A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else)

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

We present a new language pair agnostic approach to inducing bilingual vector spaces from non-parallel data without any other resource in a bootstrapping fashion. The paper systematically introduces and describes all key elements of the bootstrapping procedure: (1) starting point or seed lexicon, (2) the confidence estimation and selection of new dimensions of the space, and (3) convergence. We test the quality of the induced bilingual vector spaces, and analyze the influence of the different components of the bootstrapping approach in the task of bilingual lexicon extraction (BLE) for two language pairs. Results reveal that, contrary to conclusions from prior work, the seeding of the bootstrapping process has a heavy impact on the quality of the learned lexicons. We also show that our approach outperforms the best performing fully corpus-based BLE methods on these test sets.

1 Introduction

Bilingual lexicons serve as an indispensable source of knowledge for various cross-lingual tasks such as cross-lingual information retrieval (Lavrenko et al., 2002; Levow et al., 2005) or statistical machine translation (Och and Ney, 2003). Additionally, they are a crucial component in cross-lingual knowledge transfer, where the knowledge about utterances in one language may be transferred to another. The utility of the transfer or annotation projection by means of bilingual lexicons has already been proven in various tasks such as semantic role labeling (Padó and Lapata, 2009; van der Plas et al., 2011), parsing (Zhao et al., 2009; Durrett et al., 2012; Täckström et al., 2013b), POS tagging (Yarowsky and Ngai, 2001; Das and Petrov, 2011; Täckström et al., 2013a), etc.

Techniques for automatic bilingual lexicon extraction (BLE) from parallel corpora on the basis of word alignment models are well established (Och and Ney, 2003). However, due to a relative scarceness of parallel data for many language pairs and domains, alternative approaches that rely on comparable corpora have also gained much interest (e.g., Fung and Yee (1998); Rapp (1999)).

The models that rely on non-parallel data typically represent each word by a high-dimensional vector in a feature vector space, where the dimensions of the vector are its *context features*. The context features are typically words co-occurring with the word in a predefined context.[1] The similarity of two words, w_1^S given in the source language L_S with vocabulary V^S and w_2^T in the target language L_T with vocabulary V^T is then computed as $sim(w_1^S, w_2^T) = SF(cv(w_1^S), cv(w_2^T))$. $cv(w_1^S) = [sc_1^S(c_1), \ldots, sc_1^S(c_N)]$ is a context vector for w_1^S with N context features c_k, where $sc_1^S(c_k)$ denotes the score for w_1^S associated with context feature c_k (similar for w_2^T). SF is a similarity function (e.g., cosine, the Kullback-Leibler divergence, the Jaccard index) operating on the context vectors (Lee, 1999).

When operating with 2 languages, the context features cannot be compared directly. Therefore, in order to compare the feature vectors $cv(w_1^S)$ and $cv(w_2^T)$, the context features need to span a shared

[1] The context may be a document, a paragraph, a window of predefined size around each occurrence of w_i^S in $\mathcal{C}_S$, etc. For an overview, see, e.g., (Tamura et al., 2012).

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1613–1624,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

bilingual vector space. The standard way of building a bilingual vector space is to use *bilingual lexicon entries* (Rapp, 1999; Fung and Cheung, 2004; Gaussier et al., 2004) as dimensions of the space. However, there seems to be an apparent flaw in logic, since the methods assume that there exist readily available bilingual lexicons that are then used to induce bilingual lexicons! Therefore, the focus of the researchers has turned to designing BLE methods that do not rely on any external translation resources such as machine-readable bilingual lexicons and parallel corpora (Haghighi et al., 2008; Vulić et al., 2011).

In order to circumvent this issue, one line of recent work aims to *bootstrap high-quality bilingual vector spaces from a small initial seed lexicon*. The seed lexicon is constructed by harvesting identical or similarly spelled words across languages (Koehn and Knight, 2002; Peirsman and Padó, 2010), and it spans the initial bilingual vector space. *The space is then gradually enriched with new dimensions/axes during the bootstrapping procedure*. The bootstrapping process has already proven its validity in inducing bilingual lexicons for closely similar languages such as Spanish-Portuguese or Croatian-Slovene (Fišer and Ljubešić, 2011), but it still lacks further generalization to more distant language pairs.

The main goal of this paper is to shed new light on the bootstrapping approaches to bilingual lexicon extraction, and to construct a language pair agnostic bootstrapping method that is able to build high-quality bilingual vector spaces that consequently lead to high-quality bilingual lexicons for more distant language pairs where orthographic similarity is not sufficient to seed bilingual vector spaces. We aim to answer the following key questions:

- How to seed bilingual vector spaces besides using only orthographically similar words?
- Is it better to seed bilingual spaces with translation pairs/dimensions that are frequent in the corpus, and does the frequency matter at all? Does the size of the initial seed lexicon matter?
- How to enrich bilingual vector spaces with only highly reliable dimensions in order to prevent semantic drift?

With respect to these questions, the main contributions of this article are:

- We present a complete overview of the framework of bootstrapping bilingual vector spaces from non-parallel data without any additional resources. We dissect the bootstrapping process and describe all its key components.
- We introduce a new way of seeding the bootstrapping procedure that does not rely on any orthographic clues and that yields bilingual vector spaces of higher quality. We analyze the impact of different seed lexicons on the quality of induced bilingual vector spaces.
- We show that in the setting without any external translation resources, our bootstrapping approach yields lexicons that outperform the best performing corpus-based BLE methods on standard test datasets for 2 language pairs.

