Bilingual Lexicon Extraction
from Comparable Corpora and involving a Pivot Language

By

ALEXIS LINARD

Laboratoire d’Informatique de Nantes Atlantique
UNIVERSITÉ DE NANTES

Co-tutored by : Béatrice DAILLE,
Emmanuel MORIN, Professors

JUNE 30TH 2015
Abstract

Bilingual alignment was initially designed as a specific task for parallel corpora in which a corpus is the translation of the other and vice versa. Nevertheless, parallel corpora has been scarce and hard to build so far due to the need of human translators so far. In order to avoid this, many techniques now involve comparable corpora instead. Comparable corpora are easier to build and to obtain, and also cost less. But it exists a strong dependency towards the quality of those resources. What we try to demonstrate here is that involving a third language as pivot language can improve the achieved results. Our hypothesis is that we can benefit from the richness of the pivot language’s resources – English by default. Our so far achieved result show that in the case of languages poor in resources, the benefits obtained from a third rich language are interesting.

Résumé

L’alignement bilingue était à ses débuts considéré comme une tâche spécifique aux corpus parallèles, où un corpus est la traduction de l’autre, et vice versa. Cependant, de telles ressources sont difficilement accessibles. De plus, construire des corpus parallèles bilingues est une tâche coûteuse et nécessitant, jusqu’à présent, l’intervention humaine. Pour résoudre ce problème, des techniques permettent l’utilisation de corpus comparables, plus simples à obtenir. Or, lors de l’utilisation de corpus comparables, il existe une forte dépendance vis-à-vis de la qualité des ressources. Ce que nous essayons de démontrer par nos travaux est que l’intervention d’une langue pivot peut améliorer les résultats. Notre hypothèse est que l’on peut tirer les bénéfices de la richesse des ressources de cette langue pivot, par défaut l’Anglais. Nos résultats montrent que lorsque les ressources des langues source et cible sont de piètre qualité, le gain apporté par la troisième langue riche en données est intéressant.
DEDICATION AND ACKNOWLEDGEMENTS

First of all, I would like to express my warm thanks to Professors Béatrice Daille and Emmanuel Morin for proposing me an internship under their supervision. I would also like to acknowledge them for their support and guidance at LINA, and for my future projects.

I would like express my gratitude to the technical staff who often fixed problems with my computer, and gave me access to LINA's calculator.

I am deeply grateful to Mr. Damien Cram, who quickly fixed some bugs found on TermSuite and who helped me continue working efficiently.

I wish to thank Dr. Amir Hazem who gave me some scripts to evaluate corpus comparability and thus helped me save time.

Thanks to Dr. Solen Quiniou for her help during the revision of this Master's thesis, and above all to Miss Martine Bourquin for correcting my English!

Many thanks to my office colleagues, for their kindness and the good moments we've had together. Thanks to Christine who also taught me how to use the bookbinder.

Finally, my deepest gratefulness goes to Frédéric and Annette, my parents, for their unconditional love as well as their past, present and, no doubt, future enthusiastic support.

Merci !
# Table of Contents

1 Introduction 1
   1.1 Goals and Objectives ............................................. 2
   1.2 Outline of the Dissertation .................................... 2

2 Internship context 3

3 State of the Art 5
   3.1 Introduction .................................................... 5
   3.2 Bilingual lexicon extraction using Comparable Corpora .............. 6
      3.2.1 Standard Method ............................................ 6
      3.2.2 Improving the Standard Method .............................. 9
      3.2.3 Extensions of the Standard Method .......................... 9
      3.2.4 Improving Corpora Comparability .......................... 10
   3.3 Bilingual lexicon extraction using Pivot Language .................... 10

4 Approaches involving pivot language 12
   4.1 Choosing the best pivot language ................................ 12
      4.1.1 Theory of linguistic continuum ............................ 13
      4.1.2 English, by default ....................................... 13
   4.2 Pivot dictionary based approaches ................................ 14
      4.2.1 Translating context vector successively ...................... 14
      4.2.2 Transposing both source and target vectors to pivot language .......... 15
   4.3 Pivot corpora based approaches ................................ 16
      4.3.1 Successive translation ..................................... 16
      4.3.2 Using linear regression to reshape context vectors ............. 17
      4.3.3 Computing similarity over pivot corpus ...................... 18
      4.3.4 Weighting candidate translations according to the pivot ones ........ 19
      4.3.5 Pivot candidates threshold .................................. 19

5 Experiments and Results 20
   5.1 Linguistic Resources ............................................ 20
      5.1.1 Specialized Comparable Corpora ............................ 21
      5.1.2 Terminology Reference Lists ................................ 22
      5.1.3 Bilingual Dictionaries ...................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 Evaluation of the approaches</td>
<td>23</td>
</tr>
<tr>
<td>5.2.1 Experimental parameters</td>
<td>23</td>
</tr>
<tr>
<td>5.2.2 Mean Average Precision</td>
<td>23</td>
</tr>
<tr>
<td>5.3 Results Achieved for Pivot Dictionary Based Approaches</td>
<td>24</td>
</tr>
<tr>
<td>5.4 Results Achieved for Pivot Corpora Based Approaches</td>
<td>26</td>
</tr>
<tr>
<td>5.4.1 Varying the pivot candidates threshold</td>
<td>26</td>
</tr>
<tr>
<td>5.4.2 Expectation of the quality of the pivot candidates</td>
<td>26</td>
</tr>
<tr>
<td>5.4.3 Results</td>
<td>27</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>30</td>
</tr>
<tr>
<td>6.1 Contributions</td>
<td>30</td>
</tr>
<tr>
<td>6.2 Further work</td>
<td>31</td>
</tr>
<tr>
<td>A Pivot Based Approaches</td>
<td>32</td>
</tr>
<tr>
<td>B Detailed Results</td>
<td>37</td>
</tr>
<tr>
<td>C Publications</td>
<td>43</td>
</tr>
<tr>
<td>Bibliography</td>
<td>51</td>
</tr>
</tbody>
</table>
Automatic translation is a Natural Language Processing task that is more and more employed nowadays. This task lays on statistical models, themselves laying on alignment tasks. Alignment tasks can be divided in two sub-fields. Monolingual alignments, that perform words alignments to find out the different contexts in which a word can occur in a specific language, are useless for automatic translation; and bilingual alignments can extract one word in the source language and its translation in the target language of a bilingual corpora. This bilingual alignment is called bilingual lexicon extraction. At this stage, we distinguish two kinds of bilingual lexicon extractions depending on the available resources. Firstly, bilingual lexicon extraction from parallel corpora (Chen, 1993). Parallel corpora is the translation of a corpus in a source language to a target language. As the translation of a corpus can only be handmade, this type of resource is quite scarce and costly to build. Secondly, it’s possible to use comparable corpora instead (Fung, 1995; Rapp, 1995): the two corpora are not any translation of one into the other, but they must belong to the same time lapse, the same domain, and sometimes the same authors. Obviously, it really costs less in comparison to parallel corpora to obtain or automatically build such a resource. That is the reason why we have decided to focus on bilingual lexicon extraction using comparable corpora, first because building a parallel corpus in an internship’s time lapse was not realistic and also because comparable corpora are more abundant, and so focusing on them more relevant and substantiated.
1.1 Goals and Objectives

The main goal of this work is to show to which extent bilingual lexicon extraction using 2 comparable corpora in the case of poor resource language pairs can be improved using a third one. Indeed, it has been proved that the use of additional languages brings more information (Dagan and Itai, 1991; Simard, 1999). We also know that the quality of the extracted bilingual lexicon strongly depends on the quality of the resources, that is to say the corpora itself and a bilingual dictionary. Of course, a third language for which we dispose of poor resources will not be helpful at all. In this study, we bet on the potential high quality of the resources of a pivot language. The idea of involving a third language is to benefit from the lexical information contained in the resources of this additional language.

First, let’s take an example, taking French as a source language and Korean as a target language. We know that, in our case, the two comparable corpora are of medium quality, and the bilingual French-Korean dictionary is quite weak, due to the nonexistence of such a dictionary on the free market. Moreover, its manual building remains costly. As a consequence We expect a bad quality extracted lexicon. Nevertheless, we are highly confident that a language in which we dispose of a lot of resources can thwart the effect of the poor original resources. So, if we take English for instance, we have good French-English or English-Korean dictionaries as well as a good English corpus since English is probably the first language in term of work and resources in Natural Language Processing.

A second hypothesis is to take an “intermediary language” as a pivot language, no matter the quality of the resources of the pivot. By “intermediary language” we mean language tiling in the sense of language continuum theory. For example, it’s a well known fact that French and Spanish speakers have a feeling that Catalan is a sort of mix between them. Thus, an idea is benefit from language similarities.

1.2 Outline of the Dissertation

First, we will describe very briefly the context in which this internship took place. Then, we will expose the works that have been done so far, and try to bridge the gap between literature and my current issue. Next, we will show up my experiments and the results achieved depending on a battery of parameters, and finally conclude on the contributions and further work.
My Master’s final internship took place in Nantes, France, and more especially in the Laboratoire d’Informatique de Nantes Atlantique (LINA, 2015), a “software sciences and technologies” laboratory with an overall workforce of 170. It covers two research themes: distributed software architectures (DSA) and decision-making aid systems (DMAS). The DSA theme addresses in particular problems relating to software engineering and data management, in mass distribution, nomadic, communicating and adaptive architectures. The DMAS theme includes research aiming at developing algorithms and high performance software tools relating to decision-making aids, language processing, knowledge management and bio-informatics.

In order to develop these major research areas, LINA relies on ten research teams listed above:

- AeLOS (Secure architectures and software)
- ASCOLA (ASpect and COmponent LAnguage)
- AtlanMod
- COD (KnOwledge and Decision)
- COMBI (Combinatorics and Bioinformatics)
- GDD (Distributed Data Management)
- GRIM (Management and summarization of multimedia data)
- OPTI (global optimisation, multi-objective optimisation)
- TALN (Natural Language Processing)
- TASC (Constraint Programming)
LINA provides calculation servers enabling each researcher to launch wide experiments. In my case, I have been given access to a server with 30 processor cores and RAM of 130GB.

The department I have been attached to for my research work was TALN, as my work belongs to the field of Natural Language Processing.

TALN team’s leader is Béatrice Daille – who is also one of the two advisors of this Master’s thesis – and the team’s research areas are close to Analysis and Discovery on the one hand, and Alignment and Comparison on the other hand. The last one includes processing of multilingualism, and is the main area of expertise to which I applied for the 6 months I completed at LINA.

TALN team also provide software. We can cite *TermSuite*\(^1\) for instance. *TermSuite* is the Open Source and UIMA-based application drawn out from the European project *TTC*: Terminology Extraction, Translation Tools and Comparable Corpora. This software, as it’s described in the European project’s name, brings the possibility of doing 2 tasks in 7 languages: extracting terminology from a given corpus and extracting bilingual lexicon extraction from two comparable corpora. The languages supported by *TermSuite* are the following: English, French, German, Spanish, Latvian, Chinese and Russian.

The tasks of terminology extraction and bilingual lexicon extraction are more accurately described in the next sections.

