

BEYOND BILINGUAL: MULTI-SENSE WORD EMBEDDINGS USING MULTILINGUAL CONTEXT

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, Ratinov, and Bengio, 2010; Mikolov, Yih, and Zweig, 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. 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, Gimpel, and Livescu, 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 performance.

Recently, several approaches (Guo et al., 2014a; Šuster, Titov, and van Noord, 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, Wang, and Chan, 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, Titov, and van Noord, 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, Qiu, and Huang, 2015; Liu et al., 2015), 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, Dyer, and Hovy, 2015), which aims to infuse semantic ontologies from resources like WordNet (Miller, 1995) and Framenet (Baker, Fillmore, and Lowe, 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 recent work has attempted to remedy this by using parallel text for retrofitting sense-specific embeddings (Ettinger, Resnik, and Carpuat, 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).

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

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, DeNero, and Lin, 2012; Guo et al., 2014a; Šuster, Titov, and van Noord, 2016) with only (Šuster, Titov, and van Noord, 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 $E = \{x_1^e, \dots, x_i^e, \dots, x_{N_e}^e\}$ denote the words of the English side and $F = \{x_1^f, \dots, x_i^f, \dots, x_{N_f}^f\}$ 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 $\text{Nbr}(x, L, d)$ 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 \mathbf{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 \mathbf{x}_k^e .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:

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

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,

$$\int_{\beta} \sum_z P(y^e, y^f z, \beta \mid x^e, x^f, \alpha; \theta) d\beta$$

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,

$$P(z_x = k \mid \beta_x) = \beta_{xk} \prod_{r=1}^{k-1} (1 - \beta_{xr}), \beta_{xk} \mid \alpha \stackrel{\text{ind}}{\sim} \text{Beta}(\beta_{xk} \mid 1, \alpha), \quad k = 1, \dots \quad (2)$$

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 $P(y^e, y^f \mid z, x^e, x^f; \theta)$ into two terms, $P(y^e \mid x^e, x^f, z; \theta)P(y^f \mid x^e, x^f, z; \theta)$. Both the first and the second terms are sense-dependent, and each factors as,

$$P(y \mid x^e, x^f, z = k; \theta) \propto \Psi(x^e, z = k, y) \Psi(x^f, y) = \exp(\mathbf{y}^T \mathbf{x}_k^e) \exp(\mathbf{y}^T \mathbf{x}^f) = \exp(\mathbf{y}^T (\mathbf{x}_k^e + \mathbf{x}^f)),$$

²We 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.

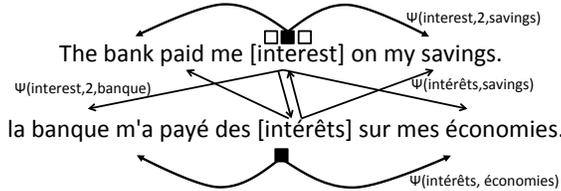


Figure 2: The aligned pair (*interest,intérêt*) is used to predict monolingual and crosslingual context in both languages (see factors in eqn. (3)). We pick each sense (here 2nd) vector for *interest*, to perform weighted update. We only model polysemy in English.

where x_k^e is the embedding corresponding to the k^{th} sense of the word x^e , and y is either y^e or y^f . The factor $\Psi(x^e, z = k, y)$ use the corresponding sense vector in a skip-gram-like formulation. This results in total of 4 factors,

$$P(y^e, y^f | z, x^e, x^f; \theta) \propto \Psi(x^e, z, y^e) \Psi(x^f, y^f) \Psi(x^e, z, y^f) \Psi(x^f, y^e) \quad (3)$$

See Figure 2 for illustration of each factor. This modeling approach is reminiscent of (Luong, Pham, and Manning, 2015), who jointly learned embeddings for two languages l_1 and l_2 by optimizing a joint objective containing 4 skip-gram terms using the aligned pair (x^e, x^f) – two predicting monolingual contexts $l_1 \rightarrow l_1, l_2 \rightarrow l_2$, and two predicting crosslingual contexts $l_1 \rightarrow l_2, l_2 \rightarrow l_1$.

