Crosslingual Distributed Representations for Translation Lexicon Induction

Abstract

Using distributed representations of words instead of treating them as atomic units have been shown to alleviate data sparsity problems common to natural language processing tasks. Their primary appeal is that they can be induced from cheap and plentiful unannotated data. Recent work induces distributed representations for the crosslingual setting, in which some parallel data is also available, and makes them particularly relevant for low-resource machine translation. In this work, we induce crosslingual distributed representations for a set of truly low resource languages and English, and use them for translation lexicon induction. We demonstrate them to be dramatically more informative than the standard vector-space approach, which uses the same learning signals.

1 Introduction

Inducing translations lexicons from monolingual data has a long history in natural language processing literature (**?**). These techniques are usually motivated by their use in statistical machine translation, especially for low-resource languages where sufficient amounts of expensive parallel is unavailable, so that other resources must be used to induce and score word and phrase translations. Recently, these techniques have been shown effective for the full phrase-base MT pipeline (Koehn et al., 2003), in which most of the parameters were estimated from monolingual data (Klementiev et al., 2012a) instead of bitexts.

Using distributed representations of words instead of treating them as atomic units have been shown to alleviate data sparsity problems common to natural language processing tasks. Their primary appeal is that they can be induced from a large cheap unannotated corpus. Recent work (Klementiev et al., 2012b) proposes to induce distributed representations for the crosslingual setting, in which some parallel data is also available. Semantically similar words in the induced embedding end up "close" to one another irrespective of the language. This set-up is particularly relevant to a realistic low-resource machine translation set-up: along with plentiful monolingual data, small amounts of parallel data are typically also available or could be annotated cheaply (Post et al., 2012).

In this paper, we propose to use crosslingual distributed representations for translation lexicon induction for a truly low-resource translation setting. First, we follow the setup of (Klementiev et al., 2012b) and induce crosslingual embedding for English-Tamil, English-Bengali, and English-Hindi. Unlike their experiments on English-German, our language pairs have relatively little available parallel data. However, we show that that the induced representations are still informative for lexicon induction. Next, we use a small set of translation pairs and induce a distance metric over the embedding specifically for lexicon induction. Finally, we compare our results with a variant of the standard vector-space technique (**?**), which uses contextual information and a bilingual dictionary to induce translation lexicons. While

it makes use of the same set of signals as the distributed representation approach, it represents words with large (on the order of the vocabulary size) heuristically induced feature vectors.

In sum, the main contributions of this work are:

- We begin by inducing crosslingual distributed representations for three pairs of languages: English-Tamil, English-Bengali, and English-Hindi. We follow the recent work of Klementiev et al. (2012b), however, our set-up is truly low resource: each pair has a relatively small amount of parallel data.

- We use the induced representations for the task of translation lexicon induction. With a small set of translations extracted from parallel data, we learn a metric over the induced embedding, and use it to select translations for a large vocabulary.

- We experimentally demonstrate dramatic performance improvements over the standard vector-space based approach, which uses the same set of signals to induce translations.

2 Crosslingual Distributed Representations

We begin with a brief overview of the cross-lingual distributed representation setup of Klementiev et al. (2012b); we use features based on these representations in our translation lexicon induction experiments in Section **??**.

Their approach induces the *same* embedding for words of both languages so that semantically similar words end up "close" to each other irrespective of the language. They use large unannotated monolingual corpora to simultaneously induce representations for words within each language and parallel data to bring them together across languages. The intuition for their approach to crosslingual representation induction comes from the multitask learning setup of Cavallanti et al. (2010). The goal of multitask learning (MTL) is to learn a set of related tasks jointly exploiting learning signals across the tasks. In MTL terms, when inducing crosslingual representations, Klementiev et al. (2012b) treat each word w in languages' vocabularies as an individual task. Tasks related to w are then defined as its possible translations in the other language. They extract sets of related tasks and the "degree of relatedness" between them from co-occurrence statistics in a parallel corpus.

