Beyond Bilingual: Multi-sense Word Embeddings using Multilingual Context

Shyam Upadhyay[1] **Kai-Wei Chang**[2] **Matt Taddy**[3] **Adam Kalai**[3] **James Zou**[4]

[1]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2]University of Virginia, Charlottesville, VA, USA
[3]Microsoft Research, Cambridge, MA, USA
[4]Stanford University, Stanford, CA, USA
upadhya3@illinois.edu, kw@kwchang.net
{taddy,adum}@microsoft.com, jamesyzou@gmail.com

Abstract

Word embeddings, which represent a word as a point in a vector space, have become ubiquitous to several NLP tasks. A recent line of work uses bilingual (two languages) corpora to learn a different vector for each sense of a word, by exploiting crosslingual signals to aid sense identification. We present a multi-view Bayesian non-parametric algorithm which improves multi-sense word embeddings by (a) using multilingual (i.e., more than two languages) corpora to significantly improve sense embeddings beyond what one achieves with bilingual information, and (b) uses a principled approach to learn a variable number of senses per word, in a data-driven manner. Ours is the first approach with the ability to leverage multilingual corpora efficiently for multi-sense representation learning. Experiments show that multilingual training significantly improves performance over monolingual and bilingual training, by allowing us to combine different parallel corpora to leverage multilingual context. Multilingual training yields comparable performance to a state of the art monolingual model trained on five times more training data.

1 Introduction

Word embeddings (Turian et al., 2010; Mikolov et al., 2013, *inter alia)* represent a word as a point in a vector space. This space is able to capture semantic relationships: vectors of words with similar meanings have high cosine similarity (Turney, 2006; Turian et al., 2010). Use of embeddings as features has been shown to benefit several NLP tasks and serve as good initializations for deep architectures ranging from dependency parsing (Bansal et al., 2014) to named entity recognition (Guo et al., 2014b).

Although these representations are now ubiquitous in NLP, most algorithms for learning word-embeddings do not allow a word to have different meanings in different contexts, a phenomenon known as polysemy. For example, the word *bank* assumes different meanings in financial (eg. "bank pays interest") and geographical contexts (eg. "river bank") and which cannot be represented adequately with a single embedding vector. Unfortunately, there are no large sense-tagged corpora available and such polysemy must be inferred from the data during the embedding process.

I got high **interest** on my savings from the bank.	Je suis un grand **[intérêt]** sur mes économies de la banque.	我得到了我的储蓄从银行高**[利息]**。
My **interest** lies in History.	Mon **[intérêt]** réside dans l'Histoire.	我的**[兴趣]**在于历史。

Figure 1: **Benefit of Multilingual Information (beyond bilingual)**: Two different senses of the word "interest" and their translations to French and Chinese (word translation shown in **[bold]**). While the surface form of both senses are same in French, they are different in Chinese.

Several attempts (Reisinger and Mooney, 2010; Neelakantan et al., 2014; Li and Jurafsky, 2015) have been made to infer multi-sense word representations by modeling the sense as a latent variable in a Bayesian non-parametric framework. These approaches rely on the "one-sense per collocation" heuristic (Yarowsky, 1995), which assumes that presence of nearby words correlate with the sense of the word of interest. This heuristic provides only a weak signal for sense identification, and such algorithms require large amount of training data to achieve competitive perfor-

Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 101–110,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

mance.

Recently, several approaches (Guo et al., 2014a; Šuster et al., 2016) propose to learn multi-sense embeddings by exploiting the fact that different senses of the same word may be translated into different words in a foreign language (Dagan and Itai, 1994; Resnik and Yarowsky, 1999; Diab and Resnik, 2002; Ng et al., 2003). For example, *bank* in English may be translated to *banc* or *banque* in French, depending on whether the sense is financial or geographical. Such bilingual distributional information allows the model to identify which sense of a word is being used during training.

