Paraphrasing Out-of-Vocabulary Words with Word Embeddings and Semantic Lexicons for Low Resource Statistical Machine Translation

Chenhui Chu[1] **and Sadao Kurohashi**[2]
[1]Japan Science and Technology Agency
[2]Graduate School of Informatics, Kyoto University
E-mail: chu@pa.jst.jp, kuro@i.kyoto-u.ac.jp

Abstract

Out-of-vocabulary (OOV) word is a crucial problem in statistical machine translation (SMT) with low resources. OOV paraphrasing that augments the translation model for the OOV words by using the translation knowledge of their paraphrases has been proposed to address the OOV problem. In this paper, we propose using word embeddings and semantic lexicons for OOV paraphrasing. Experiments conducted on a low resource setting of the OLYMPICS task of IWSLT 2012 verify the effectiveness of our proposed method.

Keywords: Paraphrasing, Out-of-Vocabulary Word, Word Embedding, Semantic Lexicon

1. Introduction

In statistical machine translation (SMT) (Koehn et al., 2003), because translation knowledge is acquired from parallel data, the quality and quantity of parallel data are crucial. However, except for a few language pairs, such as English-French, English-Arabic, English-Chinese and several European language pairs, parallel data remains a scarce resource. Moreover, even for these language pairs, the available domains are limited.

The scarceness of parallel corpora makes the coverage of the translation model low, which leads to high out-of-vocabulary (OOV) word rates when conducting translation (Callison-Burch et al., 2006). Even we have parallel corpora in sufficient size in one domain, this OOV problem occurs when the domain shifts. Irvine et al. (2013a) showed that SMT performance decreases significantly when using a system trained on one domain to translate texts in different domains mainly because of OOVs.

As one of the ways to address the OOV problem, paraphrasing has been proposed (Callison-Burch et al., 2006; Marton et al., 2009; Razmara et al., 2013). That is augmenting the translation model for the OOV words by using the translation knowledge of their paraphrases in the translation model. Previous studies use paraphrases generated by bilingual pivoting (Callison-Burch et al., 2006), distributional similarity (Marton et al., 2009), and graph propagation (Razmara et al., 2013), which suffer from high computational complexity. In this study, we propose using word embeddings (Mikolov et al., 2013) to address this problem. We also propose using semantic lexicons including WordNet (Miller, 1995), FrameNet (Baker et al., 1998), and the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) for paraphrasing. In addition, we apply a method to combine these two types of paraphrases (Faruqui et al., 2015), which achieves further improvements in SMT.

2. Paraphrasing Out-of-Vocabulary Words

In this paper, we study on phrase based SMT (Koehn et al., 2003). Figure 1 shows an overview of our proposed method. We first construct a phrase table based on unsupervised word alignments, containing phrase pairs together with their feature scores. From the development and test sets, we extract OOV words that do not exist in the phrase table.

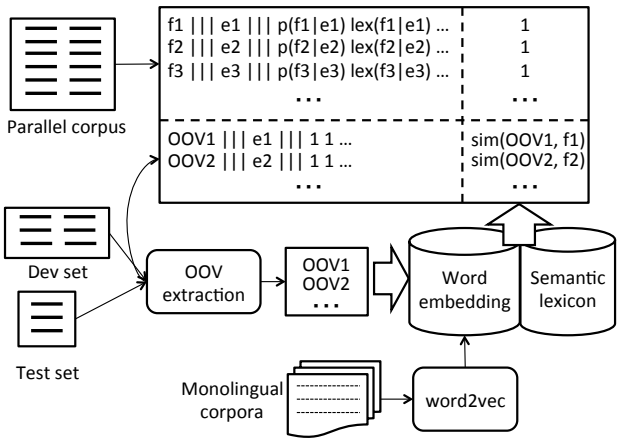

Figure 1: Overview of our proposed method.

Using the word embeddings trained on monolingual data and semantic lexicons, we obtain a list of paraphrases for each OOV word. For each paraphrase existing in the phrase table, we append a new entry for the OOV word to the phrase table, with the corresponding target translation. The similarity between the OOV word and the paraphrase is added as a new feature in the phrase table.[1] This similarity will be tuned along with the other features in a log-linear framework. Following (Razmara et al., 2013), we set the value of this newly introduced feature for original entries in the phrase table to 1. Similarly, the values of original feature scores in the phrase table are set to 1 for the new entries. This newly produced phrase table is then used for tuning and decoding the test sentences.

2.1. Word Embeddings

There are many publicly available word embedding toolkits. Among which we chose the word2vec tool (Mikolov et al., 2013),[2] because of its efficiency and wide uses in nat-

[1]Note that when we use semantic lexicons for paraphrasing, this similarity is set to 1 consistently.

