Improve Statistical Machine Translation with Context-Sensitive Bilingual Semantic Embedding Model

Haiyang Wu[1] **Daxiang Dong**[1] **Wei He**[1] **Xiaoguang Hu**[1] **Dianhai Yu**[1]
Hua Wu[1] **Haifeng Wang**[1] **Ting Liu**[2]
[1] Baidu Inc., No. 10, Shangdi 10th Street, Beijing, 100085, China
[2] Harbin Institute of Technology, Harbin, China
wuhaiyang,dongdaxiang,hewei,huxiaoguang,yudianhai,
wu_hua,wanghaifeng@baidu.com
tliu@ir.hit.edu.cn

Abstract

We investigate how to improve *bilingual embedding* which has been successfully used as a feature in phrase-based *statistical machine translation* (SMT). Despite bilingual embedding's success, the *contextual information*, which is of critical importance to translation quality, was ignored in previous work. To employ the contextual information, we propose a simple and memory-efficient model for learning bilingual embedding, taking both the source phrase and context around the phrase into account. Bilingual translation scores generated from our proposed bilingual embedding model are used as features in our SMT system. Experimental results show that the proposed method achieves significant improvements on large-scale Chinese-English translation task.

1 Introduction

In Statistical Machine Translation (SMT) system, it is difficult to determine the translation of some phrases that have ambiguous meanings.For example, the phrase "结果 jieguo" can be translated to either "results", "eventually" or "fruit", depending on the context around it. There are two reasons for the problem: First, the length of phrase pairs is restricted due to the limitation of model size and training data. Another reason is that SMT systems often fail to use contextual information in source sentence, therefore, phrase sense disambiguation highly depends on the language model which is trained only on target corpus.

To solve this problem, we present to learn context-sensitive bilingual semantic embedding. Our methodology is to train a supervised model where labels are automatically generated from phrase-pairs. For each source phrase, the aligned target phrase is marked as the positive label whereas other phrases in our phrase table are treated as negative labels. Different from previous work in bilingual embedding learning(Zou et al., 2013; Gao et al., 2014), our framework is a supervised model that utilizes contextual information in source sentence as features and make use of phrase pairs as weak labels. Bilingual semantic embeddings are trained automatically from our supervised learning task.

Our learned bilingual semantic embedding model is used to measure the similarity of phrase pairs which is treated as a feature in decoding. We integrate our learned model into a phrase-based translation system and experimental results indicate that our system significantly outperform the baseline system. On the NIST08 Chinese-English translation task, we obtained 0.68 BLEU improvement. We also test our proposed method on much larger web dataset and obtain 0.49 BLEU improvement against the baseline.

2 Related Work

Using vectors to represent word meanings is the essence of vector space models (VSM). The representations capture words' semantic and syntactic information which can be used to measure semantic similarities by computing distance between the vectors. Although most VSMs represent one word with only one vector, they fail to capture homonymy and polysemy of word. Huang et al. (2012) introduced global document context and multiple word prototypes which distinguishes and uses both local and global context via a joint training objective. Much of the research focus on the task of inducing representations for single languages. Recently, a lot of progress has

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 142–146,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

been made at representation learning for bilingual words. Bilingual word representations have been presented by Peirsman and Padó (2010) and Sumita (2000). Also unsupervised algorithms such as LDA and LSA were used by Boyd-Graber and Resnik (2010), Tam et al. (2007) and Zhao and Xing (2006). Zou et al. (2013) learn bilingual embeddings utilizes word alignments and monolingual embeddings result, Le et al. (2012) and Gao et al. (2014) used continuous vector to represent the source language or target language of each phrase, and then computed translation probability using vector distance. Vulić and Moens (2013) learned bilingual vector spaces from non-parallel data induced by using a seed lexicon. However, none of these work considered the word sense disambiguation problem which Carpuat and Wu (2007) proved it is useful for SMT. In this paper, we learn bilingual semantic embeddings for source content and target phrase, and incorporate it into a phrase-based SMT system to improve translation quality.