Learning. Learning involves maximizing the log-likelihood,

$$P(y^e, y^f | x^e, x^f; \alpha, \theta) = \int_{\beta} \sum_z P(y^e, y^f, z, \beta | x^e, x^f, \alpha; \theta) d\beta$$

Let $q(z, \beta) = q(z)q(\beta)$ where $q(z) = \prod_i q(z_i)$ and $q(\beta) = \prod_{w=1}^V \prod_{k=1}^T \beta_{wk}$ be the fully factorized variational approximation of the true posterior $P(z, \beta | y^e, y^f, x^e, x^f, \alpha)$, where V is the size of english vocabulary, and T is the maximum number of senses for any word. The optimization problem solves for $\theta, q(z)$ and $q(\beta)$ using the stochastic variational inference technique (Hoffman et al., 2013) similar to (Bartunov et al., 2016) (refer for details).

The resulting learning algorithm is shown as Algorithm 1. The first for-loop (line 1) updates the English sense vectors using the crosslingual and monolingual contexts. First, the expected sense distribution for the current English word w is computed using the current estimate of $q(\beta)$ (line 4). The sense distribution is updated (line 7) using the combined monolingual and crosslingual contexts (line 5) and re-normalized (line 8). Using the updated sense distribution $q(\beta)$'s sufficient statistics is re-computed (line 9) and the global parameter θ is updated (line 10) as follows,

$$\theta \leftarrow \theta + \rho_t \nabla_{\theta} \sum_{k|z_{ik} > \epsilon} \sum_{y \in y_c} z_{ik} \log p(y|x_i, k, \theta) \quad (4)$$

Note that in the above sum, a sense participates in a update only if its probability exceeds a threshold $\epsilon (=0.001)$. The final model retains sense vectors whose sense probability exceeds the same threshold. The last for-loop (line 11) jointly optimizes the foreign embeddings using English context with the standard skip-gram updates.

Disambiguation. Similar to (Bartunov et al., 2016), we can disambiguate the sense for the word x^e given a monolingual context y^e as follows,

$$P(z | x^e, y^e) \propto P(y^e | x^e, z; \theta) \sum_{\beta} P(z | x^e, \beta) q(\beta) \quad (5)$$

Although the model trains embeddings using both monolingual and crosslingual context, we only use monolingual context at test time. We found that so long as the model has been trained with multilingual context, it performs well in sense disambiguation on new data even if it contains only monolingual context. A similar observation was made by (Šuster, Titov, and van Noord, 2016).

4 MULTILINGUAL EXTENSION

Bilingual distributional signal alone may not be sufficient as polysemy may survive translation in the second language. Unlike existing approaches, we can easily incorporate multilingual distributional

Algorithm 1 Pseudocode of Learning Algorithm

Input: parallel corpus $E = \{x_1^e, \dots, x_i^e, \dots, x_{N_e}^e\}$ and $F = \{x_1^f, \dots, x_i^f, \dots, x_{N_f}^f\}$ and alignments $A_{e \rightarrow f}$ and $A_{f \rightarrow e}$, Hyper-parameters α and T , window sizes d, d' .

Output: $\theta, q(\beta), q(\mathbf{z})$

- 1: **for** $i = 1$ to N_e **do** ▷ update english vectors
- 2: $w \leftarrow x_i^e$
- 3: **for** $k = 1$ to T **do**
- 4: $z_{ik} \leftarrow \mathbb{E}_{q(\beta_w)}[\log p(z_i = k | x_i^e)]$
- 5: $y_c \leftarrow \text{Nbr}(x_i^e, E, d) \cup \text{Nbr}(x_i^f, F, d') \cup \{x_i^e\}$ where $x_i^f = A_{e \rightarrow f}(x_i^e)$
- 6: **for** y in y_c **do**
- 7: SENSE-UPDATE(x_i^e, y, z_i)
- 8: Renormalize z_i using softmax
- 9: Update suff. stats. for $q(\beta)$ like (Bartunov et al., 2016)
- 10: Update θ using eq. (4)
- 11: **for** $i = 1$ to N_f **do** ▷ jointly update foreign vectors
- 12: $y_c \leftarrow \text{Nbr}(x_i^f, F, d) \cup \text{Nbr}(x_i^e, E, d') \cup \{x_i^e\}$ where $x_i^e = A_{f \rightarrow e}(x_i^f)$
- 13: **for** y in y_c **do**
- 14: SKIP-GRAM-UPDATE(x_i^f, y)
- 15: **procedure** SENSE-UPDATE(x_i, y, z_i)
- 16: $z_{ik} \leftarrow z_{ik} + \log p(y | x_i, k, \theta)$