[2]https://code.google.com/p/word2vec/

ural language processing. In particular, we used the skip-gram, which inputs each current word to a log-linear classifier with a continuous projection layer, and predicts its context words within a certain window.
Once the word embeddings were trained, we calculate the cosine similarity between the vector of an OOV word and the vectors of the other words in the vocabulary, and obtain a ranked list of paraphrases for the OOV word. From the list, we kept the top 40 paraphrases for paraphrasing.

2.2. Semantic Lexicons

In our study, we use three types of semantic lexicons:

- WordNet (Miller, 1995): WordNet is a structured semantic lexical database of English. In WordNet, the main relation among words is synonym (word level paraphrase). Words with super/subordinate relations are also linked with hyponym/hypernym labels. In our experiments, we compared two settings
  - WordNet synonyms: only using the words in the synonym relation
  - WordNet all: using the words in all relations
- FrameNet (Baker et al., 1998): FrameNet is a large semantic lexical database of English, constructed based on semantic frames. Frame is a description of a type of event, relation, or entity and the participants in it, and thus two words evoking the same frame are semantically related. In our experiments, we collected the words grouped in same frames and used them for paraphrasing.
- PPDB (Ganitkevitch et al., 2013): PPDB[3] is a large paraphrase database of English, created from parallel corpora through bilingual pivoting (Callison-Burch et al., 2006). The idea of this method is that if two source phrases f_1 and f_2 are translated to the same target phrase e, we can assume that f_1 and f_2 are a paraphrase pair. PPDB is packaged in 6 sizes from S to XXXL to leverage precision and coverage, and it is also divided into to lexical and phrasal paraphrases. In our experiments, we used the lexical paraphrases in the XL size.

2.3. Combination

One problem of word embeddings is that they are learnt without supervision, which limits of the quality. To achieve better quality embeddings, applying the existing semantic lexicons to change the objective of embedding training (Yu and Dredze, 2014), and relation-specific augmentation (Chang et al., 2013) have been studied.
Here, we apply the word embedding retrofitting method (Faruqui et al., 2015), because of its independence from the embedding learning method and efficiency. This method minimizes the following objective so that the retrofitted word embedding q_i will be close to both the original embedding $\hat{q_i}$ and its neighbor q_j in the semantic lexicons:

$$\Psi(Q) = \sum_{i=1}^{n}\{\alpha_i||q_i - \hat{q_i}||^2 + \sum_{(i,j)\in E}\beta_{ij}||q_i - q_j||^2\} \quad (1)$$

where n is the vocabulary size; E denotes a relation in the semantic lexicons; α_i and β_{ij} control the relative strengths of associations, which we set both to 1 in our experiments. Taking the first derivative of Ψ with respect to one q_i vector, we arrive at the following online update by equating it to zero:

$$q_i = \frac{\sum_{j:(i,j)\in E}\beta_{ij}q_j + \alpha_i\hat{q_i}}{\sum_{j:(i,j)\in E}\beta_{ij} + \alpha_i} \quad (2)$$

Following (Faruqui et al., 2015), we run 10 iterations for this update. After retrofitting, we use the new embeddings for paraphrasing in the same manner as before.

3. Experiments

We conducted English-to-Chinese translation experiments in a low resource setting. In all our experiments, we preprocessed the data by segmenting Chinese sentences using a segmenter proposed by Chu et al. (2012a), and tokenizing English sentences.

3.1. Task

We conducted our experiments on the OLYMPICS task of IWSLT 2012 (Federico et al., 2012). The OLYMPICS task is carried out using parts of the HIT Olympic Trilingual Corpus (HIT) (Yang et al., 2006) and the Basic Travel Expression Corpus (BTEC) as an additional training corpus. The HIT corpus is a multilingual corpus that covers 5 domains (traveling, dining, sports, traffic and business) that are closely related to the Beijing 2008 Olympic Games. The HIT corpus contains around 52k sentences 2.8 million words in total. The BTEC corpus is a multilingual speech corpus containing tourism-related sentences. The BTEC corpus consists of 20k sentences including the evaluation data sets of previous IWSLT evaluation campaigns. As SMT systems nowadays are trained on millions of sentences, we consider the OLYMPICS task as a low resource setting. We processed the training corpus using sub-sentence splitting following (Chu et al., 2012b). The development and test sets have only one reference, which contain 1,050 and 998 sentences respectively. For more details of this task, please refer to (Federico et al., 2012).