\(^1\)https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite (Rocheteau and Daille, 2011)
Many works have already been done in the field of bilingual lexicon extraction. Depending on the resources at our disposal, several ways of extracting bilingual terminology exist: first, using parallel corpora. The advantage of parallel corpora is that results achieved are very good, and the translations found mainly relevant. The alternative is to use comparable corpora, which advantage is to be easier and faster to build. In opposition, the intervention of human beings to translate a text from a language to another language is mandatory in the case of parallel corpora. The state of the art concerning bilingual lexicon extraction of single word terms, sometimes involving a third language as a pivot, is presented below.

3.1 Introduction

First of all, we would like to thoroughly describe deeper what comparable corpora are. Unlike parallel corpora a corpus in a source language is not the translation of the corpus in a target language. Several features are important to take into account in this definition. We can mention the time lapse when the documents constituting the corpora were written. This feature is substantiated because languages change with time: expressions and vocabulary are not static, they constantly evolve. Comparable corpora also have to be domain-specific, as lexical contexts differ from one topic to another. A definition of what a comparable corpora is has been laid by Déjean and Gaussier (2007) and says that two corpora in two languages $l_1$ and $l_2$ are called comparables if it exists a significant subpart of corpus in $l_1$’s vocabulary, respectively $l_2$, translation of which can be found in corpus in $l_1$, respectively $l_2$. This definition is very interesting in the case of bilingual alignments, because it lays on the idea of translation expectancy.

Concerning using pivot corpora, the assumption that ”Three Languages Are Better Than Two” (Simard, 1999) lays on the hypothesis that an additional language can bring more information (Dagan and Itai, 1991). This information can be helpful to improve standard method. At the beginning, this assumption was mainly designed for statistical translation and has been fully explored. But in the past few years, it
has also been examined for bilingual lexicon extraction using parallel corpora. Nevertheless, nothing has been done using comparable corpora for such a task yet.

We can also mention language tiling (Kraif, 2013). The idea is to extract information from cross-language similarity. Kraif describes language tiling as the minimal set of pairs of languages such as:

1. each language belongs to one pair at least
2. each pair has at least one language in common with the other pair

Amongst all the language tilings, the most researched and interesting is the language tiling using the strongest associated pairs that is to say the nearest languages from a genetic point of view. The idea is that the newly formed pairs can additionally lean one on top of the others to form a more solid entity. The idea of using language tiling ie. a pivot language is the first assumption that occurred to me.

3.2 Bilingual lexicon extraction using Comparable Corpora

To perform bilingual lexicon extraction using Comparable Corpora, we have to be aware of the basis concepts in this domain. Unlike parallel corpora, we cannot rely on the translation of a corpus into the other to perform alignments and then deduce the translations of a source word into target language. The idea is to lay on the first-order affinities of a word, that is to say its co-occurrences (Grefenstette, 1994a). The main existing approach based on this assumption is the standard approach presented below, with the improvements done so far.

3.2.1 Standard Method

The first and most relevant reference in the world of bilingual lexicon extraction using Comparable Corpora for single word terms is probably Fung and McKeown (1997). Their work has led to a method enabling going without parallel corpora. Many implementations have been designed in order to do so (Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2007). The method schematized on Figure 3.1 – also called Standard Method – is closely based on the notion of context vectors. A context vector is, for a given word (e.g. the head of the context vector), the representation of its co-occurrences and the number of occurrences found within a given window ($\pm k$ words), the whole thing in the offset of an entire corpus. In the Standard Method, context vectors are calculated both in source and target language corpora. Then, thanks to a bilingual dictionary, source context vectors are translated into target language. The similarity between the translated context vectors for a given source word to translate and all target context vectors lead to the creation of a list of ranked candidate translations. The rank is function of the similarity between context vectors so that the closer they are, the better the ranked translation is. All those steps are detailed hereinafter.

**Computation of context vectors** There are many ways of computing context vectors, depending on several metrics. But first, let’s focus on what a context vector is. A context vector $\vec{w}$ is the representation of the lexical context of a given word $w$. The lexical context is represented by all the co-occurrences $ct_1 \ldots ct_i$
of the given word \( w \). By co-occurrences we mean the sequence of words or terms that co-occurs more often than expected. In the case of the standard approach:

1. for each word \( w \) in the corpus, co-occurrences \( ct_1 \ldots ct_i \) are computed. Several metrics are computed according to Table 3.1

2. according to Table 3.1 (Morin and Hazem, 2014), the value \( \text{assoc}(ct) \) associated to each co-occurrence \( ct \) of the given word \( w \) of the corpus is computed. At this stage, many ways exist to compute it. The most used measures are Mutual Information (Dunning, 1993) and Log Likelihood (Fano and Hawkins, 1961) presented below. This association score is also a way of normalizing context vectors.

\[
\begin{array}{c|cc}
\text{word} & \text{co-occurring term} & \neg \text{co-occurring term} \\
\hline
a & \text{cooc}(\text{word, co-occurring term}) & b = \text{cooc}(\text{word, } \neg \text{co-occurring term}) \\
\neg \text{word} & c = \text{cooc}(\neg \text{word, co-occurring term}) & d = \text{cooc}(\neg \text{word, } \neg \text{co-occurring term}) \\
\end{array}
\]

Table 3.1: Contingency table.

\[
\text{Mutual Information (ct)} = \log\frac{a}{(a+b)(a+c)}
\]

Log Likelihood (ct) = \( a \log(a) + b \log(b) + c \log(c) + d \log(d) + (a + b + c + d) \log(a + b + c + d) - (a + b) \log(a + b) - (a + c) \log(a + c) - (b + d) \log(b + d) - (c + d) \log(c + d) \)
A schematized result of the computation of context vectors would be the following:

<table>
<thead>
<tr>
<th>$\bar{w}_1$</th>
<th>$w_1$</th>
<th>$ct_1 \mid assoc(ct_1)$</th>
<th>...</th>
<th>$ct_i \mid assoc(ct_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{w}_2$</td>
<td>$w_2$</td>
<td>$ct_1 \mid assoc(ct_1)$</td>
<td>...</td>
<td>$ct_i \mid assoc(ct_i)$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{w}_j$</td>
<td>$w_j$</td>
<td>$ct_1 \mid assoc(ct_1)$</td>
<td>...</td>
<td>$ct_i \mid assoc(ct_i)$</td>
</tr>
</tbody>
</table>

Figure 3.2: Terminology of a corpus represented by its context vectors.

**Translation of context vectors**  The second step is to translate the tail of context vectors into target language. By tail, we mean all the co-occurring terms $ct_1 ... ct_i$ of the given word $w$ and their associated value $assoc(ct) = \text{Mutual Information (ct)}$ or $assoc(ct) = \text{Log Likelihood (ct)}$, in opposition to the head of the context vector, that is the word $w$ itself. The head is not translated. To transfer context vectors into a target language, a bilingual source/target dictionary is mandatory. Nevertheless, exceptions can occur during this operation: indeed, the usual case is that for one co-occurring source term $ct_1$ to translate, it exists one translation $ct_{ktrans}$ according to the bilingual dictionary. In this case, $assoc(ct_{ktrans}) = assoc(ct_k)$

\[
\bar{w} \quad w \quad ct_1 \mid assoc(ct_1) \quad ... \quad ct_i \mid assoc(ct_i) \\
\downarrow \\
\bar{w}_{trans} \quad w \quad ct_{ktrans} \mid assoc(ct_{ktrans}) \quad ... \quad ct_{i_{trans}} \mid assoc(ct_{i_{trans}})
\]

Figure 3.3: Transferring context vectors.

In case a co-occurring source term $ct_k$, we have to translate, does not have any translation according to the bilingual dictionary, it is then, not transferred into the translated context vector $\bar{w}_{trans}$.

Finally, if many translations according to the bilingual dictionary are possible, the association score is function of the frequency of the term in the target language:

\[
assoc(ct_{ktrans}) = \frac{assoc(ct_k) \times freq(ct_{ktrans})}{\sum_j freq(ct_{jtrans})}
\]
**Similarity computation** The last step is, for one source word $w_{source}$ for which we want to find the translation in target language, to compute the similarity between its context vector $\vec{w}_{source}$ and all target context vectors $\vec{w}_{target_1}, \ldots, \vec{w}_{target_j}$. Such a similarity can be computed according to Cosine Similarity (Salton and Lesk, 1968) or Weighted Jaccard Distance (Grefenstette, 1994b).

\[
\text{Cosine} (\vec{w}_s, \vec{w}_t) = \frac{\sum_{k} \text{assoc}(w_{sk}) \times \text{assoc}(w_{tk})}{\sqrt{\sum_{k} \text{assoc}(w_{sk})^2} \sqrt{\sum_{k} \text{assoc}(w_{tk})^2}}
\]

\[
\text{Jaccard} (\vec{w}_s, \vec{w}_t) = \frac{\sum_{k} \min(\text{assoc}(w_{sk}), \text{assoc}(w_{tk}))}{\sum_{k} \max(\text{assoc}(w_{sk}), \text{assoc}(w_{tk}))}
\]

After computing the similarities, the target words are ranked by the similarity computed in function of the context vector of the source word. The best ranked target words are the words that have the highest similarity score, and thus are sorted out as possible candidate translations.

### 3.2.2 Improving the Standard Method

In order to improve the results achieved by the standard method, researchers looked into unbalanced specialized comparable corpora (Morin and Hazem, 2014), and concluded that we can involve a language with strong and numerous resources and use them all without having to balance the size of the two comparable corpora. This work confirms the validity of using an unbalanced pivot corpus in order to fully use the information it contains to deal with the lack of data of the source or the target language corpus.

Chiao et al. (2004) showed that it’s possible to reverse the computation of candidate translations in the sense that it can be helpful to rerank the candidate translations. The idea is, for a given source word to translate, we obtain many candidate translations in target language. Those candidate translations are then realigned thanks to the same process they were found with but in a reverse way. If the original source word belongs to the candidate translations of the candidate translation, then the candidate translation under test has high chances to be the right translation of the source word. Thus, it is reranked so that its coming within the first obtained candidate translations.

Finally, some works have also been done looking upstream of the standard method to see whether context vectors can be reshaped thanks to statistical models (Hazem and Morin, 2013). The idea of this kind of process is, in the case of short corpora, to strengthen the representativity of the computed context vectors by using data from an external bigger corpus.

### 3.2.3 Extensions of the Standard Method

Déjean and Gaussier (2007) showed up that bilingual lexicon extraction from comparable corpora can be improved thanks to interlanguage similarity. First, they remind us the distributional semantic in the case of comparable corpora. The main assumptions are the followings:

1. if two words have similar distributions, then they are semantically related.
2. words which distributions are the most similar to a given word are, with a high probability, semantically related to this word.
3. when normalized distributions of the words from language $l_1$ are very similar to the distribution of a given word in $l_2$, they then have a high probability of being a right translation.

Hence they use a bilingual thesaurus. Thanks to it, they bridge the gap between many languages and make a correspondence between conceptual classes of a given word. The protocol is the following:

1. getting the nearest conceptual classes
2. compute the closest path between each pair of the nearest classes

Candidate translations are thus of better quality, because the use of a bilingual thesaurus is a way to refine the protocol based on word semantic.