2 Boostrapping Bilingual Vector Spaces: A General Overview

This section presents the complete bootstrapping procedure that starts with an initial seed lexicon which spans the initial bilingual vector space, while as the output in each iteration of the procedure it produces an updated bilingual vector space that can be used to extract a bilingual lexicon.

2.1 General Framework

We assume that we are solely in possession of a (non-parallel) bilingual corpus $\mathcal{C}$ that is composed of a sub-corpus $\mathcal{C}_S$ given in the source language L_S, and a sub-corpus $\mathcal{C}_T$ in the target language L_T. All word types that occur in $\mathcal{C}_S$ constitute a set V^S. All word types in $\mathcal{C}_T$ constitute a set V^T. The goal is to build a bilingual vector space using only corpus $\mathcal{C}$.

Assumption 1. *Dimensions of the bilingual vector space are one-to-one word translation pairs*. For instance, dimensions of a Spanish-English space are pairs like *(perro, dog)*, *(ciencia, science)*, etc. The *one-to-one constraint* (Melamed, 2000), although not valid in general, simplifies the construction of the bootstrapping procedure. $\mathcal{Z}$ denotes the set of translation pairs that are the dimensions of the space.

Computing cross-lingual word similarity in a bilingual vector space. Now, assume that our bilingual vector space consists of N one-to-one word translation pairs $c_k = (c_k^S, c_k^T), k = 1, \ldots, N$. For each word $w_i^S \in V^S$, we compute the similarity of that word with each word $w_j^T \in V^T$ by computing the similarity between their context vectors $cv(w_i^S)$ and $cv(w_j^T)$, which are actually their representations in the N-dimensional bilingual vector space.

The cross-lingual similarity is computed following the standard procedure (Gaussier et al., 2004):
(1) For each source word $w_i^S \in V^S$, build its N-dimensional context vector $cv(w_i^S)$ that consists of association scores $sc_k^S(c_k^S)$, that is, we compute the strength of association with the "source" part of each dimension c_k that constitutes the N-dimensional bilingual space. The association is dependent on the co-occurrence of w_i^S and c_k^S in a predefined context. Various functions such as the log-likelihood ratio (LLR) (Rapp, 1999; Ismail and Manandhar, 2010), TF-IDF (Fung and Yee, 1998), or pointwise mutual information (PMI) (Bullinaria and Levy, 2007; Shezaf and Rappoport, 2010) are typically used as *weighting functions* to quantify the strength of the association.
(2) Repeat step (1) for each target word $w_j^T \in V^T$ and build context vectors $cv(w_j^T)$ that consist of scores $sc_k^T(c_k^T)$.
(3) Since c_k^S and c_k^T address the same dimension c_k in the bilingual vector space for each $k = 1, \ldots, N$, we are able to compute the similarity between $cv(w_i^S)$ and $cv(w_j^T)$ using any similarity measure such as the Jaccard index, the Kullback-Leibler or the Jensen-Shannon divergence, the cosine measure, or others (Lee, 1999; Cha, 2007).

The similarity score for two words w_i^S and w_j^T is $sim(w_i^S, w_j^T)$. For each source word w_i^S, we can build a *ranked list* $RL(w_i^S)$ that consists of all words $w_j^T \in V^T$ ranked according to their respective similarity scores $sim(w_i^S, w_j^T)$. In the similar fashion, we can build a ranked list $RL(w_j^T)$, for each target word w_j^T. We call the top scoring target word w_j^T for some source word w_i^S its *translation candidate*, and write $TC(w_i^S) = w_j^T$. Additionally, we label the ranked list $RL(w_i^S)$ that is pruned at position M as $RL_M(w_i^S)$.

Bootstrapping. The key idea of the bootstrapping approach relies on an insight that *highly reliable translation pairs* (w_1^S, w_2^T) that are encountered using the N-dimensional bilingual vector space might be added as new dimensions of the space. By adding these new dimensions, it might be possible to extract more highly reliable translation pairs that were previously not used as dimensions of the space, and the iterative procedure repeats until no new dimensions are found. The induced bilingual vector space may then be observed as a bilingual lexicon *per se*, but it may also be used to find translation equivalents for other words which are not used to span the space.

Algorithm 1: Bootstrapping a bilingual vector space

```
Input  : Bilingual corpus $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$
Initialize: (1) Obtain a one-to-one seed lexicon. The
entries from the lexicon are initial dimensions of the
space: $\mathcal{Z}_0$; (2) $s = 0$;
Bootstrap:
repeat
    1: For each $w_i^S \in V^S$: compute $RL(w_i^S)$ using $\mathcal{Z}_s$ ;
    2: For each $w_j^T \in V^T$: compute $RL(w_j^T)$ using $\mathcal{Z}_s$ ;
    3: For each $w_i^S \in V^S$ and $w_j^T \in V^T$: score each
    translation pair $(w_i^S, TC(w_i^S))$ and $(TC(w_j^T), w_j^T)$
    and add them to a pool of candidate dimensions ;
    4: Choose the best candidates from the pool and add
    them as new dimensions: $\mathcal{Z}_{s+1} \leftarrow \mathcal{Z}_s \cup \{best\}$ ;
    5: Resolve collisions in $\mathcal{Z}_{s+1}$;
    6: $s \leftarrow s + 1$ ;
until no new dimensions are found (convergence) ;
Output: One-to-one translation pairs → Dimensions of a
        bilingual vector space $\mathcal{Z}_{final}$
```

