Bilingual Word Embeddings for Phrase-Based Machine Translation

Will Y. Zou[†], **Richard Socher, Daniel Cer, Christopher D. Manning**
Department of Electrical Engineering[†] and Computer Science Department
Stanford University, Stanford, CA 94305, USA
{wzou,danielcer,manning}@stanford.edu, richard@socher.org

Abstract

We introduce bilingual word embeddings: semantic embeddings associated across two languages in the context of neural language models. We propose a method to learn bilingual embeddings from a large unlabeled corpus, while utilizing MT word alignments to constrain translational equivalence. The new embeddings significantly out-perform baselines in word semantic similarity. A single semantic similarity feature induced with bilingual embeddings adds near half a BLEU point to the results of NIST08 Chinese-English machine translation task.

1 Introduction

It is difficult to recognize and quantify semantic similarities across languages. The Fr-En phrase-pair {'*un cas de force majeure*', '*case of absolute necessity*'}, Zh-En phrase pair {'依然故我','*persist in a stubborn manner*'} are similar in semantics. If co-occurrences of exact word combinations are rare in the training parallel text, it can be difficult for classical statistical MT methods to identify this similarity, or produce a reasonable translation given the source phrase.

We introduce an unsupervised neural model to learn bilingual semantic embedding for words across two languages. As an extension to their monolingual counter-part (Turian et al., 2010; Huang et al., 2012; Bengio et al., 2003), bilingual embeddings capture not only semantic information of monolingual words, but also semantic relationships across different languages. This property allows them to define semantic similarity metrics across phrase-pairs, making them perfect features for machine translation.

To learn bilingual embeddings, we use a new objective function which embodies both monolingual semantics and bilingual translation equivalence. The latter utilizes word alignments, a natural sub-task in the machine translation pipeline. Through large-scale curriculum training (Bengio et al., 2009), we obtain bilingual distributed representations which lie in the same feature space. Embeddings of direct translations overlap, and semantic relationships across bilingual embeddings were further improved through unsupervised learning on a large unlabeled corpus.

Consequently, we produce for the research community a first set of Mandarin Chinese word embeddings with 100,000 words trained on the Chinese Gigaword corpus. We evaluate these embedding on Chinese word semantic similarity from *SemEval-2012* (Jin and Wu, 2012). The embeddings significantly out-perform prior work and pruned *tf-idf* base-lines. In addition, the learned embeddings give rise to 0.11 F1 improvement in Named Entity Recognition on the OntoNotes dataset (Hovy et al., 2006) with a neural network model.

We apply the bilingual embeddings in an end-to-end phrase-based MT system by computing semantic similarities between phrase pairs. On NIST08 Chinese-English translation task, we obtain an improvement of 0.48 BLEU from a competitive baseline (30.01 BLEU to 30.49 BLEU) with the Stanford Phrasal MT system.

2 Review of prior work

Distributed word representations are useful in NLP applications such as information retrieval (Paşca et al., 2006; Manning et al., 2008), search query expansions (Jones et al., 2006), or representing semantics of words (Reisinger et al., 2010). A number of methods have been explored to train and apply word embeddings using continuous models for language. Collobert et al. (2008) learn embeddings in an unsupervised manner through a contrastive estimation technique. Mnih and Hinton (2008), Morin and Bengio (2005) proposed efficient hierarchical continuous-space models. To systematically compare embeddings, Turian et al. (2010) evaluated improvements they bring to state-of-the-art NLP benchmarks. Huang et al. (2012) introduced global document context and multiple word prototypes. Recently, morphology is explored to learn better word representations through Recursive Neural Networks (Luong et al., 2013).

Bilingual word representations have been explored with hand-designed vector space models (Peirsman and Padó , 2010; Sumita, 2000), and with unsupervised algorithms such as LDA and LSA (Boyd-Graber and Resnik, 2010; Tam et al., 2007; Zhao and Xing, 2006). Only recently have continuous space models been applied to machine translation (Le et al., 2012). Despite growing interest in these models, little work has been done along the same lines to train bilingual distributioned word represenations to improve machine translation. In this paper, we learn bilingual word embeddings which achieve competitive performance on semantic word similarity, and apply them in a practical phrase-based MT system.

