EasyChair Preprint

№ 518

Improving Performance of NMT Using Semantic Concept of WordNet Synset

Fangxu Liu, Jinan Xu, Guoyi Miao, Yufeng Chen and Yujie Zhang

September 20, 2018

Improving Performance of NMT Using Semantic Concept of WordNet Synset

Fangxu Liu, JinAn Xu*, Gouyi Miao, Yufeng Chen, Yujie Zhang

School of Computer and Information Technology
Beijing Jiaotong University

{fangxuliu,jaxu,gymiao,chenyf,yjzhang}@bjtu.edu.cn

Abstract. Neural machine translation (NMT) has shown promising progress in recent years. However, for reducing the computational complexity, NMT typically needs to limit its vocabulary scale to a fixed or relatively acceptable size, which leads to the problem of rare word and out-of-vocabulary (OOV). In this paper, we present that the semantic concept information of word can help NMT learn better semantic representation of word and improve the translation accuracy. The key idea is to utilize the external semantic knowledge base WordNet to replace rare words and OOVs with their semantic concepts of WordNet synsets. More specifically, we propose two semantic similarity models to obtain the most similar concepts of rare words and OOVs. Experimental results on 4 translation tasks[1] show that our method outperforms the baseline RNNSearch by 2.38~2.88 BLEU points. Furthermore, the proposed hybrid method by combining BPE and our proposed method can also gain 0.39~0.97 BLEU points improvement over BPE. Experiments and analysis presented in this study also demonstrate that the proposed method can significantly improve translation quality of OOVs in NMT.

Keywords: NMT; Semantic concept of synset; Rare words; Unknown words.

1 Introduction

In the past few years, Neural Machine Translation (NMT) has made rapid progress and it has shown state-of-the-art performance [1-3]. However, for the purpose of reducing the computational complexity, NMT typically needs to limit its vocabulary scale to an appropriate size, and this leads to rare word and OOV problems. Both Sutskever et al. (2014) and Bahdanau et al. (2015) observed that sentences with high ratio of rare words tend to be translated much more poor than sentences mainly containing frequent words.

To address the rare word and OOV problems, researchers have proposed several different methods. Luong et al. [4] proposed to annotate target unknown words with

[1] We verify the effectiveness of our method on four translation tasks, including English-to-German, German-to-English, English-to-Chinese and Chinese-to-English.

positional information to track their alignments. This method utilizes the position information, but it lacks the ability of taking advantage of linguistic knowledge such as syntax and semantics. Sennrich et al. [5] and Wu et al. [3] proposed to address the rare word problem by splitting the words into sub-word units through unsupervised learning. These methods significantly alleviate the rare word problem and have been widely used in practice. However, these methods also suffer from the problem caused by the sparseness of rare words in the monolingual corpus used to train BPE or word-piece model. Thus the rare word problem still remains challenging.

In this paper, to address the rare word problem, we propose to replace the rare words and unknown words with their semantic concepts of WordNet synsets so as to better obtain their semantic information during training and testing. Different from traditional methods, our method explicitly integrates the concepts embedding of rare words and unknown words into NMT, and it can better learn the semantic representations of rare words and unknown words. An example is shown in Figure 1:

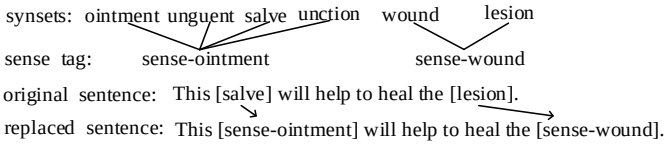

Fig. 1. An illustration of our main idea. We address rare words or OOVs of sentences in training set, and the rare words "ointment", "unction", "unguent" and "salve" are synonyms and they can be replaced and represented with their same semantic concept tag "sense-ointment". Also, rare words "wound" and "lesion" have the same treatment.

