Sub-corpora Sampling with an Application to Bilingual Lexicon Extraction

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

ABSTRACT
We propose a novel associative approach for bilingual word lexicon extraction (BLE) from parallel corpora that relies on the paradigm of data reduction instead of data augmentation. The key insight of the approach is the effective usage of sub-corpora sampling and properties of low-frequency words in the task of lexicon induction, particularly in a setting where only limited parallel data are available. Word translation pairs are extracted from many smaller sub-corpora (sampled from the original corpus) according to several frequency-based criteria of similarity. We prove the validity of our data sampling approach, and show that this method outperforms IBM Model 1 and associative methods based on similarity scores and hypothesis testing in terms of precision and F-measure in the task of lexicon extraction. Additionally, we show that our sampling-based method can learn correct word translations from fewer data.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE (CROATIAN)

Uzorkovanje Potkorpusa uz Primjenu u Ekstrakciji Dvojezičnih Rječnika

U radu se predlaže nov asocijativan pristup ekstrakciji dvojezičnih rječnika iz usporednih korpusa koji se oslanja na paradigmu smanjivanja količine podataka umjesto njezinog povećavanja. Ključna je ideja pristupa učinkovita uporaba uzorkovanja potkorpusa te svojstava niskofrekventnih riječi u zadatku indukcije rječnika, posebice u situacijama kada je na raspolaganju ograničen skup usporednih podataka. Prijevodni parovi riječi ekstrahirani su iz većeg broja manjih potkorpusa (uzorkovanih iz izvornog korpusa) temeljem nekoliko frekvencijski utemeljenih kriterija sličnosti. U radu je pokazana ispravnost našeg pristupa temeljenog na uzorkovanju potkorpusa. Pokazano je da ovaj postupak u smislu F-mjere na zadatku ekstrakcije leksikona nadmašuje IBM-ov Model 1 te asocijativne postupke temeljene na ocjenama sličnosti i testiranju hipoteze. Također je pokazano da naš postupak temeljen na uzorkovanju može naučiti ispravne prijevode riječi iz manjih količina podataka.

KEYWORDS: bilingual lexicon extraction, empirical word translation, sub-corpora sampling, data reduction, low-frequency words.

KEYWORDS IN CROATIAN: ekstrakcija dvojezičnih rječnika, empirijsko prevođenje riječi, uzorkovanje potkorpusa, smanjivanje količine podataka, niskofrekventne riječi.

Proceedings of COLING 2012: Technical Papers, pages 2721–2738,
COLING 2012, Mumbai, December 2012.

1 Introduction

Bilingual word lexicons serve as an invaluable and indispensable source of knowledge for both end users (as an aid for translators or other language specialists) and many natural language processing tasks, such as dictionary-based cross-language information retrieval (Carbonell et al., 1997; Levow et al., 2005) and statistical machine translation (Och and Ney, 2003).

In order to construct high quality bilingual lexicons for various domains, it is necessary to build such lexicons manually by hand or extract them automatically from parallel corpora. Compiling such lexicons manually is often a labor-intensive and time-consuming task, whereas parallel corpora either do not exist or are of limited size for most language pairs. Therefore the focus of the researchers has turned towards bilingual lexicon extraction (BLE) from comparable corpora (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Diab and Finch, 2000; Fung and Cheung, 2004; Morin et al., 2007; Haghighi et al., 2008; Laroche and Langlais, 2010; Andrade et al., 2010; Shezaf and Rappoport, 2010; Vulić et al., 2011; Prochasson and Fung, 2011; Vulić and Moens, 2012; Tamura et al., 2012). However, such lexicons contain a great deal of noise and, moreover, the methods for BLE from comparable corpora typically rely on seed lexicons which are again hand-built or extracted from parallel corpora.

With respect to that observation, numerous systems for various applications trained on parallel or comparable data almost exclusively rely on knowledge from bilingual lexicons extracted from parallel texts. These lexicons are usually acquired from word translation probabilities of the IBM alignment models (Brown et al., 1993; Och and Ney, 2003) or obtained by associative methods such as the log-likelihood score or the Dice coefficient. They are then used in systems for extracting parallel sentences from non-parallel corpora (Fung and Cheung, 2004; Munteanu and Marcu, 2005), bilingual sentence alignment (Moore, 2002), estimating phrase translation probabilities (Venugopal et al., 2003), extracting parallel sub-sentential fragments from non-parallel corpora (Munteanu and Marcu, 2006), word-level confidence estimation (Ueffing and Ney, 2007), sub-sentential alignment for terminology extraction (Lefever et al., 2009), cross-lingual text classification and plagiarism detection (Pinto et al., 2009) and others.

High accuracy of automatically constructed bilingual word lexicons is the top priority for these systems. Church and Mercer (1993) advocate a simple solution of collecting more data in order to utilize statistical and stochastic methods in a more effective way. However, these systems are typically faced with only limited parallel data for many language pairs and domains (Resnik and Smith, 2003).

