Bootstrapping Unsupervised Bilingual Lexicon Induction

Bradley Hauer **Garrett Nicolai** **Grzegorz Kondrak**
Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,nicolai,gkondrak}@ualberta.ca

Abstract

The task of unsupervised lexicon induction is to find translation pairs across monolingual corpora. We develop a novel method that creates seed lexicons by identifying cognates in the vocabularies of related languages on the basis of their frequency and lexical similarity. We apply bidirectional bootstrapping to a method which learns a linear mapping between context-based vector spaces. Experimental results on three language pairs show consistent improvement over prior work.

1 Introduction

The objective of bilingual lexicon induction is to find translation pairs between two languages. Specifically, we aim to pair each word in the *source vocabulary* with its translation in the *target vocabulary*. In this paper, we assume that the languages are sufficiently closely related to allow some translation pairs to be identified on the basis of orthographic similarity. Our setting is completely unsupervised: we extract the bilingual lexicons from non-parallel monolingual corpora representing the same domain. By contrast, most of the prior work depend on parallel data in the form of a small bitext (Genzel, 2005), a gold seed lexicon (Mikolov et al., 2013b), or document-aligned comparable corpora (Vulić and Moens, 2015). Other prior work assumes access to additional resources or features, such as dependency parsers (Dou and Knight, 2013; Dou et al., 2014), temporal and web-based features (Irvine and Callison-Burch, 2013), or BabelNet (Wang and Sitbon, 2014).

Our approach consists of two stages: we first create a seed set of translation pairs, and then iteratively expand the lexicon with a bootstrapping procedure. The seed set is constructed by identifying words with similar spelling (*cognates*). We filter out non-translation pairs that look similar but differ in meaning (*false friends*) by imposing a relative-frequency constraint. We then use this noisy seed lexicon to train context vectors via neural network (Mikolov et al., 2013b), inducing a cross-lingual transformation that approximates semantic similarity. Although the initial accuracy of the transformation is low, it is sufficient to identify a certain number of correct translation pairs. Adding the high-confidence pairs to the seed lexicon allows us to refine the cross-lingual transformation matrix. We proceed to iteratively expand our lexicon by alternating the two steps of translation pair identification, and transformation induction.

We conduct a series of experiments on English, French, and Spanish. The results demonstrate a substantial error reduction with respect to a word-vector-based method of Mikolov et al. (2013b), when using the same word vectors on six source-target pairs. We also improve on the results reported by Haghighi et al. (2008) with both automatically-extracted and gold seed lexicons.

2 Methods

In this section, we describe the two components of our approach: seed lexicon extraction, and lexicon expansion via bootstrapping.

2.1 Seed Lexicon Extraction

Our seed extraction algorithm is aimed at identifying cross-lingual word pairs that exhibit high orthographic similarity, and have comparable frequency, both factors being indicators of translations (Kondrak, 2013). For each language, represented by a raw monolingual corpus, we first generate the list of word types, sorted by frequency. For each of the m most frequent source word

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 619–624,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

```
function EXTRACT_SEED($m$, $p$, $d$)
    $seed \leftarrow \emptyset$
    for $i$ from 1 to $m$ do
        $s \leftarrow$ source word such that $r_s = i$
        for each target word $t$ do
            if NED$(s,t) \leq d$
               and $|r_s - r_t| \leq p$
               and $s \neq t$ then
                 $seed \leftarrow seed \cup \{(s,t)\}$
    return $seed$
```

Figure 1: The seed lexicon extraction algorithm. r_w is the frequency rank of word w.

```
function LEX_INDUCTION($k$, $c$, $m$, $p$, $d$)
    $R \leftarrow$EXTRACT_SEED$(m,p,d)$
    for $c$ iterations do
        Train source-target TM $T$ on $R$
        Train target-source TM $T'$ on $R$
        for each source word $s$ do
            $f[s] \leftarrow \arg\max(score(s,t))$
        $R \leftarrow R \cup \{$top $k$ scoring pairs$\}$
    return translation mapping $f$
```

Figure 2: The lexicon induction algorithm. The *score* function is defined in Section 2.2.

types, starting from the top of the frequency list, we find all target words that satisfy the following constraints, as described in Figure 1, with parameters established on the development set.

1. Normalized edit distance (NED) between the source and target words, which is calculated by dividing the total edit cost by the length of the longer word, is within $d = 0.25$.
2. The absolute difference between the respective frequency ranks of the two words is within $p = 100$.
3. The source and target words are not identical.