They apply this set-up to a variant of neural probabilistic language model (Bengio et al., 2003). Along with other model parameters W, these models learn a latent d-dimensional representation $c \in \mathbb{R}^{d|V|}$ of all words in a language vocabulary V and use it to estimate conditional probabilities of the next word w_t given n words preceeding it in text $\hat{P}(w_t|w_{t-n+1:t-1})$. An important property of the induced embedding c is that it captures semantic and syntactic similarity of words in a language: similar words end up "close" to each other in c. Klementiev et al. (2012b) train two neural language models for a pair of languages jointly and use the MTL set-up to ensure that the similarity property holds across languages in the induced embedding c. More formally, they optimize the following objective:

$$L(\theta^{(1,2)}) = \sum_{l=1}^{2}\sum_{t=1}^{T^{(l)}} \log \hat{P}_{\theta^{(l)}}(w_t^{(l)}|w_{t-n+1:t-1}^{(l)}) + \frac{1}{2} c^\top (A \otimes I_m) c,$$

where $\theta^{(l)} = (W^{(l)}, c)$ include neural language model parameters $W^{(l)}$ as well as the shared representation c, $\otimes$ is the Kronecker product and I_m is the identity matrix of size m.

The first summand is the log-likelihood of the texts $(w_1^{(l)}, w_2^{(l)}, \ldots w_{T^{(l)}}^{(l)})$ of length $T^{(l)}$ for each language l. This language modeling part of the objective ensures that embedding c maps similar words close to one another within each language (see Bengio et al. (2003)). The second part of the objective is the MTL regularizer ensuring that the same property also holds across the two languages languages. The interaction matrix A encodes the degree of relatedness between words and their translations. It is defined using word alignments in a parallel corpus: the more frequently a pair of words is aligned the better they fit as translations.

	Tamil	Bengali	Hindi
Monolingual	4.5m	5.9m	24.4m
Training	452k	272k	708k
OOV Rate	44%	37%	34%

Table 1: Information about our monolingual and parallel datasets for each language. Monolingual gives the millions of monolingual word tokens that we use to induce distributed word representations and baseline contextual vectors for each language. Training data gives the number of thousand of words in the source language training set provided by Post et al. (2012). OOV rates give the percent of development set word types (our test set for bilingual lexicon induction) that do not appear in the training data.

The language models are leaned jointly from unannotated texts in both languages using stochastic gradient descent. When an update is made for a representation of a word in one language, some of it is also propagated to the representations of all words related to it (i.e. it's translations).

Klementiev et al. (2012b) show that the induced embedding is very informative for crosslingual document classification, where a classifier trained with word representations as features on annotation available for one language we used in another languages directly.

In this work, we follow their setup and induce distributed crosslingual representations, learn a distance function over the embedding, and use it to select translation candidates. TODO: Separate into its own section? And say more about how we do it.

3 Additional Related Work

TODO: Include Ryan's paper on direct annotation transfer: multilingual clusters for dependency parsing

4 Experiments

4.1 Data

Post et al. (2012) used Amazon's Mechanical Turk to collect small parallel datasets for several low resource Indian languages. We use those datasets in our experiments here in order to test our models in a realistic low resource machine translation setting. In particular, we use automatic alignments over the training datasets to derive seed bilingual dictionaries, which both the vector-space contextual baseline and our model for learning distributed word representations utilize. Then, we evaluate the translations that we induce over the set of all word types in the development sets released by Post et al. (2012). This setting uses not only realistic seed bilingual dictionaries, but also a realistic diversity of source language word types in our test sets. The test sets includes words of all parts of speech, words that appear in the training data as well as unknown words, and words that have both high and low monolingual frequencies. A lot of the prior work in bilingual lexicon induction only seeks to translate high frequency words (**?**), and, in some cases, only high frequency nouns (Koehn and Knight, 2002; Haghighi et al., 2008).