However, bilingual distributional signals often do not suffice. It is common that polysemy for a word survives translation. Fig. 1 shows an illustrative example – both senses of *interest* get translated to *intérêt* in French. However, this becomes much less likely as the number of languages under consideration grows. By looking at Chinese translation in Fig. 1, we can observe that the senses translate to different surface forms. Note that the opposite can also happen (i.e. same surface forms in Chinese, but different in French). Existing crosslingual approaches are inherently bilingual and cannot naturally extend to include additional languages due to several limitations (details in Section 4). Furthermore, works like (Šuster et al., 2016) sets a fixed number of senses for each word, leading to inefficient use of parameters, and unnecessary model complexity.[1]

This paper addresses these limitations by proposing a multi-view Bayesian non-parametric word representation learning algorithm which leverages multilingual distributional information. Our representation learning framework is the first multilingual (not bilingual) approach, allowing us to utilize arbitrarily many languages to disambiguate words in English. To move to multilingual system, it is necessary to ensure that the embeddings of each foreign language are relatable to each other (i.e., they live in the same space). We solve this by proposing an algorithm in which word representations are learned *jointly* across languages, using English as a bridge. While large parallel corpora between two languages are scarce, using our approach we can concatenate multiple parallel corpora to obtain a large multilingual corpus. The parameters are estimated in a Bayesian nonparametric framework that allows our algorithm to only associate a word with a new sense vector when evidence (from either same or foreign language context) requires it. As a result, the model infers different number of senses for each word in a data-driven manner, avoiding wasting parameters.

Together, these two ideas – multilingual distributional information and nonparametric sense modeling – allow us to disambiguate multiple senses using far less data than is necessary for previous methods. We experimentally demonstrate that our algorithm can achieve competitive performance after training on a small multilingual corpus, comparable to a model trained monolingually on a much larger corpus. We present an analysis discussing the effect of various parameters – choice of language family for deriving the multilingual signal, crosslingual window size etc. and also show qualitative improvement in the embedding space.

2 Related Work

Work on inducing multi-sense embeddings can be divided in two broad categories – two-staged approaches and joint learning approaches. Two-staged approaches (Reisinger and Mooney, 2010; Huang et al., 2012) induce multi-sense embeddings by first clustering the contexts and then using the clustering to obtain the sense vectors. The contexts can be topics induced using latent topic models(Liu et al., 2015a,b), or Wikipedia (Wu and Giles, 2015) or coarse part-of-speech tags (Qiu et al., 2014). A more recent line of work in the two-staged category is that of retrofitting (Faruqui et al., 2015; Jauhar et al., 2015), which aims to infuse semantic ontologies from resources like WordNet (Miller, 1995) and Framenet (Baker et al., 1998) into embeddings during a post-processing step. Such resources list (albeit not exhaustively) the senses of a word, and by retro-fitting it is possible to tease apart the different senses of a word. While some resources like WordNet (Miller, 1995) are available for many languages, they are not exhaustive in listing all possible senses. Indeed, the number senses of a word is highly dependent on the task and cannot be pre-determined using a lexicon (Kilgarriff, 1997). Ideally, the senses should be inferred in a data-driven manner, so that new senses not listed in such lexicons can be discovered. While re-

[1] Most words in conventional English are monosemous, i.e. single sense (eg. the word *monosemous*)

cent work has attempted to remedy this by using parallel text for retrofitting sense-specific embeddings (Ettinger et al., 2016), their procedure requires creation of *sense graphs*, which introduces additional tuning parameters. On the other hand, our approach only requires two tuning parameters (prior α and maximum number of senses T).

In contrast, joint learning approaches (Neelakantan et al., 2014; Li and Jurafsky, 2015) jointly learn the sense clusters and embeddings by using non-parametrics. Our approach belongs to this category. The closest non-parametric approach to ours is that of (Bartunov et al., 2016), who proposed a multi-sense variant of the skip-gram model which learns the different number of sense vectors for all words from a large monolingual corpus (eg. English Wikipedia). Our work can be viewed as the multi-view extension of their model which leverages both monolingual and crosslingual distributional signals for learning the embeddings. In our experiments, we compare our model to monolingually trained version of their model.