The set of source-target pairs that satisfy these requirements form the seed lexicon. Note that there is no one-to-one constraint, so both source and target words may appear multiple times in the seed. The pseudo-code of the algorithm is shown in Figure 1.

2.2 Lexicon Expansion

Since our task is to find translations for each of a given set of source-language words, which we refer to as the source vocabulary, we must expand the seed lexicon to cover all such words. We adapt the approach of Mikolov et al. (2013b) for learning a linear transformation between the source and target vector spaces to enable it to function given only a small, noisy seed.

We use WORD2VEC (Mikolov et al., 2013a) to map words in our source and target corpora to n-dimensional vectors. The mapping is derived in a strictly monolingual context of both the source and target languages. While Mikolov et al. (2013b) derive the translation matrix using five thousand translation pairs obtained via Google Translate, our fully unsupervised method starts from a small and noisy seed lexicon extracted automatically with the algorithm described in Section 2.1.

Given a list of source-target translation pairs (s_i, t_i), with associated pairs of source and target vectors $(\mathbf{u_i}, \mathbf{v_i})$, we use stochastic gradient descent to learn a matrix $\mathbf{T}$ with objective $\mathbf{T} \cdot \mathbf{u_i} = \mathbf{v_i}$ for all i. In order to find a translation for a source-language word s represented by vector $\mathbf{u}$, we search for a target-language word t represented by vector $\mathbf{v}$ that minimizes the value of the cosine similarity function *sim*:

$$\mathbf{v} = \underset{\mathbf{v}' \in \text{ target word vectors}}{\text{argmin}} \mathbf{sim}(\mathbf{T} \cdot \mathbf{u},\ \mathbf{v}')$$

We use the cosine similarity $\mathbf{sim}(\mathbf{T} \cdot \mathbf{u},\ \mathbf{v})$ to calculate the confidence score for the corresponding candidate translation pair (s, t).

An important innovation of our algorithm is considering not only the fitness of t as a translation of s, but also of s as a translation of t. Distinct translation matrices are derived in both directions: source-to-target ($\mathbf{T}$) and target-to-source ($\mathbf{T}'$). We define the score of a pair (s, t) corresponding to the pair of vectors $(\mathbf{u}, \mathbf{v})$ as the average of the two cosine similarity values:

$$score(s,t) = \frac{\mathbf{sim}(\mathbf{T} \cdot \mathbf{u},\ \mathbf{v}) + \mathbf{sim}(\mathbf{T}' \cdot \mathbf{v},\ \mathbf{u})}{2}$$

Unlike Mikolov et al. (2013b), our algorithm iteratively expands the lexicon, which gradually increases the accuracy of the translation matrices. The initial translation matrices, derived from a small, noisy seed, are sufficient to identify a small number of correct translation pairs, which are added to the lexicon. The expanded lexicon is then used to derive new translation matrices, leading to more accurate translations.

In each iteration, we sort the candidate translation pairs by their current confidence scores, and add the highest-scoring k pairs to the lexicon. We exclude pairs that contain a word which is already in the lexicon. The next iteration uses the augmented lexicon to derive new translation matrices. We refer to this approach as *bootstrapping*, and continue the process for a set number of iterations, which is tuned on development data. The output of our algorithm is the set of translation pairs produced in the final iteration, with each source vocabulary word paired (not necessarily injectively) with one target vocabulary word.

3 Experiments

In this section we compare our method to two prior methods, our reimplementation of the supervised word-vector-based method of Mikolov et al. (2013b) (using the same vectors as our method), and the reported results of an EM-based method of Haghighi et al. (2008).

3.1 Data

Our experiments involve three language pairs: Spanish–French (ES–FR), English–French (EN–FR), and English–Spanish (EN–ES), which we consider in both directions. The corpora are from Europarl (Koehn, 2005; Tiedemann, 2012). In order to exclude parallel data, for each language pair, we take the first half of the source-language corpus, and the second half of the target-language corpus. (Less than 1% of sentences appear in both halves of any corpus.)

For evaluation, we require a gold-standard bilingual lexicon to decide whether a proposed source-target pair provides a correct translation of the source word. Following Dou and Knight (2013), we align the full source and target Europarl corpora with GIZA++ (Och and Ney, 2003). Since such alignments are asymmetric, we take the intersection of two alignments: source-to-target and target-to-source. The pairs of words that are aligned in both directions form our gold standard lexicon.