In order to tackle these issues, we propose a novel approach built upon the idea of *data reduction* instead of *data augmentation*. The method is directed towards extraction of only *highly reliable translation pairs* from *parallel data of limited size*. It is based on the idea of *sub-corpora sampling* from the original corpus. For instance, given an initial corpus $\mathscr{C}$ of 4 data items $\{I_1, I_2, I_3, I_4\}$, the construction of, say, a sub-corpus $SC = \{I_2, I_4\}$ may be observed as: (1) sampling items $I_2, I_4 \in \mathscr{C}$ for SC (hence the term sub-corpora sampling) or (2) removing data items I_1, I_3 from the original corpus $\mathscr{C}$, so that $SC = \mathscr{C} - \{I_1, I_3\}$ (hence the term data reduction). By reducing the size of the initial corpus, we typically decrease frequencies of the words in a newly formed sub-corpus. This simplifies the establishment of potential translation candidates, since that is now reduced to a problem of establishing reliable translational equivalence between low-frequency words. We explain the method for establishing translational equivalence based on the absolute frequency distributions of words in a sub-corpus. We exploit it in the construction of the algorithm for BLE. Moreover, each word exhibits a different distribution over items in each

newly built sub-corpus, and it is different from the fixed distribution in the original corpus. It allows us to identify different potential translation candidates in different sub-corpora and then form word translation tables by combining these evidences acquired from different sub-corpora. The key strength of the proposed algorithm is that it takes the entire initial corpus into account, regardless of its size, and at the same time it also benefits from the sampling of a vast number of different subsets/sub-corpora sampled from that initial corpus, and the evidences of potential word translation pairs coming from these sub-corpora.

In the remainder of the paper, we show that: (1) Bilingual lexicon extraction benefits from the concept of data reduction and sub-corpora sampling - the key intuitions, assumptions and the construction of the algorithm are provided in Section 2; (2) The proposed algorithm for BLE removes a lot of noise from the bilingual word lexicons by harvesting only the most accurate translation candidates, and it outscores other standard models for BLE from parallel data; (3) Due to the concept of data reduction, the proposed algorithm does not suffer from a problem of indirect associations; (4) Most importantly, the proposed algorithm outperforms other models for BLE when dealing with parallel data of limited size. The results are presented in Section 4. Finally, Section 5 lists conclusions and possible paths of future work.

2 Learning Translation Pairs Using Sub-Corpora Sampling

Section 1 has already provided a general intuition behind our method for mining translational candidates from aligned corpora. Now, we provide an in-depth description and analysis of our algorithm for bilingual word lexicon extraction. First, we explain the key reasoning that led us to our approach that relies on *data sampling*. Second, we provide the criteria for extracting translation candidates that purely rely on their distributional features, but do not employ any similarity-based measure or hypothesis testing for word association, and finally, we present our algorithm for BLE that processes words of all frequencies in an uniform way.

2.1 Why Sampling Sub-corpora?

The foundation of this work is built upon the so-called *Zipfian phenomenon* which states that, regardless of the size of a corpus, most of the distinct words occur only a small number of times. For instance, Moore (2004b) measures that in the first 500,000 English sentences taken from the Canadian Hansards data (Mihalcea and Pedersen, 2003), one finds 52,921 distinct word types, of which 60.5% occur five or fewer times, and, moreover, 32.8% occur only once. A general solution to mitigate the problem of low-frequency words is by augmenting the amount of input training data. However, that approach leads to a *chicken and egg problem* - adding more data will increase frequencies of the words already present in the corpus, and, accordingly, solve the issue of the low-frequency words, but at the same time, it will introduce many extra words, where some of them were previously out-of-vocabulary. Most of these new words will now be low-frequency words - again we observe the very same Zipfian phenomenon, and the problem of low-frequency words is still present. Therefore, we have decided to take an opposite path, where "removing" data from the initial corpus (that actually means sampling a sub-corpus with less data items from the original large corpus) and properties of low-frequency words (Moore, 2004b; Prochasson and Fung, 2011) should actually help us detect correct cross-lingual word associations. By reducing the corpus size, we also decrease frequencies of the words in the corpus. In an extreme case, when the reduced corpus consists of only one sentence, almost all words in that "corpus" will occur only once or twice. Intuitively, for words with higher frequencies, one needs to remove more data, i.e., to sample a sub-corpus of smaller size, to

bring the words down to only a few occurrences in the sub-corpus. We will show that it is easier to establish translational equivalence for low-frequency words.

2.2 Criteria for Extraction of Translation Pairs

Given is a source language S, a target language T, and a corpus $\mathscr{C}$ of N aligned item pairs $\mathscr{C} = \{(I_1^S, I_1^T), (I_2^S, I_2^T), \ldots, (I_N^S, I_N^T)\}$, where, depending on the corpus type, item pairs may be sentences, paragraphs, chunks, documents, etc. For parallel corpora, the item pairs are pairs of sentences. The goal is to extract potential translation candidates from the item-aligned set using only *internal distributional evidences*. Internal evidences, according to Kay and Röscheisen (1993), represent information derived only from the given corpora themselves. Our criteria for establishing translational equivalence between words are derived from this trivial case:

Imagine the scenario where a source word w_1^S occurs only once on the source side of the corpus $\mathscr{C}$, in a source item I_j^S. There is a target word w_2^T occurring in a target item I_j^T (which is aligned to I_j^S) and the word w_2^T also occurs only once on the target side of the corpus $\mathscr{C}$. Additionally, there does not exist another source word w_a^S such that it occurs only once on the source side of the corpus and, at the same time, exactly in the item I_j^S, and there does not exist another target word w_b^T that occurs only once on the target side of the corpus and exactly in the item I_j^T. Our key assumption is that the words w_1^S and w_2^T should then be listed as translation candidates. We can further generalize the intuition, that is, two words are extracted as translation candidates if they both satisfy the entire set of features $\mathscr{F}$, and there are no other words that satisfy this set of features.[1]

The set $\mathscr{F}$ may include various clues as features, but in our work we opt only for the internal, language-independent features that are related to the distributions of words over corpora. A source word w_1^S and a target word w_2^T are listed as potential translation candidates if they fulfil the following criteria:

