Detecting Highly Confident Word Translations from Comparable Corpora without Any Prior Knowledge

Ivan Vulić and Marie-Francine Moens
Department of Computer Science
KU Leuven
Celestijnenlaan 200A
Leuven, Belgium
{ivan.vulic,marie-francine.moens}@cs.kuleuven.be

Abstract

In this paper, we extend the work on using latent cross-language topic models for identifying word translations across comparable corpora. We present a novel precision-oriented algorithm that relies on per-topic word distributions obtained by the *bilingual LDA* (BiLDA) latent topic model. The algorithm aims at harvesting only the most probable word translations across languages in a greedy fashion, without any prior knowledge about the language pair, relying on a symmetrization process and the one-to-one constraint. We report our results for Italian-English and Dutch-English language pairs that outperform the current state-of-the-art results by a significant margin. In addition, we show how to use the algorithm for the construction of high-quality initial seed lexicons of translations.

1 Introduction

Bilingual lexicons serve as an invaluable resource of knowledge in various natural language processing tasks, such as dictionary-based cross-language information retrieval (Carbonell et al., 1997; Levow et al., 2005) and statistical machine translation (SMT) (Och and Ney, 2003). In order to construct high quality bilingual lexicons for different domains, one usually needs to possess parallel corpora or build such lexicons by hand. Compiling such lexicons manually is often an expensive and time-consuming task, whereas the methods for mining the lexicons from parallel corpora are not applicable for language pairs and domains where such corpora is unavailable or missing. Therefore the focus of researchers turned to comparable corpora, which consist of documents with partially overlapping content, usually available in abundance. Thus, it is much easier to build a high-volume comparable corpus. A representative example of such a comparable text collection is Wikipedia, where one may observe articles discussing the similar topic, but strongly varying in style, length and vocabulary, while still sharing a certain amount of main concepts (or topics).

Over the years, several approaches for mining translations from non-parallel corpora have emerged (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Déjean et al., 2002; Chiao and Zweigenbaum, 2002; Gaussier et al., 2004; Fung and Cheung, 2004; Morin et al., 2007; Haghighi et al., 2008; Shezaf and Rappoport, 2010; Laroche and Langlais, 2010), all sharing the same Firthian assumption, often called the *distributionial hypothesis* (Harris, 1954), which states that words with a similar meaning are likely to appear in similar contexts across languages. All these methods have examined different representations of word contexts and different methods for matching words across languages, but they all have in common a need for a seed lexicon of translations to efficiently bridge the gap between languages. That seed lexicon is usually crawled from the Web or obtained from parallel corpora. Recently, Li et al. (2011) have proposed an approach that improves precision of the existing methods for bilingual lexicon extraction, based on improving the comparability of the corpus under consideration, prior to extracting actual bilingual lexicons. Other methods such as (Koehn and Knight, 2002) try to design a bootstrapping algorithm based on an initial seed lexicon of translations and various lexical evidences. However, the quality of their initial seed lexicon is disputable,

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 449–459,
Avignon, France, April 23 - 27 2012.

since the construction of their lexicon is language-pair biased and cannot be completely employed on distant languages. It solely relies on unsatisfactory language-pair independent cross-language clues such as words shared across languages.

Recent work from Vulić et al.(2011) utilized the distributional hypothesis in a different direction. It attempts to abrogate the need of a seed lexicon as a prerequisite for bilingual lexicon extraction. They train a cross-language topic model on document-aligned comparable corpora and introduce different methods for identifying word translations across languages, underpinned by per-topic word distributions from the trained topic model. Due to the fact that they deal with comparable Wikipedia data, their translation model contains a lot of noise, and some words are poorly translated simply because there are not enough occurrences in the corpus. The goal of this work is to design an algorithm which will learn to harvest only the most probable translations from the per-word topic distributions. The translations learned by the algorithm then might serve as a highly accurate, precision-based initial seed lexicon, which can then be used as a tool for translating source word vectors into the target language. The key advantage of such a lexicon lies in the fact that there is no language-pair dependent prior knowledge involved in its construction (e.g., orthographic features). Hence, it is completely applicable to any language pair for which there exist sufficient comparable data for training of the topic model.

