HCCL at SemEval-2017 Task 2: Combining Multilingual Word Embeddings and Transliteration Model for Semantic Similarity

Junqing He[1,3], **Long Wu**[1,3], **Xuemin Zhao**[1], **Yonghong Yan**[1,2,3]
[1]The Key Laboratory of Speech Acoustics and Content Understanding
Institute of Acoustics, Chinese Academy of Sciences
[2]Xinjiang Laboratory of Minority Speech and Language Information Processing
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
{hejunqing,wulong,zhaoxuemin,yonghongyan}@hccl.ioa.ac.cn

Abstract

In this paper, we introduce an approach to combining word embeddings and machine translation for multilingual semantic word similarity, the task2 of SemEval-2017. Thanks to the unsupervised transliteration model, our cross-lingual word embeddings encounter decreased sums of OOVs. Our results are produced using only monolingual Wikipedia corpora and a limited amount of sentence-aligned data. Although relatively little resources are utilized, our system ranked 3rd in the monolingual subtask and can be the 6th in the cross-lingual subtask.

1 Introduction

With convenient word representation methods being proposed, word embeddings are successfully utilized in state-of-the-art systems ranging from text classification (Kim, 2014), opinion categorization (Enríquez et al., 2016), machine translation (Zou et al., 2013), to stock price prediction (Peng and Jiang, 2016) and so on.

In earlier studies, the latent semantic analysis (LSA) was introduced by Deerwester (1990). It is called topic model because terms are represented as the vectors of topics and was popularized by Landauer (1997). In 2003, researchers developed the topic model based on latent Dirichlet allocation(LDA) (Blei et al., 2003). LDA did not widely spread until the Gibbs sampling was applied to the on-line training of LDA (Hoffman et al., 2010). Another traditional distributional method, pointwise mutual information metric was proposed by Turney and Pental (2010). Recently, fast distributed embeddings like (Mikolov et al., 2013c) and GloVe (Pennington et al., 2014) are based on the assumption that the meaning of a word depends on its context. As Levy et al. (2015) pointed out, there is no significant performance difference between them.

For cross-lingual word representation, there are generally four categories: Monolingual mapping (Mikolov et al., 2013b), pseudo-cross-lingual training (Gouws and Søgaard, 2015), cross-lingual training (Hermann and Blunsom, 2014) and joint optimization (Coulmance et al., 2015). As presented in (Mogadala and Rettinger, 2016) , the joint optimization method represents the state-of-the-art level in cross-lingual text classification and translation. These methods train embeddings both on monolingual and parallel corpora by jointly optimizing the losses. However, they are rarely used in word similarity due to the unsatisfying performance.

In this task, we adopt different strategies for the two subtasks. We use word2vec for subtask1, monolingual word similarity. For the subtask2, cross-lingual word similarity, we use jointly optimized cross-lingual word representation in addition to transliteration model. We build a cross-lingual word embedding system and a special machine translation system. Our approach has the following characteristics:

- Fast and efficient. Both word2vec and the cross-lingual word embeddgings tool have impressive speed (Coulmance et al., 2015) and not need expensive annotated word-aligned data.

- Decreasing OOVs. Our translation system is featured by its transliteration model that deal with OOVs outside the parallel corpus.

We constructed a naive system and did not try out the parameters for embeddings and translation models in limited time.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 220–225,
Vancouver, Canada, August 3 - 4, 2017.

2 Our Approach

We use skip-gram word embeddings directly for monolingual subtask. For cross-lingual subtask, we use English as pivot language and train multilingual word embeddings using monolingual corpora and sentence-aligned parallel data. A translation model is also trained by our statistical machine translation system. Subsequently, we translate the words in the test set into English and look up their word embeddings. For those out of English word embeddings, we check them from original language word embeddings.

2.1 Word Embeddings

For monolingual task, we choose word2vec to generate our word representations for robustness reason. Mikolov (2013c) modeled input word embeddings $\vec{w}$ as the weights from the input layer to the projection layer and its output vector $\vec{w}_o$ as weights from the projection layer to the one-hot output layer.