1. The overall frequency of w_1^S on the source side of the corpus is equal to the overall frequency of w_2^T on the target side of the corpus.
2. The overall frequency of both words is above some minimum frequency threshold M_f.
3. w_1^S and w_2^T occur only in aligned item pairs, and with exactly the same frequency.
4. The number of aligned item pairs in which the words occur is above some minimum M_i.
5. There is no source word w_a^S such that the pair (w_a^S, w_2^T) satisfies all the previous conditions, and there is no target word w_b^T such that the pair (w_1^S, w_b^T) satisfies all the previous conditions.[1]

For instance, if the French word *pluie* occurs 4 times in the whole corpus, 2 times in item I_j^S, 1 time in item I_k^S, and 1 time in item I_l^S, and there is the English word *rain* that also occurs 4 times in total, 2 times in item I_j^T, 1 time in item I_k^T, and 1 time in item I_l^T, and there are no other words with the same frequency distribution in the corpus, we claim *(pluie, rain)* to be a pair of translation candidates.

In our work, we have opted for the listed criteria/constraints, but we are free to adjust or add more criteria if we want to boost a certain behavior of the model, that is, if we want to focus

[1] This specifies one-to-one alignment constraint, but more relaxed criteria are also possible. For instance, we could allow 2 or more target words to have the same features as a source word and then distribute partial link counts over all target candidates.

more on accuracy or on coverage of the lexicon. By imposing, for instance, stricter thresholds for M_f or M_i (e.g., accepting only candidates that occur in at least two items), we can direct the algorithm for lexicon extraction towards higher accuracy, and, vice versa, by relaxing the thresholds, we boost the coverage of the lexicon.

In summary, the proposed criteria for extraction of translation candidates are not biased towards high-frequency or low-frequency words, as they treat all words the same, trying to find potential candidates according to the defined set of features. However, in practice, the majority of the matched candidates will be low-frequency words.

2.3 The Algorithm for Lexicon Extraction

By employing the aforementioned criteria for extraction of translation candidates on the initial corpus $\mathscr{C}$, we are able to extract only a limited number of translation pairs, since distributional evidences for the large corpus $\mathscr{C}$ are fixed and unchangeable. But by sampling data from $\mathscr{C}$, we actually build a new corpus, a sub-corpus $SC \subset \mathscr{C}$ of size $K < N$, which now has a changed set of distributional evidences, which may lead to extracting additional translation candidates. The process of data reduction may be observed as a process of *sampling*, i.e., we randomly pick a subset of item pairs from $\mathscr{C}$, and build a new sub-corpus SC. We can then repeat the process, sample another sub-corpus and try to detect more translation candidate pairs.

Having the large corpus $\mathscr{C}$ of a finite size N, the number of different sub-corpora is huge, but finite. The exact number of different sub-corpora that can be sampled from $\mathscr{C}$ is $M_{\mathscr{C}} = \sum_{K=1}^{N} \binom{N}{K}$. Since we are clearly unable to process all the possible sub-corpora, we need to design a smart strategy to: (1) cover the entire initial corpus and (2) detect translation pairs for both high-frequency and low-frequency words.

2.3.1 One Sampling Round with Fixed Sub-Corpora Size

Let us fix the size of sub-corpora to some value K. We want to assure that every item pair from $\mathscr{C}$ is taken into account in at least one sub-corpus of size K. Additionally, we want to be able to repeat the procedure and obtain more different sub-corpora of the same size. The procedure is as follows:

1. Initialize: Detect the number of sub-corpora for this round: $\lfloor \frac{N}{K} \rfloor$.
2. Randomly shuffle the item pairs in $\mathscr{C}$ to obtain a permutation of the item pairs in $\mathscr{C}$.
3. Split $\mathscr{C}$ into sub-corpora of equal size K as follows:
   - For $i = 1, \ldots, \lfloor \frac{N}{K} \rfloor - 1$, assign the item pairs from position $(i-1) \cdot K + 1$ until position $i \cdot K$ to the sub-corpus SC_i.
   - Assign the remaining item pairs from position $(\lfloor \frac{N}{K} \rfloor - 1) \cdot K + 1$ until the end (position N) to the sub-corpus $SC_{\lfloor \frac{N}{K} \rfloor}$.

We build a set of $\lfloor \frac{N}{K} \rfloor - 1$ sub-corpora of size K and one sub-corpus of size $K + N \bmod K$, while, at the same time, we ensure that the complete original corpus $\mathscr{C}$ is covered. We will call the described procedure the *sampling round*. If we want to repeat the procedure and acquire another set of sub-corpora of the same size, we simply go back to Step 2 of the procedure and perform another sampling round.

2.3.2 The Final Algorithm: SampLEX

Now, we have everything set for the construction of the algorithm. In order to capture words with different frequencies, we need to vary the sub-corpora size K. With respect to the Zipf's law (Prochasson and Fung, 2011), we have decided to vary the values of K from N down to 1, where K is divided by 2 in each step of the loop (see the final algorithm). In that way, we ensure that all the words occur as low-frequency words in at least some sub-corpora of various sizes. Again, if we want to reduce frequencies of high-frequency words, we need samples of smaller sizes, so such words will typically learn its translation candidates from sampled sub-corpora consisting of only a few sentences. One pass of the algorithm from the values N to 1 is called an *iteration*.

