Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals

Ann Irvine
Center for Language and Speech Processing
Johns Hopkins University

Chris Callison-Burch*
Computer and Information Science Dept.
University of Pennsylvania

Abstract

Prior research into learning translations from source and target language monolingual texts has treated the task as an unsupervised learning problem. Although many techniques take advantage of a seed bilingual lexicon, this work is the first to use that data for supervised learning to combine a diverse set of signals derived from a pair of monolingual corpora into a single discriminative model. Even in a low resource machine translation setting, where induced translations have the potential to improve performance substantially, it is reasonable to assume access to some amount of data to perform this kind of optimization. Our work shows that only a few hundred translation pairs are needed to achieve strong performance on the bilingual lexicon induction task, and our approach yields an average relative gain in accuracy of nearly 50% over an unsupervised baseline. Large gains in accuracy hold for all 22 languages (low and high resource) that we investigate.

1 Introduction

Bilingual lexicon induction is the task of identifying word translation pairs using source and target monolingual corpora, which are often comparable. Most approaches to the task are based on the idea that words that are translations of one another have similar distributional properties across languages. Prior research has shown that contextual similarity (Rapp, 1995), temporal similarity (Schafer and Yarowsky, 2002), and topical information (Mimno et al., 2009) are all good signals for learning translations from monolingual texts.

Most prior work either makes use of only one or two monolingual signals or uses unsupervised methods (like rank combination) to aggregate orthogonal signals (Schafer and Yarowsky, 2002; Klementiev and Roth, 2006). Surprisingly, no past research has employed *supervised* approaches to combine diverse monolingually-derived signals for bilingual lexicon induction. The field of machine learning has shown decisively that supervised models dramatically outperform unsupervised models, including for closely related problems like statistical machine translation (Och and Ney, 2002).

For the bilingual lexicon induction task, a supervised approach is natural, particularly because computing contextual similarity typically requires a seed bilingual dictionary (Rapp, 1995), and that same dictionary may be used for estimating the parameters of a model to combine monolingual signals. Alternatively, in a low resource machine translation (MT) setting, it is reasonable to assume a small amount of parallel data from which a bilingual dictionary can be extracted for supervision. In this setting, bilingual lexicon induction is critical for translating source words which do not appear in the parallel data or dictionary.

We frame bilingual lexicon induction as a binary classification problem; for a pair of source and target language words, we predict whether the two are translations of one another or not. For a given source language word, we score all target language candidates separately and then rerank them. We use a variety of signals derived from source and target

*Performed while faculty at Johns Hopkins University

Proceedings of NAACL-HLT 2013, pages 518–523,
Atlanta, Georgia, 9–14 June 2013.

monolingual corpora as features and use supervision to estimate the strength of each. In this work we:

- Use the following similarity metrics derived from monolingual corpora to score word pairs: contextual, temporal, topical, orthographic, and frequency.
- For the first time, explore using supervision to combine monolingual signals and learn a discriminative model for predicting translations.
- Present results for 22 low and high resource languages paired with English and show large accuracy gains over an unsupervised baseline.

2 Previous Work

Prior work suggests that a wide variety of monolingual signals, including distributional, temporal, topic, and string similarity, may inform bilingual lexicon induction (Rapp, 1995; Fung and Yee, 1998; Rapp, 1999; Schafer and Yarowsky, 2002; Schafer, 2006; Klementiev and Roth, 2006; Koehn and Knight, 2002; Haghighi et al., 2008; Mimno et al., 2009; Mausam et al., 2010). Klementiev et al. (2012) use many of those signals to score an existing phrase table for end-to-end MT but do not learn any new translations. Schafer and Yarowsky (2002) use an unsupervised rank-combination method for combining orthographic, contextual, temporal, and frequency similarities into a single ranking.

Recently, Ravi and Knight (2011), Dou and Knight (2012), and Nuhn et al. (2012) have worked toward learning a phrase-based translation model from monolingual corpora, relying on decipherment techniques. In contrast to that work, we use a seed bilingual lexicon for supervision and multiple monolingual signals proposed in prior work.

Haghighi et al. (2008) and Daumé and Jagarlamudi (2011) use some supervision to learn how to project contextual and orthographic features into a low-dimensional space, with the goal of representing words which are translations of one another as vectors which are close together in that space. However, both of those approaches focus on only two signals, high resource languages, and frequent words (frequent nouns, in the case of Haghighi et al. (2008)). In our classification framework, we can incorporate any number of monolingual signals, including contextual and string similarity, and directly learn how to combine them.