Skip-gram Model. The skip-gram model assumes that $P(w|c) = \sigma(\vec{w} \cdot \vec{c})$, with c as the embedding of context. Then minimize the loss function which is simplified as:

$$J = \sum_{s \in C} \sum_{w \in s} \sum_{c \in s[w-l:w+l]} -\log \sigma(\vec{w} \cdot \vec{c}) \quad (1)$$

where C is the sentence set of training corpus, s means a sentence and l is the window length. σ is the sigmoid function. Negative sampling is ignored in the equation for simplification.

Trans-gram Model. With skip-gram model introduced, we now extend it to the trans-gram model (Coulmance et al., 2015) for cross-lingual task. For sentence aligned data $A_{s,t}$, where s is the source language and t is the target language, we consider the whole sentence s_t as the context of each word w_s in sentence s_s. The loss for the source language is written as:

$$J_{s,t} = \sum_{s_s \in C_s} \sum_{w_s \in s_s} \sum_{c_t \in s_t} -\log \sigma(\vec{w_s} \cdot \vec{c_t}) \quad (2)$$

The skip-gram model also adopts the negative sampling.

The skip-gram model is famous for its efficiency (Mikolov et al., 2013a). The trans-gram model is of the same computational complexity, thus has the same speed. Although the cross-lingual embeddings can be trained fast, their performance on word similarity task is unsatisfying (0.493 of correlation) with word aligned data (Luong et al., 2015). So we turn to machine translation for steady performance with assistance of these word embeddings.

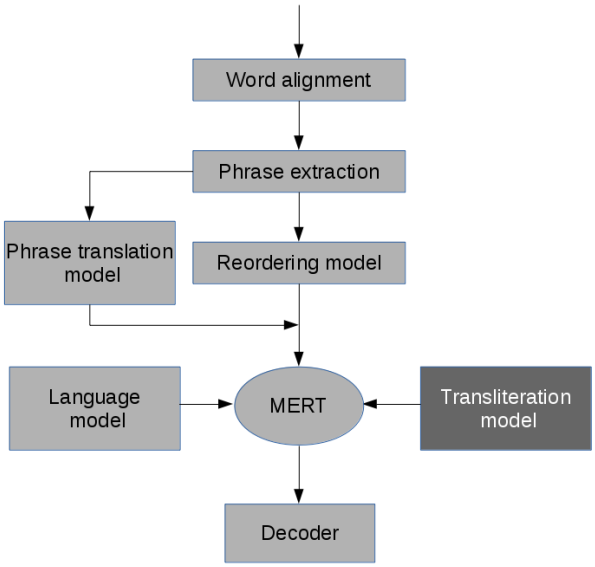

Figure 1: Framework of our translation system.

2.2 Machine Translation System

We constructed a phrase-based statistical machine translation (SMT) system with the transliteration model (TM) (Durrani et al., 2014). Our SMT system is illustrated in Figure 1. Like most of the phrased-based machine translation model, our system follow the steps which are shallow gray in the diagram. First we use GIZA++ (Och and Ney, 2003) as our aligner to align words and get lexical translation table. Then phrases are extracted and we estimate their translation scores directly and inversely by refining the word alignments heuristically. Subsequently, a distance-based bidirectional reordering model conditioned on both source and target language is built to arrange the word orders. For more details, please see (Koehn et al., 2003). Since our SMT system is a discriminative model, after all the features are captured, their weights are tuned using minimum error rate training (MERT) (Och, 2003). We choose KenLM (Heafield et al., 2013) as our language model and a stack decoder (Zens and Ney, 2008) with beam search for our system.

Transliteration model. Since the parallel corpus is of small size and the coverage of words is very limited, we apply a transliteration model to translate the OOVs. It models the character re-

lationships between words and generate words at the character level. For the word alignments with character relationship, consider a word pair (e, f), the transliteration model is defined as:

$$p_{tr}(e,f) = \sum_{a \in Align(e,f)} \prod_{j=1}^{|a|} p(q_j) \quad (3)$$

where $Align(e, f)$is the set of possible character alignment sequence, a is one of the alignment sequences, q_j is one alignment. For word pairs without character relation, it is modeled by multiplying source and target character unigram models. The whole model is defined as the combination of transliteration and non-transliteration submodel, where λ is the prior probability of non-transliteration:

$$p_{ntr}(e,f) = \prod_{i=1}^{|e|} p_E(e_i) \prod_{j=1}^{|f|} p_F(f_i) \quad (4)$$

$$p(e,f) = (1-\lambda)p_{tr}(e,f) + \lambda p_{ntr}(e,f) \quad (5)$$

The transliteration model learns the character alignment using expectation maximization (EM) over the character pairs. λ is computed in the tuning stage of the whole system.