The overview of the procedure as given by alg. 1 reveals these crucial points in the procedure: (Q1) how to provide initial dimensions of the space? (the initialization step), (Q2) how to score each translation pair, estimate their confidence, and how to choose the best candidates from the pool of candidates? (steps 3 and 4), and (Q3) how to resolve potential collisions that violate the one-to-one constraint? (step 5). We will discuss (Q1) and (Q2) in more detail later, while we resolve (Q3) following a simple heuristic as follows:

Assumption 2. *In case of collision, dimensions/pairs that are found at later stages of the bootstrapping process overwrite previous dimensions.*

The intuition here is that we expect for the quality of the space to increase at each stage of the bootstrapping process, and newer translation pairs should be more confident than the older ones. For instance, if 2 out of N dimensions of a Spanish-English bilingual space are pairs *(piedra,wall)* and *(tapia,stone)*, but then if during the bootstrapping process we extract a new candidate pair *(piedra,stone)*, we will delete the former two dimensions and add the latter.

2.2 Initializing Bilingual Vector Spaces

Seeding or initializing a bootstrapping procedure is often a critical step regardless of the actual task (McIntosh and Curran, 2009; Kozareva and Hovy, 2010), and it decides whether the complete process will end as a success or a failure. However, Peirsman and Padó (2011) argue that the initialization step is not crucial when dealing with bootstrapping bilingual vector spaces. Here, we present two different strategies of initializing the bilingual vector space.

Identical words and cognates. Previous work relies exclusively on identical and similarly spelled words to build the initial set of dimensions $\mathcal{Z}_0$ (Koehn and Knight, 2002; Peirsman and Padó, 2010; Fišer and Ljubešić, 2011). This strategy yields promising results for closely similar language pairs, but is of limited use for other language pairs.

High-frequency seeds. Another problem with using only identical words and cognates as seeds lies in the fact that many of them might be infrequent in the corpus, and as a consequence the expressiveness of a bilingual vector space might be limited. On the other hand, high-frequency words offer a lot of evidence in the corpus that could be exploited in the bootstrapping approach. In order to induce initial translation pairs, we rely on the framework of *multilingual probabilistic topic modeling* (MuPTM) (Boyd-Graber and Blei, 2009; De Smet and Moens, 2009; Mimno et al., 2009; Zhang et al., 2010), that does not require a bilingual lexicon, it operates with non-parallel data, and is able to produce highly confident translation pairs for high-frequency words (Mimno et al., 2009; Vulić and Moens, 2013).[2] Therefore, we can construct the initial seed lexicon as follows:

(1) Train a multilingual topic model on the corpus.

(2) Obtain one-to-one translation pairs using any of the MuPTM-based models of cross-lingual similarity, e.g., (Vulić et al., 2011; Vulić and Moens, 2013).

(3) Retain only *symmetric* translation pairs. This step ensures that only highly confident pairs are used as seed translation pairs.

(4) Rank translation pairs according to their frequency in the corpus and use a subset of the most frequent symmetric pairs as seeds.

2.3 Estimating Confidence of New Dimensions

Another crucial step in the bootstrapping procedure is the estimation of confidence in a translation pair/candidate dimension. Errors in the early stages of the procedure may negatively affect the learning process and even cause *semantic drift* (Riloff and Shepherd, 1999; McIntosh and Curran, 2009). We therefore impose the constraint which requires translation pairs to be *symmetric* in order to qualify as potential new dimensions of the space. In other words, given the current set of dimensions $\mathcal{Z}_s$, a translation pair (w_i^S, w_j^T) has a possibility to be chosen as a new dimension from the pool of candidate dimensions if and only if it holds: $TC(w_i^S) = w_j^T$ and $TC(w_j^T) = w_i^S$. This *symmetry constraint* should ensure a relative reliability of translation pairs.

In each iteration of the bootstrapping process, we may add all symmetric pairs from the pool of candidates as new dimensions, or we could impose additional selection criteria that quantify the degree of confidence in translation pairs. We are then able to rank the symmetric candidate translation pairs in the pool of candidates according to their confidence scores (step 3 of alg. 1), and choose only the best B candidates from the pool in each iteration (step 4) as done in (Thelen and Riloff, 2002; McIntosh and Curran, 2009; Huang and Riloff, 2012). By picking only a subset of the B most confident candidates in each iteration, we hope to further prevent a possibility of semantic drift, i.e., "poisoning" the bootstrapping process that might happen if we include incorrect translation pairs as dimensions of the space.

In this paper, we investigate 3 different confidence estimation functions:[3]

(1) **Absolute similarity score.** Confidence of a translation pair $CF(w_i^S, TC(w_i^S))$ is simply the absolute similarity value $sim(w_i^S, TC(w_i^S))$

(2) **M-Best confidence function.** It contrasts the score of the translation candidate with the average score over the first M most similar words in the ranked list. The larger the difference, the more confidence we have in the translation candidate. Given a word $w_i^S \in V^S$ and a ranked list $RL_M(w_i^S)$, the

[2] One can also use other models that are similar to MuPTM such as (Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011) to produce the initial seed lexicon, but that analysis is beyond the scope of this work.