signals in our model. For using languages l_1 and l_2 to learn multi-sense embeddings for English, we train on a concatenation of En- l_1 parallel corpus with an En- l_2 parallel corpus. This technique can easily be generalized to more than two foreign languages to obtain a large multilingual corpus.

Value of $\Psi(y^e, x^f)$. The factor modeling the dependence of the english context word y^e on foreign word x^f is crucial to performance when using multiple languages. Consider the case of using French and Spanish contexts to disambiguate the financial sense of the english word *bank*. In this case, the (financial) sense vector of *bank* will be used to predict vector of *banco* (Spanish context) and *banque* (French context). If vectors for *banco* and *banque* do not reside in the same space or are not close, the model will incorrectly assume they are different contexts to introduce a new sense for *bank*. This is precisely why the bilingual models, like that of (Šuster, Titov, and van Noord, 2016), cannot be extended to multilingual setting, as they pre-train the embeddings of second language before running the multi-sense embedding process. As a result of naive pre-training, the French and Spanish vectors of semantically similar pairs like (*banco, banque*) will lie in different spaces and need not be close. A similar reason holds for (Guo et al., 2014a), as they use a two step approach instead of joint learning.

To avoid this, the vector for pairs like *banco* and *banque* should lie in the same space and close to each other and the sense vector for *bank*. The $\Psi(y^e, x^f)$ term attempts to ensure this by using the vector for *banco* and *banque* to predict the vector of *bank*. This way, the model brings the embedding space for Spanish and French closer by using English as a bridge language during joint training. A similar idea of using English as a bridging language was used in the models proposed in (Hermann and Blunsom, 2014) and (Coulmance et al., 2015). Beside the benefit in the multilingual case, the $\Psi(y^e, x^f)$ term improves performance in the bilingual case as well, as it forces the English and second language embeddings to remain close in space.

To show the value of $\Psi(y^e, x^f)$ factor in our experiments, we ran a variant of Algorithm 1 without the $\Psi(y^e, x^f)$ factor, by only using monolingual neighborhood $\text{Nbr}(x_i^f, F)$ in line 12 of Algorithm 1. We call this variant ONE-SIDED model and the model in Algorithm 1 the FULL model.

5 EXPERIMENTAL SETUP

Parallel Corpora. We use parallel corpora in English (En), French (Fr), Spanish (Es), Russian (Ru) and Chinese (Zh) in our experiments. Corpus statistics for all datasets used in our experiments are shown in Table 1. For En-Zh, we use the FBIS parallel corpus (LDC2003E14). For En-Fr, we use the first 10M lines from the Giga-EnFr corpus released as part of the WMT shared task (Callison-Burch et al., 2011). Note that the domain from which parallel corpus has been derived can affect the final result. To understand what choice of languages provide suitable disambiguation signal,

Corpus	Source	Lines (M)	EN-Words (M)
En-Fr	Canadian & EU proc.	≈ 10	250
En-Zh	FBIS news	≈ 9.5	286
En-Es	UN proc.	≈ 10	270
En-Fr	UN proc.	≈ 10	260
En-Zh	UN proc.	≈ 8	230
En-Ru	UN proc.	≈ 10	270

Table 1: Corpus Statistics (in millions). Horizontal lines demarcate corpora from the same domain.

it is necessary to control for domain in all parallel corpora. To this end, we also used the En-Fr, En-Es, En-Zh and En-Ru sections of the MultiUN parallel corpus (Eisele and Chen, 2010). Word alignments were generated using `fast_align` tool (Dyer, Chahuneau, and Smith, 2013) in the symmetric intersection mode. Tokenization and other preprocessing were performed using `cdec`³ toolkit. Stanford Segmenter (Tseng et al., 2005) was used to preprocess the chinese corpora.