3 Algorithm and methods

3.1 Unsupervised training with global context

Our method starts with embedding learning formulations in Collobert et al. (2008). Given a context window c in a document d, the optimization minimizes the following Context Objective for a word w in the vocabulary:

$$J_{CO}^{(c,d)} = \sum_{w^r \in V_R} \max(0, 1 - f(c^w, d) + f(c^{w^r}, d)) \quad (1)$$

Here f is a function defined by a neural network. w^r is a word chosen in a random subset V_R of the vocabulary, and c^{w^r} is the context window containing word w^r. This unsupervised objective function contrasts the score between when the correct word is placed in context with when a random word is placed in the same context. We incorporate the global context information as in Huang et al. (2012), shown to improve performance of word embeddings.

3.2 Bilingual initialization and training

In the joint semantic space of words across two languages, the Chinese word '政府' is expected to be close to its English translation 'government'. At the same time, when two words are not direct translations, e.g. 'lake' and the Chinese word '潭' (deep pond), their semantic proximity could be correctly quantified.

We describe in the next sub-sections the methods to intialize and train bilingual embeddings. These methods ensure that bilingual embeddings retain their translational equivalence while their distributional semantics are improved during online training with a monolingual corpus.

3.2.1 Initialization by MT alignments

First, we use MT Alignment counts as weighting to initialize Chinese word embeddings. In our experiments, we use MT word alignments extracted with the Berkeley Aligner (Liang et al., 2006) [1]. Specifically, we use the following equation to compute starting word embeddings:

$$W_{t\text{-}init} = \sum_{s=1}^{S} \frac{C_{ts} + 1}{C_t + S} W_s \quad (2)$$

In this equation, S is the number of possible target language words that are aligned with the source word. C_{ts} denotes the number of times when word t in the target and word s in the source are aligned in the training parallel text; C_t denotes the total number of counts of word t that appeared in the target language. Finally, Laplace smoothing is applied to this weighting function.

[1] On NIST08 Zh-En training data and data from GALE MT evaluation in the past 5 years

Single-prototype English embeddings by Huang et al. (2012) are used to initialize Chinese embeddings. The initialization readily provides a set (*Align-Init*) of benchmark embeddings in experiments (Section 4), and ensures translation equivalence in the embeddings at start of training.

3.2.2 Bilingual training

Using the alignment counts, we form alignment matrices $A_{en \to zh}$ and $A_{zh \to en}$. For $A_{en \to zh}$, each row corresponds to a Chinese word, and each column an English word. An element a_{ij} is first assigned the counts of when the ith Chinese word is aligned with the jth English word in parallel text. After assignments, each row is normalized such that it sums to one. The matrix $A_{zh \to en}$ is defined similarly. Denote the set of Chinese word embeddings as V_{zh}, with each row a word embedding, and the set of English word embeddings as V_{en}. With the two alignment matrices, we define the Translation Equivalence Objective:

$$J_{TEO\text{-}en \to zh} = \|V_{zh} - A_{en \to zh} V_{en}\|^2 \quad (3)$$

$$J_{TEO\text{-}zh \to en} = \|V_{en} - A_{zh \to en} V_{zh}\|^2 \quad (4)$$

We optimize for a combined objective during training. For the Chinese embeddings we optimize for:

$$J_{CO\text{-}zh} + \lambda J_{TEO\text{-}en \to zh} \quad (5)$$

For the English embeddings we optimize for:

$$J_{CO\text{-}en} + \lambda J_{TEO\text{-}zh \to en} \quad (6)$$

During bilingual training, we chose the value of λ such that convergence is achieved for both J_{CO} and J_{TEO}. A small validation set of word similarities from (Jin and Wu, 2012) is used to ensure the embeddings have reasonable semantics. [2]

In the next sections, 'bilingual trained' embeddings refer to those initialized with MT alignments and trained with the objective defined by Equation 5. 'Monolingual trained' embeddings refer to those intialized by alignment but trained without $J_{TEO\text{-}en \to zh}$.

[2] In our experiments, $\lambda = 50$.

