Crosslingual Distributed Representations for Translation Lexicon Induction

Abstract

Using distributed representations of words instead of treating them as atomic units has been shown to alleviate data sparsity problems common to natural language processing tasks. Unlike standard contextual vectors that represent words as vectors of counts of nearby words, distributed representations are low-dimensional representations of words that are induced iteratively using the distributed representations of nearby words. Such representations can be induced from plentiful monolingual data. Recently, Klementiev et al. (2012b) made use of parallel data to induce *crosslingual* distributed representations in which the embedding that defines the representations is learned jointly across languages. We extend that work to induce crosslingual distributed representations for a set of 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 translation lexicons from monolingual data has a long history in natural language processing literature (Rapp, 1995; Fung and Yee, 1998; Schafer and Yarowsky, 2002; Koehn and Knight, 2002). These techniques are usually motivated by their use in machine translation, especially for low-resource languages where suitable training data are hard to come by. In such cases, other resources must be used to induce and score word and phrase translations. For example, Klementiev et al. (2012a) used bilingual lexicon induction techniques to estimate most of the parameters of phrase-based MT (Koehn et al., 2003) from monolingual instead of parallel data.

Using distributed representations of words instead of treating them as atomic units has been shown to alleviate data sparsity problems. Low-dimensional distributed representations for words are learned not from nearby words themselves but from the distributed representations of nearby words. These representations are induced iteratively. Recently, Klementiev et al. (2012b) induced distributed representations for the crosslingual setting using parallel data in addition to large monolingual corpora. There, the induced embedding is learned jointly over multiple languages so that the representations of semantically similar words end up "close" to one another irrespective of language. This setting is particularly relevant to low-resource machine translation, where, along with plentiful monolingual data, small amounts of parallel data are often available or could be created cheaply (Post et al., 2012).

In this paper, we propose to use crosslingual distributed representations for translation lexicon induction in a low-resource translation setting. We follow the setup of Klementiev et al. (2012b) to induce crosslingual embeddings for English-Tamil, English-Bengali, and English-Hindi. Unlike their English-German experiments, our language pairs have little available parallel data. However, we show that that the induced representations are still informative for lexicon induction. We compare our results with a variant of the standard vector-space projection technique (Fung and Yee, 1998), which uses contextual information and a bilingual dictionary to induce translation lexicons. While it makes use of the same signals as the distributed representation approach, it represents words with large (on the order of the vocabulary size) feature vectors.

In sum, the main contributions of this work are:

- We induce crosslingual distributed representations for three low-resource language pairs: English-Tamil, English-Bengali, and English-Hindi.
- We experimentally demonstrate dramatic performance improvements over the standard vector-space based approach, which uses the same 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). In Section 4 we use these representations to induce translation lexicons.

The approach described in Klementiev et al. (2012b) induces the *same* embedding for words in a pair of languages so that semantically similar words end up "close" to each other irrespective of their language. They simultaneously use large monolingual corpora to induce representations for words in each language and use parallel data to bring the representations 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 by exploiting learning signals across the tasks. In MTL terms, when inducing crosslingual representations, Klementiev et al. (2012b) treat each word w as an individual task. Tasks related to w are then defined as the set of its possible translations in the other language. They extract sets of related tasks and the "degree of relatedness" between them from an aligned parallel corpus.

They apply this set-up to a variant of a 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, $\hat{P}(w_t|w_{t-n+1:t-1})$, of the next word, w_t, given the n words preceding it in text. An important property of the induced embedding c is that it captures the 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:

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

where $\theta^{(l)} = (W^{(l)}, c)$ includes 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 is the MTL regularizer ensuring that the same property also holds across the two 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 more likely they are translations.

The language models are leaned jointly from monolingual texts in each language 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 informative for crosslingual document classification, where the classifier uses word representations as features. In this work, we follow their setup to induce distributed crosslingual representations. In all experiments, we induce 80-dimensional distributed representations for each source and target language word and use Euclidean distance to rank English words for each source language word in our test set.

3 Additional Related Work

Täckström et al. (2012) propose a related technique for inducing crosslingual word clusters, which they use for direct transfer of delexicalized parsers and named-entity recognizers. That work also uses aligned parallel data to encourage crosslingual consistency over induced word representations. Although their induced clusters are informative for these downstream applications, these representations commit to a particular clustering and thus do not handle multiple word senses, which less applicable to lexicon induction.

4 Experiments

4.1 Vector-space Contextual Baseline

We use contextual similarity, first proposed by Fung and Yee (1998) and Rapp (1999), as a baseline method for inducing translation pairs. 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. We use bag of words contexts in a window of two words (two words to the left of s_i and two words to the right). Similarly, we collect context vectors for all target language words, t_j. We use a seed bilingual dictionary 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.

4.2 Data

Post et al. (2012) used crowdsourcing 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 intersection alignments over the training datasets to inform both the baseline and our proposed model. The contextual baseline uses a seed bilingual dictionary based on the aligned training data to project context vectors (see Section 4.1), and the distributed representations learner uses them to define the interaction matrix A (see Section 2). 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 test set word types. 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).

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

Table 1: Dataset statistics 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 released 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.

We induce distributed representations and contextual vectors (see Section 4.1) using web crawl and Wikipedia monolingual data as well as each side of the training data for each language. For all languages, we subsample our English web crawl data to roughly equal the amount of monolingual source language data and use the English Wikipedia pages which have an interlanguage link to a source language page. Table 1 gives statistics about our datasets.

As an additional baseline, we evaluate the performance of the seed bilingual dictionary itself, which is based on training data alignments. 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. In a real MT setting, a grammar extracted from the same aligned data would be similarly noisy and incomplete, and, thus, we may want to induce translations for all words, even if they appear in the training data. We hope that our induced translations will preserve the correct translation pairs in the seed dictionary, forget the incorrect pairs, and induce good translations for additional words.

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

4.3 Results

Table 2 shows performance on the lexicon induction task. 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 the signal is not precise, it does seem 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. While learning an embedding, the representations of words in the preceding n-gram context (along with representations of their translations) are modified based on the next word in the training sequence. So, the induced representations tend to capture syntactic and semantic information predictive of the next word in a sequence. On the other hand, bag-of-words counts of words appearing before and after a given word are used to induce vector-space contextual representations. Thus, it is not surprising that while contextual vectors have a high accuracy in their top-100 rankings, their precision in the top-1 and top-10 is substantially worse compared to the ranking based on distributed representations, which take advantage of finer-grained contextual information. When using the induced lexicons for machine translation, we are particularly interested in good precision for highly ranked translation candidates. In the context of phrase based MT, that would mean smaller and more accurate phrase tables and better estimates of phrase pair features.

	Top-1	Top-10	Top-100
Tamil			
Align-based dict	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			
Align-based dict	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			
Align-based dict	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: Accuracy of standard contextual similarity vs new distributed representations model

5 Conclusions

In this paper, we have built upon the recent work of Klementiev et al. (2012b). We have shown that using their model for inducing crosslingual distributed word representations is effective even in the setting in which we have a limited amount of parallel data from which to inform the interaction matrix, which ensures that the induced embedding maps similar words close to one another across languages. Using this method to induce high precision translations dramatically outperforms the standard contextual vector approach.

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)*.

Nikesh Garera and David Yarowsky. 2008. Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 403–410.

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. 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)*.

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)*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada, June.