Since comparable corpora often construct a very noisy environment, it is of the utmost importance for a precision-oriented algorithm to learn when to stop the process of matching words, and which candidate pairs are surely not translations of each other. The method described in this paper follows this intuition: while extracting a bilingual lexicon, we try to rematch words, keeping only the most confident candidate pairs and disregarding all the others. After that step, the most confident candidate pairs might be used with some of the existing context-based techniques to find translations for the words discarded in the previous step. The algorithm is based on: (1) the assumption of symmetry, and (2) the one-to-one constraint. The idea of symmetrization has been borrowed from the symmetrization heuristics introduced for word alignments in SMT (Och and Ney, 2003), where the intersection heuristics is employed for a precision-oriented algorithm. In our setting, it basically means that we keep a translation pair (w_i^S, w_j^T) if and only if, after the symmetrization process, the top translation candidate for the source word w_i^S is the target word w_i^T and vice versa. The one-to-one constraint aims at matching the most confident candidates during the early stages of the algorithm, and then excluding them from further search. The utility of the constraint for parallel corpora has already been evaluated by Melamed (2000).

The remainder of the paper is structured as follows. Section 2 gives a brief overview of the methods, relying on per-topic word distributions, which serve as the tool for computing cross-language similarity between words. In Section 3, we motivate the main assumptions of the algorithm and describe the full algorithm. Section 4 justifies the underlying assumptions of the algorithm by providing comparisons with a current-state-of-the-art system for Italian-English and Dutch-English language pairs. It also contains another set of experiments which investigates the potential of the algorithm in building a language-pair unbiased seed lexicon, and compares the lexicon with other seed lexicons. Finally, Section 5 lists conclusion and possible paths of future work.

2 Calculating Initial Cross-Language Word Similarity

This section gives a quick overview of the **Cue** method, the **TI** method, and their combination, described by Vulić et al.(2011), which proved to be the most efficient and accurate for identifying potential word translations once the cross-language BiLDA topic model is trained and the associated per-topic distributions are obtained for both source and target corpora. The BiLDA model we use is a natural extension of the standard LDA model and, along with the definition of per-topic word distributions, has been presented in (Ni et al., 2009; De Smet and Moens, 2009; Mimno et al., 2009). BiLDA takes advantage of the document alignment by using a single variable that contains the topic distribution θ. This variable is language-independent, because it is shared by each of the paired bilingual comparable documents. Topics for each document are sampled from θ, from which the words are then sampled in conjugation with the vocabulary distribution ϕ

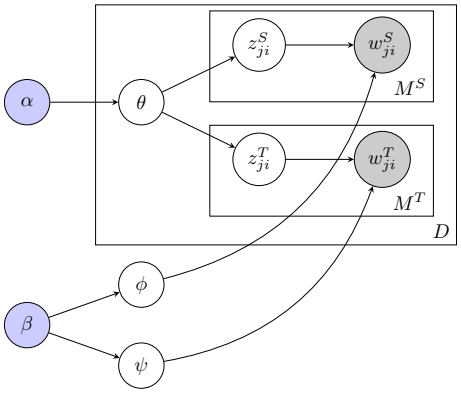

Figure 1: The bilingual LDA (BiLDA) model

(for language S) and ψ (for language T).

2.1 Cue Method

A straightforward approach to express similarity between words tries to emphasize the associative relation in a natural way - modeling the probability $P(w_2^T|w_1^S)$, i.e. the probability that a target word w_2^T will be generated as a response to a cue source word w_1^S, where the link between the words is established via the shared topic space: $P(w_2^T|w_1^S) = \sum_{k=1}^{K} P(w_2^T|z_k)P(z_k|w_1^S)$, where K denotes the number of cross-language topics.

2.2 TI Method

This approach constructs word vectors over a shared space of cross-language topics, where values within vectors are the *TF-ITF* scores (term frequency - inverse topic frequency), computed in a completely analogical manner as the *TF-IDF* scores for the original word-document space (Manning and Schütze, 1999). *Term frequency*, given a source word w_i^S and a topic z_k, measures the importance of the word w_i^S within the particular topic z_k, while *inverse topical frequency (ITF)* of the word w_i^S measures the general importance of the source word w_i^S across all topics. The final *TF-ITF* score for the source word w_i^S and the topic z_k is given by $TF-ITF_{i,k} = TF_{i,k} \cdot ITF_i$.

The *TF-ITF* scores for target words associated with target topics are calculated in an analogical manner and the standard cosine similarity is then used to find the most similar target word vectors for a given source word vector.