Nevertheless, the problem of bilingual thesauri is that they are scarce resources and they are very difficult to build.

### 3.2.4 Improving Corpora Comparability

A method to improve corpus comparability for bilingual lexicon extraction from comparable corpora has also been proposed (Li and Gaussier, 2010). To do so, they base their method on the notion of comparability and whether a corpus is poorly comparable, highly comparable, or parallel. This measure is defined as the expectation of finding for each source word its translation into the vocabulary of the target corpus. The improvement of the quality of the corpus follows two steps:

1. extraction of the highly comparable sub-part of the corpus: they have progressively extracted the sub-part with a minimum threshold of comparability from a bilingual corpus, iteratively adding new elements, if and only if the new elements are comparable enough and less comparable than the current extracted sub-part.
2. improvement of the poorly comparable sub-part of the corpus: one way is to extract the most representative words of a set of documents from the sparsely sub-part of the corpus, then translate them with a bilingual dictionary and get the documents corresponding to those keywords.

The problem of this method is that we rely on two resources: a bilingual dictionary and external corpora. Nonetheless, we will reuse the notion of comparability in some of the following sections.

### 3.3 Bilingual lexicon extraction using Pivot Language

Although we will only work with Comparable Corpora in this study, it also exists some relevant work over Parallel Corpora and statistical translation from which we drew our inspiration to develop an original method for Comparable Corpora. All the references hereinafter involve pivot languages, this is a reason they have been mentioned.

In the work done by Kwon et al. (2013), we can see that a bilingual lexicon extraction via a pivot language has been completed. Contra...
provides weighted bilingual context vectors. At this stage, we have in possession weighted bilingual context vectors from the source side, that we call \( S \). The same operation is done by the target side, and we finally obtain \( T \), context vectors from the target side. Next, the similarity between each word’s context vector in \( S \) every word’s context vector in \( T \) is computed. We finally obtain the ranked candidate translations for each word from \( S \) under test. The main advantage of doing it this way is that we do not depend on bilingual dictionaries. Nevertheless, this method is no longer applicable to comparable corpora due to the necessity of this kind of resource, and to the infeasibility of performing sub-sentential alignment over comparable corpora.

An extended method for pivot-based approach for bilingual lexicon extraction (Seo et al., 2014; Kim et al., 2015) is the next step of the work cited above. This approach not only lays on Kwon et al. (2013) but also on Déjean et al. (2002), in which the main contribution was to reduce the dependency between the initial seed dictionary using the notion of k-nearest words. The idea is the following: k-nearest words in the source corpus are computed, then translated into target language with a bilingual dictionary. At the end, a similarity score between each word in the target language is computed. In the work of Seo et al. (2014), k-nearest words are used in order to improve the quality of the candidate translations obtained. Unlike the previous work, k-nearest words are computed from the source corpus. Those k-nearest words are then used with source/pivot context vectors to build the context vectors of the k-nearest words of the source-pivot corpora. Finally, from target-pivot context vectors and context vectors of the k-nearest words of the source-pivot corpora, the similarity between each word to translate in the source corpus sharing k-nearest words in the target language is computed and the ranked candidate translations are finally obtained. The main advantage of this method is the improvement of the quality of the extracted bilingual lexicon for infrequent words, even if the improvements for frequent words are doubtful.

Although statistical translation has very little to do with bilingual lexicon extraction, the idea of using more than two languages has also been performed in this domain. In Och and Ney (2001) the idea of statistical multi-source translation is to take a text given in many source languages as input, and to translate it into a target language. The pros of this method are: first, a better disambiguation of words, because ambiguities often only happen between 2 languages; second, a better reordering of words, because the syntax between two genetically close languages has high chances to be quite similar; and finally, anaphora resolution is no longer needed, because we expect that in source languages the pronouns are already correctly translated. Thus, the procedure for one given document in N languages is, first, to align segments by size (assuming that paragraphs are alignable if they are pretty much of the same length (Gale and Church, 1993)). Secondly, with the same assumption, a phrasal alignment is performed. Then, all sentences with a doubtful alignment are filtered out. To finish, a statistical algorithm is performed. In this algorithm, given a target word, all source words are considered as statistically independent. All target words are taken into account to perform maximization. So, it’s hypothesized that there are as many target sentences as source languages. The best translation is the most probable one. Up to 6 source languages, this method will perform better.
APPRAOCHES INVOLVING PIVOT LANGUAGE

As substantiated in the paragraphs hereinbefore, the fact of searching for more information through a pivot language is probably the best way in the case of poor source and target languages. Thus, we are going to present several original ideas of implementing such method. The first methods are based on using a pivot dictionary, and the last ones are based on using pivot corpora. But first, we have decided to present several manners of choosing the best pivot language according to the languages we have to deal with: genealogically close languages or English by default.

4.1 Choosing the best pivot language

Before processing algorithms, one question arises: what is the best pivot language we might use for bilingual lexicon extraction? At this stage of linguistic history, several answers are possible. The first one is that English, as the third most spoken language in the world and the first one in business, is the language that has numerous resources. The second one is that it may exist an intermediate language between the source and target languages that could act as an interface between them. The pros and cons of the two theories are presented below.
SECTION 4. APPROACHES INVOLVING PIVOT LANGUAGE

4.1.1 Theory of linguistic continuum

Many languages are considered as very close genealogically speaking. That is particularly the case when two different native speakers try to communicate using their own languages. This phenomenon is called intercomprehension or mutual intelligibility. It often occurs when the languages are genetically related, and when they are likely to be similar in grammar, vocabulary, pronunciation, or other features. As intercomprehension across the borders, we can quote the following examples:

- Danish, Norwegian and Swedish
- Dutch with: Afrikaans, West Frisian, German (to a limited degree)
- Galician with Spanish and Portuguese

In the examples above, we can see that many languages (more than two) can be involved in the process of mutual intelligibility. In such cases, one language can play the role of umbrella language (Kloss, 1967). Let us remember the concept of “Dachsprache” or dialect continuum, when two or more different languages or dialects merge one into the other(s) without definables boundaries. This concept is valid in the case of:

- Arabic (as it exists different dialects from Moroccan Arabic to Gulf Arabic)
- Swedish-Gutnish-Elfdalian-Scanian-Danish-Norwegian-Faroe-ICelandi
- Portuguese-Galician-Asturian-Castilian-Aragonese-Catalan-Occitan-French-Italian

What we assume is that this concept can be extended to other languages that are not considered close enough to fall within the scope of direct language continuum. Of course, obtaining a sufficient amount of data for dialects is an utopia. It is also the case of translations between two languages belonging to the same linguistic continuum. For instance, we assume that it exists paths like Spanish-French-English or English-Dutch-German or French-English-German. That’s on this assumption we lay on to develop a pivot method based on language genealogy.

4.1.2 English, by default

In the case that source and target languages are geographically or genealogically distant, we assume that English is the default language to use as pivot. Moreover, the scarcity of the resources such as parallel corpora has already been proved (Martin et al., 2005). Let’s take French to Korean translation: although comparable corpora can be built between those two examples, it is a well known fact that a bilingual dictionary is lacking. The more obvious reason to choose English to bridge the gap between French and Korean – or any other exotic pair – is the quasi infinite amount of resources we can fetch from the web or the marketplace. Indeed, bilingual lexicon extraction from comparable corpora has been initially designed to work with resources found on the net, in particular the corpus itself. Or, in the case we have to deal with languages with scarce resources, bringing back the whole to English in order to counterbalance the lack of data appear as the ideal solution. As mentioned in the state-of-the-art section, getting limited by balanced corpora is completely unfounded (Morin and Hazem, 2014). Hence, the idea of using a “strong” language appears reinforced.
4.2 Pivot dictionary based approaches

The two methods below are above all derivations of the standard approach (Fung and McKeown, 1997), but are original in the sense that, unlike the standard approach, they involve another resource: a pivot dictionary. As a consequence, the initial source/target dictionary is no longer needed, whereas source/pivot and pivot/target dictionaries become compulsory instead. In this particular case, we assume that a source/target dictionary is unavailable, but that source/pivot and pivot/target dictionaries exist and are of good quality.

4.2.1 Translating context vector successively

The first method, and the most naive is to translate context vectors successively, to start with from source to pivot language, and to follow from pivot to target language. As shown in Appendix A, Figure A.1, context vectors in a source language are computed as it is usually done in the standard method. Then, the second step is to translate context vectors’ tail. Unlike the standard method in which context vectors’ tails are translated into a target language, we translate the tails into pivot language thanks to a source/pivot dictionary. The method is the same, the language is different. Next, this operation is done a second time from pivot to target with a pivot/target dictionary. So, we obtain source context vectors translated into a target language. We can say that we transferred the context vectors via a pivot language. The final step of similarity computation stays unchanged: for one source word $w_{source}$ for which we want to find the translation into a target language, we compute the similarity between its context vector transferred into pivot language $\overrightarrow{w_{source \_pivot}}$ and all target context vectors $\overrightarrow{w_{target_1}} \ldots \overrightarrow{w_{target_j}}$.

\[
\overrightarrow{w} \quad \begin{array}{c} w \mid ct_1 \mid assoc(\text{ct}_1) \mid \ldots \mid ct_i \mid assoc(\text{ct}_i) \end{array} \\
\downarrow \\
\overrightarrow{w_{pivot}} \quad \begin{array}{c} w \mid ct_{pivot} \mid assoc(\text{ct}_{pivot}) \mid \ldots \mid ct_{pivot} \mid assoc(\text{ct}_{pivot}) \end{array} \\
\downarrow \\
\overrightarrow{w_{target}} \quad \begin{array}{c} w \mid ct_{target} \mid assoc(\text{ct}_{target}) \mid \ldots \mid ct_{target} \mid assoc(\text{ct}_{target}) \end{array}
\]

Figure 4.1: Transferring context vectors successively.