3.3 Curriculum training

We train 100k-vocabulary word embeddings using curriculum training (Turian et al., 2010) with Equation 5. For each curriculum, we sort the vocabulary by frequency and segment the vocabulary by a band-size taken from {5k, 10k, 25k, 50k}. Separate bands of the vocabulary are trained in parallel using minibatch L-BFGS on the Chinese Gigaword corpus [3]. We train 100,000 iterations for each curriculum, and the entire 100k vocabulary is trained for 500,000 iterations. The process takes approximately 19 days on a eight-core machine. We show visualization of learned embeddings overlaid with English in Figure 1. The two-dimensional vectors for this visualization is obtained with t-SNE (van der Maaten and Hinton, 2008). To make the figure comprehensible, subsets of Chinese words are provided with reference translations in boxes with green borders. Words across the two languages are positioned by the semantic relationships implied by their embeddings.

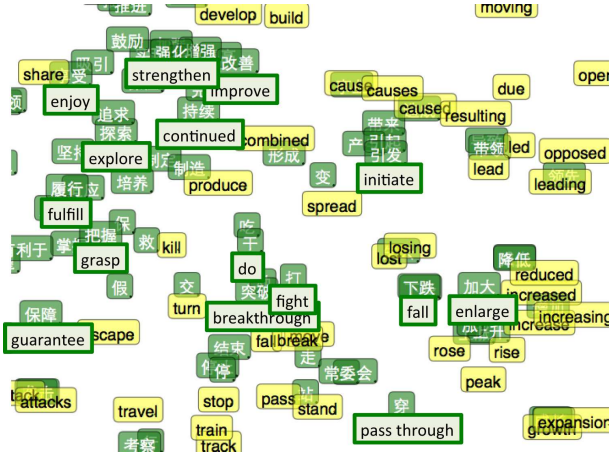

Figure 1: Overlaid bilingual embeddings: English words are plotted in yellow boxes, and Chinese words in green; reference translations to English are provided in boxes with green borders directly below the original word.

4 Experiments

4.1 Semantic Similarity

We evaluate the Mandarin Chinese embeddings with the semantic similarity test-set provided by the or-

[3] Fifth Edition. LDC catelog number *LDC2011T13*. We only exclude *cna_cmn*, the Traditional Chinese segment of the corpus.

Table 1: Results on Chinese Semantic Similarity

Method	*Sp. Corr.* ($\times$100)	*K. Tau* ($\times$100)
Prior work (Jin and Wu, 2012)		5.0
Tf-idf		
Naive tf-idf	41.5	28.7
Pruned tf-idf	46.7	32.3
Word Embeddings		
Align-Init	52.9	37.6
Mono-trained	59.3	42.1
Biling-trained	**60.8**	**43.3**

Table 2: Results on Named Entity Recognition

Embeddings	*Prec.*	*Rec.*	*F1*	*Improve*
Align-Init	0.34	0.52	0.41	
Mono-trained	0.54	0.62	**0.58**	**0.17**
Biling-trained	0.48	0.55	0.52	0.11

Table 3: Vector Matching Alignment *AER* (lower is better)

Embeddings	*Prec.*	*Rec.*	*AER*
Mono-trained	0.27	0.32	0.71
Biling-trained	0.37	0.45	**0.59**

ganizers of *SemEval-2012* Task 4. This test-set contains 297 Chinese word pairs with similarity scores estimated by humans.

The results for semantic similarity are shown in Table 1. We show two evaluation metrics: Spearman Correlation and Kendall's Tau. For both, bilingual embeddings trained with the combined objective defined by Equation 5 perform best. For pruned tf-idf, we follow Reisinger et al. (2010; Huang et al. (2012) and count word co-occurrences in a 10-word window. We use the best results from a range of pruning and feature thresholds to compare against our method. The bilingual and monolingual trained embeddings[4] out-perform pruned *tf-idf* by 14.1 and 12.6 Spearman Correlation ($\times$100), respectively. Further, they out-perform embeddings initialized from alignment by 7.9 and 6.4. Both our *tf-idf* implementation and the word embeddings have significantly higher Kendall's Tau value compared to Prior work (Jin and Wu, 2012). We verified Tau calculations with original submissions provided by the authors.