2.3 Combining the Methods

Topic models have the ability to build clusters of words which might not always co-occur together in the same textual units and therefore add extra information of potential relatedness. These two methods for automatic bilingual lexicon extraction interpret and exploit underlying per-topic word distributions in different ways, so combining the two should lead to even better results. The two methods are linearly combined, with the overall score given by:

$$Sim_{TI+Cue}(w_1^S, w_2^T) = \lambda Sim_{TI}(w_1^S, w_2^T) + (1-\lambda) Sim_{Cue}(w_1^S, w_2^T) \quad (1)$$

Both methods posses several desirable properties. According to Griffiths et al. (2007), the conditioning for the **Cue** method automatically compromises between word frequency and semantic relatedness since higher frequency words tend to have higher probability across all topics, but the distribution over topics $P(z_k|w_1^S)$ ensures that semantically related topics dominate the sum. The similar phenomenon is captured by the **TI** method by the usage of *TF*, which rewards high frequency words, and *ITF*, which assigns a higher importance for words semantically more related to a specific topic. These properties are incorporated in the combination of the methods. As the final result, the combined method provides, for each source word, a ranked list of target words with associated scores that measure the strength of cross-language similarity. The higher the score, the more confident a translation pair is. We will use this observation in the next section during the algorithm construction.

The lexicon constructed by solely applying the combination of these methods without any additional assumptions will serve as a baseline in the results section.

3 Constructing the Algorithm

This section explains the underlying assumptions of the algorithm: the **assumption of symmetry** and the **one-to-one assumption**. Finally, it provides the complete outline of the algorithm.

3.1 Assumption of Symmetry

First, we start with the intuition that the **assumption of symmetry** strengthens the confidence of a translation pair. In other words, if the most probable translation candidate for a source word w_1^S is a target word w_2^T and, vice versa, the most probable translation candidate of the target word w_2^T

is the source word w_1^S, and their *TI+Cue* scores are above a certain threshold, we can claim that the words w_1^S and w_2^T are a translation pair. The definition of the symmetric relation can also be relaxed. Instead of observing only one top candidate from the lists, we can observe top N candidates from both sides and include them in the search space, and then re-rank the potential candidates taking into account their associated *TI+Cue* scores and their respective positions in the list. We will call N the **search space depth**. Here is the outline of the re-ranking method if the search space consists of the top N candidates on both sides:

1. Given is a source word w_s^S, for which we actually want to find the most probable translation candidate. Initialize an empty list $Final_s = \{\}$ in which target language candidates with their recalculated associated scores will be stored.
2. Obtain *TI+Cue* scores for all target words. Keep only N best scoring target candidates: $\{w_{s,1}^T, \ldots, w_{s,N}^T\}$ along with their respective scores.
3. For each target candidate from $\{w_{s,1}^T, \ldots, w_{s,N}^T\}$ acquire *TI+Cue* scores over the entire source vocabulary. Keep only N best scoring source language candidates. Each word $w_{s,i}^T \in \{w_{s,1}^T, \ldots, w_{s,N}^T\}$ now has a list of N source language candidates associated with it: $\{w_{i,1}^S, w_{i,2}^S \ldots, w_{i,N}^S\}$.
4. For each target candidate word $w_{s,i}^T \in \{w_{s,1}^T, \ldots, w_{s,N}^T\}$, do as follows:
   (a) If one of the words from the associated list is the given source word w_s^S, remember: (1) the position m, denoting how high in the list the word w_s^S was found, and (2) the associated *TI+Cue* score $Sim_{TI+Cue}(w_{s,i}^T, w_{i,m}^S = w_s^S)$. Calculate:
   (i) $G_{1,i} = Sim_{TI+Cue}(w_s^S, w_{s,i}^T)/i$
   (ii) $G_{2,i} = Sim_{TI+Cue}(w_{s,i}^T, w_{i,m}^S)/m$
   Following that, calculate GM_i, the geometric mean of the values $G_{1,i}$ and $G_{2,i}$[1]: $GM_i = \sqrt{G_{1,i} \cdot G_{2,i}}$. Add a tuple $(w_{s,i}^T, GM_i)$ to the list $Final_s$.
   (b) If we have reached the end of the list for the target candidate word $w_{s,i}^T$ without finding the given source word w_s^S, and $i < N$, continue with the next word $w_{s,i+1}^T$. Do not add any tuple to $Final_s$ in this step.
5. If the list $Final_s$ is not empty, sort the tuples in the list in descending order according to their GM_i scores. The first element of the sorted list contains a word $w_{s,high}^T$, the final translation candidate of the source word w_s^S. If the list $Final_s$ is not empty, the final result of this process will be the cross-language word translation pair $(w_s^S, w_{s,high}^T)$.