We follow the experimental setup of Haghighi et al. (2008). The source and target vocabularies consist of the 2000 most frequent words from the source and target corpora, with the exception of the words that are in the seed lexicons. For each of these 2000 source words, the task is to find a translation among the 2000 target words. We define a single test set for each language pair. Over 99% of words in the source vocabulary have translations in the target vocabulary.

	Pairs	Accuracy
ES–FR	206	87.9%
EN–FR	191	80.1%
EN–ES	239	83.3%
FR–ES	214	93.0%
FR–EN	210	79.1%
ES–EN	252	88.9%

Table 1: The size and accuracy of extracted seed lexicons.

3.2 Development

We performed development exclusively on the Spanish–French language pair. Since Spanish and French are more closely related to each other than either is to English, this allows us to test how our approach generalizes to more difficult language pairs. In addition, we aim for a fair comparison to prior work, who report results on English–Spanish and English–French. We use these language pairs exclusively for testing.

Based on the results of our Spanish–French development experiments, we established the following parameter settings. The seed lexicon extraction stage considers the $m = 10,000$ most frequent source words, identifying pairs with a frequency rank difference of at most $p = 100$, and a normalized edit distance of at most $d = 0.25$. We add $k = 25$ word pairs to the lexicon in each lexicon expansion iteration. The size of word vectors is set to $n = 200$ dimensions. The number of iterations depends on the metric we wish to optimize. We perform 40 iterations to optimize accuracy, and 25 iterations to optimize precision, as discussed in the next section.

During development, we found that excluding identical word pairs from the seed lexicon improves performance, so we incorporate this restriction in our system. 57 such pairs were removed from the Spanish-French seed lexicon, with most of them being numbers and proper nouns.

Table 1 shows that our extraction method produces seed lexicons of a reasonable size and accuracy, with, on average, 219 translation pairs at 85% accuracy. Less than 5% of words in any given seed are duplicates.

3.3 Evaluation

We evaluate the induced lexicon after 40 iterations of bidirectional bootstrapping by comparing it to the lexicon after the first iteration in a single direction, which is equivalent to the method of Mikolov et al. (2013b). Following Haghighi et al. (2008), we also report the accuracy of an EDITDIST baseline method, which matches words in the source and target vocabularies. We use an implementation of the Hungarian algorithm[1] (Kuhn, 1955) to solve the minimum bipartite matching problem, where the edge cost for any source-target pair is the normalized edit distance between the two words.

The results in Table 2 show that the method of Mikolov et al. (2013b) (MIK13-Auto), represented by the first translation matrix derived on our automatically extracted the seed lexicon, performs well below the edit distance baseline. By contrast, our bootstrapping approach (Bootstrap-Auto) achieves an average accuracy of 85% on the six datasets.

3.4 Unidirectional Scoring

In order to quantify the importance of our innovation of employing translation matrices in both directions, we also performed lexicon induction experiments in a unidirectional, source-to-target setting. The results show a consistent drop in accuracy on all language pairs. Error analysis reveals that this is caused by an increase in the number of incorrect translation pairs being added to the lexicon during bootstrapping, which negatively affects the quality of the resulting translation matrices.

The accuracy on English–French is particularly low (2.3%), which indicates that the unidirectional approach completely breaks down when the initial seed set contains fewer than 200 pairs. Too many incorrect translation pairs are added in the early stages, a problem the method never recovers from. In fact, when the size of the EN–ES seed is artificially reduced to the same size as the EN–FR seed (191 pairs), unidirectional scoring results in 1.2% accuracy, vs. 82% with bidirectional scoring. These results demonstrate that our innovation of bidirectional scoring makes the method more robust against smaller seed lexicons, allowing good results to be attained where previously proposed unidirectional scoring would fail.

	ES–FR	EN–FR	EN–ES
EDITDIST	47.2	36.4	34.7
MIK13-Auto	15.2	8.5	16.1
Bootstrap-Auto	89.4	79.4	82.0

	FR–ES	FR–EN	ES–EN
EDITDIST	46.9	36.8	35.0
MIK13-Auto	19.5	3.4	21.7
Bootstrap-Auto	89.4	83.5	84.5

Table 2: Accuracy of induced translation lexicons (in per cent correct).

