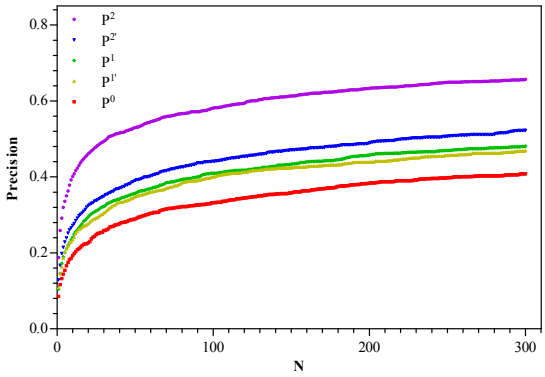
In order to measure the performance of the lexicons extracted, we follow the common practice by dividing the bilingual dictionary into 2 parts: 10% of the English words (3,338 words) together with their translations are randomly chosen and used as the evaluation set, the remaining words being used to compute the similarity of context vectors. English words not present in $\mathcal{P}_e$ or with no translation in $\mathcal{P}_f$ are excluded from the evaluation set. For each English word in the evaluation set, all the French words in $\mathcal{P}_f$ are then ranked according to their similarity with the English word. Precision and recall are then computed on the first N translation candidate lists. The precision 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). The recall is the proportion of correct translations found in the lists to all the translations in the corpus. This evaluation procedure has been used in previous studies and is now standard.

3.2.2 Results and Analysis

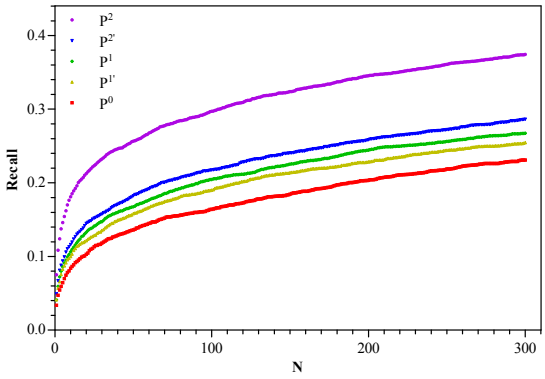
In a first series of experiments, bilingual lexicons were extracted from the corpora obtained by our approach ($\mathcal{P}^1$ and $\mathcal{P}^2$), the corpora obtained by the approach described in (Li and Gaussier, 2010) ($\mathcal{P}^{1'}$ and $\mathcal{P}^{2'}$) and the original corpus $\mathcal{P}^0$, with the fixed N value set to 20. Table 1 displays the results obtained. Each of the last two columns "$\mathcal{P}^1 > \mathcal{P}^0$" and "$\mathcal{P}^2 > \mathcal{P}^0$" contains the absolute and the relative difference (in %) w.r.t. $\mathcal{P}^0$. As one can note, the best results (in bold) are obtained from the corpora $\mathcal{P}^2$ built with the method we have described in this paper. The lexicons extracted from the enhanced corpora are of much higher quality than the ones obtained from the original corpus . For instance, the increase of the precision is 6.9% (30.5% relatively) in $\mathcal{P}^1$ and 23.5% (104.0% relatively) in $\mathcal{P}^2$, compared with $\mathcal{P}^0$. The difference is more remarkable with $\mathcal{P}^2$, which is obtained from a large external corpus $\mathcal{P}^2_T$. Intuitively, one can expect to find, in larger corpora, more documents related to a given corpus, an intuition which seems to be confirmed by our results. One can also notice, by comparing $\mathcal{P}^2$ and $\mathcal{P}^{2'}$ as well as $\mathcal{P}^1$ and $\mathcal{P}^{1'}$, a remarkable improvement when considering our approach and the early

methodology.

Intuitively, the value N plays an important role in the above experiments. In a second series of experiments, we let N vary from 1 to 300 and plot the results obtained with different evaluation measure in Figure 1. In Figure 1(a) (resp. Figure 1(b)), the x-axis corresponds to the values taken by N, and the y-axis to the precision (resp. recall) scores for the lexicons extracted on each of the 5 corpora $\mathcal{P}^0$, $\mathcal{P}^{1\prime}$, $\mathcal{P}^{2\prime}$, $\mathcal{P}^1$ and $\mathcal{P}^2$. A clear fact from the figure is that both the precision and the recall scores increase according to the increase of the N values, which coincides with our intuition. As one can note, our method consistently outperforms the previous work and also the original corpus on all the values considered for N.



(a) Precision



(b) Recall

Figure 1: Performance of bilingual lexicon extraction from different corpora with varied N values from 1 to 300. The five lines from the top down in each subfigure are corresponding to the results for $\mathcal{P}^2$, $\mathcal{P}^{2\prime}$, $\mathcal{P}^1$, $\mathcal{P}^{1\prime}$ and $\mathcal{P}^0$ respectively.

4 Discussion

As previous studies on bilingual lexicon extraction from comparable corpora radically differ on resources used and technical choices, it is very difficult to compare them in a unified framework (Laroche and Langlais, 2010). We compare in this section our method with some ones in the same vein (i.e. enhancing bilingual corpora prior to extracting bilingual lexicons from them). Some works like (Munteanu et al., 2004) and (Munteanu and Marcu, 2006) propose methods to extract parallel fragments from comparable corpora. However, their approach only focuses on a very small part of the original corpus, whereas our work aims at preserving most of the vocabulary of the original corpus.

We have followed here the general approach in (Li and Gaussier, 2010) which consists in enhancing the quality of a comparable corpus prior to extracting information from it. However, despite this latter work, we have shown here a method which ensures homogeneity of the obtained corpus, and which finally leads to comparable corpora of higher quality. In turn such corpora yield better bilingual lexicons extracted.

Acknowledgements

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

References

Lisa Ballesteros 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.

Hervé Déjean, 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.

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, Hong Kong.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international con-*

ference on Computational linguistics, pages 414–420, Montreal, Quebec, Canada.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *CoNLL 09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 129–137, Boulder, Colorado.

E. Gaussier, 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.

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 2010)*, pages 617–625, Beijing, China, August.

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

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

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.

Dragos Stefan Munteanu, 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.

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

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

Xavier Robitaille, 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.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 98–107, Uppsala, Sweden.

Kun Yu 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.