The results achieved are presented in Tables 5.5 and 5.6 column $P_1$, as $P_1$ refers to the pivot dictionary based approach done by translating context vectors successively via a pivot language.
4.2.2 Transposing both source and target vectors to pivot language

The second method based on pivot dictionaries consists in translating both source and target context vectors to pivot language. Thus, the operation of computing similarity occurs in the vectorial space of the pivot language. As shown in Appendix A and more accurately on Figure A.2, context vectors in source language are computed as it is usually done in the standard method. Then, the second step is to translate context vectors’ tail. Unlike the standard method – but similarly to the first pivot method presented above – we translate the tails into a pivot language thanks to a source/pivot dictionary. This operation is also done from the target side. All context vectors of the target terminology – ie. for each word in the target language – are translated into pivot language thanks to a target/pivot dictionary. At this stage, we have processed the source context vector of the word to translate, we have transferred it into pivot language, as well as all target context vectors. The next and last operation is to compute the similarity between the source context vector transferred into a pivot language \( \overrightarrow{w}_{source,pivot} \) and all target context vectors transferred into a pivot language \( \overrightarrow{w}_{target,pivot 1}, \ldots, \overrightarrow{w}_{target,pivot j} \).

\[
\overrightarrow{w}_s \quad \overrightarrow{w}_s \quad ct_1 \mid \text{assoc}(ct_1) \quad \ldots \quad ct_i \mid \text{assoc}(ct_i) \\
\downarrow \\
\overrightarrow{w}_{spivot} \quad \overrightarrow{w}_s \quad ct_{1,pivot} \mid \text{assoc}(ct_{1,pivot}) \quad \ldots \quad ct_{i,pivot} \mid \text{assoc}(ct_{i,pivot})
\]

**Similarity Computation**

\[
\overrightarrow{w}_{t1,pivot} \quad \overrightarrow{w}_{t1} \quad ct_{1,pivot} \mid \text{assoc}(ct_{1,pivot}) \quad \ldots \quad ct_{i,pivot} \mid \text{assoc}(ct_{i,pivot}) \\
\overrightarrow{w}_{t2,pivot} \quad \overrightarrow{w}_{t2} \quad ct_{2,pivot} \mid \text{assoc}(ct_{2,pivot}) \quad \ldots \quad ct_{i,pivot} \mid \text{assoc}(ct_{i,pivot}) \\
\ldots \\
\overrightarrow{w}_{tj,pivot} \quad \overrightarrow{w}_{tj} \quad ct_{j,pivot} \mid \text{assoc}(ct_{j,pivot}) \quad \ldots \quad ct_{j,pivot} \mid \text{assoc}(ct_{j,pivot})
\]

\[
\overrightarrow{w}_{t1} \quad \overrightarrow{w}_{t1} \quad ct_1 \mid \text{assoc}(ct_1) \quad \ldots \quad ct_i \mid \text{assoc}(ct_i) \\
\overrightarrow{w}_{t2} \quad \overrightarrow{w}_{t2} \quad ct_1 \mid \text{assoc}(ct_1) \quad \ldots \quad ct_i \mid \text{assoc}(ct_i) \\
\ldots \\
\overrightarrow{w}_{tj} \quad \overrightarrow{w}_{tj} \quad ct_1 \mid \text{assoc}(ct_1) \quad \ldots \quad ct_i \mid \text{assoc}(ct_i)
\]

Figure 4.2: Transposing source and target terminologies to pivot language.
A graphical representation of the context vectors mentioned above is given on Figure 4.3. The results we obtained are presented in Tables 5.5 and 5.6 column $P_2$, as $P_2$ refers to the pivot dictionary based approach and done by transposing both source and target context vectors into a pivot language.

### 4.3 Pivot corpora based approaches

Unlike the two previous methods laying on pivot dictionaries, the use of a pivot corpus and its terminology consists in betting on the high quality of this pivot corpus. As a consequence, we use only English and French to prove this assumption, as we know that English resources are used by default, and French ones are rich enough for our experiments as in our case they are of high quality. Nevertheless, source/pivot and pivot/target dictionaries are still required to perform those techniques.

#### 4.3.1 Successive translation

One naive idea (see Appendix A, Figure A.3) is to reuse the standard method and to apply it first, performing source/pivot translation and then pivot/target translation. The main advantage of this method is to use the pivot language terminology in order to perform the pivot/target standard method. For each translation into a pivot language that we have obtained from the source/pivot standard approach, its context vector is retrieved from the pivot terminology. Finally, with this information from the pivot corpus and a pivot/target dictionary, the standard approach is applied and candidate translations into a target language are obtained.

![Figure 4.3: Successive translation.](image-url)
As the results achieved for this method depend on the results performed by the source/pivot side, this method is capable of conveying a lot of noise, especially as the bad results achieved from source/pivot are transferred to the pivot/target side. We, indeed, believe that finding translations via the translations performed by the standard method through the pivot language, hoping to increase the results, is not realistic.

Even so, this method has been implemented. The results achieved are presented in Tables 5.8 and 5.9 as $P_3$ refers to the pivot corpus based approach and done by successive translation. We also present the results obtained in the case of using balanced and unbalanced pivot corpora (Table B.2) and reshaping source and target context vectors thanks to linear regression function described in the next section (Table B.3). Statistics about balanced or unbalanced corpus are in Section 5.1.1.

### 4.3.2 Using linear regression to reshape context vectors

In the case of bilingual lexicon extraction from specialized comparable corpora, and in our study in particular, we expect source and target corpora to be undersized in comparison with the pivot corpus. But, we assume that general resources are available. As a result, we used a strategy which consists in making co-occurrence counts more discriminant by linear regression (Hazem and Morin, 2013). Linear regression models the relationship between the co-occurrence distribution of words in a small corpus, and the co-occurrence distribution of words in a large corpus. The linear regression function to apply to the number of co-occurrences of the small corpus is presented as follows:

$$Y = \beta_1 X + \beta_0$$

where $\beta_1$ and $\beta_0$ are the parameters to estimate

In other words, for each co-occurrence count in the small corpus, its new value is:

$$f(x) = \beta_1 x + \beta_0$$

where $x$ is the value in the small corpus

Hereinafter, we see that $Y_{language}$ corresponds to the linear regression function learned from the bigger corpus. Data about this corpus are presented in Section 5.1.1. $MT$ stands for the Mobile Technology corpus and $WE$ for the Wind Energy one.

$$MT = \begin{cases} Y_{DE} = 0.884X + 1.069 \\ Y_{EN} = 3.563X + 1.599 \\ Y_{ES} = 1.694X + 1.470 \\ Y_{FR} = 2.703X + 1.615 \end{cases} \quad WE = \begin{cases} Y_{DE} = 0.921X + 1.056 \\ Y_{EN} = 3.845X + 1.753 \\ Y_{ES} = 1.760X + 1.517 \\ Y_{FR} = 3.412X + 1.695 \end{cases}$$

17
4.3.3 Computing similarity over pivot corpus

This proposal of method is just a sort of mix between Pivot 2 (presented in Section 4.2.2) and Pivot 3 (presented in Section 4.3.1). Indeed, we suggest a strategy where we could at the same time transpose the target terminology to the pivot language, and use not only source/pivot/target language dictionaries but also benefit from the information brought by a pivot corpus. As a consequence, and as shown in Appendix A, Figure A.4, the standard method is applied from source to pivot language. At this stage, candidate translations in a pivot language are retrieved. Thus, context vectors of pivot translations are taken from the pivot corpus. Then, the similarity between pivot translations and all target context vectors translated into a pivot language are computed as presented in Figure 4.4.

\[ w_s \]

\[ \begin{array}{cccc}
   w_{p1} & w_{p2} & \cdots & w_{pi} \\
   \bar{w}_{p1} & \bar{w}_{p2} & \cdots & \bar{w}_{pi} \\
\end{array} \]

**Retrieving context vectors from pivot corpus**

**SIMILARITY COMPUTATION**

\[ \begin{array}{cccc}
   \bar{w}_{t1_{pivot}} & \bar{w}_{t2_{pivot}} & \cdots \\
   \cdots & \cdots & \cdots \\
   \bar{w}_{t1_{pivot}} & \bar{w}_{t2_{pivot}} & \cdots \\
\end{array} \]

\( \text{target context vectors transferred into pivot language} \)

Figure 4.4: Similarity over pivot corpus.

The results achieved are presented in Tables 5.8 and 5.8 column \( P_4 \), as \( P_4 \) refers to the so-called approach. We also present the results obtained in the case of using balanced and unbalanced pivot corpora (Table B.2) and reshaping source and target context vectors thanks to linear regression function described in the next section (Table B.3).
4.3.4 Weighting candidate translations according to the pivot ones

In the description of the two pivot corpora based approaches described above (Successive Translation and Computing Similarity over Pivot Corpus), we did not mention how given candidate translations in a target language were related to their corresponding pivot word. In our first experiments, we ended up with translations associated with a given similarity score in a target language. However, we did not make links with the pivot translation. Here is a short example to illustrate the problem:

Given $w_s$ the source word to translate, we get 3 pivot translations $w_{p_1}$, $w_{p_2}$ and $w_{p_3}$ with similarity scores of respectively 0.8, 0.15 and 0.01. For each pivot word, we achieve the following target translations and similarity scores: $w_{t_{11}}(0.2)$, $w_{t_{12}}(0.1)$ and $w_{t_{13}}(0.05)$ for $w_{p_1}$; $w_{t_{21}}(0.7)$, $w_{t_{22}}(0.4)$ and $w_{t_{23}}(0.003)$ for $w_{p_2}$; $w_{t_{31}}(0.15)$, $w_{t_{32}}(0.002)$ and $w_{t_{33}}(0.001)$ for $w_{p_3}$. Finally, the ranked translations are: $T = \{w_{t_{21}}, w_{t_{22}}, w_{t_{11}}, w_{t_{12}}, w_{t_{22}} \ldots\}$. The problem is that the weight of the pivot candidates is not taken into account. That it to say, even if a candidate translation in pivot is the good one and has been sorted out with a strong expectancy, it is a real issue because it is not highlighted.

In order to avoid this phenomenon, we decided to weight the similarity scores of the target translations according to their associated pivot word, thanks to the following formulas:

$$\text{PROD} = \text{sim}(p) \times \text{sim}(t)$$
$$\text{MEAN} = \frac{\text{sim}(p) + \text{sim}(t)}{2}$$

where $\text{sim}(p)$ stands for the pivot word associated similarity score, $\text{sim}(t)$ for the similarity score of the target translation before the operation, $\text{PROD}$ or $\text{MEAN}$ for the final score fixed for the translation whether using the product or the mean of the similarity scores.

Experiments are shown in Section 5.4.3 to show whether the sum or the product of similarities gets better.

4.3.5 Pivot candidates threshold

Using pivot translations in the case of pivot corpora based approaches meant asking the question of the number of candidate translations to take into account. Depending on the similarity computation, the number of candidate translations is important due to the noise brought by the bad ones. In the case of not weighting the candidates according to the pivot ones; the more numerous the pivot translations are, the more we expect the results to decrease. Otherwise, the results may improve the more pivot translations we take. Although bad pivot translations exist, the weight of their corresponding target translations is not so important, thus the noise brought will have less side-effects.

Experiments are shown in Section 5.4.1 to illustrate to which extent the number of pivot translations can affect the results.
In this section, we will principally outline the resources we have used for the experiments. In addition, we will give details on the evaluation of our methods, and we will present and give an interpretation to the results achieved so far. All the experiments took place in the context of the European Project TTC TermSuite (Rocheteau and Daille, 2011). We can also mention that a component of TermSuite already implemented the standard method, as a consequence, we just had to overwrite it – without modifying the original parameters – to produce our own algorithms. All the results shown have been computed thanks to the Curiosiphi calculation server at LINA.

5.1 Linguistic Resources

As shown in the previous sections, the standard method involve comparable corpora and also a bilingual dictionary. During my experiments, we have used 2 different comparable corpora (the first one about Wind Energy, and the second one about Mobile Technologies), on which different terminology reference lists are based. Those terminology reference lists consist in a way of testing the developed methods. Each terminology reference list contains a word in a source language – in principle belonging to the source language corpus terminology – and its translation in a target language. Each terminology reference list is, as a consequence, specific to a corpus and a pair of source-target languages.
5.1.1 Specialized Comparable Corpora

**Wind Energy corpus**  The first comparable corpus we used during our experiments is the *Wind Energy corpus*. This corpus is part of the deliverable D-2.5 from the European project TTC TermSuite\(^1\). It has been built from a crawl of webpages using many keywords related to the wind energy field. The comparable corpus is composed of documents in 7 languages: German, English, Spanish, French, Latvian, Russian and Chinese. Among them, we particularly looked into German, English, Spanish and French for my experiments.