We will call this symmetrization process the **symmetrizing re-ranking**. It attempts at pushing the correct cross-language synonym to the top of the candidates list, taking into account both the strength of similarities defined through the *TI+Cue* scores in both directions, and positions in ranked lists. A blatant example depicting how this process helps boost precision is presented in Figure 2. We can also design a thresholded variant of this procedure by imposing an extra constraint. When calculating target language candidates for the source word w_s^S in Step 2, we proceed further only if the first target candidate scores above a certain threshold P and, additionally, in Step 3, we keep lists of N source language candidates for only those target words for which the first source language candidate in their respective list scored above the same threshold P. We will call this procedure the **thresholded symmetrizing re-ranking**, and this version will be employed in the final algorithm.

3.2 One-to-one Assumption

Melamed (2000) has already established that most source words in parallel corpora tend to translate to only one target word. That tendency is modeled by the **one-to-one assumption**, which constrains each source word to have at most one translation on the target side. Melamed's paper reports that this bias leads to a significant positive impact on precision and recall of bilingual lexicon extraction from parallel corpora. This assumption should also be reasonable for many types of comparable corpora such as Wikipedia or news corpora, which are topically aligned or cover similar themes. We

[1] Scores $G_{1,i}$ and $G_{2,i}$ are structured in such a way to balance between positions in the ranked lists and the *TI+Cue* scores, since they reward candidate words which have high *TI+Cue* scores associated with them, and penalize words if they are found lower in the list of potential candidates.

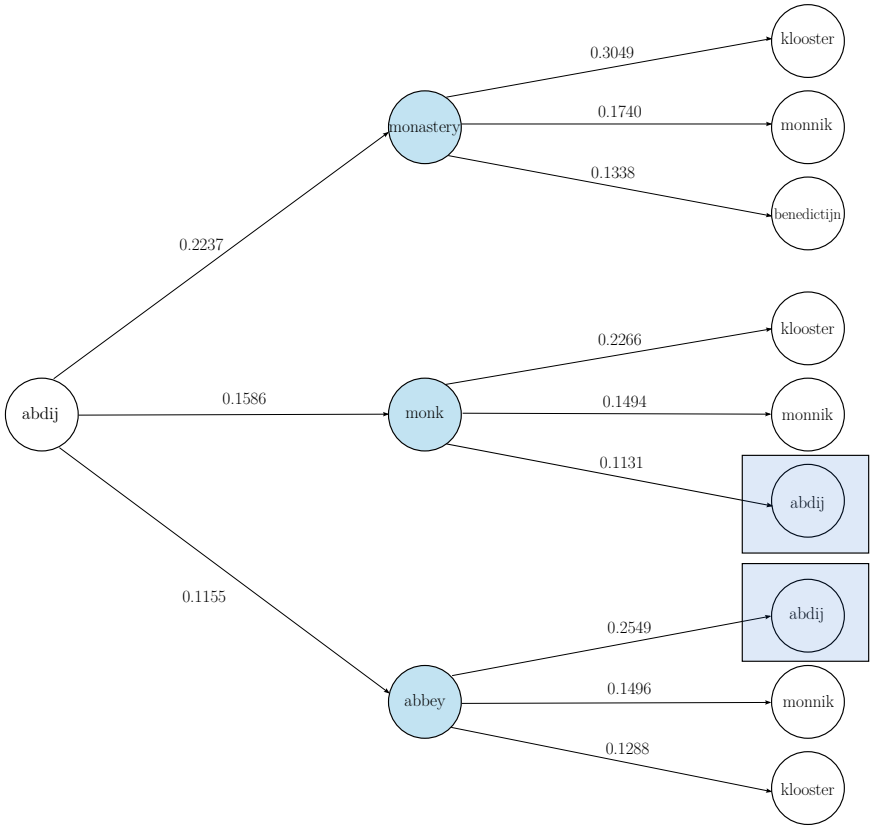

Figure 2: An example where the assumption of symmetry and the one-to-one assumption clearly help boost precision. If we keep top $N_c = 3$ candidates from both sides, the algorithm is able to detect that the correct Dutch-English translation pair is *(abdij, abbey)*. The *TI+Cue* method without any assumptions would result with an indirect association *(abdij, monastery)*. If only the one-to-one assumption was present, the algorithm would greedily learn the correct direct association *(monastery, klooster)*, remove those words from their respective vocabularies and then again result with another indirect association *(abdij, monk)*. By additionally employing the assumption of symmetry with the re-ranking method from Subsection 3.1, the algorithm correctly learns the translation pair *(abdij, abbey)*. Correct translation pairs *(klooster, monastery)* and *(monnik, monk)* are also obtained. Again here, the pair *(monnik, monk)* would not be obtained without the one-to-one assumption.

will prove that the assumption leads to better precision scores even for bilingual lexicon extraction from such comparable data. The intuition behind introducing this constraint is fairly simple. Without the assumption, the similarity scores between source and target words are calculated independently of each other. We will illustrate the problem arising from the independence assumption with an example.