More specifically, during training, for the rare words in English side, we first collect their synonyms using WordNet [6], and annotate the rare words with the most similar semantic concepts. Then this new bilingual corpus with rare words replaced will be used to train a NMT model. To get the most similar semantic concepts of rare words, two models were proposed: 1) a RNN LM-based similarity model to compute the similarity on continuous space; 2) a statistical LM-based similarity model to compute the similarity on discrete space. During testing, we determine the detailed method according to the factor that English is the source or target language: 1) If English is the source language, we first replace the rare words with their semantic concept tags, and then the sentences are translated by the trained NMT model; 2) If English is the target language, the target sequence of words generated from the decoder of NMT may contain some semantic concept tags of rare words. We use attention mechanism and the bilingual phrase table to restore the semantic concept tags and get the final translation. Figure 2 illustrates the processing of our method.

Experiments show that our method can improve performance by up to 1.79 and 1.4 BLEU points over PosUnk [4] on the WMT 14 translation tasks of English-to-German and German-to-English, respectively. It can also outperform the PosUnk system by up to 1.32 and 1.03 BLEU points on English-to-Chinese and Chinese-to-English tasks, respectively. Furthermore, the proposed hybrid method by combining BPE and our method can also gain 0.39~0.97 BLEU points improvement over BPE.

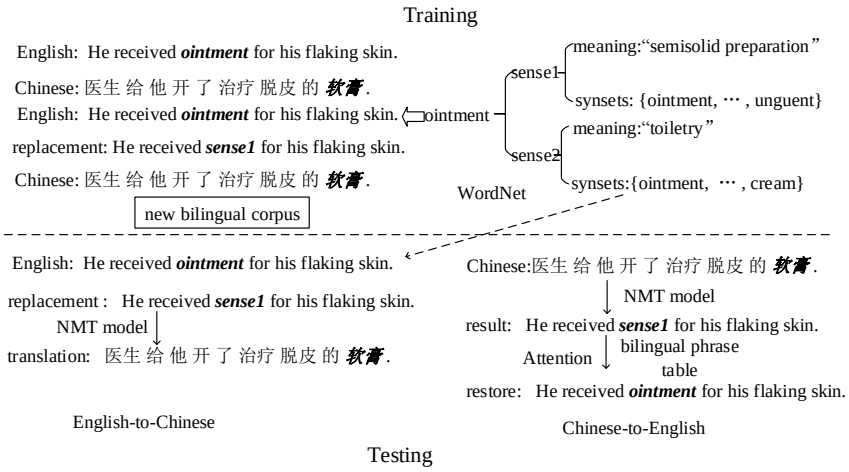

Fig. 2. An example of processing rare words for training and testing

2 Neural Machine Translation and Impact of Rare Words

This section will briefly introduce the NMT method, and analyze the impact of the rare words on NMT.

2.1 Neural Machine Translation

Attention-based encoder-decoder framework [1] is used in most of the state-of-the-art NMT models. The encoder consists of a bidirectional recurrent neural network (Bi-RNN), which reads a source sequence X(x_1,⋯,x_t) and generates a sequence of forward hidden states $\vec{h} = (\overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_n})$ and a sequence of backward hidden states $\overleftarrow{h} = (\overleftarrow{h_1}, \overleftarrow{h_2}, \dots, \overleftarrow{h_n})$. We obtain the annotation h_i of each source word x_i by concatenating the forward hidden state $\vec{h}$ and the backward hidden state $\overleftarrow{h}$; then they are calculated using two RNNs from left-to-right and right-to-left, respectively, as follows:

$$\vec{h_i} = f_{RNN}(x_i, \vec{h}_{i-1}) \tag{1}$$

$$\overleftarrow{h_i} = f_{RNN}(x_i, \overleftarrow{h}_{i-1}) \tag{2}$$

The decoder consists of a RNN, an attention network and a logical regression network. At each time step i, the probability $p(y_i \mid y_{<i}, \theta)$ is computed as follows:

$$\mathrm{p}(y_i | y_{<i}, \theta) = \mathrm{g}(s_i, y_{i-1}, c_i) \tag{3}$$

where the hidden state s_i is generated based on the previous hidden state s_{i-1}, the previous predicted word y_{i-1}, and the context vector c_i:

$$s_i = f_{RNN}(y_{i-1}, s_{i-1}, c_i) \tag{4}$$

where c_i is calculated as a weighted sum of the source annotations.

$$c_i = \sum_{k=1}^{n} \alpha_{jk} h_k \tag{5}$$

A detailed description can be found in Bahdanau et al. [1].