3 Context-Sensitive Bilingual Semantic Embedding Model

We propose a simple and memory-efficient model which embeds both contextual information of source phrases and aligned phrases in target corpus into low dimension. Our assumption is that high frequent words are likely to have multiple word senses; therefore, top frequent words are selected in source corpus. We denote our selected words as focused phrase. Our goal is to learn a bilingual embedding model that can capture discriminative contextual information for each focused phrase. To learn an effective context sensitive bilingual embedding, we extract context features nearby a focused phrase that will discriminate focused phrase's target translation from other possible candidates. Our task can be viewed as a classification problem that each target phrase is treated as a class. Since target phrases are usually in very high dimensional space, traditional linear classification model is not suitable for our problem. Therefore, we treat our problem as a ranking problem that can handle large number of classes and optimize the objectives with scalable optimizer stochastic gradient descent.

3.1 Bilingual Word Embedding

We apply a linear embedding model for bilingual embedding learning. Cosine similarity between bilingual embedding representation is considered as score function. The score function should be discriminative between target phrases and other candidate phrases. Our score function is in the form:

$$f(\mathbf{x}, \mathbf{y}; \mathbf{W}, \mathbf{U}) = cos(\mathbf{W}^T\mathbf{x}, \mathbf{U}^T\mathbf{y}) \quad (1)$$

where $\mathbf{x}$ is contextual feature vector in source sentence, and $\mathbf{y}$ is the representation of target phrase, $\mathbf{W} \in R^{|\mathbf{X}|\times k}, \mathbf{U} \in R^{|\mathbf{Y}|\times k}$ are low rank matrix. In our model, we allow $\mathbf{y}$ to be bag-of-words representation. Our embedding model is memory-efficient in that dimensionality of $\mathbf{x}$ and $\mathbf{y}$ can be very large in practical setting. We use $|\mathbf{X}|$ and $|\mathbf{Y}|$ means dimensionality of random variable $\mathbf{x}$ and $\mathbf{y}$, then traditional linear model such as max-entropy model requires memory space of $O(|\mathbf{X}||\mathbf{Y}|)$. Our embedding model only requires $O(k(|\mathbf{X}| + |\mathbf{Y}|))$ memory space that can handle large scale vocabulary setting. To score a focused phrase and target phrase pair with $f(\mathbf{x}, \mathbf{y})$, context features are extracted from nearby window of the focused phrase. Target words are selected from phrase pairs. Given a source sentence, embedding of a focused phrase is estimated from $\mathbf{W}^T\mathbf{x}$ and target phrase embedding can be obtained through $\mathbf{U}^T\mathbf{y}$.

3.2 Context Sensitive Features

Context of a focused phrase is extracted from nearby window, and in our experiment we choose window size of 6 as a focused phrase's context. Features are then extracted from the focused phrase's context. We demonstrate our feature extraction and label generation process from the Chinese-to-English example in figure 1. Window size in this example is three. Position features and Part-Of-Speech Tagging features are extracted from the focused phrase's context. The word *fruit*

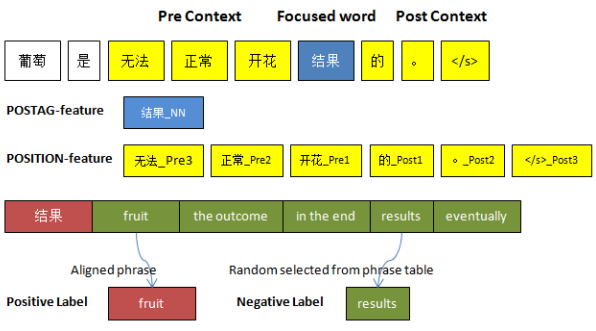

Figure 1: Feature extraction and label generation

is the aligned phrase of our focused phrase and is treated as positive label. The phrase *results* is a randomly selected phrase from phrase table results of 结果. Note that feature window is not well defined near the beginning or the end of a sentence. To conquer this problem, we add special padding word to the beginning and the end of a sentence to augment sentence.