Language	#Words	Language	#Words
Nepali	0.4	Somali	0.5
Uzbek	1.4	Azeri	2.6
Tamil	3.7	Albanian	6.5
Bengali	6.6	Welsh	7.5
Bosnian	12.9	Latvian	40.2
Indonesian	21.8	Romanian	24.1
Serbian	25.8	Turkish	31.2
Ukrainian	37.6	Hindi	47.4
Bulgarian	49.5	Polish	104.5
Slovak	124.3	Urdu	287.2
Farsi	710.3	Spanish	972

Table 1: Millions of monolingual web crawl and Wikipedia word tokens

3 Monolingual Data and Signals

3.1 Data

Throughout our experiments, we seek to learn how to translate words in a given source language into English. Table 1 lists our languages of interest, along with the total amount of monolingual data that we use for each. We use web crawled time-stamped news articles to estimate temporal similarity, Wikipedia pages which are inter-lingually linked to English pages to estimate topic similarity, and both datasets to estimate frequency and contextual similarity. Following Irvine et al. (2010), we use pairs of Wikipedia page titles to train a simple transliterator for languages written in a non-Roman script, which allows us to compute orthographic similarity for pairs of words in different scripts.

3.2 Signals

Our definitions of orthographic, topic, temporal, and contextual similarity are taken from Klementiev et al. (2012), and the details of each may be found there. Here, we give briefly describe them and give our definition of a novel, frequency-based signal.

Orthographic We measure orthographic similarity between a pair of words as the normalized[1] edit distance between the two words. For non-Roman script languages, we transliterate words into the Roman script before measuring orthographic similarity.

Topic We use monolingual Wikipedia pages to estimate topical signatures for each source and target

[1]Normalized by the average of the lengths of the two words

language word. Signature vectors are the length of the number of inter-lingually linked source and English Wikipedia pages and contain counts of how many times the word appears on each page. We use cosine similarity to compare pairs of signatures.

Temporal We use time-stamped web crawl data to estimate temporal signatures, which, for a given word, are the length of the number of time-stamps (dates) and contain counts of how many times the word appears in news articles with the given date. We use a sliding window of three days and use cosine similarity to compare signatures. We expect that source and target language words which are translations of one another will appear with similar frequencies over time in monolingual data.

Contextual We score monolingual contextual similarity by first collecting context vectors for each source and target language word. The context vector for a given word contains counts of how many times words appear in its context. We use bag of words contexts in a window of size two. We gather both source and target language contextual vectors from our web crawl data and Wikipedia data (separately).

Frequency Words that are translations of one another are likely to have similar relative frequencies in monolingual corpora. We measure the frequency similarity of two words as the absolute value of the difference between the logs of their relative monolingual corpus frequencies.

4 Supervised Bilingual Lexicon Induction

4.1 Baseline

Our unsupervised baseline method is based on ranked lists derived from each of the signals listed above. For each source word, we generate ranked lists of English candidates using the following six signals: Crawls Context, Crawls Time, Wikipedia Context, Wikipedia Topic, Edit distance, and Log Frequency Difference. Then, for each English candidate we compute its mean reciprocal rank[2] (MRR) based on the six ranked lists. The baseline ranks English candidates according to the MRR scores. For evaluation, we use the same test sets, accuracy metric, and correct translations described below.

[2]The MRR of the jth English word, e_j, is $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_{ij}}$, where N is the number of signals and $rank_{ij}$ is e_j's rank according to signal i.

4.2 Supervised Approach

In addition to the monolingual resources described in Section 3.1, we have a bilingual dictionary for each language, which we use to project context vectors and for supervision and evaluation. For each language, we choose up to $8,000$ source language words among those that occur in the monolingual data at least three times and that have at least one translation in our dictionary. We randomly divide the source language words into three equally sized sets for training, development, and testing. We use the training data to train a classifier, the development data to choose the best classification settings and feature set, and the test set for evaluation.

For all experiments, we use a linear classifier trained by stochastic gradient descent to minimize squared error[3] and perform 100 passes over the training data.[4] The binary classifiers predict whether a pair of words are translations of one another or not. The translations in our training data serve as positive supervision, and the source language words in the training data paired with random English words[5] serve as negative supervision. We used our development data to tune the number of negative examples to three for each positive example. At test time, after scoring all source language words in the test set paired with all English words in our candidate set,[6] we rank the English candidates by their classification scores and evaluate accuracy in the top-k translations.

4.3 Features

Our monolingual features are listed below and are based on raw similarity scores as well as ranks:

- Crawls Context: Web crawl context similarity score
- Crawls Context RR: reciprocal rank of crawls context

[3]We tried using logistic rather than linear regression, but performance differences on our development set were very small and not statistically significant.

[4]We use `http://hunch.net/~vw/` version 6.1.4, and run it with the following arguments that affect how updates are made in learning: –exact adaptive norm –power t 0.5

[5]Among those that appear at least five times in our monolingual data, consistent with our candidate set.