Suppose we have an Italian word *arcipelago*, and we would like to detect its correct English translation (*archipelago*). However, after the *TI+Cue* method is employed, and even after the symmetrizing re-ranking process from the previous step is used, we still acquire a wrong translation candidate pair *(arcipelago, island)*. Why is that so? The word (*arcipelago*) (or its translation) and the acquired translation (*island*) are semantically very close, and therefore have similar distributions over cross-language topics, but *island* is a much more frequent term. The *TI+Cue* method concludes that two words are potential translations whenever their distributions over cross-language topics are much more similar than expected by chance. Moreover, it gives a preference to more frequent candidates, so it will eventually end up learning an **indirect association**[2] between words *arcipelago* and *island*. The one-to-one assumption should mitigate the problem of such indirect associations if we design our algorithm in such a way that it learns the most confident **direct associations**[2] first:

[2]A direct association, as defined in (Melamed, 2000), is an association between two words (in this setting found by the *TI+Cue* method) where the two words are indeed mutual translations. Otherwise, it is an indirect association.

1. Learn the correct direct association pair *(isola, island)*.
2. Remove the words *isola* and *island* from their respective vocabularies.
3. Since *island* is not in the vocabulary, the indirect association between *arcipelago* and *island* is not present any more. The algorithm learns the correct direct association *(arcipelago, archipelago)*.

3.3 The Algorithm

3.3.1 One-Vocabulary-Pass

First, we will provide a version of the algorithm with a fixed threshold P which completes only one pass through the source vocabulary. Let V^S denote a given source vocabulary, and let V^T denote a given target vocabulary. We need to define several parameters of the algorithm. Let N_0 be the initial maximum search space depth for the thresholded symmetrizing re-ranking procedure. In Figure 2, the current depth N_c is 3, while the maximum depth might be set to a value higher than 3. The algorithm with the fixed threshold P proceeds as follows:

1. Initialize the maximum search space depth $N_M = N_0$. Initialize an empty lexicon *L*.
2. For each source word $w_s^S \in V^S$ do:
   (a) Set the current search space depth $N_c = 1$.[3]
   (b) Perform the thresholded symmetrizing re-ranking procedure with the current search space set to N_c and the threshold P. If a translation pair $(w_s^S, w_{s,high}^T)$ is found, go to the Sub-step 2(d).
   (c) If a translation pair is not found, and $N_c < N_M$, increment the current search space $N_c = N_c + 1$ and return to the previous Sub-step 2(b). If a translation pair is not found and $N_c = N_M$, return to Step 2 and proceed with the next word.
   (d) For the found translation pair $(w_s^S, w_{s,high}^T)$, remove words w_s^S and $w_{s,high}^T$ from their respective vocabularies: $V^S = V^S - \{w_s^S\}$ and $V^T = V^T - \{w_{s,high}^T\}$ to satisfy the one-to-one constraint. Add the pair $(w_s^S, w_{s,high}^T)$ to the lexicon L.

We will name this procedure the **one-vocabulary-pass** and employ it later in an iterative algorithm with a varying threshold and a varying maximum search space depth.

3.3.2 The Final Algorithm

Let us now define P_0 as the initial threshold, let P_f be the threshold at which we stop decreasing the value for threshold and start expanding our maximum search space depth for the thresholded symmetrizing re-ranking, and let dec_p be a value for which we decrease the current threshold in each step. Finally, let N_f be the limit for the maximum search space depth, and N_M denote the current maximum search space depth. The final algorithm is given by:

1. Initialize the maximum search space depth $N_M = N_0$ and the starting threshold $P = P_0$. Initialize an empty lexicon L_{final}.
2. Check the stopping criterion: If $N_M > N_f$, go to Step 5, otherwise continue with Step 3.
3. Perform the *one-vocabulary-pass* with the current values of P and N_M. Whenever a translation pair is found, it is added to the lexicon L_{final}. Additionally, we can also save the threshold and the depth at which that pair was found.
4. Decrease P: $P = P - dec_p$, and check if $P < P_f$. If still not $P < P_f$, go to Step 3 and perform the *one-vocabulary-pass* again. Otherwise, if $P < P_f$ and there are still unmatched words in the source vocabulary, reset P: $P = P_0$, increment N_M: $N_M = N_M + 1$ and go to Step 2.
5. Return L_{final} as the final output of the algorithm.