**Mobile Technologies corpus**  The second comparable corpus we used during our experiments is the *Mobile Technologies corpus*. It as also been built by crawling the web, and is also part of the deliverable D-2.5 from the European project TTC TermSuite.

The number of documents of each corpus in each languages is presented in Table 5.1 hereinafter.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th># documents</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind Energy</td>
<td>EN</td>
<td>38</td>
<td>313.954</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>12</td>
<td>314.551</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>46</td>
<td>453.958</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>33</td>
<td>358.602</td>
</tr>
<tr>
<td>Mobile Technologies</td>
<td>EN</td>
<td>34</td>
<td>303.972</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>20</td>
<td>437.505</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>31</td>
<td>474.534</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>61</td>
<td>474.316</td>
</tr>
</tbody>
</table>

Table 5.1: Corpora composition.

We also had at our disposal an unbalanced version of the *Wind Energy corpus* for English and French languages. This stronger version, in term of resources, is very helpful in the case of the pivot corpus, because we want to demonstrate that a strongly supported pivot language can improve the extracted lexicon. The size of the unbalanced *Wind Energy corpus* is of about 700k and 800k words for French and English respectively.

**Newspaper corpus**  In order to reshape context vectors (procedure of Section 4.3.2), electronic publications – composed of a set of documents for a total amount of 10 million words for each language – of the following newspapers have been used: *Le Monde* (1994) for French, *Los Angeles Times* (1994) for English, *EFE* (1994) for Spanish and *Der Spiegel* (1994) for German.

French, English, Spanish and German documents have been pre-processed using TTC TermSuite. The operations done during pre-processing are the following: tokenization, part-of-speech tagging and lemmatization. Moreover, function words such as articles, pronouns, adpositions, conjunctions or auxiliary verbs for instance, and also hapaxes have been removed.

\(^1\)https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite
5.1.2 Terminology Reference Lists

In order to evaluate the output of the different approaches, terminology reference lists have been built from each corpus in each language. The reference term lists compilation task were part of the deliverable D-3.2 from the European project TTC TermSuite (Loginova et al., 2012). As the terminology reference lists are monolingual, we had to manually compute the bilingual terminology reference list. The number of Single Word Terms belonging to reference term lists depending on the language and corpus is presented in Table 5.2.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>EN</th>
<th>FR</th>
<th>ES</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WE</td>
<td>48</td>
<td>58</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>MT</td>
<td>52</td>
<td>58</td>
<td>60</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 5.2: Number of references evaluated.

From the references lists described above, we computed the Maximal Reachable Recall ($R_{MAX}$) that reflects the best result we can achieve for a given pair of source-target language for a given corpus. Indeed, Terminology Reference Lists have been built over bigger corpora than the ones directly used for the experiments. $R_{MAX}$ is the best score we can achieve given those two information, and knowing the number of source-target words we are able to find or not in the terminologies.

What we can infer from the different values of $R_{MAX}$, depending on the pairs of languages, is that many words belonging to the reference list are probably hapaxes or infrequent words. Hence, they have been filtered out during pre-treatments. Nevertheless, $R_{MAX}$ is a way of taking this effect into account, and to know that we cannot expect a 100% accuracy for this task. The values of $R_{MAX}$ are presented with the results achieved and for the sake of comparison in Tables 5.5 to 5.9 and B.1 to B.3.

5.1.3 Bilingual Dictionaries

In order to perform bilingual lexicon extraction from comparable corpora, a bilingual dictionary is mandatory. Nevertheless, we only disposed of French/English, French/Spanish and French/German dictionaries from the ELRA catalogue\(^2\). The dictionaries are generalist, and contains few terms related to the Wind Energy and Mobile Technologies domains.

To obtain the missing dictionaries, the French/English, French/Spanish and French/German have been reversed to obtain English/French, Spanish/French and German/French dictionaries. For the others, they have been built by triangularization from the ones above. The number of entries of each dictionary is presented in Table 5.3.

\(^2\)http://catalog.elra.info/
### 5.2 Evaluation of the approaches

#### 5.2.1 Experimental parameters

##### 5.2.1.1 Comparability measure between two context vectors

**Cosine similarity**  
The first comparability measure between two context vectors is obviously Cosine similarity (Salton and Lesk, 1968).

\[
\text{Cosine } (\overrightarrow{w_s}, \overrightarrow{w_t}) = \frac{\sum_k \text{assoc}(w_{sk}) \cdot \text{assoc}(w_{tk})}{\sqrt{\sum_k \text{assoc}(w_{sk})^2 \cdot \sum_k \text{assoc}(w_{tk})^2}}
\]

**Jaccard distance**  
The other comparability measure between two context vectors is the Jaccard distance (Grefenstette, 1994b).

\[
\text{Jaccard } (\overrightarrow{w_s}, \overrightarrow{w_t}) = \frac{\sum_k \min(\text{assoc}(w_{sk}), \text{assoc}(w_{tk}))}{\sum_k \max(\text{assoc}(w_{sk}), \text{assoc}(w_{tk}))}
\]

##### 5.2.1.2 Filtering hapaxes or not

While pre-processing a corpus, a very important question is to decide whether to filter hapaxes or not, hapaxes being words that only occur once in the whole corpus. The big disadvantage of keeping hapaxes in the corpus (or words only occurring under a certain threshold) is that they consist in noisy data in the sense that they are barely representative. Furthermore, their context vectors are not a proof of representativity, as the more a word occurs, the more the context vector is rich in term of data. Our commitment is to assume that filtering out such words – despite the risk of seeing $R_{MAX}$ and the results collapse – will permit to remain more accurate.

#### 5.2.2 Mean Average Precision

Let’s recall what precision is.

\[
P = \frac{\text{number of relevant documents retrieved}}{\text{number of retrieved documents}}
\]
Mean Average Precision (abbreviated MAP) derives from Average Precision (AP). AP measures the quality of the system at all recall levels by averaging the precision for a single query:

\[ \text{AP} = \frac{1}{\text{RDN}} \times \sum_{k=1}^{\text{RDN}} \left( \text{Precision at rank of } k^{th} \text{ relevant document} \right) \]

where RDN is the number of relevant documents in the collection.

Mean Average Precision (MAP) is the mean of Average Precision over all queries. Most frequently, arithmetic mean is used over the query set.

In our case, as we consider a unique translation as valid, and as we only take a look at the first 20 given candidate translations, we can simplify the MAP formula. We can see that it is equal to Mean Reciprocal Rank (Voorhees, 1999) in this particular case.

\[ \text{MAP} = \text{MRR} = \frac{1}{\# \text{terms to evaluate}} \times \sum_{k=1}^{\# \text{terms to evaluate}} \left( \frac{1}{\text{rank of correct translation}} \right) \]

### 5.3 Results Achieved for Pivot Dictionary Based Approaches

The MAP achieved for both Pivot Dictionary Based Approaches is shown in Tables 5.5 and 5.6 for Wind Energy and Mobile Technologies corpora respectively. We present, for the sake of comparison, the results achieved by the standard method (Std.), method transferring context vectors successively (P1) and the method transposing context vectors to pivot language (P2). We also give additional information, such as the best result achievable according to the reference lists and the words belonging to the filtered corpus (RMAX) and corpora comparability C (Li and Gaussier, 2010).

The parameters for the results shown are:

1. Log Likelihood for computation and normalization of context vectors.
2. Cosine Similarity for similarity computation between context vectors.

The comparison between results achieved using Cosine Similarity or Weighted Jaccard Distance is shown in Appendix B, Table B.1.

According to the results, we can see that there is a strong correlation between the improvements achieved by pivot based approaches and corpus comparability. We only improve the quality of the extracted bilingual lexicon in the case of poorly comparable corpora, respectively ≤ 65.76% and ≤ 66.52% for Wind Energy and Mobile Technologies corpora. For instance, we increase the MAP from 0.268 to 0.390 and 0.374 in the case of translation from English to Spanish for the Wind Energy corpus, and from 0.126 to 0.355 and 0.347 for German to Spanish via French for the Mobile Technologies corpus.
### Table 5.5: Results for Pivot Dictionary Based Approaches for Wind Energy Corpus.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Pivot</th>
<th>Std.</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$R_{\text{MAX}}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.268</td>
<td>0.390</td>
<td>0.374</td>
<td>0.646</td>
<td>65.76%</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.119</td>
<td>0.232</td>
<td>0.233</td>
<td>0.491</td>
<td></td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.158</td>
<td>0.125</td>
<td>0.215</td>
<td>0.458</td>
<td>66.21%</td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.200</td>
<td></td>
</tr>
<tr>
<td>FR-DE</td>
<td>EN</td>
<td>0.056</td>
<td>0.118</td>
<td>0.132</td>
<td>0.418</td>
<td>77.63%</td>
</tr>
<tr>
<td>DE-FR</td>
<td>EN</td>
<td>0.038</td>
<td>0.028</td>
<td>0.028</td>
<td>0.151</td>
<td></td>
</tr>
<tr>
<td>FR-ES</td>
<td>EN</td>
<td>0.366</td>
<td>0.150</td>
<td>0.176</td>
<td>0.528</td>
<td>82.36%</td>
</tr>
<tr>
<td>ES-FR</td>
<td>EN</td>
<td>0.210</td>
<td>0.103</td>
<td>0.117</td>
<td>0.357</td>
<td></td>
</tr>
</tbody>
</table>

For fairly comparable corpora ($\leq 68% \leq C \leq 80\%)$, results stay unchanged in comparison with the standard approach. Finally, for highly comparable corpora ($C > 80\%$) the quality of the extracted lexicon via a pivot language gets worse.

### Table 5.6: Results for Pivot Dictionary Based Approaches for Mobile Technologies Corpus.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Pivot</th>
<th>Std.</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$R_{\text{MAX}}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.000</td>
<td>0.041</td>
<td>0.097</td>
<td>0.273</td>
<td>44.24%</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.000</td>
<td>0.045</td>
<td>0.027</td>
<td>0.273</td>
<td></td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.000</td>
<td>0.018</td>
<td>0.018</td>
<td>0.218</td>
<td></td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.000</td>
<td>0.018</td>
<td>0.018</td>
<td>0.218</td>
<td></td>
</tr>
</tbody>
</table>

The interpretation we suggest is the following: given two corpora, $S$ in source language, $T$ in target and a bilingual dictionary source/target $\mathcal{D}$, the comparability is function of $S$, $T$, $\mathcal{D}_{S/T}$. Hence, a low comparability measure can be due to a poor expectation of finding the translation in target language for each source word in the corpus because the two corpora are not lexically close enough, or because the dictionary is weak. We assume this second option, and this is how we substantiate the pivot dictionary.
based approaches. Thus, the use of source/pivot $D_{S/P}$ and pivot/target $D_{P/T}$ dictionary can artificially improve the comparability and enhance the extracted lexicon.

Of course, we do not pretend that our methods can make enhancements with an initially very highly comparable corpora since the use of pivot dictionaries will introduce more noise than it will bring additional information.