[6]All English words appearing at least five times in our monolingual data. In practice, we further limit the set to those that occur in the top-1000 ranked list according to at least one of our signals.

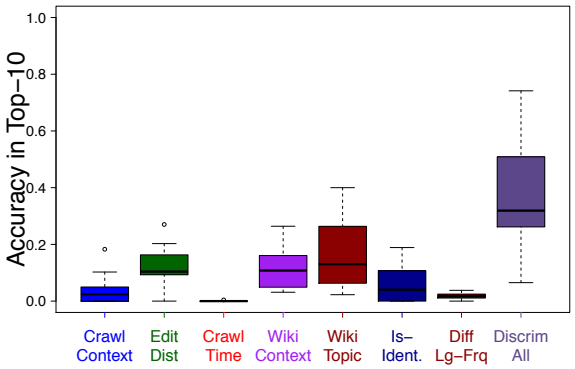

Figure 1: Each box-and-whisker plot summarizes performance on the development set using the given feature(s) across all 22 languages. For each source word in our development sets, we rank all English target words according to the monolingual similarity metric(s) listed. All but the last plot show the performance of individual features. *Discrim-All* uses supervised data to train classifiers for each language based on all of the features.

- Crawls Time: Web crawl temporal similarity score
- Crawls Time RR: reciprocal rank of crawls time
- Edit distance: normalized (by average length of source and target word) edit distance
- Edit distance RR: reciprocal rank of edit distance
- Wiki Context: Wikipedia context similarity score
- Wiki Context RR: recip. rank of wiki context
- Wiki Topic: Wikipedia topic similarity score
- Wiki Topic RR: recip. rank of wiki topic
- Is-Identical: source and target words are the same
- Difference in log frequencies: Difference between the logs of the source and target word monolingual frequencies
- Log Freqs Diff RR: reciprocal rank of difference in log frequencies

We train classifiers separately for each source language, and the learned weights vary based on, for example, corpora size and the relatedness of the source language and English (e.g. edit distance is informative if there are many cognates). In order to use the trained classifiers to make top-k translation predictions for a given source word, we rank candidates by their classification scores.

4.4 Feature Evaluation and Selection

After training initial classifiers, we use our development data to choose the most informative subset of features. Figure 1 shows the top-10 accuracy on the development data when we use individual features

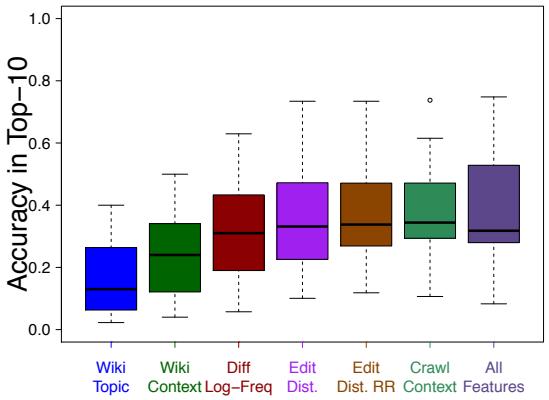

Figure 2: Performance on the development set goes up as features are greedily added to the feature space. Mean performance is slightly higher using this subset of six features (second to last bar) than using all features (last bar).

to predict translations. Top-10 accuracy refers to the percent of source language words for which a correct English translation appears in the top-10 ranked English candidates. Each box-and-whisker plot summarizes performance over the 22 languages. We don't display reciprocal rank features, as their performance is very similar to that of the corresponding raw similarity score. It's easy to see that features based on the Wikipedia topic signal are the most informative. It is also clear that training a supervised model to combine all of the features (the last plot) yields performance that is dramatically higher than using any individual feature alone.

Figure 2, from left to right, shows a greedy search for the best subset of features among those listed above. Again, the Wikipedia topic score is the most informative stand-alone feature, and the Wikipedia context score is the most informative second feature. Adding features to the model beyond the six shown in the figure does not yield additional performance gains over our set of languages.

4.5 Learning Curve Analysis

Figure 3 shows learning curves over the number of positive training instances. In all cases, the number of randomly generated negative training instances is three times the number of positive. For all languages, performance is stable after about 300 correct translations are used for training. This shows that our supervised method for combining signals requires only a small training dictionary.