[3] A symmetrized version of the confidence functions is computed as the geometric mean of source-to-target and target-to-source confidence scores.

average score of the best M words is computed as:

$$sim_M(w_i^S) = \frac{1}{M} \sum_{w_j^T \in RL_M(w_i^S)} sim(w_i^S, w_j^T)$$

The final confidence score is then:

$$CF(w_i^S, TC(w_i^S)) = sim(w_i^S, TC(w_i^S)) - sim_M(w_i^S)$$

(3) **Entropy-based confidence function.** We adapt the well-known entropy-based confidence (Smith and Eisner, 2007; Tu and Honavar, 2012) to this particular task. First, we need to define a distribution:

$$p(w_j^T | w_i^S) = \frac{e^{sim(w_i^S, w_j^T)}}{\sum_{w_l^T \in V^T} e^{sim(w_i^S, w_l^T)}}$$

The confidence function is then minus the entropy of the probability distribution p:

$$CF(w_i^S, TC(w_i^S)) = \sum_{w_l^T \in V^T} p(w_l^T | w_i^S) \log p(w_l^T | w_i^S)$$

3 Experimental Setup

Data collections. We investigate our bootstrapping approach on the BLE task for 2 language pairs: Spanish-English (ES-EN) and Italian-English (IT-EN), and work with the following corpora previously used by Vulić and Moens (2013): (i) a collection of $13,696$ Spanish-English Wikipedia article pairs (**Wiki-ES-EN**), (ii) $18,898$ Italian-English Wikipedia article pairs (**Wiki-IT-EN**).[4]

Following (Koehn and Knight, 2002; Haghighi et al., 2008; Prochasson and Fung, 2011; Vulić and Moens, 2013), we use TreeTagger (Schmid, 1994) for POS-tagging and lemmatization of the corpora, and then retain only nouns that occur at least 5 times in the corpus. We record the lemmatized form when available, and the original form otherwise. Our final vocabularies consist of $9,439$ Spanish nouns and $12,945$ nouns for ES-EN, and $7,160$ Italian nouns and $9,116$ English nouns for IT-EN.

Ground truth. The goal of the BLE task is to extract a bilingual lexicon of one-to-one translations. In order to test the quality of bilingual vector spaces induced by our bootstrapping approach, we evaluate it on standard 1000 ground truth one-to-one translation pairs built for the Wiki-ES-EN and Wiki-IT-EN datasets (Vulić and Moens, 2013). Note that we do not explicitly test the bilingual vector space as a bilingual lexicon, but rather its ability to find semantically similar words and translations also for words that are not used as dimensions of the space (see sect. 2.1).

Evaluation metrics. We measure the performance on the BLE task using a standard *Top M* accuracy (Acc_M) metric. It denotes the number of source words w_i^S from ground truth translation pairs whose list $RL_M(w_i^S)$ contains the correct translation according to our ground truth over the total number of ground truth translation pairs (*=1000*) (Gaussier et al., 2004; Tamura et al., 2012).[5] Additionally, we report the *mean reciprocal rank (MRR)* scores (Voorhees, 1999) for some experimental runs.

Multilingual topic model. We utilize a straightforward multilingual extension of the standard Blei et al.'s LDA model (Blei et al., 2003) called *bilingual LDA* (Mimno et al., 2009; Ni et al., 2009; De Smet and Moens, 2009). BiLDA training follows the procedure from (Vulić and Moens, 2013), that is, the training method is Gibbs sampling with the number of topics set to $K = 2000$. Hyperparameters of the model are set to standard values (Steyvers and Griffiths, 2007): $\alpha = 50/K$ and $\beta = 0.01$.

Building initial seed lexicons. To produce the lists of one-to-one translation pairs that are used as seeds for the bootstrapping approach (see sect. 2.2), we experiment with the *TopicBC* and the *ResponseBC* methods from (Vulić and Moens, 2013), which are the MuPTM-based models of cross-lingual semantic similarity that obtain the best results in the BLE task on these datasets. In short, the *TopicBC* method computes the similarity of two words according to the similarity of their conditional topic distributions (Griffiths et al., 2007; Vulić et al., 2011) using

[4] Vulić and Moens (2013) also worked with Dutch-English (NL-EN), but we have decided to leave out the results obtained for that language pair due to space constraints, high similarity between the two languages, and the fact that the results obtained for that language pair are qualitatively similar to the results we report for ES-EN and IT-EN. Hence including the results for NL-EN would not contribute to the paper with any new important insight and conclusion.

[5] We can build a one-to-one bilingual lexicon by harvesting one-to-one translation pairs $(w_i^S, TC(w_i^S))$, and the quality of that lexicon is best reflected in the Acc_1 score.

the Bhattacharyya coefficient (BC) (Kazama et al., 2010) as the similarity function. *ResponseBC* is a second-order similarity method. It first computes initial similarity scores between all words cross-lingually and monolingually using the cross-lingual topical space and, in the second step, it computes the similarity between 2 words as the similarity between their word vectors that now contain the initial word-to-word similarity scores with all source and target words. The similarity function is again BC.

We use these models of similarity as a black box to acquire seeds for the bootstrapping approach, but we encourage the interested reader to find more details about the methods in the relevant literature. These two models also serve as our *baseline models*, and our goal is to test whether we are able to obtain bilingual lexicons of higher quality using bootstrapping that starts from the output of these models.