2.2 Impact of Rare Words

As discussed in the introduction part, the rare word problem has two major negative effects: first, treating all the unknown words as the same *unk* symbol undermines the semantic integrity of the sentences; second, the sparseness of rare words makes it difficult to learn better representation from training data.

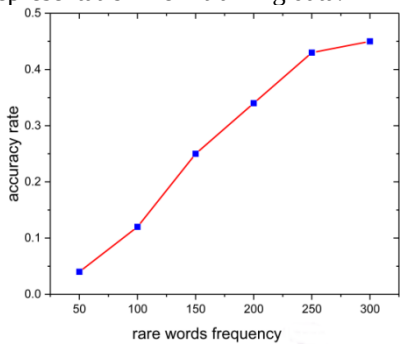

Fig. 3. Frequency vs. Accuracy

To illustrate the impact of rare words to translation, we designed the following experiment: A thousand sentences were extracted from the United Nations parallel corpus English-to-Chinese translation data set. And then these sentences were translated by NMT model. Finally, we manually analyzed the translation accuracy of words with different frequencies. The results in Figure 3 show that the words with higher frequency tend to be better translated, while the words with lower frequency typically tend to be incorrectly translated.

3 Methodology

The analysis in the above section shows that the rare words have a considerable influence on the translation performance. So we propose to replace rare words in training and testing data with their semantic concepts by employing the semantic knowledge resource WordNet [6]. Specifically, we design two strategies of semantic similarity model to obtain the most similar semantic concepts of the rare words:

- RNN LM-based model, which computes the similarity on continuous space.
- Statistical LM-based model, which calculates the similarity on discrete space.

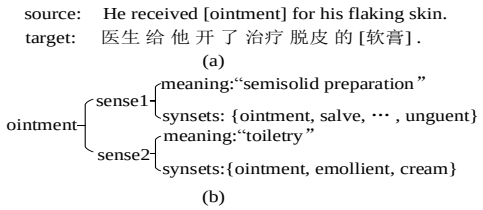

Fig. 4. An example of illustrating our method. (a)A parallel sentence pair. (b)The semantic concepts of "ointment" in WordNet.

To illustrate our method, we introduce a parallel sentence pair as an example, as shown in Figure 4. (a). Note that the brackets indicate the rare word of the source side and its counterpart of the target side. From WordNet, we can get the semantic concepts of "ointment", as shown in Figure 4. (b). It shows how the semantic concepts of word "ointment" are organized in WorNet [6].

3.1 WordNet

WordNet [6] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Words in the same set of semantic concepts have the same meaning and typically can be used in the same context. These concepts are separately organized into four networks, and each semantic concept corresponds to a synonym set. The semantic concepts are connected by various relationships, such as synonymy, hypernym and hyponymy. In this paper, we leverage synonymy to process the rare words and OOVs.

3.2 Model 1: RNN LM-based Similarity Model

Our first model seeks to employ sense embedding and neural language model to obtain the most similar semantic concepts of rare words.

We illustrate our model 1 with a concrete example in Figure 4. To get the most similar semantic concept of the rare word "ointment", we first collect all the synsets of the word "ointment" from WordNet, and each synset expresses a distinct semantic concept. In this case, the collected synsets are described: synset_1={ ointment, … , unguent } and synset_2={ ointment, cream, emollient }, and the two synsets correspond semantic concepts sense_1 and sense_2, respectively. Then we construct the semantic concept embeddings of the word "ointment" based on its synsets, namely synset_1 and synset_2. Specifically, the vector of sense_1 can be expressed as a weighted average of all embeddings of words in synset_1. Formally, the vector representation of semantic concept 'sense$_i$' is calculated as follows:

$$vec(sense_i) = \frac{vec('word_1') + \cdots + vec('word_n')}{n} \tag{6}$$

where word_j (1≤j≤n) denotes a word in $synset_i$, n is the number of words in $synset_i$, and the embedding of word_j is learned from LSTM-RNN language model [7] that trained on large scale monolingual corpus.