3 Experiments

3.1 Implementation

Word representations based on different corpus may have a significant gap on the performance. Larger corpus typically generate better word embeddings. But we only use the shared corpus for comparison.

Data. We use the benchmark monolingual Wikipedia and Europarl copora in the task description (Camacho-Collados et al., 2017) as our data. Especially, we only utilize the EN-DE, EN-ES, EN-it, EN-FA parallel data for translation and cross-lingual embedding training, where `EN`: English, `DE`: German, `FA`: Farsi, `ES`: Spanish, `IT`: Italian.

Preprocessing. For Wikipedia data, we first filter out the stop words using the list from RANKS NL[1]. Then we clean up digits and normalize the marks. Empty lines and web tags are deleted further. For parallel data, we just filter out the stop words and normalize the marks. Parallel data are split with 99% as training set and 1% as develop set for tuning in translation system.

[1]http://www.ranks.nl/stopwords

similarity score. We use the cosine distance of two embeddings as the similarity score of a word pair. Its range is [-1,1].

3.2 Monolingual Experiments

We conduct an experiment on English word embeddings to see the performance of our vectors. We use phrasing and positional context when training. The phrasing is to extract phrased based on co-occurence and the threshold is 400. Positional context treats the same word in different position as different words. Our monolingual embeddings are trained with 500 dimension, 5 iterations, 15 negative samples, `win=5` and `mincount=10`. We use similary part of WordSim353 (Agirre et al., 2009), MEN (Bruni et al., 2012) , M.Turk (Radinsky et al., 2011), Rare Words (Luong et al., 2013) and SimLex (Hill et al.) as test sets, which contain 203, 3000, 287, 2034 and 999 word pairs respectively. The results of our embeddings and in (Levy et al., 2015) of the same window size without phrasing and positional context are listed in Table 1.

The performance of the submitted systems (extra resources are used) including ours (in bold) and RUFINO (the other system uses the same corpus) on all languages are listed in Table 2.

3.3 Cross-lingual Experiments

In the cross-lingual word similarity subtask each word pair is composed by words in different languages. This subtask consists of ten cross-lingual word similarity datasets: EN-DE, EN-ES, EN-FA, EN-IT, DE-ES, DE-FA, DE-IT, ES-FA, ES-IT, and FA-IT. We define the OOVs as the words that can either be found in parallel data or word embeddings. In this subtask, due to the limited amount of parallel data, OOVs occupy a large proportion in the test sets. We show the statistics of OOVs in test sets before, after transliteration model and their final counts after looking up cross-lingual word embeddings in Table 3.

In subtask 2, for the sake of limited time, we did not use phrasing and positional context like in subtask1. For phrases in test sets, we sum up the vectors of all word in the phrase as its embedding. The results of random embeddings that equal to random guess without any semantics, correct results of our system and the top system (Luminoso2) are listed in Table 4.

	WordSim353s		MEN		M.Turk		RareWords		SimLex	
correlation	sp	pr	sp	pr	sp	pr	sp	pr	sp	pr
our embeddings	.814	.800	.769	.756	.650	.684	.444	.416	.436	.435
(Levy et al., 2015)	.772	-	.772	-	.663	-	.454	-	.403	-

Table 1: Performance of English word embeddings on different test sets. `sp` is short for Spearman correlation, `pr` is short for Pearson correlation.

