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 also available or could be annotated cheaply (**?**).

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 relatively small amount of parallel data.
- We apply the induced 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 the these representations in our translation lexicon induction experiments in Section 4.

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 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.

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 Vector-space Cntextual Baseline

4.2 Results

Table 1 shows performance on the lexicon induction task. The alignment dictionary score 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 alignment based dictionary alone is so low speaks to how noisy the alignments are and how limited the training data is. The old contextual score uses the same dictionary based on the intersection alignments over the training data for each language to project context vectors. The distributed representations use an interaction matrix defined also by the intersection alignments over the training data for each language. Both models use the same tokenization of all of the monolingual data that we have available for each language, which is taken from web crawls and Wikipedia. Evaluation is over *all word types* in the development set for each language.

	Top-1	Top-10	Top-100
Tamil			
Intersection Train Dict	6.70	9.58	9.60
Old-Contextual	2.32	8.38	25.44
Distrib Rep L2 Dist	15.50	17.77	20.44
Distrib Rep Learn Dist			
Bengali			
Intersection Train Dict	8.60	11.39	11.39
Old-Contextual	3.91	12.39	30.53
Distrib Rep L2 Dist	24.01	25.86	28.01
Distrib Rep Learn Dist			
Hindi			
Intersection Train Dict	13.51	18.38	18.38
Old-Contextual	5.22	14.72	34.31
Distrib Rep L2 Dist	33.93	37.64	42.00
Distrib Rep Learn Dist			

Table 1: 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.

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, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Pro-*

ceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL/HLT), Vancouver, Canada.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 320–322.