The parameters of the algorithm model its behavior. Typically, we would like to set P_0 to a high value, and N_0 to a low value, which makes our constraints strict and narrows our search space, and consequently, extracts less translation pairs in the first steps of the algorithm, but the set of those translation pairs should be highly accurate. Once it is not possible to extract any more pairs with such strict constraints, the algorithm re-

[3]The intuition here is simple – we are trying to detect a direct association as high as possible in the list. In other words, if the first translation candidate for the source word $isola$ is the target word $island$, and, vice versa, the first translation candidate for the target word $island$ is $isola$, we do not need to expand our search depth, because these two words are the most likely translations.

laxes them by lowering the threshold and expanding the search space by incrementing the maximum search space depth. The algorithm may leave some of the source words unmatched, which is also dependent on the parameters of the algorithm, but, due to the one-to-one assumption, that scenario also occurs whenever a target vocabulary contains more words than a source vocabulary.

The number of operations of the algorithm also depends on the parameters, but it mostly depends on the sizes of the given vocabularies. The complexity is $O(|V^S||V^T|)$, but the algorithm is computationally feasible even for large vocabularies.

4 Results and Discussion

4.1 Training Collections

The data used for training of the models is collected from various sources and varies strongly in theme, style, length and its comparableness. In order to reduce data sparsity, we keep only lemmatized non-proper noun forms.

For Italian-English language pair, we use 18,898 Wikipedia article pairs to train BiLDA, covering different themes with different scopes and subtopics being addressed. Document alignment is established via interlingual links from the Wikipedia metadata. Our vocabularies consist of 7,160 Italian nouns and 9,116 English nouns.

For Dutch-English language pair, we use 7,602 Wikipedia article pairs, and 6,206 Europarl document pairs, and combine them for training.[4] Our final vocabularies consist of 15,284 Dutch nouns and 12,715 English nouns.

Unlike, for instance, Wikipedia articles, where document alignment is established via interlingual links, in some cases it is necessary to perform document alignment as the initial step. Since our work focuses on Wikipedia data, we will not get into detail with algorithms for document alignment. An IR-based method for document alignment is given in (Utiyama and Isahara, 2003; Munteanu and Marcu, 2005), and a feature-based method can be found in (Vu et al., 2009).

4.2 Experimental Setup

All our experiments rely on BiLDA training with comparable data. Corpora and software for BiLDA training are obtained from Vulić et al. (2011). We train the BiLDA model with 2000 topics using Gibbs sampling, since that number of topics displays the best performance in their paper. The linear interpolation parameter for the combined *TI+Cue* method is set to $\lambda = 0.1$.

The parameters of the algorithm, adjusted on a set of 500 randomly sampled Italian words, are set to the following values in all experiments, except where noted different: $P_0 = 0.20$, $P_f = 0.00$, $dec_p = 0.01$, $N_0 = 3$, and $N_f = 10$.

The initial ground truth for our source vocabularies has been constructed by the freely available *Google Translate* tool. The final ground truth for our test sets has been established after we have manually revised the list of pairs obtained by *Google Translate*, deleting incorrect entries and adding additional correct entries. All translation candidates are evaluated against this benchmark lexicon.

4.3 Experiment I: Do Our Assumptions Help Lexicon Extraction?

With this set of experiments, we wanted to test whether both the assumption of symmetry and the one-to-one assumption are useful in improving precision of the initial *TI+Cue* lexicon extraction method. We compare three different lexicon extraction algorithms: (1) the basic *TI+Cue* extraction algorithm (**LALG-BASIC**) which serves as the baseline algorithm[5], (2) the algorithm from Section 3, but without the one-to-one assumption (**LALG-SYM**), meaning that if we find a translation pair, we still keep words from the translation pair in their respective vocabularies, and (3) the complete algorithm from Section 3 (**LALG-ALL**). In order to evaluate these lexicon extraction algorithms for both Italian-English and Dutch-English, we have constructed a test set of 650 Italian nouns, and a test set of 1000 Dutch nouns of high and medium frequency. Precision scores for both language pairs and for all lexicon extraction algorithms are provided in Table 1.

Based on these results, it is clearly visible that both assumptions our algorithm makes are valid

[4] In case of Europarl, we use only the evidence of document alignment during the training and do not benefit from the parallelness of the sentences in the corpus.