We can detect potential translation candidates in many different sub-corpora (of various sizes). Additionally, we should assign more weight to translation pairs that fulfil the strict criteria in sub-corpora of larger size K. For instance, if we detect that two words have identical frequency distributions and have fulfilled all the criteria from Subsection 2.2 in a sub-corpus consisting of a few millions items, that evidence should be more important than detecting that the two words could be extracted from a sub-corpus comprising only a few sentences. Thus, for each potential translation pair t_{ij} we assign a corresponding overall score $c_{t_{ij}}$. If we detect that the two words that form the translation pair t_{ij} could be extracted from a sub-corpus of size K, we update the score $c_{t_{ij}} := c_{t_{ij}} + 1 \cdot weight_K$, where $weight_K = \lfloor \frac{N}{K} \rfloor$. This way we assign more importance when the pairs are extracted from larger sub-corpora. For instance, if we detect that two words from the potential translation pair t_{ij} are extracted as translation candidates from the original corpus $\mathscr{C}$, then $K = N$ and $c_{t_{ij}} := c_{t_{ij}} + 1$.

The final algorithm is as follows:

1. **Input:** The initial large corpus $\mathscr{C}$ of size N.
2. **Initialize:** (1) Define the criteria for extraction of translation candidates; (2) Initialize an empty lexicon L. Each entry in the lexicon L will have the following form: $(t_{ij}, c_{t_{ij}})$, where t_{ij} denotes the extracted translation pair consisting of a source word w_i^S and a target word w_j^T, while $c_{t_{ij}}$ is a variable that denotes the score for the translation pair t_{ij}.
3. **Set** initial sub-corpora size: $K := N$.
4. **Perform** one *sampling round* with the current sub-corpora size set to K (Subsection 2.3.1). We obtain $\lfloor \frac{N}{K} \rfloor$ different sub-corpora: $SC_1, \ldots, SC_{\lfloor \frac{N}{K} \rfloor}$, all of size K except the last one (its size is always $K + N \bmod K$).
5. **Extract** translation pairs from all sub-corpora obtained in Step 4. If a translation pair t_{ij} is already present in the lexicon L, update the score $c_{t_{ij}} := c_{t_{ij}} + 1 \cdot weight_K$ for that pair. Otherwise, add the translation t_{ij} to L and set its current score $c_{t_{ij}} := 1 \cdot weight_K$.
6. **Set** new sub-corpora size: $K := \lfloor K/2 \rfloor$.
7. **If** $K > 0$, **go** to Step 4. Otherwise, we have reached the end of one *iteration* and we need to **check** the stopping criteria (**go** to Step 8).
8. **Check** the stopping criteria: **if** no new translation pairs were extracted after the end of one whole iteration **or** we have reached the maximum or the predefined number of iterations **or** timeout, **go** to Step 9. Otherwise, **go** to Step 3.
9. **Output:** The lexicon L.

We will call this procedure the **SampLEX** algorithm. The proposed algorithm exhibits only one possible strategy for mining translation pairs from sub-corpora. For instance, we could

opt for another strategy when deciding how to change the size of sub-corpora, skip already processed sub-corpora, remodel the criteria for extraction from Section 2.2, change stopping criteria, or employ a procedure for the sub-corpora sampling different from the one presented in Subsection 2.3.1. However, our main goal is to propose a general framework for lexicon extraction when the data sampling approach is employed, where other researchers could design their own algorithms built upon the same idea.

2.3.3 Properties of the Algorithm

Reducing corpora size provides several benefits. First, establishing associations between translation candidates is much easier when we deal with low-frequency words - we reduce our problem to a *binary decision problem*. According to the specified criteria for extraction, two words are simply considered to be a translation pair, or they are not. By employing the criteria that rely on raw frequency counts as distributional evidences, we remove the need of an association measure based on hypothesis testing such as the G^2 statistic (Dunning, 1993; Agresti, 2002) or a similarity-based measure such as the Dice coefficient (Dice, 1945), which are often unreliable when dealing with low-frequency words (Manning and Schütze, 1999).

The *SampLEX* algorithm is *symmetric* and *non-directional*. The final output of the algorithm provides translation pairs along with their counts obtained after training. We can easily transform them into word translation probabilities to build word translation tables similar to those of IBM Model 1. Since the algorithm is symmetric, we can obtain both source-to-target and target-to-source word translation probabilities after the algorithm run is completed:

$$P(w_2^T|w_1^S) = \frac{c_{t_{12}}}{\sum_j c_{t_{1j}}} \qquad P(w_1^S|w_2^T) = \frac{c_{t_{12}}}{\sum_j c_{t_{j2}}} \tag{1}$$

Surprisingly, another modeling advantage lies in *randomness* when selecting sub-corpora. Namely, if we detect that two words constantly co-occur in aligned items randomly sampled from the large corpus, regardless of the surrounding context, it actually strengthens the confidence that those two words really constitute a translation pair. During the sampling procedure, sentences are moved from their "natural" surrounding of other sentences (the context in this case) and the new sub-corpus is built by randomly taking sentences from the entire corpus. If the same translational equivalence between the two words is encountered in more different sub-corpora, it also raises the *significance* of that equivalence. Also, by building sub-corpora of smaller sizes from the original large corpus, we perform an implicit *disambiguation* - a word occurring only once or twice in a small sub-corpus cannot bear more meanings in that sub-corpus, although it might have more meanings in the large superset corpus.

3 Experimental Setup

In this section, we present datasets used for training, training setup of the SampLEX method and state-of-the-art models for bilingual lexicon extraction from parallel data often used in real-life applications.

3.1 Training

3.1.1 Training Collections

We work with Europarl data (Koehn, 2005) for Dutch-English and Italian-English language pairs, retrieved from the website[2] of the OPUS project (Tiedemann, 2009). We use subsets of the corpora, comprising the first 300,000 sentence pairs. For Dutch-English, there are 76,762 unique Dutch words, and 37,138 unique English words. For Italian-English, there are 68,710 unique Italian words and 37,391 unique English words. The unbalance between the number of unique vocabulary words is mostly due to a richer morphological system in Italian and the noun compounding phenomenon in Dutch.