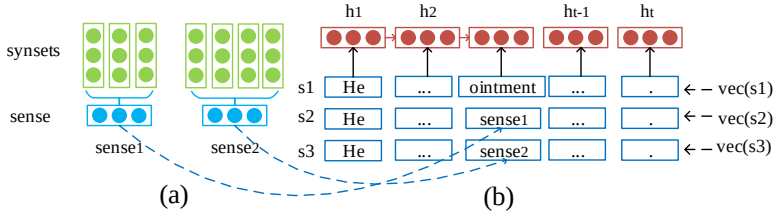

Fig. 5. An example of selecting the most similar semantic concept

As illustrated in Figure 5, we determine the most similar semantic concept of the rare word "ointment". More specifically, word embedding of "ointment" in sentence s1 is replaced with sense_1 and sense_2 and thus two new sentences s2 and s3 are generated. We define a set of sentences, S={ s2, s3…}, which contain the generated sentences. Then we compute the similarity between sentence s and sentence s' by means of the cosine distance of vec(s) and vec(s′), where vec(s) and vec(s′) are the embeddings of sentence s and sentence s', and s is the original sentence (i.e. s1), s' is from set S. The similarity formula is described as follows:

$$sim(s, s') = \cos(vec(s), vec(s')) \tag{7}$$

where we use the last hidden state (i.e. h_t in Figure 5) as the sentence vector.

Finally, the semantic concept, which embedded in the sentence s' with the highest similarity, is chosen as the most similar semantic concept w_{sense}, as follows:

$$s_{w_{sense}} = \arg\max_{s' \in S} sim(s, s') \tag{8}$$

After obtaining the most similar semantic concept of a rare word, we replace the rare word with its most similar semantic concept and get a sentence with the semantic concept tag (i.e. "He received [sense_1] for his flaking skin.", where sense_1 is the semantic concept calculated by formula 8.) .

3.3 Model 2: Statistical LM-based Similarity Model

Similar to Model 1, our Model 2 first collect all the synsets of the word "ointment" from WordNet (i.e. synset_1 and synset_2 in section 3.2), and each synset denote a distinct semantic concept. Unlike Model 1, we seek to determine the most similar semantic concept of the rare word "ointment" by computing the score of semantic concepts. Specifically, for each synset, we construct a list of sentences, which is described as $S = \{s_1, s_2, \ldots, s_n\}$, by replacing the rare word with each word of the synset. Then we compute the score of the semantic concept with the following equation:

$$\text{score}(sense_i) = \frac{lm(s_1)+lm(s_2)+\cdots+lm(s_n)}{n} \tag{9}$$

where n is the number of sentences in S, $\mathrm{lm}(s_i)$ is the statistical language model (3-gram) score of the sentence s_i $(1 \leq i \leq n)$, and the statistical language model is trained on large scale monolingual corpus.

Finally, we choose the semantic concept with the highest score as the most similar semantic concept:

$$w_{sense} = \arg\max_{1 \leq i \leq n} score(sense_i) \tag{10}$$

3.4 Integrating Semantic Concepts into NMT

In this section, we seek to explicitly integrate semantic concepts in NMT by replacing rare words with their corresponding semantic concept tags. We describe our work in detail in training and testing phases. Note that, we only handle the words with frequency less than *M* in the training data and in the scope of WordNet [6]. And the effects of different threshold *M* is compared in section 4.2.

During training, for each sentence pair of the parallel corpus, we replace the rare words in the English side with the most similar semantic concept tags by applying model 1 or model 2. And then the new generated parallel corpus is fed into NMT for training. We also need to learn word level alignment for sentence pairs in the bilingual corpus. As a byproduct, a lexical translation table can be derived from the aligned bilingual corpus. Our method can also combine with BPE: first we replace the rare words with the most similar semantic concept tags, and the remaining rare words will be split into sub-word units by BPE.

During testing, if English is the source language, the rare words in the source sentence will be first replaced with their corresponding semantic concept tags, and then the new sentence will be translated by the trained NMT model. If a target word e_j in the translation aligns to a semantic concept tag of the source side, we will restore e_j to the translation of the original source word via the lexical translation table and attention mechanism, where the lexical translation table is derived from the aligned bilingual corpus and it provides word level alignment for sentence pairs, and the attention mechanism provides a kind of soft alignment that helps to find the corresponding source word for each target word. In another case, if English is the target language, first the trained NMT model will translate the test sentence and get the translation that may contains concept tags. Then we track the source words of the concept tags by attention mechanism. With the help of the lexical translation table and the language models (3-gram), the concept tags will be restored to the translation of the source words and get the final translation result.