5.4 Results Achieved for Pivot Corpora Based Approaches

5.4.1 Varying the pivot candidates threshold

In order to get an idea of the right number of pivot candidates to accept, we computed the mean rank of the right translation obtained by the standard method in the case of any source language to French or English. In order to get this metric, we made the mass of the overall similarity scores vary from 10% to 90%.

According to graphs in Figure 5.1, we can deduce that the mean rank of the good translation in pivot language is from 1 to 2. Nevertheless, we can take into account more than 2 candidate translations, because more information can be brought. In this case, we do not expect to significantly improve the results.

5.4.2 Expectation of the quality of the pivot candidates

The MAP of the pivot candidates achieved is a good indicator on how we can pretend to improve the original source/target alignment. Indeed, if the pivot translations obtained are not the good ones for the corresponding source word, we have poor chances to get the right final target translation.

![Image of Figure 5.1](image.png)

Figure 5.1: Mean rank of good translation in function of the mass of the overall similarity scores.
According to Table 5.7, we can see that we are highly confident in improving the results of the standard method in the cases of English to any target language via French and French to any target language via English. On the other hand, we are fairly confident for Spanish to any target language via French. Finally, we do not expect to improve the results at all in the other cases (German to target language via French or English and Spanish to target language via English).

### 5.4.3 Results

The MAP achieved for both Pivot Dictionary Based Approaches is shown in Tables 5.8 and 5.9 for Wind Energy and Mobile Technologies corpora respectively. We present, for the sake of comparison, the results achieved by the standard method (Std.), method transferring context vectors successively ($P_3$) and the method transposing context vectors to pivot language ($P_4$). As in the previous section, we also give additional information, such as the best result achievable according to the reference lists and the words belonging to the filtered corpus ($R_{MAX}$) and corpora comparability $C$.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Pivot</th>
<th>Std.</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$R_{MAX}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.268</td>
<td><strong>0.339</strong></td>
<td><strong>0.340</strong></td>
<td>0.646</td>
<td>65.76%</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.119</td>
<td><strong>0.212</strong></td>
<td><strong>0.212</strong></td>
<td>0.491</td>
<td>66.21%</td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.158</td>
<td>0.126</td>
<td>0.127</td>
<td><strong>0.458</strong></td>
<td></td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.200</td>
<td></td>
</tr>
<tr>
<td>FR-DE</td>
<td>EN</td>
<td>0.056</td>
<td><strong>0.147</strong></td>
<td><strong>0.147</strong></td>
<td>0.418</td>
<td>77.63%</td>
</tr>
<tr>
<td>DE-FR</td>
<td>EN</td>
<td>0.038</td>
<td>0.019</td>
<td>0.019</td>
<td><strong>0.151</strong></td>
<td></td>
</tr>
<tr>
<td>FR-ES</td>
<td>EN</td>
<td>0.366</td>
<td>0.172</td>
<td>0.173</td>
<td>0.528</td>
<td>82.36%</td>
</tr>
<tr>
<td>ES-FR</td>
<td>EN</td>
<td>0.210</td>
<td>0.088</td>
<td>0.082</td>
<td><strong>0.357</strong></td>
<td></td>
</tr>
<tr>
<td>ES-DE</td>
<td>FR</td>
<td>0.000</td>
<td><strong>0.018</strong></td>
<td><strong>0.020</strong></td>
<td>0.273</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>0.000</td>
<td><strong>0.032</strong></td>
<td><strong>0.032</strong></td>
<td></td>
<td>44.24%</td>
</tr>
<tr>
<td>DE-ES</td>
<td>FR</td>
<td>0.001</td>
<td><strong>0.018</strong></td>
<td><strong>0.018</strong></td>
<td>0.218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>0.001</td>
<td><strong>0.018</strong></td>
<td><strong>0.018</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.8: Results for Pivot Corpora Based Approaches for Wind Energy Corpus.

The parameters for the results shown are:

1. Log Likelihood for computation and normalization of context vectors.
2. Cosine similarity for similarity computation between context vectors.
3. The use of balanced corpora.
4. No context vectors reshaping with linear regression.
5. 5 candidate translations used in pivot language.
6. Weighting the candidate translations according to the *PROD* measure (product of the similarity scores of the pivot and target translations).

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Pivot</th>
<th>Std.</th>
<th>(P_3)</th>
<th>(P_4)</th>
<th>(R_{MAX})</th>
<th>(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.445</td>
<td>0.528</td>
<td>0.530</td>
<td>0.882</td>
<td>66.52%</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.193</td>
<td>0.239</td>
<td>0.239</td>
<td>0.533</td>
<td>68.95%</td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.622</td>
<td>0.210</td>
<td>0.191</td>
<td>0.896</td>
<td>68.95%</td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.074</td>
<td>0.046</td>
<td>0.045</td>
<td>0.455</td>
<td>80.06%</td>
</tr>
<tr>
<td>FR-DE</td>
<td>EN</td>
<td>0.053</td>
<td>0.088</td>
<td>0.088</td>
<td>0.597</td>
<td>82.02%</td>
</tr>
<tr>
<td>DE-FR</td>
<td>EN</td>
<td>0.034</td>
<td>0.034</td>
<td>0.034</td>
<td>0.432</td>
<td>82.02%</td>
</tr>
<tr>
<td>FR-ES</td>
<td>EN</td>
<td>0.514</td>
<td>0.319</td>
<td>0.322</td>
<td>0.807</td>
<td>82.02%</td>
</tr>
<tr>
<td>ES-FR</td>
<td>EN</td>
<td>0.238</td>
<td>0.166</td>
<td>0.165</td>
<td>0.552</td>
<td>82.02%</td>
</tr>
<tr>
<td>ES-DE</td>
<td>FR</td>
<td>0.001</td>
<td>0.050</td>
<td>0.050</td>
<td>0.500</td>
<td>44.02%</td>
</tr>
<tr>
<td>EN</td>
<td>0.001</td>
<td>0.042</td>
<td>0.041</td>
<td>0.500</td>
<td>44.02%</td>
<td></td>
</tr>
<tr>
<td>DE-ES</td>
<td>FR</td>
<td>0.126</td>
<td>0.376</td>
<td>0.376</td>
<td>0.585</td>
<td>44.02%</td>
</tr>
<tr>
<td>EN</td>
<td>0.126</td>
<td>0.180</td>
<td>0.180</td>
<td>0.585</td>
<td>44.02%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.9: Results for Pivot Corpora Based Approaches for Mobile Technologies Corpus.

The comparison between the achieved results, using different parameters (balanced/unbalanced pivot corpus, reshaping or not context vectors) is shown in Appendix B, Tables B.2 and B.3. The influence of the number of pivot candidates and the way of taking them into account is shown in Figures B.1 to B.6.

According to the results presented and Tables 5.8, 5.9, B.2, B.3 and Figures B.1 to B.6, we can see that the use of a pivot corpus as intermediary language can be very helpful. First, the way of taking into account the weight of pivot candidates has important repercussions on the final results. By computing the mean or the product of the similarity scores of pivot and corresponding target translations, we significantly succeed in putting aside the effect of the noise brought by the incorrect pivot candidates. We have shown that the number of pivot candidates is important: in the case of weighting the target candidate translations by the corresponding pivot candidates, we highlighted that a good threshold should be of 5 pivot candidates. Moreover, we can also see that, as expected in Section 5.4.2, we particularly succeeded in improving the results in the case of English to Spanish (and vice versa) and French to German. Finally, the use of an unbalanced pivot corpus has been substantiated especially in the case of English-Spanish and Spanish-English, where we significantly improved the quality of the extracted bilingual lexicon. However, reshaping source and target context vectors involving linear regression when dealing with an unbalanced pivot corpus did not made improvements: thus, we conclude that trying to increase the representativity of source and target context vectors to project them onto a stronger pivot corpus is not helpful at all.
We can also comment that other experiments have been performed – but not presented in this report – in the way of using a pivot corpus:

1. Computing the mean context vector of all the context vectors of the candidate translations obtained in pivot language. Then, the similarity computation is done between this mean context vector and all context vectors in target language.

2. Applying a reverse computation of the translations obtained in pivot language: if the source word if found when performing pivot to source translation, then the pivot translation under test is kept in the list.

Nevertheless, all the ideas experimented above became unsuccessful, in the sense that they improve or degrade the results only a little bit.
The validity of using a pivot language for bilingual lexicon extraction from specialized comparable corpora has been proved by this work. Let us recap all the contributions brought by this work, that is to say pivot dictionary and pivot corpora based approaches, and the potential further works and improvements still to be achieved.

6.1 Contributions

We have presented two pivot based approaches for bilingual lexicon extraction from comparable specialized corpora. Both of them lay on pivot dictionaries. We have shown that the bilingual lexicon extraction depends on the quality of the resources. Furthermore, we have also demonstrated that the problem can be fixed involving a third strongly supported language such as English for instance. We also carried out that the enhancements are function of the comparability of the corpora. These first experiments have shown that using a pivot language can make improvements in the case of poorly comparable initial corpora. Because our results have managed to characterize and explain when the use of an additional language helps improve the standard method, we decided to write an article we published at the 8th Workshop on Building and Using Comparable Corpora (BUCC ’15).

We have also presented two other approaches, based on pivot corpora. At this stage of our research work, the only way of exploiting a pivot corpus consists in retrieving pivot translations performing a standard method between source and pivot languages. Then, we get information about the pivot translations from the pivot corpus.
The problem of these pivot corpora based methods, as described in this report, is that a standard approach is performed from source to pivot language. As it is not 100% accurate, the pivot translations are obviously noisy, and this noise has repercussions on the continuation of our proposed algorithms. In order to thwart this effect, we have proposed several measures in order to take into account the weight of the pivot candidates. We have also tried to use an unbalanced pivot corpus to reinforce the main hypothesis consisting in saying that we have at our disposal a strong pivot language. In this case, the pivot stronger corpus brings more lexical information and, in some pairs of languages, the results are significantly improved.

6.2 Further work

In future works, we think that trying to involve a bilingual or trilingual thesaurus can be a good idea. Nevertheless, having access to such a resource will have to be fixed first.

Finally, we think that as single word terms, multi word terms alignments can be successful using pivot based approaches, especially via pivot dictionaries. In the case of this kind of resources, we recon that experiments have to be done and have to adapt compositional methods to the use of a pivot language.
In this appendix, we present the 4 schemes of the different methods we have implemented and of which I show the results. The first approach is presented in Figure A.1 and it is the method involving a pivot dictionary and consisting in translating context vectors successively. The second approach, in Figure A.2, is the method transposing source and target context vectors to pivot language. The third approach, in Figure A.3, consists in using a pivot corpus to perform successive translation. Finally, in Figure A.4, the method for computing similarity over the pivot corpus.
Figure A.1. Pivot dictionary based approach #1 (translating context vectors successively).
Figure A.2. Pivot dictionary based approach #2 (transposing source and target context vectors to pivot language).
SECTION A. PIVOT BASED APPROACHES

Figure A.3. Pivot corpora based approach #3 (successive translation).
FIGURE A.4. Pivot corpora based approach #4 (computing similarity over the pivot corpus).
In this appendix I present the results achieved for the described methods in further details. In Table B.1, we present the results of the standard method (Standard), method transferring context vectors successively ($P_1$) and the method transposing context vectors to pivot language ($P_2$). We also give additional information, such as the best result achievable according to the reference lists and the words belonging to the filtered corpus ($R_{MAX}$) and corpora comparability (Comparability). For each corpus (WE for Wind Energy and MT for Mobile Technologies), the MAP has been computed using Log Likelihood for context vectors computation, and Cosine Similarity or Weighted Jaccard Distance (column Similarity).