Since we also want to test and evaluate the behavior of our system in a setting where only limited parallel data are present, we construct additional subsets of Europarl data consisting of the first 2,000, 10,000 and 50,000 sentence pairs from the corpora.

3.1.2 Training Setup of the SampLEX Method

Parameter values are set to the same values for all training datasets. We set $M_f = M_i = 0$, which means that all words that occur in a sub-corpus at least once may be extracted. By setting some higher thresholds M_f and M_i, we could move the algorithm towards extracting lexicons of higher accuracy, but lower coverage. We stop our training procedure for *SampLEX* after 1000 iterations for all corpora. The SampLEX algorithm converges quickly - many translations are found in the first few iterations. However, having more iterations implies obtaining more different evidences from different sub-corpora and assigning more significance for the extracted candidates (see Subsection 2.3.3). Therefore, we have decided to use 1000 iterations for safety. Other stopping criteria are also possible (see Step 8 in Subsection 2.3.2).

3.2 State-of-the-Art Models for BLE

In order to evaluate the performance of our *SampLEX* algorithm for bilingual lexicon extraction, we compare it with other models that constitute state-of-the-art for BLE, and are often used in real-life applications (see Section 1).

3.2.1 IBM Model 1

Our first baseline is IBM Model 1 (Brown et al., 1993) for word alignment, which is a purely lexical model, i.e, the only set of parameters employed by the model are word translation probabilities. We omit the exact generative story for IBM Model 1, but the curious reader may find all the details in (Brown et al., 1993) or (Och and Ney, 2003). Word translation probability $P(w_2^T|w_1^S)$ denotes a probability that a source word w_1^S generates a target word w_2^T. These probabilities can then be used to decide upon translational equivalence between words and to build bilingual lexicons from parallel texts.[3] That makes it comparable to our *SampLEX* model, which can also output word translation probabilities (Equation 1). IBM Model 1 is used in many systems as a primary tool for bilingual lexicon extraction from parallel data (e.g., Venugopal et al. (2003), Munteanu and Marcu (2005), Munteanu and Marcu (2006), Lefever et al. (2009)). We use standard GIZA++ settings and train IBM Model 1 with 5 iterations

[2] http://opus.lingfil.uu.se/Europarl3.php

[3] We have also tried to use word translation probabilities from the higher order IBM Models, but we have not detected any major difference in results on the task of bilingual word lexicon extraction.

(**IBM1-i5**) and 20 iterations (**IBM1-i20**) of the EM algorithm, as often found in the literature (Och and Ney, 2003; Moore, 2004a).

3.2.2 The Dice Coefficient

Another baseline model is a similarity-based model relying on the Dice coefficient (**DICE**):

$$DICE(w_1^S, w_2^T) = \frac{2 \cdot C(w_1^S, w_2^T)}{C(w_1^S) + C(w_2^T)} \tag{2}$$

where $C(w_1^S, w_2^T)$ denotes the co-occurrence count of words w_1^S and w_2^T in the aligned items from the corpus. $C(w_1^S)$ and $C(w_2^T)$ denote the count of w_1^S on the source side of the corpus, and the count of w_2^T on the target side of the corpus, respectively. The Dice coefficient was used as an associative method for word alignment by Och and Ney (2003), Tiedemann (2003) used it as one associative clue for his clue-based word alignment, and Melamed (2000) used it to measure the strength of translational equivalence.

3.2.3 Log-Likelihood Ratio

Another associative model that we use is based on the log-likelihood-ratio (**LLR**), that is derived from the G^2 statistic (Dunning, 1993). LLR is a more appropriate hypothesis testing method for detecting word associations from limited data than the χ^2 test (Manning and Schütze, 1999) and was previously used as an effective tool for automatically constructing bilingual lexicons (Melamed, 2000; Moore, 2001; Munteanu and Marcu, 2006). Its definition is easily explained on the basis of a contigency table (Kilgarriff, 2001; Padó and Lapata, 2007), which is a four-cell matrix for each pair of words (w_1^S, w_2^T) (see Table 1).

	w_1^S	$\neg w_1^S$
w_2^T	k	l
$\neg w_2^T$	m	n

Table 1: The contigency table for a pair of words (w_1^S, w_2^T).

The contingency table records that source word w_1^S and target word w_2^T co-occur in k aligned item/sentences pairs, and w_1^S occurs in m aligned pairs in which w_2^T is not present. Similarly, w_2^T occurs in l aligned pairs in which w_1^S is not present, and n is the number of aligned pairs that involve neither w_1^S nor w_2^T. The final formula for the log-likelihood ratio is then defined as:

$$\begin{aligned} LLR(w_1^S, w_2^T) = G^2(k,l,m,n) = 2(&k\log k + l\log l + m\log m + n\log n \\ &-(k+l)\log(k+l) - (k+m)\log(k+m) \\ &-(l+n)\log(l+n) - (m+n)\log(m+n) \\ &+(k+l+m+n)\log(k+l+m+n)) \end{aligned} \tag{3}$$

High LLR scores can indicate either a positive association or a negative one (Moore, 2004b). Since we expect translation pairs to be positively associated, we impose an additional constraint: $P(w_1^S, w_2^T) > P(w_1^S) \cdot P(w_2^T)$, where $P(w_1^S, w_2^T) = \frac{k}{k+l+m+n}$, $P(w_1^S) = \frac{k+m}{k+l+m+n}$ and $P(w_2^T) = \frac{k+l}{k+l+m+n}$. This constraint keeps only positively associated words in the lists of potential translation candidates.