Word Sense Induction (WSI). We evaluate our approach on word sense induction task. In this task, we are given several sentences showing usages of the same word, and are required to cluster all sentences which use the same sense (Nasiruddin, 2013). The predicted clustering is then compared against a provided gold clustering. Note that WSI is a harder task than Word Sense Disambiguation (WSD)(Navigli, 2009), as unlike WSD, this task does not involve any supervision or explicit human knowledge about senses of words. We use the disambiguation approach in eq. (5) to predict the sense given the word and four context words.

To allow for fair comparison with earlier work, we use the same benchmark datasets as (Bartunov et al., 2016) – Semeval-2007, 2010 and Wikipedia Word Sense Induction (WWSI). We report Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) in the experiments, as ARI is a more strict and precise metric than F-score and V-measure.

Parameter Tuning. For fairness, we used five context words on either side to update each English word-vectors in all the experiments. In the monolingual setting, all five words are English; in the multilingual settings, we used four neighboring English words plus the one foreign word aligned to the word being updated ($d = 4, d' = 0$ in Algorithm 1). We also analyze effect of varying d' .

We tune the parameters α and T by maximizing the log-likelihood of a held out English text.⁴ The parameters were chosen from the following values $\alpha = \{0.05, 0.1, \dots, 0.25\}$, $T = \{5, 10, \dots, 30\}$. All models were trained for 10 iteration with a decaying learning rate of 0.025, decayed to 0. Unless otherwise stated, all embeddings are 100 dimensional.

Under various choice of α and T , we identify only about 10-20% polysemous words in the vocabulary using monolingual training and 20-25% polysemous using multilingual training. It is evident using the non-parametric prior has led to substantially more efficient representation compared to previous methods with fixed number of senses per word.

6 EXPERIMENTAL RESULTS

We performed extensive experiments to evaluate the benefit of leveraging bilingual and multilingual information during training. We also analyze how the different choices of language family (i.e. using more distant vs more similar languages) affect performance of the embeddings.

Word Sense Induction Results. The results for WSI are shown in Table 2. MONO refers to the AdaGram model of (Bartunov et al., 2016) trained on the English side of the parallel corpus. In all cases, the MONO model is outperformed by ONE-SIDED and FULL models, showing the benefit of using crosslingual signal in training. Best performance is attained by the multilingual model (En-FrZh), showing value of multilingual signal. The value of $\Psi(y^e, x^f)$ term is also verified by the fact that the ONE-SIDED model performs worse than the FULL model.

We can also compare (unfairly to FULL model) to the best results described in (Bartunov et al., 2016), which achieved ARI scores of 0.069, 0.097 and 0.286 on the three datasets respectively after

³github.com/redpony/cdec

⁴first 100k lines from the En-Fr Europarl (Koehn, 2005)

Setting	S-2007	S-2010	WWSI	avg. ARI	SCWS
En-Fr					
MONO	.044	.064	.112	.073	41.1
ONE-SIDED	.054	.074	.116	.081	41.9
FULL	.055	.086	.105	.082	41.8
En-Zh					
MONO	.054	.074	.073	.067	42.6
ONE-SIDED	.059	.084	.078	.074	45.0
FULL	.055	.090	.079	.075	41.7
En-FrZh					
MONO	.056	.086	.103	.082	47.3
ONE-SIDED	.067	.085	.113	.088	44.6
FULL	.065	.094	.120	.093	41.9

Table 2: Results on word sense induction (left four columns) in ARI and contextual word similarity (last column) in percent correlation. Language pairs are separated by horizontal lines. Best results shown in **bold**.