Weighting and similarity functions. We have experimented with different families of weighting (e.g., PMI, LLR, TF-IDF, chi-square) and similarity functions (e.g., cosine, Dice, Kullback-Leibler, Jensen-Shannon) (Lee, 1999; Turney and Pantel, 2010). In this paper, we present results obtained by *positive pointwise mutual information (PPMI)* (Niwa and Nitta, 1994) as a weighting function, which is a standard choice in vector space semantics (Turney and Pantel, 2010), and (combined with cosine) yields the best results over a group of semantic tasks according to (Bullinaria and Levy, 2007). We use a smoothed version of PPMI as presented in (Pantel and Lin, 2002; Turney and Pantel, 2010). Again, based on the results reported in the relevant literature (Bullinaria and Levy, 2007; Laroche and Langlais, 2010; Turney and Pantel, 2010), we opt for the cosine similarity as a standard choice for SF. We have also experimented with different window sizes ranging from 3 to 15 in both directions around the pivot word, but we have not detected any major qualitative difference in the results and their interpretation. Therefore, all results reported in the paper are obtained by setting the window size to 6.

4 Results and Discussion

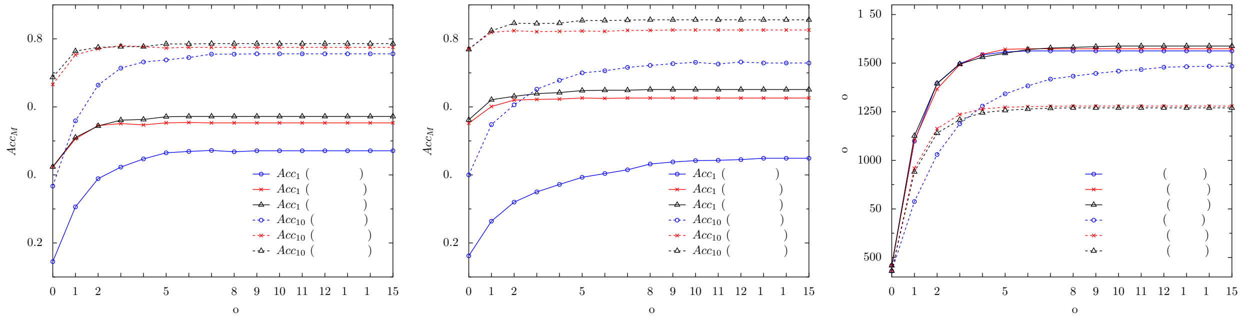
4.1 Are Seeds Important?

In recent work, Peirsman and Padó (2010; 2011) report that "the size and quality of the (seed) lexicon are not of primary importance given that the bootstrapping procedure effectively helped filter out incorrect translation pairs and added more newly identified mutual nearest neighbors." According to their findings, (1) noisy translation pairs are corrected in later stages of the bootstrapping process, since the quality of bilingual vector spaces gradually increases, (2) the size of the seed lexicon does not matter since the bootstrapping approach is able to learn translation pairs that were previously not present in the seed lexicon. Additionally, they do not provide any insight whether the frequency of seeds in the corpus influences the quality of induced bilingual vector spaces. In this paper, we question these claims with a series of BLE experiments.

All experiments conducted in this section do not rely on any extra confidence estimation except for the symmetry constraint, that is, in each step we enrich the bilingual vector space with all new symmetric translation pairs (see alg. 1 and sect. 2.3).

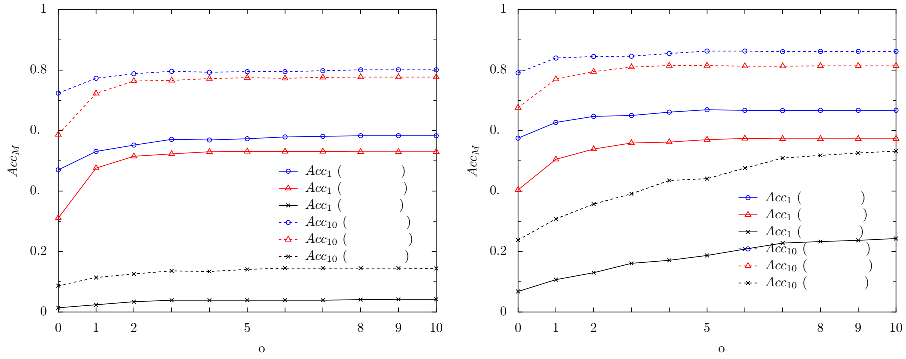
Exp. I: Same size, different seedings? The goal of this experiment is to test whether the quality of seeds plays an important role in the bootstrapping approach. We experiment with 3 different seed lexicons: (1) Following (Peirsman and Padó, 2010; Fišer and Ljubešić, 2011), we harvest identically spelled words across 2 languages and treat them as one-to-one translations. This procedure results in 459 seed translation pairs for ES-EN, and 431 pairs for IT-EN (SEED-ID), (2) We obtain symmetric translation pairs using the *TopicBC* method (see sect. 3) and use 459 pairs that have the highest frequency in the Wiki-ES-EN corpus as seeds for ES-EN (similarly 431 pairs for IT-EN) (SEED-TB), (3) As in (2), but we now use the *ResponseBC* method to acquire seeds (SEED-RB). The frequency of a one-to-one translation pair is simply computed as the geometric mean of the frequencies of words that constitute the translation pair.