In Figures B.1 to B.6, we present the results achieved for the methods transferring context vectors successively ($P_3$) and computing similarity over pivot corpus ($P_4$) using balanced corpora, Log Likelihood for context vectors computation, and Cosine Similarity for the similarity computation. In those figures, we can see how the number of candidate translations in pivot language makes vary the results. The figures also show how the way of taking into account the weight of pivot candidates affect the final MAP ($MAX$ for not taking the weight into account, $PROD$ and $MEAN$ for computing the product or the mean of the pivot and target similarity scores).

In Table B.2, we present the differences of results achieved using balanced corpora ($P_{3B}$ and $P_{4B}$) or an unbalanced pivot corpus ($P_{3U}$ and $P_{4U}$). The parameters are: Log Likelihood for computation and normalization of context vectors; Cosine similarity for similarity computation between context vectors; no context vectors reshaping with linear regression; 5 candidate translations used in pivot language and weighted according to the $PROD$ measure.

Finally, in Table B.3, we present the differences between using an unbalanced pivot corpus ($P_{3U}$ and $P_{4U}$) and an unbalanced pivot corpus with source and target context vectors reshaped using linear regression parameters presented in Section 4.3.2 ($P_{3UR}$ and $P_{4UR}$). The parameters are the same as for $P_{3U}$ and $P_{4U}$ described above.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Comparability</th>
<th>Lang.</th>
<th>Pivot</th>
<th>Similarity</th>
<th>Standard</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$R_{MAX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EN-ES</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.268</td>
<td>0.390</td>
<td>0.374</td>
</tr>
<tr>
<td>65.76%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.284</td>
<td>0.394</td>
<td>0.379</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES-EN</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.119</td>
<td>0.232</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.161</td>
<td>0.236</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EN-DE</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.158</td>
<td>0.125</td>
<td>0.215</td>
</tr>
<tr>
<td>66.21%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.159</td>
<td>0.128</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-EN</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR-DE</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.056</td>
<td>0.118</td>
<td>0.132</td>
</tr>
<tr>
<td>77.63%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.055</td>
<td>0.118</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-FR</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.038</td>
<td>0.028</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.038</td>
<td>0.028</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR-ES</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.366</td>
<td>0.150</td>
<td>0.170</td>
</tr>
<tr>
<td>82.36%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.336</td>
<td>0.148</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES-FR</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.210</td>
<td>0.103</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.216</td>
<td>0.108</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES-DE</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.000</td>
<td>0.041</td>
<td>0.097</td>
</tr>
<tr>
<td>44.24%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.018</td>
<td>0.039</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.045</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EN</td>
<td></td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.018</td>
<td>0.039</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.018</td>
<td>0.039</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-ES</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-FR</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FR-ES</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.514</td>
<td>0.275</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.510</td>
<td>0.274</td>
<td>0.289</td>
</tr>
<tr>
<td>82.02%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.238</td>
<td>0.207</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES-FR</td>
<td>EN</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.238</td>
<td>0.207</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.283</td>
<td>0.220</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES-DE</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.001</td>
<td>0.058</td>
<td>0.087</td>
</tr>
<tr>
<td>44.02%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.058</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>0.033</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EN</td>
<td></td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.000</td>
<td>0.058</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td>0.033</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.058</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE-ES</td>
<td>FR</td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.126</td>
<td>0.355</td>
<td>0.347</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.132</td>
<td>0.376</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.126</td>
<td>0.189</td>
<td>0.179</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EN</td>
<td></td>
<td>Cosine</td>
<td>Jaccard</td>
<td>0.126</td>
<td>0.355</td>
<td>0.347</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.132</td>
<td>0.376</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.126</td>
<td>0.189</td>
<td>0.179</td>
</tr>
</tbody>
</table>

Table B.1: Results achieved for pivot dictionary based approaches
SECTION B. DETAILED RESULTS

Figure B.1: MAP for $P_3$ in function of the number of pivot candidates (MAX method).

Figure B.2: MAP for $P_4$ in function of the number of pivot candidates (MAX method).
SECTION B. DETAILED RESULTS

Figure B.3: MAP for $P_3$ in function of the number of pivot candidates ($PROD$ method).

Figure B.4: MAP for $P_4$ in function of the number of pivot candidates ($PROD$ method).
**SECTION B. DETAILED RESULTS**

**Figure B.5:** MAP for $P_3$ in function of the number of pivot candidates (*MEAN* method).

**Figure B.6:** MAP for $P_4$ in function of the number of pivot candidates (*MEAN* method).
Table B.2: Results for Pivot Corpora Based Approaches for Wind Energy Corpus using unbalanced pivot corpus.

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Pivot</th>
<th>Std.</th>
<th>$P_{3B}$</th>
<th>$P_{3U}$</th>
<th>$P_{4B}$</th>
<th>$P_{4U}$</th>
<th>$R_{MAX}$</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.268</td>
<td>0.339</td>
<td>0.441</td>
<td>0.340</td>
<td>0.456</td>
<td>0.646</td>
<td>65.76%</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.119</td>
<td>0.212</td>
<td>0.270</td>
<td>0.212</td>
<td>0.270</td>
<td>0.491</td>
<td>66.21%</td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.158</td>
<td>0.126</td>
<td>0.042</td>
<td>0.127</td>
<td>0.068</td>
<td>0.458</td>
<td>66.21%</td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.200</td>
<td>77.63%</td>
</tr>
<tr>
<td>FR-DE</td>
<td>EN</td>
<td>0.056</td>
<td>0.147</td>
<td>0.147</td>
<td>0.147</td>
<td>0.147</td>
<td>0.418</td>
<td>82.36%</td>
</tr>
<tr>
<td>DE-FR</td>
<td>EN</td>
<td>0.038</td>
<td>0.019</td>
<td>0.019</td>
<td>0.019</td>
<td>0.019</td>
<td>0.151</td>
<td></td>
</tr>
<tr>
<td>FR-ES</td>
<td>EN</td>
<td>0.366</td>
<td>0.172</td>
<td>0.160</td>
<td>0.173</td>
<td>0.163</td>
<td>0.528</td>
<td></td>
</tr>
<tr>
<td>ES-FR</td>
<td>EN</td>
<td>0.210</td>
<td>0.088</td>
<td>0.089</td>
<td>0.082</td>
<td>0.088</td>
<td>0.357</td>
<td></td>
</tr>
<tr>
<td>ES-DE</td>
<td>FR</td>
<td>0.000</td>
<td>0.018</td>
<td>0.003</td>
<td>0.020</td>
<td>0.018</td>
<td>0.273</td>
<td>44.24%</td>
</tr>
<tr>
<td>EN</td>
<td>0.000</td>
<td>0.032</td>
<td>0.030</td>
<td>0.032</td>
<td>0.029</td>
<td>0.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE-ES</td>
<td>FR</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.218</td>
<td></td>
</tr>
<tr>
<td>EN</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.3: Results for Pivot Corpora Based Approaches for Wind Energy Corpus using unbalanced pivot corpus and source and target context vectors reshaped or not.
In this section, I present the article published under this internship. The paper entitled *Attempting to Bypass Alignment from Comparable Corpora via Pivot Language* aims at putting forward the first two methods based on pivot dictionaries. It has been submitted and accepted at the 8th Workshop on Building and Using Comparable Corpora (BUCC '15) in Beijing, China and it will be presented on July, 30th 2015.
Attesting to Bypass Alignment from Comparable Corpora via Pivot Language

Alexis Linard Béatrice Daille Emmanuel Morin
Université de Nantes, LINA UMR CNRS 6241
2 rue de la Houssinière, BP 92208
44322 Nantes Cedex 03, France
firstname.lastname@univ-nantes.fr

Abstract
Alignment from comparable corpora usually involves two languages, one source and one target language. Previous works on bilingual lexicon extraction from parallel corpora demonstrated that more than two languages can be useful to improve the alignments. Our works have investigated to which extent a third language could be interesting to bypass the original alignment. We have defined two original alignment approaches involving pivot languages and we have evaluated over four languages and two pivot languages in particular. The experiments have shown that in some cases the quality of the extracted lexicon has been enhanced.

1 Introduction
The main goal of this work is to investigate to which extent bilingual lexicon extraction using comparable corpora can be improved using a third language when dealing with poor resource language pairs. Indeed, the quality of the result of the extracted bilingual lexicon strongly depends on the quality of the resources, that is to say the corpora and a general language bilingual dictionary. In this study, we stress the key role of the potential high quality resources of the pivot language (Chiao and Zweigenbaum, 2004; Morin and Prochasson, 2011; Hazem and Morin, 2012). The idea of involving a third language is to benefit from the lexical information conveyed by the additional language. We also assume that in the case of not so usual language pairs the two comparable corpora are of medium quality, and the bilingual dictionary seems weak, due to the nonexistence of such a dictionary. We expect as a consequence a bad quality of the extracted lexicon. Nevertheless, we are highly confident that a language for which we have of a lot of resources can thwart the effect of the poor original resources. English is probably the first language in term of work and resources in Natural Language Processing, hence it can appear as a good candidate as pivot language.

The paper is organized as follows: we give a short overview of bilingual lexicon extraction standard method in Section 2. Our proposed approaches are described in Section 3. The resources we have used are presented in Section 4 and experimental results in Section 5. Finally, we expose further works and improvements in Sections 6 and 7.

2 Bilingual Lexicon Extraction
Initially designed for parallel corpora (Chen, 1993), and due to the scarcity of this kind of resources (Martin et al., 2005), bilingual lexicon extraction then tried to deal with comparable corpora instead (Fung, 1995; Rapp, 1995). An algorithm using comparable corpora is the standard method (Fung and McKeown, 1997) closely based on the notion of context vectors. Many implementations have been designed in order to do so (Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2010). A context vector \( w \) is, for a given word \( w \), the representation of its contexts \( ct_1 \ldots ct_i \) and the number of occurrences found in the window of a corpus. In this approach, context vectors are calculated both in source and target languages corpora. They are also normalized according to association scores. Then, thanks to a seed dictionary, source context vectors are transferred into target language. The similarity between the translated context vector \( W \) for a given source word \( w \) to translate and all target context vectors \( t \) lead to the creation of a list of ranked candidate translations. The rank is function of the similarity between context vectors so that the closer they are, the better the ranked translation is.

Research in this field aims at improving the
quality of the extracted lexicon. For instance, we can cite the use of a bilingual thesaurus (Déjean et al., 2002), implication of predictive methods for word co-occurrence counts (Hazem and Morin, 2013) or the use of unbalanced corpora (Morin and Hazem, 2014). Among them, and in the case of comparable corpora, we can denote that none looked into pivot-language approaches.