Train Setting	S-2007		S-2010		WWSI		Avg. ARI	
	En-FrEs	En-RuZh	En-FrEs	En-FrEs	En-FrEs	En-RuZh	En-FrEs	En-RuZh
(1) MONO	.035	.033	.046	.049	.054	.049	.045	.044
(2) ONE-SIDED	.044	.044	.055	.063	.062	.057	.054	.055
(3) FULL	.046	.040	.056	.070	.068	.069	.057	.059
(3) - (1)	.011	.007	.010	.021	.014	.020	.012	.015

Table 3: Effect (in ARI) of language family distance on WSI task. Best results for each column is shown in **bold**. The improvement from MONO to FULL is also shown as (3) - (1). Note that this is not comparable to results in Table 2, as we use a different training corpus to control for the domain.

training 300 dimensional embeddings on English Wikipedia ($\approx 100M$ lines). Note that, as WWSI was derived from Wikipedia, training on Wikipedia gives AdaGram model an undue advantage, resulting in high ARI score on WWSI. Nevertheless, even in the unfair comparison, it noteworthy that on S-2007 and S-2010, we can achieve comparable performance (0.067 and 0.094) with multilingual training to a model trained on almost 5 times more data and higher (300) dimensional embeddings.

Contextual Word Similarity Results. For completeness, we report correlation scores on Stanford contextual word similarity dataset (SCWS) (Huang et al., 2012) in Table 2. The task requires computing similarity between two words given their contexts. While the bilingually trained model outperforms the monolingually trained model, surprisingly the multilingually trained model does not perform well on SCWS. We believe this may be due to our parameter tuning strategy.⁵

Effect of Language Family Distance. Intuitively, choice of language can affect the result from crosslingual training as some languages may provide better disambiguation signals than others. We performed a systematic set of experiment to evaluate whether we should choose languages from a closer family (Indo-European languages) or farther family (Non-Indo European Languages) as training data alongside English.⁶ To control for domain here we use the MultiUN corpus. We use En paired with Fr and Es as Indo-European languages, and English paired with Ru and Zh for representing Non-Indo-European languages.

From Table 3, we see that using Non-Indo European languages yield a slightly higher average improvement in WSI task than using Indo-European languages. This suggests that using languages from a distance family aids better disambiguation. Our findings echo those of (Resnik and Yarowsky, 1999), who found that the tendency to lexicalize different senses of an English word differently in a second language correlated with language distance.

Effect of Window Size. Figure 3d shows the effect of increasing the crosslingual window (d') on the average ARI on the WSI task for the En-Fr and En-Zh models. While increasing the window size improves the average score for En-Zh model, the score for the En-Fr model goes down. This suggests that it might be beneficial to have a separate window parameter per language. This also

⁵Most works tune directly on the test dataset for Word Similarity tasks (Faruqui et al., 2016)

⁶(Šuster, Titov, and van Noord, 2016) compared different languages but did not control for domain.