[5] We have also tested whether LALG-BASIC outperforms a method modeling direct co-occurrence, that uses cosine to detect similarity between word vectors consisting of TF-IDF scores in the shared document space (Cimiano et al., 2009). Precision using that method is significantly lower, e.g. 0.5538 vs. 0.6708 of LALG-BASIC for Italian-English.

LEX Algorithm	Italian-English	Dutch-English
LALG-BASIC	0.6708	0.6560
LALG-SYM	0.6862	0.6780
LALG-ALL	**0.7215**	**0.7170**

Table 1: Precision scores on our test sets for the 3 different lexicon extraction algorithms.

and contribute to better overall scores. Therefore in all further experiments we will use the *LALG-ALL* extraction algorithm.

4.4 Experiment II: How Does Thresholding Affect Precision?

The next set of experiments aims at exploring how precision scores change while we gradually decrease threshold values. The main goal of these experiments is to detect when to stop with the extraction of translation candidates in order to preserve a lexicon of only highly accurate translations. We have fixed the maximum search space depth $N_0 = N_f = 3$. We used the same test sets from Experiment I. Figure 3 displays the change of precision in relation to different threshold values, where we start harvesting translations from the threshold $P_0 = 0.2$ down to $P_f = 0.0$. Since our goal is to extract as many correct translation pairs as possible, but without decreasing the precision scores, we have also examined what impact this gradual decrease of threshold also has on the number of extracted translations. We have opted for the F_β measure (van Rijsbergen, 1979):

$$F_\beta = (1 + \beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (2)$$

Since our task is precision-oriented, we have set $\beta = 0.5$. $F_{0.5}$ measure values precision as twice as important as recall. The $F_{0.5}$ scores are also provided in Figure 3.

4.5 Experiment III: Building a Seed Lexicon

Finally, we wanted to test how many accurate translation pairs our best scoring LALG-ALL algorithm is able to acquire from the entire source vocabulary, with very high precision still remaining paramount. The obtained highly-precise seed lexicon then might be employed for an additional bootstrapping procedure similar to (Koehn and Knight, 2002; Fung and Cheung, 2004) or simply for translating context vectors as in (Gaussier et al., 2004).

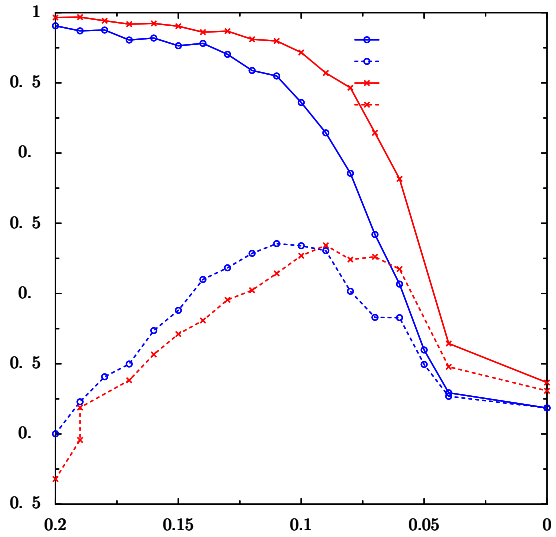

Figure 3: Precision and $F_{0.5}$ scores in relation to threshold values. We can observe that the algorithm retrieves only highly accurate translations for both language pairs while the threshold goes down from value 0.2 to 0.1, while precision starts to drop significantly after the threshold of 0.1. $F_{0.5}$ scores also reach their peaks within that threshold region.

If we do not know anything about a given language pair, we can only use words shared across languages as lexical clues for the construction of a seed lexicon. It often leads to a low precision lexicon, since many false friends are detected.

For Italian-English, we have found 431 nouns shared between the two languages, of which 350 were correct translations, leading to a precision of 0.8121. As an illustration, if we take the first 431 translation pairs retrieved by LALG-ALL, there are 427 correct translation pairs, leading to a precision of 0.9907. Some pairs do not share any orthographic similarities: *(uccello, bird), (tastiera, keyboard), (salute, health), (terremoto, earthquake)* etc.