Fig. 1(a) and 1(b) display the progress of the same bootstrapping procedure using the 3 different seed lexicons. We derive several interesting conclusions: (i) *Regardless of the actual choice of the seeding method, the bootstrapping process proves its validity and utility* since we observe that the quality of induced bilingual vector spaces increases over time for all 3 seeding methods. The bootstrapping procedure converges quickly. The increase is especially



(a) Spanish to English (b) Italian to English (c) # Dimensions over iterations

Figure 1: Results with 3 different seeding methods as starting points of the bootstrapping process: (i) identical words only (SEED-ID), (ii) the *TopicBC* method (SEED-TB), (iii) the *ResponseBC* method (SEED-RB). (a) Acc_M scores for ES-EN; (b) Acc_M scores for IT-EN; (c) the number of dimensions in the space with the 3 different seeding methods in each iteration for ES-EN and IT-EN. The bootstrapping procedure typically converges after a few iterations.



(a) Spanish to English (b) Italian to English

Figure 2: Results on the BLE task with SEED-RB when using seed translation pairs of different frequency: (i) high-frequency (HF-SEED), (ii) medium-frequency (MF-SEED), (iii) low-frequency (LF-SEED).

prominent in the first few iterations, when the approach learns more new dimensions (see fig. 1(c)). (ii) *The seeding method is important.* A bootstrapping approach that starts with a better seed lexicon is able to extract bilingual lexicons of higher quality as reflected in Acc_1 scores. Although the bootstrapping approach seems more beneficial when dealing with noisier seed lexicons (226% increase in terms of Acc_1 for ES-EN and 177% increase for IT-EN when starting with SEED-ID, compared to 35% increase for ES-EN, and 15% for IT-EN with SEED-RB), when starting from a noisy seed lexicon such as SEED-ID the method is unable to reach the same level of performance. Starting with SEED-ID, the approach is able to recover noisy dimensions from an initial bilingual vector space, but it is still unable to match the results that are obtained when starting from a better initial space (e.g., SEED-RB).

(iii) SEED-RB produces slightly better results than SEED-TB (e.g., the final Acc_1 of 0.649 for SEED-RB compared to 0.626 for SEED-TB for IT-EN, and 0.572 compared to 0.553 for ES-EN). This finding is in line with the results reported in (Vulić and Moens, 2013) where *ResponseBC* proved to be a more robust and a more effective method when applied to the BLE task directly. In all further experiments we use *ResponseBC* to acquire seed pairs, i.e., the seeding method is SEED-RB.

Exp. II: Does the frequency of seeds matter? In the next experiment, we test whether the frequency of seeds in the corpus plays an important role in the bootstrapping process. The intuition is that by using highly frequent and highly confident translation pairs the bootstrapping method has more reliable clues that help extract new dimensions in subsequent iterations. On the other hand, low-frequency pairs (although potentially correct one-to-one translations) do not occur in the corpus and in the contexts of other words frequently enough, and are therefore not sufficient to extract reliable new dimensions of the space.

To test the hypothesis, we again obtain all symmetric translation pairs using *ResponseBC* and then sort them in descending order based on their frequency in the corpus. In total, we retrieve a sorted list of 2031 symmetric translation pairs for ES-EN, and 1689 pairs for IT-EN. Following that, we split the list in 3 parts of equal size: (i) the top third comprises translation pairs with the highest frequency in the corpus (HF-SEED), (ii) the middle third comprises pairs of "medium" frequency (MF-SEED), (iii) the bottom third are low-frequency pairs (LF-SEED). We then use these 3 different seed lexicons of equal size to seed the bootstrapping approach. Fig. 2(a) and 2(b) show the progress of the bootstrapping process using these 3 seed lexicons. We again observe several interesting phenomena:

(i) *High-frequency seed translation pairs are better seeds*, and that finding is in line with our hypothesis. Although the bootstrapping approach again displays a positive trend regardless of the actual choice of seeds (we observe an increase even when using LF-SEED), high-frequency seeds lead to better overall results in the BLE task. Besides its high presence in contexts of other words, another advantage of high-frequency seed pairs is the fact that an initial similarity method will typically acquire more reliable translation candidates for such words (Pekar et al., 2006). For instance, 89.5% of ES-EN pairs in HF-SEED are correct one-to-one translations, compared to 65.1% in MF-SEED, and 44.3% in LF-SEED.

(ii) The difference in results between HF-SEED and MF-SEED is more visible in Acc_1 scores. Although both seed lexicons for all test words provide ranked lists which contain words that exhibit some semantic relation to the given word, the reliability and the frequency of translation pairs are especially important for detecting the relation of cross-lingual word synonymy, that is, the translational equivalence that is exploited in building one-to-one bilingual lexicons.

Exp. III: Does size matter? The following experiment investigates whether bilingual vector spaces may be effectively bootstrapped from small high-quality seed lexicons, and if larger seed lexicons necessarily lead to bilingual vector spaces of higher quality as reflected in BLE results. We again retrieve a sorted list of symmetric translation pairs as in Exp. II. Following that, we build seed lexicons of various sizes by retaining only the first N pairs from the list, where we vary N from 200 to 1400 in steps of 200. We also use the entire sorted list as a seed lexicon (*All*), and compare the results on the BLE task with the results obtained by applying the *ResponseBC* and *TopicBC* methods directly (Vulić and Moens, 2013). The results are summarized in tables 1 and 2. We observe the following:

(i) If we provide a seed lexicon with sufficient entries, the bootstrapping procedure provides comparable results regardless of the seed lexicon size, although results tend to be higher for larger seed lexicons (e.g., compare results when starting with 600 and 1200 lexicon entries). When starting with the size of 600, the bootstrapping approach is able to find dimensions that were already in the seed lexicon of size 1200. The consequence is that, although bootstrapping with a smaller seed lexicon displays a slower start (see the difference in results at iteration 0), the performances level after convergence.