4.2 Named Entity Recognition

We perform NER experiments on OntoNotes (v4.0) (Hovy et al., 2006) to validate the quality of the Chinese word embeddings. Our experimental setup is the same as Wang et al. (2013). With embeddings, we build a naive feed-forward neural network (Collobert et al., 2008) with 2000 hidden neurons and a sliding window of five words. This naive setting, without sequence modeling or sophisticated join optimization, is not competitive with state-of-the-art (Wang et al., 2013). Table 2 shows that the bilingual embeddings obtains 0.11 F1 improvement, lagging monolingual, but significantly better than *Align-Init* (as in Section3.2.1) on the NER task.

4.3 Vector matching alignment

Translation equivalence of the bilingual embeddings is evaluated by naive word alignment to match word embeddings by cosine distance.[5] The Alignment Error Rates (AER) reported in Table 3 suggest that bilingual training using Equation 5 produces embeddings with better translation equivalence compared to those produced by monolingual training.

4.4 Phrase-based machine translation

Our experiments are performed using the Stanford Phrasal phrase-based machine translation system (Cer et al., 2010). In addition to NIST08 training data, we perform phrase extraction, filtering and phrase table learning with additional data from GALE MT evaluations in the past 5 years. In turn, our baseline is established at 30.01 BLEU and reasonably competitive relative to NIST08 results. We use NIST06 as the tuning set [6], and apply Minimum Error Rate Training (MERT) (Och, 2003) to tune the decoder.

In the phrase-based MT system, we add one feature to bilingual phrase-pairs. For each phrase, the word embeddings are averaged to obtain a feature vector. If a word is not found in the vocabulary, we disregard and assume it is not in the phrase; if no

[4] Due to variations caused by online minibatch L-BFGS, we take embeddings from five random points out of last 10^5 minibatch iterations, and average their semantic similarity results.

[5] This is evaluated on 10,000 randomly selected sentence pairs from the MT training set.

[6] Updated to clarify the decoder tuning procedure.

Table 4: NIST08 Chinese-English translation *BLEU*

Method	*BLEU*
Our baseline	30.01
Embeddings	
Random-Init Mono-trained	30.09
Align-Init	30.31
Mono-trained	30.40
Biling-trained	**30.49**

word is found in a phrase, a zero vector is assigned to it. We then compute the cosine distance between the feature vectors of a phrase pair to form a semantic similarity feature for the decoder.

Results on NIST08 Chinese-English translation task are reported in Table 4[7]. An increase of 0.48 BLEU is obtained with semantic similarity with bilingual embeddings. The increase is modest, just surpassing a reference standard deviation 0.29 BLEU Cer et al. (2010)[8] evaluated on a similar system. We intend to publish further analysis on statistical significance of this result as an appendix. From these suggestive evidence in the MT results, random initialized monolingual trained embeddings add little gains to the baseline. Bilingual initialization and training seem to be offering relatively more consistent gains by introducing translational equivalence.

5 Conclusion

In this paper, we introduce bilingual word embeddings through initialization and optimization constraint using MT alignments The embeddings are learned through curriculum training on the Chinese Gigaword corpus. We show good performance on Chinese semantic similarity with bilingual trained embeddings. When used to compute semantic similarity of phrase pairs, bilingual embeddings improve NIST08 end-to-end machine translation results by just below half a BLEU point. This implies that semantic embeddings are useful features for improving MT systems. Further, our results offer suggestive evidence that bilingual word embeddings act as high-quality semantic features and embody bilingual translation equivalence across languages.

[7]We report case-insensitive BLEU

[8]With 4-gram BLEU metric from Table 4

Acknowledgments

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) Broad Operational Language Translation (BOLT) program through IBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, or the US government. We thank John Bauer and Thang Luong for helpful discussions.

