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.

We can also compare (unfairly to our 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 training 300 dimensional embeddings on English Wikipedia (≈ 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. In comparison, our model did not train on English Wikipedia, and uses 100 dimensional embeddings. 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 using higher (300) dimensional embeddings.

6.2 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]

6.3 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 improvement on an average 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 senses of an English word differently in a second language, correlated with language distance.

6.4 Effect of Window Size.

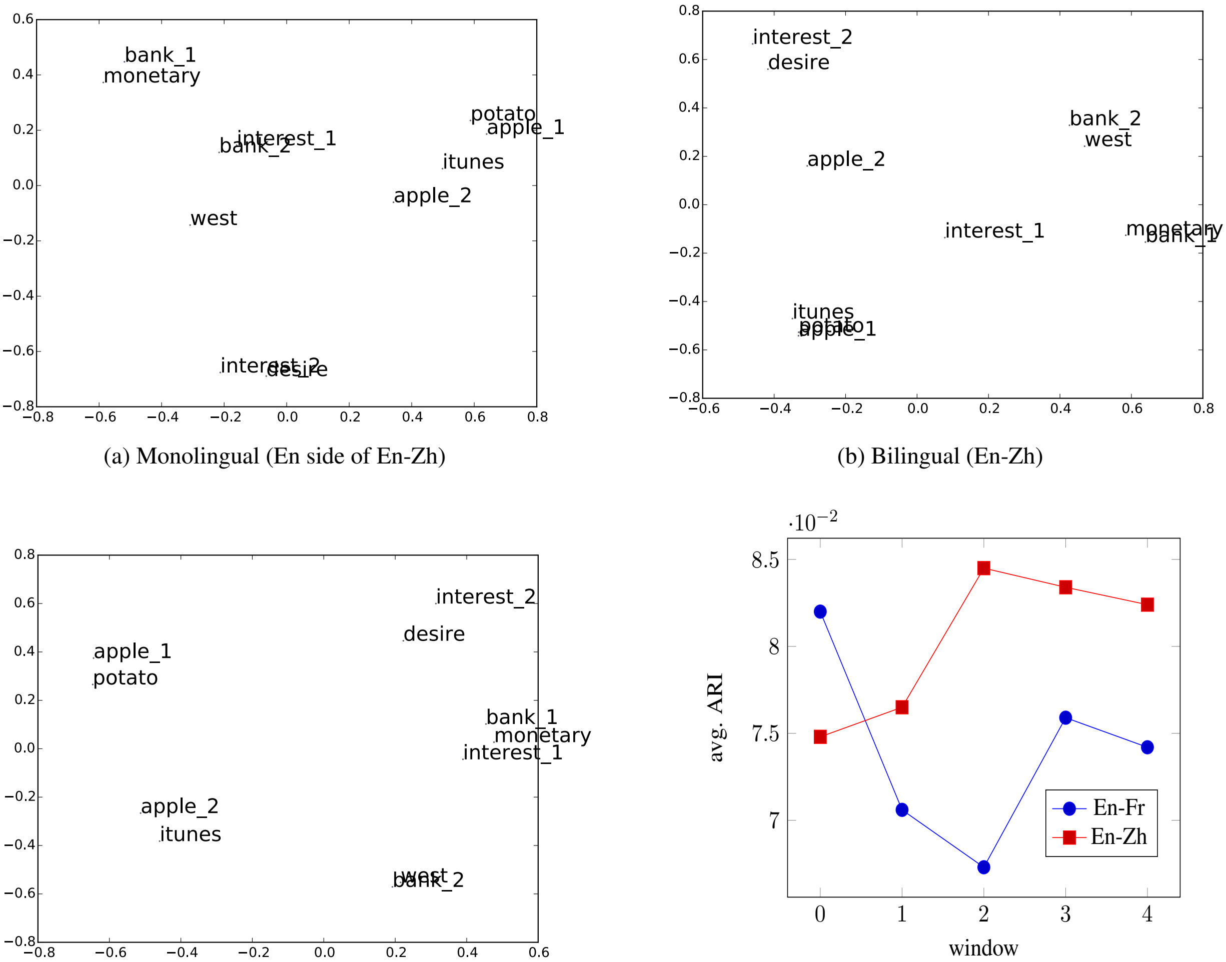
Figure 3d shows the effect of increasing the crosslingual window (d') on the average ARI on the WSI task for the En-Fr and En-Zh models. While increasing the window size improves the average score for En-Zh model, the score for the En-Fr model goes down. This suggests that it might be beneficial to have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

7 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 observation can be

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

[6] (Šuster et al., 2016) compared different languages but did not control for domain.



(a) Monolingual (En side of En-Zh) (b) Bilingual (En-Zh)

(c) Multilingual (En-FrZh) (d) Window size v.s. avg. ARI

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.

made for the senses of *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_2*. 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.

8 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. We also analyzed the effect of various parameters such as choice of language family and cross-lingual window size on the performance. While we focused on improving the embedding of English words in this work, the same algorithm could learn better multi-sense embedding for other languages. Exciting avenues for future research include extending our approach to model polysemy in foreign language. The sense vectors can then be aligned across languages, to generate a multilingual Wordnet like resource, in a completely unsupervised manner thanks to our joint training paradigm.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *ACL*.

Mohit Bansal, John DeNero, and Dekang Lin. 2012. Unsupervised translation sense clustering. In *NAACL*.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *ACL*.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. *AISTATS* .

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *WMT Shared Task*.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *EMNLP*.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational linguistics* .

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *LREC*.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016. Retrofitting sense-specific word vectors using parallel text. In *NAACL*.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *1st RepEval Workshop*.

Thomas S Ferguson. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics* .

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014a. Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014b. Revisiting embedding features for simple semi-supervised learning. In *EMNLP*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Distributed Representations without Word Alignment. In *ICLR*.

Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *JMLR* .

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* .

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *NAACL*.

Kazuya Kawakami and Chris Dyer. 2015. Learning to represent words in context with multilingual supervision. *ICLR Workshop* .

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities* .

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. volume 5, pages 79–86.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? *EMNLP* .

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015a. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *IJCAI*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015b. Topical word embeddings. In *AAAI*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Workshop on Vector Space Modeling for NLP*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *NAACL*.