3.3 Evaluation Methodology

3.3.1 Lists of Ground Truth Translation Pairs

In order to evaluate the BLE models, we have designed a set of ground truth translations - we have randomly sampled a set of Dutch content words that occur in the full corpus comprising 300, 000 sentences. Following that, we have used the *Google Translate* tool plus an additional annotator to translate those words to English. The annotator has manually revised the lists and has kept only words that have their corresponding translation in the English vocabulary. In order to build a one-to-one ground truth dataset of translations, only one possible translation has been annotated as correct. In case when more than 1 translation is possible, the annotator has marked as correct the translation that occurs more frequently in the English Europarl data. Finally, we have come up with a set of 1001 ground truth one-to-one translation pairs. We have followed the same procedure for Italian-English and have also constructed a set of 1001 ground truth translation pairs.[4]

3.3.2 Evaluation Metrics

All the methods under consideration actually retrieve ranked lists of translation candidates. Let us keep only the first translation candidate from each ranked list, and build a non-probabilistic lexicon of one-to-one word translations: L_e. Assuming that we now have a set G of ground truth one-to-one word translation pairs, we can evaluate the quality of our lexicon with respect to the ground truth set G. We use standard precision, recall and F-Measure ($\beta = 1$) scores as our evaluation metrics:

$$Prec_{L_e,G} = \frac{|L_e \cap G|}{|L_e|} \qquad Rec_{L_e,G} = \frac{|L_e \cap G|}{|G|} \qquad F_{L_e,G} = (1+\beta^2)\frac{Prec_{L_e,G} \cdot Rec_{L_e,G}}{\beta^2 \cdot Prec_{L_e,G} + Rec_{L_e,G}}$$

Since sometimes a word has more than one correct translation (e.g., Dutch word *verklaring* can be translated as both *statement* and *declaration*), and the current evaluation setting cannot capture that phenomenon, we also evaluate the quality of the lexicon in a more lenient setting, where, instead of performing the hard cut-off, i.e., instead of keeping only the top candidate from the ranked list, we keep the ranked list of all the candidates from the list and calculate the *mean reciprocal rank* (MRR) (Voorhees, 1999). For a source word w_i^S, $rank(w_i^S)$ denotes the rank of its correct translation (as provided by the set of ground truth translation pairs) within the retrieved list of potential translation candidates. MRR of the lexicon is then defined by the following formula:

$$MRR_{L_e,G} = \frac{1}{|L_e|} \sum_{w_i^S \in L_e} \frac{1}{rank(w_i^S)} \tag{4}$$

4 Results and Discussion

We conduct several experiments to measure the quality of the lexicon constructed using the *SampLEX* algorithm: (1) we evaluate the lexicon obtained by SampLEX using the full corpus of 300, 000 sentences, and compare its accuracy with the accuracy of baseline systems from Section 3.2 trained on the same corpus, (2) after performing the error analysis, we carry out another set of experiments that prove that the *SampLEX* algorithm, due to its modeling

[4] We will make the datasets publicly available.

properties, alleviates the problem of *indirect associations* and, finally, (3) we test our lexicon in a setting where only limited parallel data are available and show that the *SampLEX*-based lexicon outperforms other bilingual word lexicons in that setting in terms of quality provided by the F-measure and precision scores.

4.1 Experiment I: Testing the Quality of the Lexicon in Terms of Precision

Unlike our baseline state-of-the-art systems for BLE, the *SampLEX* algorithm does not assure the full coverage of the source vocabulary, as it does not necessarily build ranked lists of translation candidates for all the words observed during training. However, our claim is that translation pairs obtained by *SampLEX* are of higher quality than those obtained by the baseline systems. Therefore, with this experiment we want to answer the following question: "Are translation pairs obtained by the SampLEX algorithm really more accurate than translation pairs obtained by other methods?". In order to answer that question, we calculate the precision and MRR scores on our ground truth datasets for Italian-English and Dutch-English, where all the BLE methods have been trained on the full 300,000 datasets. The obtained scores are presented in Tables 2 and 3.

	Dutch-English				
	IBM1-i5	**IBM1-i20**	**DICE**	**LLR**	**SampLEX**
Prec(300k)	0.7113	0.7023	0.6963	0.7662	**0.8221**
MRR(300k)	0.8196	0.8045	0.7767	0.8542	**0.9069**

Table 2: Precision and MRR scores for all models trained on the first 300,000 sentences of Dutch-English Europarl data, and evaluated on the sets of 1001 ground truth translation pairs for Dutch-English.

	Italian-English				
	IBM1-i5	**IBM1-i20**	**DICE**	**LLR**	**SampLEX**
Prec(300k)	0.7912	0.7752	0.7932	0.8361	**0.8771**
MRR(300k)	0.8781	0.8588	0.8494	0.8945	**0.9250**

Table 3: Precision and MRR scores for all models trained on the first 300,000 sentences of Italian-English Europarl data, and evaluated on the sets of 1001 ground truth translation pairs for Italian-English.

As previously shown by Moore (2004a), LLR serves as a better associative method than the Dice coefficient for the word alignment task. We obtain the same finding for bilingual lexicon extraction. Additionally, the model based on LLR is also better than IBM Model 1 when applied for BLE. Munteanu and Marcu (2006) drew the same conclusion, and they used the LLR-based lexicon in their system when a higher precision of the lexicon was paramount. However, the results reveal that the quality of the lexicon obtained by the *SampLEX* algorithm is superior to the LLR-lexicon *in terms of precision* and, consequently, to all other evaluated lexicons.