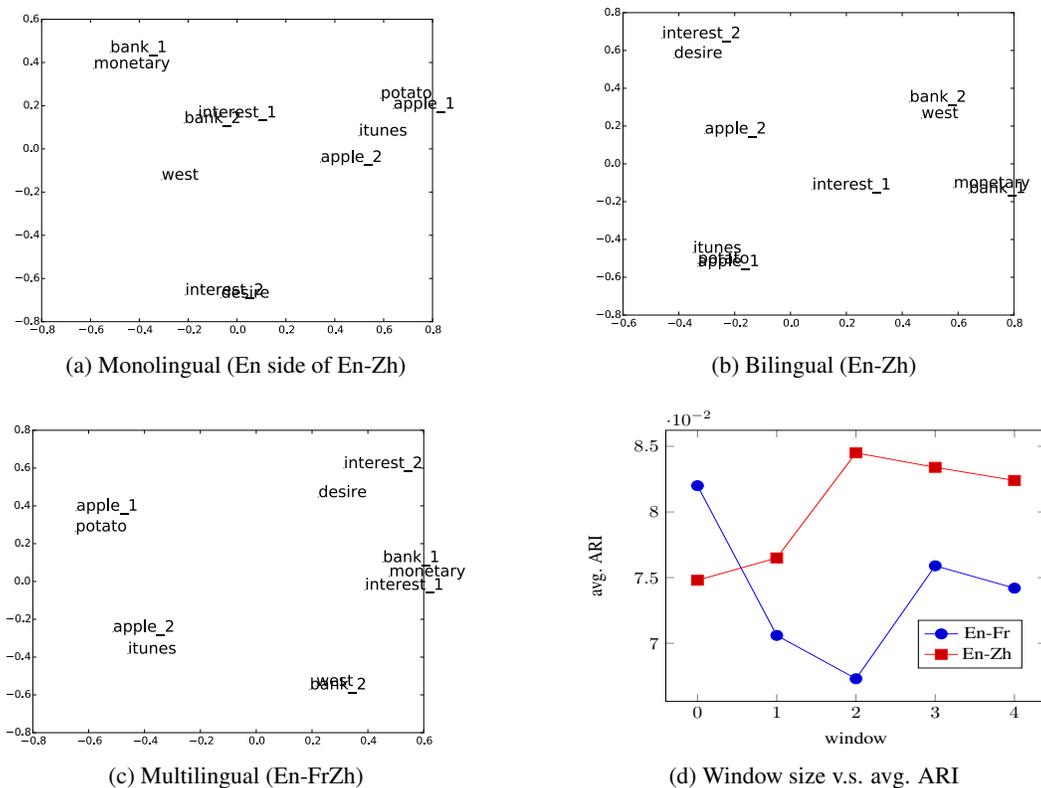


Figure 3: **Qualitative:** PCA plots for the vectors for $\{apple, bank, interest, itunes, potato, west, monetary, desire\}$ with multiple sense vectors for *apple*, *interest* and *bank* obtained using monolingual (3a), bilingual (3b) and multilingual (3c) training. **Window Tuning:** Figure 3d shows tuning window size for En-Zh and En-Fr.

aligns with the observation earlier that different language families have different suitability (bigger crosslingual context from a distant family helped) and requirements for optimal performance.

Qualitative Illustration. As an illustration for the effects of multilingual training, Figure 3 shows PCA plots for 11 sense vectors for 9 words using monolingual, bilingual and multilingual models. From Fig 3a, we note that with monolingual training the senses are poorly separated. Although the model infers two senses for *bank*, the two senses of *bank* are close to financial terms, suggesting their distinction was not recognized. The same can be observed about *apple*. In Fig 3b, with bilingual training, the model infers two senses of *bank* correctly, and two sense of *apple* become more distant. The model can still improve eg. pulling *interest* towards the financial sense of *bank*, and pulling *itunes* towards *apple*. Finally, in Fig 3c, all senses of the words are more clearly clustered, improving over the clustering of Fig 3b. The senses of *apple*, *interest*, and *bank* are well separated, and are close to sense-specific words, showing the benefit of multilingual training.

7 CONCLUSION

We presented a multi-view, non-parametric word representation learning algorithm which can leverage multilingual distributional information. Our approach effectively combines the benefits of crosslingual training and Bayesian non-parametrics. Ours is the first multi-sense representation learning algorithm capable of using multilingual distributional information efficiently, by combining several parallel corpora to obtained a large multilingual corpus. Our experiments show how this multi-view approach learns high-quality embeddings using substantially less data and parameters than prior state-of-the-art. While we focused on improving the embedding of English words here, the same algorithm could learn better multi-sense embedding for Chinese, for instance. Exciting avenues for future research include extending our approach to model polysemy in foreign language. The sense vectors can then be aligned across languages (thanks to our joint training paradigm), to generate a multilingual Wordnet like resource, in a completely unsupervised manner.