3.3 Parameter Learning

To learn model parameter $\mathbf{W}$ and $\mathbf{U}$, we apply a ranking scheme on candidates selected from phrase table results of each focused phrase. In particular, given a focus phrase w, aligned phrase is treated as positive label whereas phrases extracted from other candidates in phrase table are treated as negative label. A max-margin loss is applied in this ranking setting.

$$I(\Theta) = \frac{1}{m}\sum_{i=1}^{m}(\delta - f(x_i, y_i; \Theta) - f(x_i, y_i'; \Theta))+ \quad (2)$$

Where $f(\mathbf{x_i}, \mathbf{y_i})$ is previously defined, $\Theta = \{W, U\}$ and + means max-margin hinge loss. In our implementation, a margin of $\delta = 0.15$ is used during training. Objectives are minimized through stochastic gradient descent algorithm. For each randomly selected training example, parameters are updated through the following form:

$$\mathbf{\Theta} := \mathbf{\Theta} - \alpha\frac{\partial l(\mathbf{\Theta})}{\partial \mathbf{\Theta}} \quad (3)$$

where $\Theta = \{\mathbf{W}, \mathbf{U}\}$. Given an instance with positive and negative label pair $\{\mathbf{x}, \mathbf{y}, \mathbf{y}'\}$, gradients of parameter $\mathbf{W}$ and $\mathbf{U}$ are as follows:

$$\frac{\partial l(\mathbf{W}, \mathbf{U})}{\partial \mathbf{W}} = qs\mathbf{x}(\mathbf{W}^T\mathbf{x})^T - pqs^3\mathbf{x}(\mathbf{U}^T\mathbf{y}) \quad (4)$$

$$\frac{\partial l(\mathbf{W}, \mathbf{U})}{\partial \mathbf{U}} = qs\mathbf{y}(\mathbf{U}^T\mathbf{y})^T - pqs^3\mathbf{y}(\mathbf{W}^T\mathbf{x}) \quad (5)$$

Where we set $p = (\mathbf{W}^T\mathbf{x})^T(\mathbf{U}^T\mathbf{y})$, $q = \frac{1}{||\mathbf{W}^T\mathbf{x}||_2}$ and $s = \frac{1}{||\mathbf{U}^T\mathbf{y}||_2}$. To initialize our model parameters with strong semantic and syntactic information, word vectors are pre-trained independently on source and target corpus through word2vec (Mikolov et al., 2013). And the pre-trained word vectors are treated as initial parameters of our model. The learned scoring function $f(\mathbf{x}, \mathbf{y})$ will be used during decoding phase as a feature in log-linear model which we will describe in detail later.

4 Integrating Bilingual Semantic Embedding into Phrase-Based SMT Architectures

To incorporate the context-sensitive bilingual embedding model into the state-of-the-art Phrase-Based Translation model, we modify the decoding so that context information is available on every source phrase. For every phrase in a source sentence, the following tasks are done at every node in our decoder:

- Get the focused phrase as well as its context in the source sentence.
- Extract features from the focused phrase's context.
- Get translation candidate extracted from phrase pairs of the focused phrase.
- Compute scores for any pair of the focused phrase and a candidate phrase.

We get the target sub-phrase using word alignment of phrase, and we treat NULL as a common target word if there is no alignment for the focused phrase. Finally we compute the matching score for source content and target word using bilingual semantic embedding model. If there are more than one word in the focus phrase, then we add all score together. A penalty value will be given if target is not in translation candidate list. For each phrase in a given SMT input sentence, the Bilingual Semantic score can be used as an additional feature in log-linear translation model, in combination with other typical context-independent SMT bilexicon probabilities.

5 Experiment

Our experiments are performed using an in-house phrase-based system with a log-linear framework. Our system includes a phrase translation model, an n-gram language model, a lexicalized reordering model, a word penalty model and a phrase penalty model, which is similar to Moses (Koehn et al., 2007). The evaluation metric is BLEU (Papineni et al., 2002).