Nevertheless, the idea of involving a pivot language for translation tasks is not recent. Biligual lexicon extraction from parallel corpora has already been improved via the use of an intermediary language (Kwon et al., 2013; Seo et al., 2014; Kim et al., 2015), so does statistical translation (Simard, 1999; Och and Ney, 2001). Those works lay on the assumption that another language brings additional information (Dagan and Itai, 1991).

### 3 Alignment Approaches with Pivot Language

In this paper, we present two original approaches which derive from the standard method and involve a third language. We assume that the bilingual dictionary is unavailable or of low quality, but that the source/pivot and pivot/target dictionaries are much better.

#### 3.1 Transferring Context Vectors Successively

The first method, and the most naive is to translate context vectors successively, to start with from source to pivot language, and to follow from pivot to target language. Hence, the context vectors in the source language are computed as it is usually done in the standard method. Then, the second step is to transfer them into the pivot language thanks to a source/pivot dictionary. This operation is done a second time from pivot to target language with a pivot/target dictionary in order to obtain source context vectors translated into target language. We can say that we transferred the context vectors via a pivot language. Finally, the last step of similarity computation stays unchanged: for one source word $w$ for which we want to find the translation in target language, we compute the similarity between its context vector transferred successively $w$ and all target context vectors $t$. This method is presented in Figure 1.

#### 3.2 Transposing Context Vectors to Pivot Language

The second method based on pivot dictionaries consists in translating both source and target context vectors into pivot language. Thus, the operation of computing similarity occurs in the vectorial space of the pivot language. In order to do so, the context vector of a word in source language to translate is computed as it is usually done in the standard method. The second step is to transfer the source and target context vectors into the pivot language using source/pivot and target/pivot dictionaries. At this stage, we gather in the pivot language the translated source and all target context vectors. The next and last operation is to compute the similarity between the source context vector transferred into pivot language $w$ and all target context vectors transferred into pivot language $t$. This method is presented in Figure 2.

### 4 Multilingual Resources

In this paper, we perform translation-candidate extraction from all pairs of languages from/to En-
glish, French, German and Spanish and involving English or French as the pivot language. The use of those pivot languages in particular is motivated by two factors: first, English, because it is the language by default we have of in a quasi infinite amount of data, and last, French, because we know that our resources (corpus and dictionaries) are of good quality.

4.1 Comparable Corpora

The first comparable corpus we used during our experiments is the Wind Energy corpus\textsuperscript{1}. It was built from a crawl of webpages using many keywords related to the wind energy field. The comparable corpus is composed of documents in 7 languages, among others German, English, Spanish and French. The second comparable corpus we used is the Mobile Technologies corpus\textsuperscript{1}. It was also built by crawling the web. Both of them were composed of 300k to 470k words in each language.

4.2 Bilingual Dictionaries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>600k</td>
<td>26k</td>
<td>240k</td>
<td>100k</td>
<td>170k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of entries in each dictionary.

In order to perform bilingual lexicon extraction from comparable corpora, a bilingual dictionary was mandatory. Nevertheless, we only have of French/English, French/Spanish and French/German dictionaries from the ELRA catalogue\textsuperscript{2}. These dictionaries were generalist, and contained few terms related to the Wind Energy and Mobile Technologies domains. So, the French/English, French/Spanish and French/German were reversed to obtain English/French, Spanish/French and German/French dictionaries. As for the others, they were built by triangulation from the ones above (see Table 1). As a consequence, we expect those dictionaries to be very mediocre.

4.3 Reference Lists

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>FR</th>
<th>ES</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WE</td>
<td>48</td>
<td>58</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>MT</td>
<td>52</td>
<td>58</td>
<td>60</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 2: Number of SWT in reference lists.

In order to evaluate the output of the different approaches, terminology reference lists were built from each corpus in each language (Loginova et al., 2012). Depending on the corpus and the language, the lists were composed of 48 to 88 single word terms (abbreviated SWT – see Table 2).

5 Experiments and Results

Pre-processing French, English, Spanish and German documents were pre-processed using TTC TermSuite (Rocheteau and Daille, 2011)\textsuperscript{3}. The operations done during pre-processing were the following: tokenization, part-of-speech tagging and lemmatization. Moreover, function words and hapaxes had been removed.

\textsuperscript{1}http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html
\textsuperscript{2}http://catalog.elra.info/
\textsuperscript{3}https://logiciels.lina.univ-nantes.fr/redmine/projects
Context vectors In order to compute and normalize context vectors, the value $a(ct)$ associated to each co-occurrence $ct$ of a given word $w$ in the corpus was computed. Such value could be computed thanks to Log Likelihood (Fano and Hawkins, 1961) or Mutual Information (Dunning, 1993) for instance. Among them we chose Log Likelihood as its representativity is the most accurate (Bordag, 2008). Context vectors were computed by TermSuite, as one of its components performed this operation.

Similarity measures The so-called similarity could be computed according to Cosine similarity (Salton and Lesk, 1968) or Weighted Jaccard Distance (Grefenstette, 1994). We decided to only present the results achieved using Cosine similarity. The differences between them in term of Mean Reciprocal Rank (MRR) were insignificant.

$$\text{Cosine}(\mathbf{w}, \mathbf{t}) = \frac{\sum_k a(w_k) a(t_k)}{\sqrt{\sum_k a(w_k)^2} \sqrt{\sum_k a(t_k)^2}}$$

Evaluation metrics In order to evaluate our approaches, we used Mean Reciprocal Rank (Voorhees, 1999). The strength of this metric is that it takes into account the rank of the candidate translations. Hereinafter, the MRR defined as follows ($t$ stands for the terms to evaluate and $r_t$ for the rank achieved by the system for the good translation of $t$):

$$MRR = \frac{1}{|t|} \times \sum_{k=1}^{|t|} \left( \frac{1}{r_{t_k}} \right)$$

Results The MRR achieved for both approaches is shown in Table 3 for Wind Energy and Mobile Technologies corpora respectively. We present, for the sake of comparison, the results achieved by the standard method (Std.), method transferring context vectors successively ($P_1$) and the method transposing context vectors to pivot language ($P_2$). We also give additional information, such as the best achievable result according to the reference lists and the words belonging to the filtered corpus ($R_{MAX}$) and corpora comparability $C$ (Li and Gaussier, 2010).

The corpus comparability metric consists in the expectation of finding the translation in target language for each source word in the corpus. Thereon, it is a good way of measuring the distributional symmetry between two corpora and given a dictionary. We can also notice that the Maximum Recall $R_{MAX}$ is quite low for some pairs of languages: this is due to the high number of hapaxes belonging to the reference lists that were filtered out during pre-processing.

According to the results, we can see that there is a strong correlation between the improvements achieved by pivot based approaches and corpus comparability. We have improved the quality of the extracted bilingual lexicon only in the case of poorly comparable corpora, respectively $\leq 65.76\%$ and $\leq 66.52\%$ for Wind Energy and Mobile Technologies corpora. For instance, we have increased the MRR from 0.268 to 0.390 and 0.374 in the case of translation from English to Spanish for the Wind Energy corpus, and from 0.126 to 0.355 and 0.347 for German to Spanish via French for the Mobile Technologies corpus.

| Lang. Pivot | Std. | $P_1$ | $P_2$ | $R_{MAX}$ | C
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind Energy</td>
<td>Mobile Technologies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN-ES</td>
<td>FR</td>
<td>0.268</td>
<td>0.390</td>
<td>0.374</td>
<td>0.646</td>
</tr>
<tr>
<td>ES-EN</td>
<td>FR</td>
<td>0.119</td>
<td>0.232</td>
<td>0.233</td>
<td>0.491</td>
</tr>
<tr>
<td>EN-DE</td>
<td>FR</td>
<td>0.158</td>
<td>0.125</td>
<td>0.215</td>
<td>0.458</td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.200</td>
</tr>
<tr>
<td>FR-DE</td>
<td>EN</td>
<td>0.056</td>
<td>0.118</td>
<td>0.132</td>
<td>0.418</td>
</tr>
<tr>
<td>DE-FR</td>
<td>EN</td>
<td>0.038</td>
<td>0.028</td>
<td>0.028</td>
<td>0.157</td>
</tr>
<tr>
<td>FR-ES</td>
<td>EN</td>
<td>0.366</td>
<td>0.150</td>
<td>0.176</td>
<td>0.528</td>
</tr>
<tr>
<td>ES-FR</td>
<td>EN</td>
<td>0.210</td>
<td>0.103</td>
<td>0.117</td>
<td>0.357</td>
</tr>
<tr>
<td>ES-DE</td>
<td>FR</td>
<td>0.000</td>
<td>0.041</td>
<td>0.097</td>
<td>0.179</td>
</tr>
<tr>
<td>DE-ES</td>
<td>FR</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
<td>0.218</td>
</tr>
<tr>
<td>DE-EN</td>
<td>FR</td>
<td>0.001</td>
<td>0.018</td>
<td>0.018</td>
<td>0.218</td>
</tr>
</tbody>
</table>

Table 3: MRR achieved for pivot dictionary based approaches.
(≤ 68% ≤ C ≤ 80%), results remain unchanged in comparison with the standard approach. Finally, for highly comparable corpora (C > 80%) the quality of the extracted lexicon gets worse.

The interpretation we suggest is the following: given two corpora, S in source language, T in target and a bilingual dictionary source/target D, the comparability is function of S, T, D. Therefore, a low comparability measure can be due to a poor expectation of finding the translation in target language for each source word in the corpus because the two corpora are not lexically close enough, or because the dictionary is weak. We checked this second option, and this is how we substantiate the pivot dictionary based approaches. Thus, the use of source/pivot D_{S/P} and pivot/target D_{P/T} dictionary can artificially improve the comparability and enhance the extracted lexicon. We have also remarked that the coverage of dictionaries is an important factor: a large dictionary is better than a shorter.

Of course, we do not pretend that our methods can compare with an initially very highly comparable corpora since the use of pivot dictionaries will introduce more noise than it will bring additional information.

7 Conclusion

We have presented two pivot based approaches for bilingual lexicon extraction from comparable specialized corpora. Both of them lay on pivot dictionaries. We have shown that the bilingual lexicon extraction depends on the quality of the resources. Furthermore, we have also demonstrated that the problem can be fixed involving a third strongly supported language such as English for instance. We have also carried out that the enhancements are function of the comparability of the corpora. These first experiments have shown that using a pivot language can make improvements in the case of poorly comparable initial corpora.

In future works, we will try to benefit from the information brought by an unbalanced pivot corpus. Unlike this article in which we have only looked into pivot dictionaries in order to increase the comparability of the source and target corpora, we think that the next step is to reshape context vectors with a pivot corpus. In addition, we will see whether linear regression models to reshape context vectors can make improvements or not.

Acknowledgments

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

References


Ido Daagan and Alon Itai. 1991. Two languages are more informative than one. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pages 130–137, Berkeley, California, USA.


Jae-Hoon Kim, Hong-Seok Kwon, and Hyeong-Won Seo. 2015. Evaluating a pivot-based approach for bilingual lexicon extraction. *Computational Intelligence and Neuroscience*, 2015.