4 Experiments

We evaluate the effectiveness of the proposed method on English-to-German, German-to-English, English-to-Chinese and Chinese-to-English translation tasks. Translation quality is measured by the BLEU metric [8].

4.1 Experimental Settings

We performed experiments with attention-based RNNSearch [1] system on corpus extracted from the shared translation task of WMT 2014[2] (German ↔ English) and the United Nations parallel corpus v1.0 [9] (Chinese ↔ English). For German ↔ English translation, the training set contained about 3.5 million sentence pairs. For Chinese ↔ English, the training set contained about 3 million sentence pairs. Further, we used GIZA++ to obtain the alignment information from the same bilingual data. We trained a neural language model on monolingual data that contained about 10 million English sentences extracted from the WMT 2014. In addition, the statistical language model was trained with SRILM [10]. We used WMT newstest2013 as our development set, and reported results on newstest2012 and newstest2014 for German ↔ English. Besides, we used the open development set and the test set provided in the UN parallel corpus for Chinese ↔ English.

The hyperparameters were set as follows: the number of the hidden units was 512 for both the encoder and decoder, and the word embedding dimension was 512 for all source and target words. The parameters in the network were updated with the adadelta algorithm, a minibatch size of 64, and reshuffled the training set between epochs. We used a beam size of 10 for the beam search, with probabilities normalized by the sentence length. The dropout method was used at the readout layer, and the dropout rate was set to 0.5.

4.2 Preliminary Experiments

A preliminary experiment was performed to determine the threshold *M* and the vocabulary size *V* mentioned in section 3.4. We experimented on English-to-German translation task and chose WMT newstest2013 as the test set. We set different threshold *M* and vocabulary size *V* during the experiment, and recorded the number of unknown words. The experimental results are shown in Table 1 and Table 2:

Table 1. unknown words of different M and V in newstest2013

V \ M	baseline	250	200	100	50
30000	2363	2203	2259	2281	2309
40000	1920	1714	1742	1783	1861
50000	1661	1521	1546	1604	1627

Table 2. BLEU scores (%) of different M and V

V \ M	baseline	250	200	100	50
30000	16.52	**17.26**	17.03	16.80	16.61
40000	17.69	18.37	**18.46**	18.12	17.87
50000	18.91	19.48	**19.52**	19.28	19.20

[2] http://www.statmt.org/wmt14

As seen from Table 1, when the threshold *M* is 250 and the vocabulary size *V* is 50000, newstest2013 contains fewer unknown words. Similarly, the results in Table 2 show that, when *M* is 200, we can gain a better BLEU score.

Based on the results of Table 1 and Table 2, for reducing the computational complexity, the threshold *M* is set to 200 and the vocabulary size is set to 40000 in our comparative experiments.

4.3 Comparative Experiments and Main Results

There are 8 different systems in our comparative experiments:

1. RNNSearch: Our baseline NMT system with improved attention mechanism.
2. PosUnk: The system, proposed by Luong et al. [4], annotated target *unk* as *unk-k*, where *k* indicates the position information of the source word.
3. w2v&lm&restore: The system, proposed by Li et al. [11], replaced the unknown words with the similar in-vocabulary words.
4. wn&lm&restore: The system, proposed by Li et al. [12], replaced the unknown words of source language (English) with the similar in-vocabulary words using WordNet.
5. ours-model1&restore: Our system, presented in section 3.2, replaced rare words and unknown words with their semantic concept tags using our model1.
6. ours-model2&restore: Our system, presented in section 3.3, replaced rare words and unknown words with their semantic concept tags using our model2.
7. BPE: The system, proposed by Sennrich et.al [5], decorated rare words with subword units.
8. ours-model1&BPE: We first address rare words using the proposed ours-model1&restore, and the remaining rare words are processed by BPE.

The remaining unknown words in system 3 ~ system 6 (w2v&lm&restore, wn&lm&restore, ours-model1&restore, ours-model2&restore) were processed by the method of PosUnk. We use the same dataset and network parameters to train the comparison model, as seen in section 4.1.