5.1 Data set

We test our approach on LDC corpus first. We just use a subset of the data available for NIST OpenMT08 task[1] . The parallel training corpus

[1]LDC2002E18, LDC2002L27, LDC2002T01, LDC2003E07, LDC2003E14, LDC2004T07, LDC2005E83, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E24, LDC2006E26, LDC2006E34, LDC2006E86, LDC2006E92, LDC2006E93, LDC2004T08(HK_News, HK_Hansards)

Method	OpenMT08	WebData
	BLEU	BLEU
Our Baseline	26.24	29.32
LOC	26.78**	29.62*
LOC+POS	26.82**	29.81*

Table 1: Results of lowercase BLEU on NIST08 task. LOC is the location feature and POS is the Part-of-Speech feature * or ** equals to significantly better than our baseline($\rho < 0.05$ or $\rho < 0.01$, respectively)

contains 1.5M sentence pairs after we filter with some simple heuristic rules, such as sentence being too long or containing messy codes. As monolingual corpus, we use the XinHua portion of the English GigaWord. In monolingual corpus we filter sentence if it contain more than 100 words or contain messy codes, Finally, we get monolingual corpus containing 369M words. In order to test our approach on a more realistic scenario, we train our models with web data. Sentence pairs obtained from bilingual website and comparable webpage. Monolingual corpus is gained from some large website such as WiKi. There are 50M sentence pairs and 10B words monolingual corpus.

5.2 Results and Analysis

For word alignment, we align all of the training data with GIZA++ (Och and Ney, 2003), using the grow-diag-final heuristic to improve recall. For language model, we train a 5-gram modified Kneser-Ney language model and use Minimum Error Rate Training (Och, 2003) to tune the SMT. For both OpenMT08 task and WebData task, we use NIST06 as the tuning set, and use NIST08 as the testing set. Our baseline system is a standard phrase-based SMT system, and a language model is trained with the target side of bilingual corpus. Results on Chinese-English translation task are reported in Table 1. Word position features and part-of-speech tagging features are both useful for our bilingual semantic embedding learning. Based on our trained bilingual embedding model, we can easily compute a translation score between any bilingual phrase pair. We list some cases in table 2 to show that our bilingual embedding is context sensitive.

Contextual features extracted from source sentence are strong enough to discriminate different word senses. And we also observe from the word “结果 jieguo” that Part-Of-Speech Tagging features are effective in discriminating target phrases.

Source Sentence	4 Nearest Neighbor from bilingual embedding
只有稳定的社会环境，投资者**才能**踏踏实实地做生意。(Investors can only get down to business in a stable social environment)	will be, can only, will, can
在比赛与交往中，中国残疾人显示了非凡的体育**才能**。(In competitions, the Chinese Disabled have shown extraordinary athletic abilities)	skills, ability, abilities, talent
在哥国的自然环境下，葡萄是无法正常开花**结果**的。(In the natural environment of Costa Rica, grapes do not normally yield fruit.)	fruit, outcome of, the outcome, result
结果，东区区议会通过一项议案。(As a result, Eastern District Council passed a proposal)	in the end, eventually, as a result, results

Table 2: Top ranked focused phrases based on bilingual semantic embedding

6 Conlusion

In this paper, we proposed a context-sensitive bilingual semantic embedding model to improve statistical machine translation. Contextual information is used in our model for bilingual word sense disambiguation. We integrated the bilingual semantic model into the phrase-based SMT system. Experimental results show that our method achieves significant improvements over the baseline on large scale Chinese-English translation task. Our model is memory-efficient and practical for industrial usage that training can be done on large scale data set with large number of classes. Prediction time is also negligible with regard to SMT decoding phase. In the future, we will explore more features to refine the model and try to utilize contextual information in target sentences.

Acknowledgments

We thank the three anonymous reviewers for their valuable comments, and Niu Gang and Wu Xianchao for discussions. This paper is supported by 973 program No. 2014CB340505.

References

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 45–55, Cambridge, MA, October. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *Proc. ACL*.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July. Association for Computational Linguistics.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montréal, Canada, June. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics, Volume 29, Number 1, March 2003*. Computational Linguistics, March.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California, June. Association for Computational Linguistics.

Eiichiro Sumita. 2000. Lexical transfer using a vector-space model. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, August.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual-lsa based lm adaptation for spoken language translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 520–527, Prague, Czech Republic, June. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, Atlanta, Georgia, June. Association for Computational Linguistics.

Bing Zhao and Eric P. Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 969–976, Sydney, Australia, July. Association for Computational Linguistics.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, October. Association for Computational Linguistics.