3.5 Comparison to Haghighi et al. (2008)

Unlike most of the previous work on lexicon induction, our method is fully unsupervised, with no dependency on additional resources or tools. One other unsupervised method is that of Haghighi et al. (2008), who learn translation probabilities through a complex generative model known as matching canonical correlation analysis (MCCA). Although most of their experiments are semi-supervised, they report results obtained on English–Spanish with a version named "MCCA-Auto", which starts from an automatically-extracted seed lexicon. Since we have no access to their implementation, we attempt to re-create their experimental setup and adopt their evaluation metrics, making two accommodations in order to compare to the results reported in the original paper.

The first accommodation is the use of precision and recall, rather than accuracy, to evaluate the lexicons. After ranking the returned pairs by their score, the precision at a given point in the list is the percentage of the translation pairs that are correct, while the recall at a point is the percentage of the maximum possible number of translation pairs. Haghighi et al. (2008) chose to report precision values at four levels of recall: 0.1, 0.25, 0.33, and 0.5, as well as the best $F1$ measure achieved at any point. Unlike accuracy, point-wise precision assigns variable importance to the output translation pairs depending on their relative system score. In order to optimize the performance of our algorithm on the development set with respect to point-wise precision, we reduce the number of bootstrapping iterations to 25. The other parameters remain unchanged.

The second accommodation involves the restriction of the source and target vocabularies to

[1] *https://metacpan.org/pod/Algorithm::Munkres*

EN–ES	$p_{0.10}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	best F_1
EditDist	**99.0**	87.3	60.4	n/a	43.6
MCCA-Auto	91.2	90.5	91.8	77.5	61.7
Bootstrap-Auto	96.1	**95.9**	**93.2**	**84.9**	**67.9**
MCCA	91.4	94.3	92.3	89.7	63.7
Bootstrap	**96.6**	**95.6**	**93.6**	**89.9**	**73.7**

EN–FR	$p_{0.10}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	best F_1
EditDist	**99.0**	90.2	72.3	n/a	46.5
Bootstrap-Auto	93.0	**92.6**	**90.5**	**81.9**	**68.4**
MCCA	94.5	89.1	88.3	78.6	61.9
Bootstrap	**95.7**	**93.6**	**90.6**	**85.7**	**72.8**

Table 3: Comparison to the reported results of Haghighi et al. (2008) on EN–ES (upper table) and EN–FR (lower table). The best results are in bold.

the 2000 most frequent *nouns*. We consider a word to be a noun if it is tagged as such by TreeTagger (Schmid, 1994; Schmid, 1999). As in all of our experiments, we ensure that there is no overlap between the seed lexicon and the source and target test vocabularies.

Table 3 shows the results on English–Spanish and English–French. The upper rows contain fully-unsupervised results. The lower rows contain results obtained with the seed sets extracted directly from the gold standard lexicons by selecting the most frequent source language words. We make sure that both types of the seed sets are of equal size for each language pair. The precision of the EditDist baseline is the highest at 10% recall, but drops rapidly at the higher levels of recall. The variants of our method with both automatically-extracted (Bootstrap-Auto) and gold seed sets (Bootstrap) achieve higher precision than the corresponding variants of MCCA at all recall points, as well as higher best F_1 scores.

4 Conclusion

We have presented a bidirectional bootstrapping method for bilingual lexicon induction between related languages, which requires only a monolingual corpus in each language, with no assumptions of alignment or parallelism. We have demonstrated improvements over prior work and a strong baseline on three language pairs. The method has the potential to be applied across low-resource languages.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, Alberta Innovates – Technology Futures, and Alberta Advanced Education.

References

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1668–1676, Seattle, Washington, USA, October. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 557–565, Doha, Qatar, October. Association for Computational Linguistics.

Dmitriy Genzel. 2005. Inducing a multilingual dictionary from a parallel multitext in related languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 875–882, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the*

Eighth Workshop on Statistical Machine Translation, pages 262–270, Sofia, Bulgaria, August. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, September.

Grzegorz Kondrak. 2013. Word similarity, cognation, and translational equivalence. In Lars Borin and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, volume 265, pages 375–386. De Gruyter Mouton.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Workshop Proceedings at the International Conference on Learning Representations*, Scottsdale, AZ. 12 pages.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*. Technical report.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–49.

Helmut Schmid. 1994. Part-of-speech tagging with neural networks. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, pages 172–176, Kyoto, Japan, August.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Beijing, China, July. Association for Computational Linguistics.

Haoxing Wang and Laurianne Sitbon. 2014. Multilingual lexical resources to detect cognates in non-aligned texts. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 14–22, Brisbane, Australia, November.