	EN	DE	IT	FA	ES
Luminoso2	.789	.700	.741	.503	.743
Luminoso1	.788	.693	.738	.501	.740
HCCL	**.687**	**.594**	**.651**	**.436**	**.701**
NASARI	.682	.514	.596	.405	.600
RUFINO1	.656	.539	.476	.360	.549
...			...		
hjpwhuer	.0	.024	.048	.0	.0

Table 2: Results on subtask1.

	before TM	after TM		final
EN-DE	117	85	-27.4%	31
EN-ES	71	46	-35.2%	11
EN-IT	72	51	-29.2%	11
EN-FA	120	68	-43.3%	27
DE-ES	166	11	-33.1%	31
DE-IT	156	110	-29.5%	27
DE-FA	190	124	-34.7%	27
ES-IT	119	80	-32.8%	8
ES-FA	153	88	-42.5%	23
IT-FA	155	96	-38.1%	25

Table 3: Counts of OOVs after each steps.

	random	HCCL	Luminoso2
EN-DE	.083	.484	.763
EN-ES	.022	.554	.761
EN-IT	.040	.427	.776
EN-FA	.074	.493	.598
DE-ES	.031	.408	.728
DE-IT	.035	.303	.741
DE-FA	.056	.361	.567
ES-IT	.039	.350	.753
ES-FA	.034	.420	.627
IT-FA	.014	.303	.604
GLOBAL	.053	.464	.754

Table 4: Results on subtask2.

4 Results

Compared with the results in (Levy et al., 2015), our embeddings have an improvement of 4.2% on WordSim353s and 3.3% on SimLex while have a slight decline of 0.3% on MEN, 1.3% on M.Turk and 1.0% on RareWords. Thus phrasing and positional context fail to benefit word embeddings on some test sets. It is also concluded that the embeddings we trained are comparable.

Table 2 shows that our system is ranked 3rd and behave steadily better than RUFINO for subtask1. With phrasing and positional context, Word2vec can achieve satisfying performance.

As we can see in Table 3, up to 43.3% of OOVs are significantly reduced , which are generated at the character level with transliteration model and proved to be real words. It is revealed that our transliteration model can saliently reduce OOVs.

Our cross-lingual system was ranked 8th in official results because of using mismatched data. We rerun our model using the correct data and our true results (will be mentioned in task description paper) listed in Table 4 can rank the 6th. It can be seen that our results for subtask2 are much better than that of the random embeddings, which is equal to guess blindly. However, the gap between the best system and ours is significant. Not enough parallel data and training epochs for non-English embeddings may account for this.

5 Conclusion

For mono-lingual subtask, we train word2vec based word embeddings with positional context and phrasing. For cross-lingual subtask, we built a cross-lingual word representation model and statistical machine translation system with an unsupervised transliteration model, which can greatly translate OOVs. We are the only team that uses the benchmark corpus and achieve good performance on both subtasks. But in global ranking for open resources, there is much space for improvement, i.e. using more iterations, resources and advanced models.

Acknowledgments

We genuinely appreciate Omer Levy for his advice on the monolingual subtask.

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 61271426, 11504406, 11590770, 11590771, 11590772, 11590773, 11590774), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500, XDA06040603), National 863 Program (No. 2015AA016306), National 973 Program (No. 2013CB329302) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 19–27.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 136–145.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of SemEval*. Vancouver, Canada.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1109–1113. http://aclweb.org/anthology/D15-1131.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Association for Computational Linguistics, Gothenburg, Sweden, pages 148–153. http://www.aclweb.org/anthology/E14-4029.

Fernando Enríquez, José A Troyano, and Tomás López-Solaz. 2016. An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications* 66:1–6.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1386–1390. http://www.aclweb.org/anthology/N15-1157.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 58–68. http://www.aclweb.org/anthology/P14-1006.

Felix Hill, Roi Reichart, and Anna Korhonen. ???? Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4):665–695.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., pages 856–864. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1746–1751. http://www.aclweb.org/anthology/D14-1181.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 48–54.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. pages 151–159.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*. pages 104–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 692–702. http://www.aclweb.org/anthology/N16-1083.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pages 160–167. https://doi.org/10.3115/1075096.1075117.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29(1):19–51.

Yangtuo Peng and Hui Jiang. 2016. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *Proceedings of NAACL-HLT*. pages 374–379.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '11, pages 337–346. https://doi.org/10.1145/1963405.1963455.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37:141–188.

Richard Zens and Hermann Ney. 2008. Improvements in dynamic programming beam search for phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*. pages 195–205.

Y. Will Zou, Richard Socher, Daniel Cer, and D. Christopher Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1393–1398. http://aclweb.org/anthology/D13-1141.