Table 3. English→German BLEU scores (%) of different systems

System	13(dev)	12	14	Average
RNNSearch	17.69	15.22	15.55	16.15
PosUnk	18.80	16.23	16.69	17.24
w2v&lm&restore	19.87	17.35	17.68	18.30
wn&lm&restore	20.02	17.73	17.98	18.57
ours-model2&restore	**19.92**	**18.01**	**18.13**	**18.68**
ours-model1&restore	**20.25**	**18.39**	**18.45**	**19.03**
BPE	21.37	19.33	19.29	19.99
ours-model1&BPE	**21.98**	**19.82**	**19.66**	**20.48**

On English-to-German translation task, as seen in Table 3, our method outperforms the RNNSearch by 2.88 BLEU points and surpasses the other three traditional methods (PosUnk, w2v&lm&restore, wn&lm&restore) by 0.46~1.79 BLEU points. And the results show that ours-model1& BPE (our method combined with BPE) can achieve 0.49 BLEU points improvement over BPE. Also, the results in Table 4 show

that ours-model1&restore and ours-model1& BPE can achieve improvement on English-to-Chinese translation task. Generally, our approach can effectively improve translation performance by processing rare words of the source side.

Table 4. English→Chinese BLEU scores (%) of different systems.

System	UN(dev)	UN(test)	Average
RNNSearch	34.03	34.60	34.31
PosUnk	35.41	35.95	35.68
w2v&lm&restore	36.49	36.63	36.56
wn&lm&restore	36.83	36.74	36.78
ours-model2&restore	**36.97**	**36.89**	**36.93**
ours-model1&restore	**37.05**	**36.96**	**37.00**
BPE	37.93	37.89	37.91
ours-model1& BPE	**38.34**	**38.27**	**38.30**

Table 5. German→English BLEU scores (%) of different systems

System	13(dev)	12	14	Average
RNNSearch	22.51	19.64	20.04	20.73
PosUnk	23.69	20.86	21.11	21.88
w2v&lm&restore	24.58	21.21	22.10	22.63
ours-model2&restore	**24.60**	**21.42**	**21.93**	**22.65**
ours-model1&restore	**25.26**	**21.98**	**22.62**	**23.28**
BPE	26.23	23.89	25.15	25.09
ours-model1&BPE	**26.92**	**24.47**	**26.11**	**25.83**

Table 6. Chinese→English BLEU scores (%) of different systems.

System	UN(dev)	UN(test)	Average
RNNSearch	41.33	41.86	41.60
PosUnk	42.58	43.32	42.95
ours-model2&restore	43.04	43.96	43.50
w2v&lm&restore	43.10	44.06	43.58
ours-model1&restore	**43.55**	**44.41**	**43.98**
BPE	43.80	44.75	44.28
ours-model1& BPE	**45.01**	**45.50**	**45.25**

On German-to-English translation task, the results in Table 5 show that our method outperforms the baseline RNNSearch by 2.55 BLEU points. It also surpasses the systems PosUnk and w2v&lm&restore by 1.4 and 0.65 BLEU points, respectively. And the results show that ours-model1&BPE (our method combined with BPE) can gain 0.74 BLEU points improvement over BPE. In addition, the results in Table 6 show that ours-model1&restore and ours-model1&BPE can also achieve improvement on Chinese-to-English translation task. In general, our approach can effectively improve translation performance by processing rare words of the target side.

4.4 Analysis

Analysis of translation performance

From the results in section 4.3, our proposed method can achieve significant improvement and outperform the traditional methods [4,11,12]. Different from previous work, we replace rare words and unknown words with similar semantic concepts, and explicitly integrate semantic concepts into NMT. Thus, the concept embeddings of rare words and unknown words are used when training neural networks, which enhance the ability of NMT models to learn better semantic representation of rare words and unknown words.

Currently, BPE is an effective method which splits the rare words into sub-word units. However, fine-grained sub-word units lead to a certain degree of loss of semantic information. Unlike BPE, our work focuses on enhancing NMT model by obtaining accurate semantic information of rare words. Intuitively, our method and BPE are complementary to each other. The results in section 4.3 shows that the hybrid method by integrating BPE and our proposed method can achieve significant improvement on all translation tasks.