Incorporating crosslingual distributional information is a popular technique for learning word embeddings, and improves performance on several downstream tasks (Faruqui and Dyer, 2014; Guo et al., 2016; Upadhyay et al., 2016). However, there has been little work on learning multi-sense embeddings using crosslingual signals (Bansal et al., 2012; Guo et al., 2014a; Šuster et al., 2016) with only (Šuster et al., 2016) being a joint approach. (Kawakami and Dyer, 2015) also used bilingual distributional signals in a deep neural architecture to learn context dependent representations for words, though they do not learn separate sense vectors.

3 Model Description

Let $E = \{x^e_1, .., x^e_i, .., x^e_{N_e}\}$ denote the words of the English side and $F = \{x^f_1, .., x^f_i, .., x^f_{N_f}\}$ denote the words of the foreign side of the parallel corpus. We assume that we have access to word alignments $A_{e\rightarrow f}$ and $A_{f\rightarrow e}$ mapping words in English sentence to their translation in foreign sentence (and vice-versa), so that x^e and x^f are aligned if $A_{e\rightarrow f}(x^e) = x^f$.

We define Nbr(x, L, d) as the neighborhood in language L of size d (on either side) around word x in its sentence. The English and foreign neighboring words are denoted by y^e and y^f, respectively. Note that y^e and y^f need not be translations of each other. Each word x^f in the foreign vocabulary is associated with a dense vector $\boldsymbol{x}^f$ in $\mathbb{R}^m$, and each word x^e in English vocabulary admits at most T sense vectors, with the k^{th} sense vector denoted as $\boldsymbol{x}^e_k$.[2] As our main goal is to model multiple senses for words in English, we do not model polysemy in the foreign language and use a single vector to represent each word in the foreign vocabulary.

We model the joint conditional distribution of the context words y^e, y^f given an English word x^e and its corresponding translation x^f on the parallel corpus:

$$P(y^e, y^f \mid x^e, x^f; \alpha, \theta), \tag{1}$$

where θ are model parameters (i.e. all embeddings) and α governs the hyper-prior on latent senses.

Assume x^e has multiple senses, which are indexed by the random variable z, Eq. (1) can be rewritten,

$$\int_\beta \sum\nolimits_z P(y^e, y^f z, \beta \mid x^e, x^f, \alpha; \theta) d\beta$$

where β are the parameters determining the model probability on each sense for x^e (i.e., the weight on each possible value for z). We place a Dirichlet process (Ferguson, 1973) prior on sense assignment for each word. Thus, adding the word-x subscript to emphasize that these are word-specific senses,

$$P(z_x = k \mid \beta_x) = \beta_{xk} \prod\nolimits_{r=1}^{k-1} (1 - \beta_{xr}) \tag{2}$$

$$\beta_{xk} \mid \alpha \overset{ind}{\sim} Beta(\beta_{xk} \mid 1, \alpha), \;\; k = 1, \ldots. \tag{3}$$

That is, the potentially infinite number of senses for each word x have probability determined by the sequence of independent *stick-breaking weights*, β_{xk}, in the constructive definition of the DP (Sethuraman, 1994). The hyper-prior concentration α provides information on the number of senses we expect to observe in our corpus.

After conditioning upon word sense, we decompose the context probability,

$$P(y^e, y^f \mid z, x^e, x^f; \theta) =$$
$$P(y^e \mid x^e, x^f, z; \theta) P(y^f \mid x^e, x^f, z; \theta).$$

[2] We also maintain a context vector for each word in the English and Foreign vocabularies. The context vector is used as the representation of the word when it appears as the context for another word.