(ii) Regardless of the seed lexicon size, the bootstrapping approach is valuable. It consistently improves the quality of the induced bilingual vector space, and consequently, the quality of bilingual lexicons extracted using that vector space. The positive impact is more prominent for smaller seed lexicons, i.e., we observe an increase of 78% for ES-EN when starting with only 200 seed pairs, compared to an increase of 15% when starting with 800 seed pairs, and 10% when starting with 1400 seed pairs.

(iii) The bootstrapping approach outperforms *ResponseBC* and *TopicBC* in terms of Acc_1 and MRR scores for both language pairs when the seed lexicon provides a sufficient number of entries. However, in terms of Acc_{10}, *TopicBC* and *ResponseBC* still exhibit comparable (for IT-EN) or even better (ES-EN) results. Both *TopicBC* and *ResponseBC* are MuPTM-based methods that, due to MuPTM properties, model the similarity of two words at the level of documents as contexts, while the bootstrapping approach is a window-based approach that narrows down the context to a local neighborhood of a word. The MuPTM-based models are better suited to detect a general *topical* similarity of words, and

Iteration:	**0**			**2**			**5**			**10**		
Seed lexicon	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}
200(→1617)	0.274	0.352	0.525	0.446	0.534	0.713	0.481	0.569	0.753	0.488	0.576	0.752
400(→1563)	0.416	0.499	0.663	0.518	0.602	0.774	0.542	0.620	0.787	0.545	0.625	0.788
600(→1554)	0.459	0.539	0.707	0.550	0.630	0.787	0.573	0.650	0.803	0.578	0.654	0.802
800(→1582)	0.494	0.572	0.728	0.548	0.631	0.799	0.563	0.644	0.802	0.567	0.646	0.806
1000(→1636)	0.516	0.591	0.744	0.563	0.644	0.805	0.578	0.656	0.813	0.581	0.658	0.817
1200(→1740)	0.536	0.613	0.764	0.586	0.661	0.804	0.588	0.664	0.812	0.591	0.667	0.814
1400(→1888)	0.536	0.620	0.776	0.583	0.659	0.808	0.589	0.666	0.815	0.588	0.666	0.818
All-2031(→2437)	0.543	0.625	0.785	0.589	0.667	0.813	0.597	0.675	0.818	**0.599**	**0.677**	0.820
TopicBC	0.433	0.576	0.843	–	–	–	–	–	–	–	–	–
ResponseBC	0.517	0.635	**0.891**	–	–	–	–	–	–	–	–	–

Table 1: ES-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the bilingual space after the bootstrapping procedure converges. The seeding method is SEED-RB.

Iteration:	**0**			**2**			**5**			**10**		
Seed lexicon	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{10}
200(→1255)	0.394	0.469	0.703	0.515	0.595	0.757	0.548	0.621	0.782	0.555	0.628	0.787
400(→1265)	0.546	0.618	0.757	0.623	0.690	0.831	0.639	0.704	0.840	0.644	0.709	0.844
600(→1309)	0.585	0.657	0.798	0.653	0.718	0.856	0.664	0.726	0.859	**0.672**	0.734	0.862
800(→1365)	0.602	0.672	0.813	0.657	0.723	0.857	0.663	0.726	0.865	0.665	0.730	0.867
1000(→1416)	0.616	0.688	0.828	0.629	0.706	0.853	0.636	0.709	0.857	0.642	0.714	0.861
1200(→1581)	0.628	0.700	0.840	0.655	0.724	0.869	0.664	0.733	0.877	0.668	**0.736**	**0.883**
1400(→1749)	0.626	0.701	0.851	0.654	0.727	0.867	0.656	0.728	0.867	0.661	0.733	0.874
All-1689(→2008)	0.616	0.695	0.850	0.643	0.716	0.860	0.653	0.724	0.862	0.654	0.726	0.866
TopicBC	0.578	0.667	0.834	–	–	–	–	–	–	–	–	–
ResponseBC	0.622	0.729	0.882	–	–	–	–	–	–	–	–	–

Table 2: IT-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the bilingual space after the bootstrapping procedure converges. The seeding method is SEED-RB.

are therefore not always able to push the real cross-lingual synonyms higher in the ranked list of semantically similar words, while the window-based bootstrapping approach is better tailored to model the relation of cross-lingual synonymy, i.e., to extract one-to-one translation pairs (as reflected in Acc_1 scores). A similar conclusion for monolingual settings is drawn by Baroni and Lenci (2010).

(iv) Since our bootstrapping approach utilizes *ResponseBC* or *TopicBC* as a preprocessing step, it is obvious that the approach leads to an increased complexity. On top of the initial complexity of *ResponseBC* and *TopicBC*, the bootstrapping method requires $|V^S||V^T|$ comparisons at each iteration, but given the fact that each $w_i^S \in V^S$ may be processed independently of any other $w_j^S \in V^S$ in each iteration, the bootstrapping method is trivially parallelizable. That makes the method computationally feasible even for vocabularies larger than the ones reported in the paper.