4.2 Experiment II: Investigating Indirect Associations

When examining the results, we have detected that one advantage of our SampLEX algorithm is due to its mitigating the phenomenon of the so-called indirect associations. Indirect associations,

as defined by Melamed (2000), are associations between words that have a tendency to co-occur much more often than expected by chance, but are not mutual translations. Lexicon extraction models unaware of the indirect associations tend to give translational preference to higher-frequency words. Considering the fact that one key assumption of our model is sub-corpora sampling that causes decreasing frequencies of words in the obtained sub-corpora from which translation pairs are learned, our model should successfully mitigate the problem of indirect associations. Indeed, during the error analysis, we have detected that both IBM Model 1 and LLR provide a wrong translation of the Dutch word *beschouwen (consider)*, since both models retrieve the English word *as* as the first translation candidate (due to a very high frequency of the collocation *consider as*). Other examples of the same type include the Dutch word *integreren (integrate)* which is translated as *into*, *betwijfelen (doubt)* which is translated as *whether*, or an Italian example of the verb *entrare (enter)* which is translated as *into*. Our BLE model, on the other hand, provides correct translations for all these examples. Dagan et al. (1993) noted that collocates often tend to cause confusion among algorithms for bilingual lexicon extraction. More examples include the Dutch word *opinie (opinion)*, translated as *public* by IBM Model 1 and LLR (due to a high frequency of the collocation *public opinion*), the Dutch word *cirkels (circles)*, translated as *concentric*, or the Italian word *pensionabile (pensionable)*, translated as *age*. All these examples are again correctly translated by our model for lexicon extraction.

In order to test the hypothesis that our lexicon extraction model does not suffer from the problem of learning indirect associations, we have conducted a small experiment. For the purpose of the evaluation, we have constructed a small dataset of 219 Italian verbs in first person plural of the present tense. We have also constructed the set of ground truth translations in the same way as in Subsection 3.3.1. These verbs are easy to extract because they all have the same suffix *-iamo* (e.g., the verb *respiriamo*, meaning *(we) breathe*). If the problem of indirect associations for a lexicon extraction method is prominent, the English word *we* will appear as the first translation for many of these verbs, instead of the word that really bears the content of the verb (e.g., *breathe*). Table 4 shows precision and MRR scores for the lexicon extraction models evaluated on this toy dataset. As expected, due to its modeling property related to the

	IBM1-i5	**DICE**	**LLR**	**SampLEX**
Prec(300k)	0.4475	0.4201	0.6119	**0.8584**
MRR(300k)	0.5575	0.5108	0.7300	**0.9140**

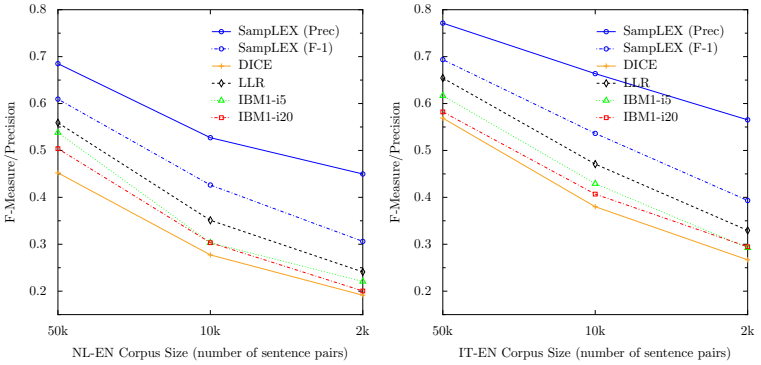
Table 4: Precision and MRR scores on our evaluation set consisting of Italian *-iamo* verbs (present tense, first person plural).

reduction of word frequencies, our BLE model does not suffer from the problem of indirect associations like other models. That property eventually has a positive impact on precision and MRR scores and the overall quality of the lexicons obtained by our SampLEX algorithm.

4.3 Experiment III: Experiments with a Limited Amount of Parallel Data

In a real-life situation, one often possesses only limited parallel data (e.g., terminology texts from special, very narrow domains and sub-domains). With this final set of experiments we test the performance of all the models for lexicon extraction in such a setting with limited parallel data. To simulate the shortage of data, we have extracted three additional corpora of smaller sizes by selecting the first 2,000, 10,000 and 50,000 sentence pairs from our Dutch-English and Italian-English Europarl data. From our initial ground truth set (Subsection 3.3.1), we

have only kept the words that occur at least once in the respective corpora as ground truth for evaluations (e.g., there are 444 words in the ground truth dataset for the corpus consisting of the first Dutch-English 2,000 sentence pairs, and 931 words for the corpus consisting of the first 50,000 Dutch-English sentence pairs). Our question is now: "Are lexicons extracted by SampLEX really of better quality than lexicons obtained by other methods when dealing with parallel corpora of limited size?" As mentioned before, the *SampLEX* algorithm does not have a property to provide results in a form of ranked lists for the entire source vocabulary, but we claim that *SampLEX* is directed towards extracting only highly reliable and precise candidates which, consequently, leads to lexicons of a higher quality. That claim is again supported by the findings presented in Figure 1(a) for Dutch-English, and in Figure 1(b) for Italian-English.



(a) Dutch to English (b) Italian to English

Figure 1: Precision and F-Measure scores over parallel corpora of different size (2k, 10k and 50k aligned sentence pairs). Since *SampLEX* does not necessarily obtain the lists of translations for all words in a vocabulary, its precision scores are different than its F-measure scores. For all other models within this evaluation setting, it is valid: *Precision=Recall=F-measure*.