3.2. Settings

For decoding, we used the state-of-the-art phrase based SMT toolkit Moses (Koehn et al., 2007) with default options. We trained a 5-gram language model on the Chinese side of the parallel corpus using the SRILM toolkit[4] with interpolated Kneser-Ney discounting, and used it for all the experiments. Tuning was performed by minimum error rate training (MERT) (Och, 2003), and it was re-run for every experiment.
The skip-gram model (Mikolov et al., 2013) was trained on the English Gigaword version 5.0,[5] with the word2vec tool. We removed the punctuations in the corpus, obtaining about $3.95B$ tokens with a vocabulary size of $854k$. The context window size was set to 5, and the vector size was set to 200.

[3] http://www.cis.upenn.edu/~ccb/ppdb/

[4] http://www.speech.sri.com/projects/srilm

[5] LDC2011T07

Method	Dev (OOV%)	Test (OOV%)
Baseline	12.96 (10.47%)	10.38 (13.47%)
Word2vec	14.02 (2.00%)	10.87† (4.49%)
WordNet synonyms	13.20 (9.10%)	10.64 (12.03%)
Word2vec retrofitted by WordNet synonyms	13.94 (1.97%)	10.90† (4.40%)
WordNet all	13.04 (8.26%)	10.28 (11.25%)
Word2vec retrofitted by WordNet all	14.00 (1.96%)	10.94† (4.40%)
FrameNet	13.01 (9.84%)	10.45 (12.58%)
Word2vec retrofitted by FrameNet	14.07 (2.00%)	11.09‡ (4.39%)
PPDB	13.72 (3.26%)	10.50 (3.40%)
Word2vec retrofitted by PPDB	**14.36** (1.36%)	**11.18**‡ (4.43%)

Table 1: Translation results evaluated on BLEU-4 scores and OOV rates ("†" and "‡" denote that the result is significantly better than "Baseline" at $p < 0.05$ and $p < 0.01$ respectively).

3.3. Results

Translation results using different methods are shown in Table 1. "Baseline" denotes the system without OOV paraphrasing; "Word2vec" denotes the system paraphrased with the word embeddings obtained by word2vec; "WordNet synonyms", "WordNet all", "FrameNet" and "PPDB" denote the systems paraphrased with different semantic lexicons. "Word2vec retrofitted *" denote the systems paraphrased with the word2vec word embeddings retrofitted by different semantic lexicons. They were evaluated on BLEU-4 scores and OOV rates. The OOV rate was the percentage of the OOV words out of the total number of source words in the development/test sets. The significance test was performed using the bootstrap resampling method proposed by Koehn (2004).

We can see that the OOV rate of the Baseline system is high, because of the small size of the parallel corpus for training. Both Word2vec and semantic lexicons decrease the OOV rate, and thus improve the MT performance. Although Word2vec is unsupervised learnt, it shows better results than the semantic lexicons that are either manual created or collected with supervised data. The reason for this is the lower coverage of the semantic lexicons compared to Word2vec, leading to lower OOV decreases. The combination of Word2vec and the semantic lexicons by retrofitting outperforms either method, because of the quality improvement of the word embeddings. Word2vec retrofitted by PPDB achieved the best performance. We believe the reason for this is the higher coverage of PPDB compared to the other semantic lexicons, leading to more improvement of the word embeddings.

To further understand the reason of the improvement, we analyzed the translation difference between Baseline and Word2vec retrofitted by PPDB. Figure 2 shows two translation examples. In example 1, the Baseline system failed to translate the word "booked". Although "booked" does not exist in the training corpus, the word "booking" appears several times. By paraphrasing "booked" to "booking", Word2vec retrofitted by PPDB successfully translated it. The "Briggs" does not exist in the corpus either. However, Word2vec retrofitted by PPDB incorrectly paraphrased it to "Smith", leading to this incorrect translation. In example 2, the word "noontime" is paraphrased to "lunchtime" and translated as it is. Although, the meaning of "noontime" and "lunchtime" are slightly different, it still helps understanding. Based on our analysis, most improvements belong to the "booked" case, while a few improvements belong to the "noontime" case. There are also many "Briggs" cases that the paraphrases are incorrect, but basically they would not hurt the qualities of the translations as they were OOVs in the Baseline system.