4.2 Is Confidence Estimation Important?

According to the results from tables 1 and 2, regardless of the seed lexicon size, the bootstrapping approach does not suffer from semantic drift, i.e., if we seed the process with high-quality symmetric translation pairs, it is able to recover more pairs and add them as new dimensions of the bilingual vector space. However, we also study the influence of applying different confidence estimation functions on top of the symmetry constraint (see sect 2.3), but we do not observe any improvement in the BLE results, regardless of the actual choice of a confidence estimation function. The only observed phenomenon, as illustrated by fig. 3, is the *slower convergence rate* when setting the parameter B to lower values.

The symmetry constraint alone seems to be sufficient to prevent semantic drift, but it might also be a too strong and a too conservative assumption, since only a small portion of all possible translation pairs is used to span the bilingual vector space (for instance,

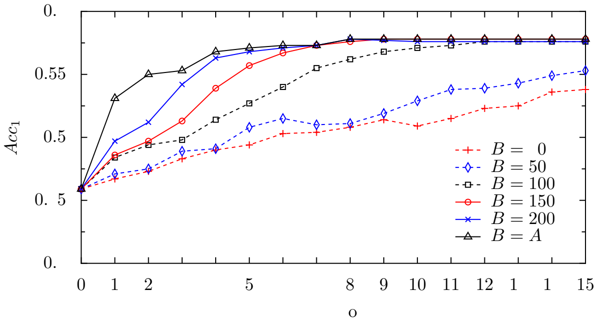


Figure 3: The effect of learning rate B on bootstrapping. Language pair: ES-EN, seed lexicon: SEED-RB with 600 pairs, confidence function: symmetrized M-Best.

when starting with 600 entries for ES-EN, the final bilingual vector space consists of only 1554 pairs, while the total number of ES nouns is 9439). One line of future work will address the construction of bootstrapping algorithms that also enable the usage of highly reliable asymmetric pairs as dimensions, and the confidence estimation functions might have a more important role in that setting.

5 Conclusion

We have presented a new bootstrapping approach to inducing bilingual vector spaces from non-parallel data, and have shown the utility of the induced space in the BLE task. The approach is fully corpus-based and, unlike previous work, it does not rely on the availability of machine-readable translation dictionaries or predefined concept categories. We have systematically described, analyzed and evaluated all key components of the bootstrapping approach. Results reveal that, contrary to conclusions from prior work, the initialization of the bilingual vector space is especially important. We have presented a novel approach to initializing the bootstrapping procedure, and have shown that better results in the BLE task are obtained by starting from seed lexicons that comprise highly reliable high-frequent translation pairs. The bootstrapping framework presented in the paper is completely language pair independent, which makes it effectively applicable to any language pair.

In future work, we will investigate other models of similarity besides *TopicBC* and *ResponseBC* (e.g, the method from (Haghighi et al., 2008)) that could be used as preliminary models for constructing an initial bilingual vector space. Furthermore, we plan to study other confidence functions and explore if asymmetric translation candidates could also contribute to the bootstrapping method. Finally, we also plan to test the robustness of our fully corpus-based bootstrapping approach by porting it to more language pairs.

Acknowledgments

We would like to thank the anonymous reviewers for their useful suggestions. This 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

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of UAI*, pages 75–82.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*, pages 600–609.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of ACL-HLT*, pages 407–412.

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.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of EMNLP-CoNLL*, pages 1–11.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of RANLP*, pages 125–131.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP*, pages 57–63.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING*, pages 414–420.

Éric 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 ACL*, 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 ACL*, pages 771–779.

Ruihong Huang and Ellen Riloff. 2012. Bootstrapped training of event extraction classifiers. In *Proceedings of EACL*, pages 286–295.

Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of COLING*, pages 481–489.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A Bayesian method for robust estimation of distributional similarities. In *Proceedings of ACL*, pages 247–256.

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

Zornitsa Kozareva and Eduard H. Hovy. 2010. Not all seeds are equal: Measuring the quality of text mining seeds. In *Proceedings of NAACL-HLT*, pages 618–626.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of COLING*, pages 617–625.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of SIGIR*, pages 175–182.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of ACL*, pages 25–32.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41(3):523–547.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of ACL*, pages 396–404.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP*, pages 880–889.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of WWW*, pages 1155–1156.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of COLING*, pages 304–309.

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

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of KDD*, pages 613–619.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of NAACL-HLT*, pages 921–929.

Yves Peirsman and Sebastian Padó. 2011. Semantic relations in bilingual lexicons. *ACM Transactions on Speech and Language Processing*, 8(2):article 3.

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

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of ACL*, pages 1327–1335.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*, pages 519–526.

Ellen Riloff and Jessica Shepherd. 1999. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *Natural Language Engineering*, 5(2):147–156.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of ACL*, pages 98–107.

David A. Smith and Jason Eisner. 2007. Bootstrapping feature-rich dependency parsers with entropic priors. In *Proceedings of EMNLP-CoNLL*, pages 667–677.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013a. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of ACL*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013b. Target language adaptation of discriminative transfer parsers. In *Proceedings of NAACL-HLT*, pages 1061–1071.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of EMNLP-CoNLL*, pages 24–36.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of EMNLP*, pages 214–221.

Kewei Tu and Vasant Honavar. 2012. Unambiguity regularization for unsupervised learning of probabilistic grammars. In *Proceedings of EMNLP-CoNLL*, pages 1324–1334.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of ACL-HLT*, pages 299–304.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of TREC*, pages 77–82.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of NAACL-HLT*, pages 106–116.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL-HLT*, pages 479–484.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*, pages 200–207.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of ACL*, pages 1128–1137.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of ACL*, pages 55–63.