References

A. Klementiev, I. Titov and B. Bhattarai. 2012. Inducing Crosslingual Distributed Representation of Words. *COLING.*

Y. Bengio, J. Louradour, R. Collobert and J. Weston. 2009. Curriculum Learning. *ICML.*

Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research.*

Y. Bengio and Y. LeCunn. 2007. Scaling learning algorithms towards AI. *Large-Scale Kernal Machines.*

J. Boyd-Graber and P. Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation. *EMNLP.*

D. Cer, M. Galley, D. Jurafsky and C. Manning. 2010. Phrasal: A Toolkit for Statistical Machine Translation with Facilities for Extraction and Incorporation of Arbitrary Model Features. *In Proceedings of the North American Association of Computational Linguistics - Demo Session (NAACL-10).*

D. Cer, C. Manning and D. Jurafsky. 2010. The Best Lexical Metric for Phrase-Based Statistical MT System Optimization. *NAACL.*

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *ICML.*

G. Foster and R. Kuhn. 2009. Stabilizing minimum error rate training. *Proceedings of the Fourth Workshop on Statistical Machine Translation.*

M. Galley, P. Chang, D. Cer, J. R. Finkel and C. D. Manning. 2008. NIST Open Machine Translation 2008 Evaluation: Stanford University's System Description. *Unpublished working notes of the 2008 NIST Open Machine Translation Evaluation Workshop.*

S. Green, S. Wang, D. Cer and C. Manning. 2013. Fast and adaptive online training of feature-rich translation models. *ACL.*

G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath

and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine.*

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel. 2006. OntoNotes: the 90% solution. *NAACL-HLT.*

E. H. Huang, R. Socher, C. D. Manning and A. Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. *ACL.*

P. Jin and Y. Wu. 2012. SemEval-2012 Task 4: Evaluating Chinese Word Similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics.*

R. Jones. 2006. Generating query substitutions. *In Proceedings of the 15th international conference on World Wide Web.*

P. Koehn, F. J. Och and D. Marcu. 2003. Statistical Phrase-Based Translation. *HLT.*

H. Le, A. Allauzen and F. Yvon 2012. Continuous space translation models with neural networks. *NAACL.*

P. Liang, B. Taskar and D. Klein. 2006. Alignment by agreement. *NAACL.*

M. Luong, R. Socher and C. Manning. 2013. Better word representations with recursive neural networks for morphology. *CONLL.*

L. van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research.*

A. Maas and R. E. Daly and P. T. Pham and D. Huang and A. Y. Ng and C. Potts. 2011. Learning word vectors for sentiment analysis. *ACL.*

C. Manning and P. Raghavan and H. Schtze. 2008. Introduction to Information Retrieval. *Cambridge University Press, New York, NY, USA.*

T. Mikolov, M. Karafiat, L. Burget, J. Cernocky and S. Khudanpur. 2010. Recurrent neural network based language model. *INTERSPEECH.*

T. Mikolov, K. Chen, G. Corrado and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781v1.*

A. Mnih and G. Hinton. 2008. A scalable hierarchical distributed language model. *NIPS.*

F. Morin and Y. Bengio. 2005. Hierarchical probabilistic neural network language model. *AISTATS.*

F. Och. 2003. Minimum error rate training in statistical machine translation. *ACL.*

M. Paşca, D. Lin, J. Bigham, A. Lifchits and A. Jain. 2006. Names and similarities on the web: fact extraction in the fast lane. *ACL.*

Y. Peirsman and S. Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. *ACL.*

J. Reisinger and R. J. Mooney. 2010. Multi-prototype vector-space models of word meaning. *NAACL.*

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv., 34:1-47, March.*

R. Socher, J. Pennington, E. Huang, A. Y. Ng and C. D. Manning. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. *EMNLP.*

R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. *NIPS.*

E. Sumita. 2000. Lexical transfer using a vector-space model. *ACL.*

Y. Tam, I. Lane and T. Schultz. 2007. Bilingual-LSA based LM adaptation for spoken language translation. *ACL.*

S. Tellex and B. Katz and J. Lin and A. Fernandes and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Search and Development in Information Retrieval, pages 41-47. ACM Press.*

J. Turian and L. Ratinov and Y. Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. *ACL.*

M. Wang, W. Che and C. D. Manning. 2013. Joint Word Alignment and Bilingual Named Entity Recognition Using Dual Decomposition. *ACL.*

K. Yamada and K. Knight. 2001. A Syntax-based Statistical Translation Model. *ACL.*

B. Zhao and E. P. Xing 2006. BiTAM: Bilingual topic AdMixture Models for word alignment. *ACL.*