Following Koehn and Knight (2002), we have also employed simple transformation rules for the adoption of words from one language to another. The rules specific to the Italian-English translation process that have been employed are: (R1) if an Italian noun ends in $-ione$, but not in $-zione$, strip the final e to obtain the corresponding English noun. Otherwise, strip the suffix $-zione$, and append $-tion$; (R2) if a noun ends in $-ia$, but not in $-zia$ or $-fia$, replace the suffix $-ia$ with $-y$. If a noun ends in $-zia$, replace the suffix with $-cy$ and if a noun ends in $-fia$, replace

	Italian-English			Dutch-English		
Lexicon	**# Correct**	**Precision**	$F_{0.5}$	**# Correct**	**Precision**	$F_{0.5}$
LEX-1	350	0.8121	0.1876	898	0.8618	0.2308
LEX-2	766	0.8938	0.3473	**1376**	0.9011	**0.3216**
LEX-LALG	**782**	**0.8958**	**0.3524**	1106	**0.9559**	0.2778
LEX-1+LEX-LALG	1070	0.8785	0.4290	1860	0.9082	0.3961
LEX-R+LEX-LALG	1141	**0.9239**	0.4548	1507	**0.9642**	0.3500
LEX-2+LEX-LALG	**1429**	0.8926	**0.5102**	**2261**	0.9217	**0.4505**

Table 2: A comparison of different lexicons. For lexicons employing our LALG-ALL algorithm, only translation candidates that scored above the threshold $P = 0.11$ have been kept.

it with $-phy$. Similar rules have been introduced for Dutch-English: the suffix $-tie$ is replaced by $-tion$, $-sie$ by $-sion$, and $-teit$ by $-ty$.

Finally, we have compared the results of the following constructed lexicons:

- A lexicon containing only words shared across languages (**LEX-1**).
- A lexicon containing shared words and translation pairs found by applying the language-specific transformation rules (**LEX-2**).
- A lexicon containing only translation pairs obtained by the LALG-ALL algorithm that score above a certain threshold P (**LEX-LALG**).
- A combination of the lexicons LEX-1 and LEX-LALG (**LEX-1+LEX-LALG**). Non-matching duplicates are resolved by taking the translation pair from LEX-LALG as the correct one. Note that this lexicon is completely language-pair independent.
- A lexicon combining only translation pairs found by applying the language-specific transformation rules and LEX-LALG (**LEX-R+LEX-LALG**).
- A combination of the lexicons LEX-2 and LEX-LALG, where non-matching duplicates are resolved by taking the translation pair from LEX-LALG if it is present in LEX-1, and from LEX-2 otherwise (**LEX-2+LEX-LALG**).

According to the results from Table 2, we can conclude that adding translation pairs extracted by our LALG-ALL algorithm has a major positive impact on both precision and coverage. Obtaining results for two different language pairs proves that the approach is generic and applicable to any other language pairs. The previous approach relying on work from Koehn and Knight (2002) has been outperformed in terms of precision and coverage. Additionally, we have shown that adding simple translation rules for languages sharing same roots might lead to even better scores (LEX-2+LEX-LALG). However, it is not always possible to rely on such knowledge, and the usefulness of the designed LALG-ALL algorithm really comes to the fore when the algorithm is applied on distant language pairs which do not share many words and cognates, and word translation rules cannot be easily established. In such cases, without any prior knowledge about the languages involved in a translation process, one is left with the linguistically unbiased LEX-1+LEX-LALG lexicon, which also displays a promising performance.

5 Conclusions and Future Work

We have designed an algorithm that focuses on acquiring and keeping only highly confident translation candidates from multilingual comparable corpora. By employing the algorithm we have improved precision scores of the methods relying on per-topic word distributions from a cross-language topic model. We have shown that the algorithm is able to produce a highly reliable bilingual seed lexicon even when all other lexical clues are absent, thus making our algorithm suitable even for unrelated language pairs. In future work, we plan to further improve the algorithm and use it as a source of translational evidence for different alignment tasks in the setting of non-parallel corpora.

Acknowledgments

The research has been carried out in the framework of the TermWise Knowledge Platform (IOF-KP/09/001) funded by the Industrial Research Fund K.U. Leuven, Belgium.

References

Jaime G. Carbonell, Jaime G. Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, Danny Lee, Yiming Frederking, Robert E, Ralf D. Geng, and Yiming Yang. 1997. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 708–714.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–5.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1513–1518.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining*, pages 57–64.

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

Mona T. Diab and Steve Finch. 2000. A statistical translation model using comparable corpora. In *Proceedings of the 6th Triennial Conference on Recherche d'Information Assistée par Ordinateur (RIAO)*, pages 1500–1508.

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

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 414–420.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 526–533.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 771–779.

Zellig S. Harris. 1954. Distributional structure. *Word 10*, (23):146–162.

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

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 617–625.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547.

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

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26:221–249.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889.

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

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

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference*, pages 1155–1156.

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

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual*

Meeting of the Association for Computational Linguistics, pages 519–526.

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

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth.

Thuy Vu, Ai Ti Aw, and Min Zhang. 2009. Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 843–851.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479–484.