REFERENCES

- Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *ACL*.
- Bansal, M.; DeNero, J.; and Lin, D. 2012. Unsupervised translation sense clustering. In *NAACL*.
- Bansal, M.; Gimpel, K.; and Livescu, K. 2014. Tailoring continuous word representations for dependency parsing. In *ACL*.
- Bartunov, S.; Kondrashkin, D.; Osokin, A.; and Vetrov, D. 2016. Breaking sticks and ambiguities with adaptive skip-gram. *AISTATS*.
- Callison-Burch, C.; Koehn, P.; Monz, C.; and Zaidan, O. F. 2011. Findings of the 2011 workshop on statistical machine translation. In *WMT Shared Task*.
- Coulmance, J.; Marty, J.-M.; Wenzek, G.; and Benhalloum, A. 2015. Trans-gram, fast cross-lingual word-embeddings. In *EMNLP*.
- Dagan, I., and Itai, A. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational linguistics*.
- Diab, M., and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*.
- Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.
- Eisele, A., and Chen, Y. 2010. MultiUN: A multilingual corpus from united nation documents. In *LREC*.
- Ettinger, A.; Resnik, P.; and Carpuat, M. 2016. Retrofitting sense-specific word vectors using parallel text. In *NAACL*.
- Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *EACL*.
- Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL*.
- Faruqui, M.; Tsvetkov, Y.; Rastogi, P.; and Dyer, C. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *1st RepEval Workshop*.
- Ferguson, T. S. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*.
- Guo, J.; Che, W.; Wang, H.; and Liu, T. 2014a. Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*.
- Guo, J.; Che, W.; Wang, H.; and Liu, T. 2014b. Revisiting embedding features for simple semi-supervised learning. In *EMNLP*.
- Guo, J.; Che, W.; Yarowsky, D.; Wang, H.; and Liu, T. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI*.
- Hermann, K. M., and Blunsom, P. 2014. Multilingual Distributed Representations without Word Alignment. In *ICLR*.
- Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. W. 2013. Stochastic variational inference. *JMLR*.
- Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*.
- Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of classification*.
- Jauhar, S. K.; Dyer, C.; and Hovy, E. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *NAACL*.
- Kawakami, K., and Dyer, C. 2015. Learning to represent words in context with multilingual supervision. *ICLR Workshop*.
- Kilgariff, A. 1997. I don't believe in word senses. *Computers and the Humanities*.

- Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 79–86.
- Li, J., and Jurafsky, D. 2015. Do multi-sense embeddings improve natural language understanding? *EMNLP*.
- Liu, Y.; Liu, Z.; Chua, T.-S.; and Sun, M. 2015. Topical word embeddings. In *AAAI*.
- Liu, P.; Qiu, X.; and Huang, X. 2015. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *IJCAI*.
- Luong, T.; Pham, H.; and Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. In *Workshop on Vector Space Modeling for NLP*.
- Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *NAACL*.
- Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*.
- Nasiruddin, M. 2013. A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. *arXiv preprint arXiv:1310.1425*.
- Navigli, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*.
- Neelakantan, A.; Shankar, J.; Passos, A.; and McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*.
- Ng, H. T.; Wang, B.; and Chan, Y. S. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *ACL*.
- Qiu, L.; Cao, Y.; Nie, Z.; Yu, Y.; and Rui, Y. 2014. Learning word representation considering proximity and ambiguity. In *AAAI*.
- Reisinger, J., and Mooney, R. J. 2010. Multi-prototype vector-space models of word meaning. In *NAACL*.
- Resnik, P., and Yarowsky, D. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *NLE*.
- Sethuraman, J. 1994. A constructive definition of dirichlet priors. *Statistica sinica*.
- Tseng, H.; Chang, P.; Andrew, G.; Jurafsky, D.; and Manning, C. 2005. A conditional random field word segmenter for sighthan bakeoff 2005. In *Proc. of SIGHAN*.
- Turian, J.; Ratinov, L.; and Bengio, Y. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.
- Upadhyay, S.; Faruqui, M.; Dyer, C.; and Roth, D. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *ACL*.
- Šuster, S.; Titov, I.; and van Noord, G. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *NAACL*.
- Wu, Z., and Giles, C. L. 2015. Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *AAAI*.
- Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*.