We induce distributed representations and contextual vectors (see Section 4.2) using both web crawl and Wikipedia monolingual data as well as the training datasets for each language. For all languages, we subsample our English web crawl and Wikipedia data to roughly equal the amount of monolingual source language data. Table 1 gives statistics about our datasets. In all experiments, we induce 80-dimensional distributed representations for each source and target language word.

As one baseline comparison, we evaluate the performance of the seed bilingual dictionary itself, which is based on training data alignments that informed our baseline and proposed models. For this baseline, we rank English candidates according to the number of times each was aligned to a given source word in the training set. There is overlap between our seed bilingual dictionary and our test set because we would like to simulate a real MT setting. The seed bilingual dictionary is based on word alignments, so it is likely to be noisy and incomplete. That is, in a real MT setting, we may have a need to induce translations for all words, even if they appear in the seed bilingual dictionary. Because of this experimental design, we use the seed bilingual dictionary itself as a baseline comparison.

4.2 Vector-space Contextual Baseline

We use contextual similarity, as first proposed by Fung and Yee (1998) and Rapp (1999), as a baseline comparison. Under this measure, for each source language word s_i, we collect a vector of counts, c_{s_i}, of how many times each source language word appears in the context of word s_i. The size of c_{s_i} is the size of the source language vocabulary. In our experiments, we use bag of words contexts in a window size of two (two words to the left and s_i and two words to the right). Similarly, we collect context vectors for all target language words, t_j. We use the seed dictionary defined by the aligned training set to project the source language context vectors into the space of the target language context vectors. We use cosine similarity to compare all pairwise contextual vectors and then rank English words for each source language word in our test set.

4.3 Results

We use the translations derived by automatically aligning our development sets in order to evaluate our induced translations. Because our datasets are small and the alignments over them are sparse and noisy, we supplemented the alignments-based dictionaries with existing digital bilingual dictionaries for each language in order to improve the coverage of our gold standard dictionary.

Table 2 shows performance on the lexicon induction task. The seed dictionary baseline is the performance of the dictionary derived from the intersection alignments over the training data alone, which is used as supervision to both the old contextual scorer and the distributed representations learner. The fact that the accuracy using the seed dictionary alone is so low speaks to how noisy the alignments are and how limited the training data is. The context baseline uses the same seed dictionary to project context vectors. Its accuracy in the top-1 and top-10 ranked words is very low, but its top-100 accuracy is quite high. The bag of words context vectors are noisy and the corresponding signal that words are translations of one another is weak. Although it is not precise, it does seem to be the case that most translation pairs tend to have somewhat similar context vectors. The translations that we induce using our distributed representations far outperform both baselines in their top-1 and top-10 rankings. Not surprisingly, high precision...

	Top-1	Top-10	Top-100
Tamil			
Seed dict baseline	6.70	9.58	9.60
Context baseline	2.32	8.38	25.44
Distrib Rep L2 Dist	15.50	17.77	20.44
Bengali			
Seed dict baseline	8.60	11.39	11.39
Context baseline	3.91	12.39	30.53
Distrib Rep L2 Dist	24.01	25.86	28.01
Hindi			
Seed dict baseline	13.51	18.38	18.38
Context baseline	5.22	14.72	34.31
Distrib Rep L2 Dist	33.93	37.64	42.00

Table 2: Comparison of performance of old definition of contextual similarity with new distributed representations model

5 Conclusions

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3:1137–1155.

Giovanni Cavallanti, Nicoló Cesa-bianchi, and Claudio Gentile. 2010. Linear algorithms for online multitask classification. *Journal of Machine Learning Research (JMLR)*, 11:2901–2934.

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)*.

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012a. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, April. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012b. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bombay, India, December.

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

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL/HLT)*, Vancouver, Canada.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June. Association for Computational Linguistics.

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)*.