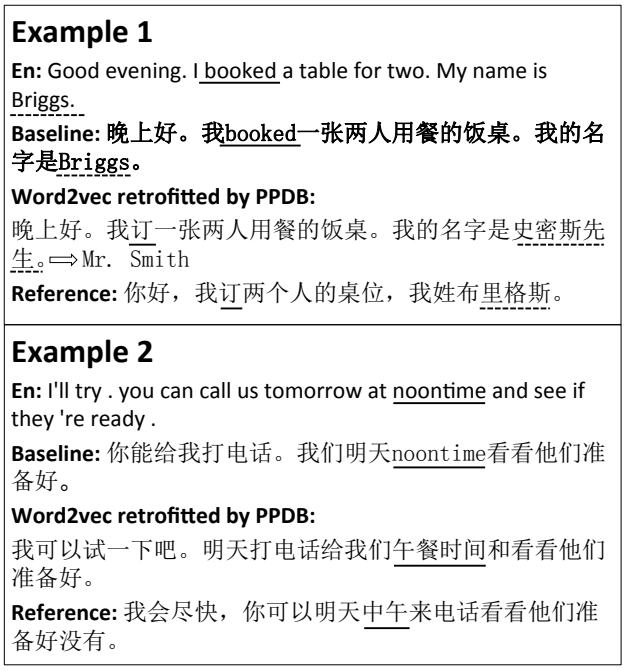

Figure 2: Translation examples of Baseline and Word2vec retrofitted by PPDB.

4. Related Work

Previous studies have proposed different paraphrase generation methods for OOV paraphrasing. Callison-Burch et al. (2006) use paraphrases generated by bilingual pivoting. In our experiments, we compared the extension of their method, which is the PPDB setting in Section 3. Marton et al. (2009) firstly used distributional similarity to generated paraphrases for OOVs. Razmara et al. (2013) extended the method of (Marton et al., 2009) via graph propagation. The

drawback of both (Marton et al., 2009) and (Razmara et al., 2013) is that they suffer from high computational complexity. The monolingual data used in their experiments were only tens of millions of tokens. We address this drawback with word embeddings (Mikolov et al., 2013), making the paraphrase generation scalable to monolingual data with billions of tokens. Moreover, we propose using semantic lexicons together with word embeddings for OOV paraphrasing.

Bilingual lexicon extraction (BLE) (Chu et al., 2014) is another common method to address the OOV problem. BLE extracts the translations for the OOVs from comparable corpora, which are a set of monolingual corpora that describe roughly the same topic in different languages. Daume III et al. (2011) firstly proposed BLE with canonical correlation analysis for the OOV problem. Irvine et al. (2013b) proposed a monolingual marginal matching BLE method for this. Irvine et al. (2013) extracted the translations for the OOV words using a supervised method. All these studies could be complemented with our study.

5. Conclusion

In this paper, we addressed the OOV problem for low resource SMT by paraphrasing with word embeddings and semantic lexicons. Experimental results verify the effectiveness of our proposed method.

As future work, we plan to try other word embeddings methods, and develop more advanced methods for combining with the semantic lexicons. Moreover, we plan to conduct experiments on language pairs that have scarce parallel corpora, to complement the experiments in this paper that were conducted on a language pair that lacks of parallel corpora in a particular domain.

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June. Association for Computational Linguistics.

Chang, K.-W., Yih, W.-t., and Meek, C. (2013). Multi-relational latent semantic analysis. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1612, Seattle, Washington, USA, October. Association for Computational Linguistics.

Chu, C., Nakazawa, T., Kawahara, D., and Kurohashi, S. (2012a). Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 35–42, Trento, Italy, May.

Chu, C., Nakazawa, T., and Kurohashi, S. (2012b). Ebmt system of kyoto university in olympics task at iwslt 2012. In *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT 2012)*, pages 96–101, Hong Kong, China, December.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2014)*, pages 8404:2:296–309, Kathmandu, Nepal, April. Springer Lecture Notes in Computer Science (LNCS).

Daume III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA, June. Association for Computational Linguistics.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June. Association for Computational Linguistics.

Federico, M., Cettolo, M., Bentivogli, L., Paul, M., and Stüker, S. (2012). Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Irvine, A. and Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria, August. Association for Computational Linguistics.

Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. (2013a). Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*, 1:429–440.

Irvine, A., Quirk, C., and Daumé III, H. (2013b). Monolingual marginal matching for translation model adaptation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1088, Seattle, Washington, USA, October. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (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*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Fed-

erico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore, August. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Miller, G. A. (1995). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Och, F. J. (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.

Razmara, M., Siahbani, M., Haffari, R., and Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yang, M., Jiang, H., Zhao, T., and Li, S., (2006). *Construct Trilingual Parallel Corpus on Demand*, volume 4274, chapter Lecture Notes in Computer Science, pages 760–767. Chinese Spoken Language Processing.

Yu, M. and Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland, June. Association for Computational Linguistics.