Analysis of *unk* translation

To have a further insight into the translation quality of our method, we manually analyzed the effect of the translation of *unk* in English-to-Chinese and Chinese-to-English translation tasks. The results are shown in Table 7:

Table 7. : Translation accuracy of unknown words, where the test set provided in the UN parallel corpus, contains 4000 sentences. "*": indicates that the number of *unk* to be processed and it is equal to the number of *unk* appearing in RNNSearch. "-": indicates that the method cannot process unknown words of Chinese.

	English-to-Chinese		Chinese-to-English	
System	total(unk)	correct	total(unk)	correct
RNNSearch	941	103(11%)	1179	230(19%)
PosUnk	*	205(22%)	*	336(28%)
w2v&lm&restore	*	272(29%)	*	379(32%)
wn&lm&restore	*	285(30%)	-	-
ours-model1&restore	*	320(34%)	*	414(35%)
BPE	*	375(39%)	*	506(42%)
ours-model1&BPE	*	417(44%)	*	554(46%)

From the data in Table 7, on English-to-Chinese and Chinese-to-English translation tasks, one observes that in general more unknown words can be correctly translated in our improved system than in the baseline system. Accordingly, replacing unknown words with their semantic concepts can better maintain the integrity of the sentences and effectively improve the translation accuracy of the unknown words.

Analysis of the applicability scope of our method

Since our approach resorts to the external semantic knowledge resources (i.e. WordNet [6]), our study is suited to translation tasks that contain languages supported by semantic knowledge resources. More importantly, our research is a further attempt to integrate semantic knowledge into NMT and the experimental results demonstrate that the proposed method can effectively improve the translation accuracy.

5 Conclusion

Rare words problem is a major factor that affect the performance of NMT. To address this problem, we propose to use the external semantic knowledge base WordNet to replace rare words and unknown words with their semantic concepts of WordNet synsets, and explicitly integrate the concept embeddings of rare words into NMT, which can help NMT model better capture the semantic information of rare words. Experiments on 4 translation tasks show that the proposed method can significantly improve the translation quality. Further analysis shows that our method is able 1) to better capture the word sense of rare words, 2) to improve the translation accuracy of unknown words, 3) to combine BPE and achieve a certain improvement over BPE.

In future work, we will explore further strategies to integrate semantic concepts into NMT. Additionally, we also hope to explore a general way to solve the problem that the acquisition of semantic information is limited by semantic resources (i.e. WordNet [6]).

Acknowledgments. The research work has been supported by the National Nature Science Foundation of China (Contract 61370130, 61473294 and 61502149), and Beijing Natural Science Foundation under Grant No. 4172047, and the Fundamental Research Funds for the Central Universities (2015JBM033), and the International Science and Technology Cooperation Program of China under grant No. 2014DFA11350.

References

1. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. Computer Science, 2014.
2. Vaswani, Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. CoRR abs/1706.03762.
3. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation [J]. 2016.
4. Luong M T, Sutskever I, Le Q V, et al. Addressing the Rare Word Problem in Neural Machine Translation[J]. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Veterinary Medicine, 2014, 27(2):82-86.
5. Sennrich, Sennrich, Rico, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. Computer Science 2015.
6. Miller G A. WordNet: a lexical database for English [J]. Communications of the Acm, 1995, 38(11):39--41.
7. Palangi, Hamid, Palangi H, Deng L, Shen Y. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Transactions on Audio Speech & Language Processing 24.4(2015) pages 694-707.
8. Kishore Papineni, Salim Roukos, et al. Bleu: a method for auto matice valuation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July2002.

9. Ziemski, Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. The United Nations Parallel Corpus, Language Resources and Evaluation (LREC'16), Portorož, Slovenia, May 2016 (Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016)
10. Stolcke A. Srilm --- An Extensible Language Modeling Toolkit[C]// International Conference on Spoken Language Processing. 2002:901--904.
11. Li X, Zhang J, Zong C. Towards zero unknown word in neural machine translation[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2016:2852-2858.
12. Li S, Xu J, Miao G, et al. A Semantic Concept Based Unknown Words Processing Method in Neural Machine Translation[M]// Natural Language Processing and Chinese Computing. 2018.
13. Koehn, Philipp Koehn et al., 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the ACL-2007 Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
14. Luong, Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412-1421, Lisbon, Portugal, September 2015.