We have also performed an additional experiment to test whether the translation candidates for Dutch and Italian words that happen to be retrieved by the *SampLEX* algorithm still display better overall precision and MRR scores than the translation candidates for the same Dutch and Italian words obtained by the other methods. If that is not true, we could use *SampLEX* only to extract source words for which a translation might be found, but the particular translation for each extracted word could then be obtained by some other method. However, it is not the case, as the results in Tables 5 and 6 reveal. As noted in the literature (Manning and Schütze, 1999), we observe that, of all the baseline models for BLE, LLR suffers the least from data sparsity, but still performs worse than our method.

Since *SampLEX* is built on the concept of data sampling, the criteria for extracting translation candidates and the whole training process inherently remain the same when working with parallel corpora of limited size. However, it is natural that the results decrease when the size of the large corpus $\mathscr{C}$ decreases. The more data we possess, the more sub-corpora we can sample, which finally provides better chances to extract correct translation pairs. We could say

	Dutch-English				
	IBM1-i5	**IBM1-i20**	**DICE**	**LLR**	**SampLEX**
Prec(2k)	0.3668	0.3624	0.3319	0.4323	**0.4498**
MRR(2k)	0.4206	0.4199	0.3968	0.4498	**0.4836**
Prec(10k)	0.4266	0.4306	0.3682	0.4889	**0.5272**
MRR(10k)	0.5071	0.5039	0.4513	0.5587	**0.5848**
Prec(50k)	0.6180	0.5952	0.5295	0.6621	**0.6850**
MRR(50k)	0.7067	0.6901	0.6183	0.7182	**0.7429**

Table 5: Precision and MRR scores for all models trained on the subsets of different sizes (2k, 10k and 50k sentences) from Dutch-English Europarl data. Only candidates retrieved by *SampLEX* have been taken into account for this evaluation.

	Italian-English				
	IBM1-i5	**IBM1-i20**	**DICE**	**LLR**	**SampLEX**
Prec(2k)	0.5087	0.5174	0.4348	0.5521	**0.5652**
MRR(2k)	0.5798	0.5778	0.4897	**0.6113**	0.6079
Prec(10k)	0.5978	0.5846	0.5011	0.6461	**0.6637**
MRR(10k)	0.6556	0.6489	0.5709	0.6914	**0.7014**
Prec(50k)	0.7129	0.6966	0.6381	0.7578	**0.7714**
MRR(50k)	0.7926	0.7847	0.7186	0.8064	**0.8278**

Table 6: Precision and MRR scores for all models trained on the subsets of different sizes (2k, 10k and 50k sentences) from Italian-English Europarl data. Only candidates retrieved by *SampLEX* have been taken into account for this evaluation.

that *SampLEX* takes the best of both worlds - it benefits from the idea of data reduction, yet it provides better scores when more input data are available.

5 Conclusions and Future Work

In this paper, we have proposed a statistical framework for the construction of a bilingual word lexicon built upon the idea of sampling many smaller sub-corpora from an initial larger item-aligned corpus.

The *SampLEX* algorithm for bilingual lexicon extraction presented in the paper is directed towards extraction of only highly reliable word translation pairs. After comparisons with other models for BLE from parallel data, we have proved that *SampLEX* builds word lexicons of higher accuracy and overall quality as revealed by the F-measure and precision scores, which is especially important in a setting where only a limited amount of parallel data is available. The proposed framework allows for many further experimentations and possible applications. The description of the framework provided in the paper is generic - it is language-independent and applicable to any corpus that provides some sort of alignment (at the sentence, paragraph or document level). In future work, we plan to design algorithms for mining word translations from comparable corpora based on the similar idea of sub-corpora sampling.

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, KU Leuven, Belgium.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition*. Wiley.

Andrade, D., Nasukawa, T., and Tsujii, J. (2010). Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 19–27.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Carbonell, J. G., Yang, J. G., Frederking, R. E., Brown, R. D., Geng, Y., Lee, D., Frederking, Y., E, R., Geng, R. D., and Yang, Y. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 708–714.

Church, K. W. and Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pp. 1–8.

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

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fung, P. and Cheung, P. (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 (EMNLP)*, pp. 57–63.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, pp. 414–420.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 771–779.

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

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit 2005*, pp. 79–86.

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 617–625.

Lefever, E., Macken, L., and Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 496–504.

Levow, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

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

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–10.

Moore, R. C. (2001). Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the Workshop on Data-Driven Methods in Machine Translation*, pp. 1–8.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (AMTA)*, pp. 135–144.

Moore, R. C. (2004a). Improving IBM word-alignment Model 1. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 518–525.

Moore, R. C. (2004b). On log-likelihood-ratios and the significance of rare events. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 333–340.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 664–671.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 81–88.

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

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pinto, D., Civera, J., Barrón-Cedeòo, A., Juan, A., and Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *Journal of Algorithms*, 64(1):51–60.

Prochasson, E. and Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 1327–1335.

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

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 519–526.

Resnik, P. and Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29:349–380.

Shezaf, D. and Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 98–107.

Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 24–36.

Tiedemann, J. (2003). Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 339–346.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing (RANLP)*, pp. 237–248.

Ueffing, N. and Ney, H. (2007). Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

Venugopal, A., Vogel, S., and Waibel, A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 319–326.

Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of the Eighth TExt Retrieval Conference (TREC-8)*.

Vulić, I., De Smet, W., and Moens, M.-F. (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 (ACL-HLT)*, pp. 479–484.

Vulić, I. and Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 449–459.