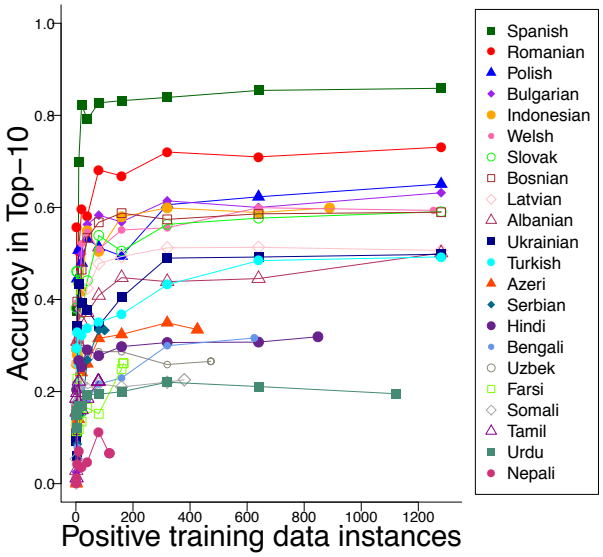

Figure 3: Learning curves over number of positive training instances, up to 1250. For some languages, 1250 positive training instances are not available. In all cases, evaluation is on the development data and the number of negative training instances is three times the number of positive. For all languages, performance is fairly stable after about 300 positive training instances.

5 Results

We use a model based on the six features shown in Figure 2 to score and rank English translation candidates for the test set words in each language. Table 2 gives the result for each language for the MRR baseline and our supervised technique. Across languages, the average top-10 accuracy using the MRR baseline is 30.4, and the average using our technique is 43.9, a relative improvement of about *44%*. We did not attempt a comparison with more sophisticated unsupervised rank aggregation methods. However, we believe the improvements we observe drastically outweigh the expected performance differences between different rank aggregation methods. Figure 4 plots the accuracies yielded by our supervised technique versus the total amount of monolingual data for each language. An increase in monolingual data tends to improve accuracy. The correlation isn't perfect, however. For example, performance on Urdu and Farsi is relatively poor, despite the large amounts of monolingual data available for each. This may be due to the fact that we have large web crawls for those languages, but their Wikipedia datasets, which tend to provide a strong topic signal, are relatively small.

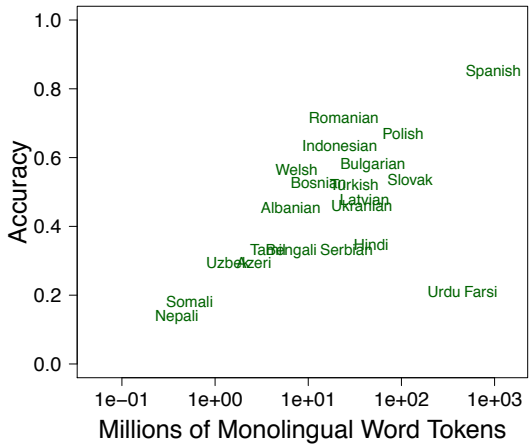

Figure 4: Millions of monolingual word tokens vs. Lexicon Induction Top-10 Accuracy

Lang	MRR	Supv.	Lang	MRR	Supv.
Nepali	11.2	13.6	Somali	16.7	18.1
Uzbek	23.2	29.6	Azeri	16.1	29.4
Tamil	28.4	33.3	Albanian	32.0	45.3
Bengali	19.3	32.8	Welsh	36.1	56.4
Bosnian	32.6	52.8	Latvian	29.6	47.7
Indonesian	41.5	63.5	Romanian	53.3	71.6
Serbian	29.0	33.3	Turkish	31.4	52.1
Ukrainian	29.7	46.0	Hindi	18.2	34.6
Bulgarian	40.2	57.9	Polish	47.4	67.1
Slovak	34.6	53.5	Urdu	13.2	21.2
Farsi	10.5	21.1	Spanish	74.8	85.0

Table 2: Top-10 Accuracy on test set. Performance increases for all languages moving from the baseline (*MRR*) to discriminative training (*Supv*).

6 Conclusions

On average, we observe relative gains of more than 44% over an unsupervised rank-combination baseline by using a seed bilingual dictionary and a diverse set of monolingual signals to train a supervised classifier. Using supervision for bilingual lexicon induction makes sense. In some cases a dictionary is already assumed for computing contextual similarity, and, in the remaining cases, one could be compiled easy, either automatically, e.g. Haghighi et al. (2008), or through crowdsourcing, e.g. Irvine and Klementiev (2010) and Callison-Burch and Dredze (2010). We have shown that only a few hundred translation pairs are needed to achieve good performance. Our framework has the additional advantage that any new monolingually-derived similarity metrics can easily be added as new features.

7 Acknowledgements

This material is based on research sponsored by DARPA under contract HR0011-09-1-0044 and by the Johns Hopkins University Human Language Technology Center of Excellence. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Hal Daumé, III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Ann Irvine and Alexandre Klementiev. 2010. Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, and Jeff Bilmes. 2010. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174:619–637, June.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Charles Schafer. 2006. *Translation Discovery Using Diverse Similarity Measures*. Ph.D. thesis, Johns Hopkins University.