A Semantic Concept Based Unknown Words Processing Method in Neural Machine Translation

Shaotong Li*, Jinan Xu, Guoyi Miao, Yujie Zhang, Yufeng Chen

School of Computer and Information Technology
Beijing Jiaotong University

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

Abstract. The problem of unknown words in neural machine translation (NMT), which not only affects the semantic integrity of the source sentences but also adversely affects the generating of the target sentences. The traditional methods usually replace the unknown words according to the similarity of word vectors, these approaches are difficult to deal with rare words and polysemous words. Therefore, this paper proposes a new method of unknown words processing in NMT based on the semantic concept of the source language. Firstly, we use the semantic concept of source language semantic dictionary to find the candidate in-vocabulary words. Secondly, we propose a method to calculate the semantic similarity by integrating the source language model and the semantic concept network, to obtain the best replacement word. Experiments on English to Chinese translation task demonstrate that our proposed method can achieve more than 2.6 BLEU points over the conventional NMT method. Compared with the traditional method based on word vector similarity, our method can also obtain an improvement by nearly 0.8 BLEU points.

Keywords: NMT; Unknown words; Semantic dictionary; Semantic concept

End-to-end NMT is a kind of machine translation method proposed in recent years[1-3]. Most of the NMT systems are based on the encoder-decoder framework, the encoder encodes the source sentence into a vector, and the decoder decodes the vector into the target sentence. Compared with the traditional statistical machine translation (SMT), NMT has many advantages. Firstly, NMT automatically models the sentences and learns the features. The problems of feature selection in SMT are very well solved. Secondly, NMT almost does not require any knowledge of the language field. Thirdly, NMT does not rely on large-scale phrase tables, rule tables and language models, so the space occupied by parameters will be greatly reduced.

Despite the above advantages, NMT still has the problem of unknown words which is caused by the limited vocabulary scale. In order to control the temporal and spatial expenses of the model, NMT usually uses small vocabularies in the source side and the target side[4]. The words that are not in the vocabulary are unknown words, which will be replaced by an UNK symbol. A feasible method to solve this problem is to find out the substitutes in-vocabulary words of the unknown words. Li et al. proposed a replacement method based on word vector similarity [4], the unknown words are replaced by the synonyms in the vocabulary through the cosine distance of the word vector and the language model. However, there are some unavoidable problems with this approach. Firstly, the vectors of rare words are difficult to train; Secondly, the trained word vectors cannot express various semantics of the polysemous words and cannot adapt to the replacement of the polysemous words in different contexts.

To solve these problems, this paper proposes an unknown words processing method which combines the semantic concept of the source language. This method uses WordNet's semantic concept network to seek the near-synonyms of the unknown words as candidate replacement words, and calculate the semantic similarity by integrating the source language model and the semantic concept network to select the best alternative words to replace the unknown words.

Experiments on Chinese to English translation tasks demonstrate that our proposed method can achieve more than 2.6 BLEU points over the baseline system, and also outperform the traditional method based on word vector similarity by nearly 0.8 BLEU points.

The main contributions of this paper are shown as follows:

- An external semantic dictionary is integrated to solve the problem of unknown words in NMT.
- The semantic concept in WordNet is used to obtain the replacement word, which can solve the problem of rare words and polysemous words better.
- A semantic similarity calculation method by integrating the source language models and semantic concept network is proposed. It ensures that the replacement words are

close to the unknown words in semantic level, and the purpose of keeping the semantics of the source sentence can be achieved as much as possible.

1 NMT and the Problem of Unknown Words

This section will introduce NMT and the impact of the unknown words on NMT.

1.1 Neural Machine Translation

Most of the proposed NMT systems are based on the encoder-decoder framework and attention mechanism.

The encoder consists of a bidirectional recurrent neural network (Bi-RNN), which can read a sequence $X(x_1,\ldots,x_t)$ and generate a sequence of forward hidden states（$\overrightarrow{h_1},\cdots,\overrightarrow{h_\tau}$） and a sequence of backward hidden states （$\overleftarrow{h_t},\cdots,\overleftarrow{h_1}$）. We obtain the annotation h_i for each source word x_i by concatenating the forward hidden state $\overrightarrow{h_i}$ and the backward hidden state $\overleftarrow{h_i}$.

The decoder consists of a recurrent neural network (RNN), an attention network and a logical regression network. At each time step i, the RNN generates the hidden state s_i based on the previous hidden state s_{i-1}, the previous predicted word y_{i-1}, and the context vector c_i which is calculated as a weighted sum of the source annotations by the attention network. Then the logical regression network predicts the target word y_i.

1.2 The Problem of Unknown Words

When predicting the target word at each time step, it is necessary to generate the probability of all the words in the target vocabulary. Therefore, the output dimension of the logical regression network is equal to the target vocabulary size, the total computational complexity will grow almost proportional to the vocabulary size. So train the model with the whole vocabulary is infeasible, which leads to the problem of unknown words caused by the limitation of the vocabulary size.

In NMT system, the unknown words mainly lead to two problems: Firstly, the NMT model is difficult to learn the representation and the appropriate translation of the unknown words, the parameter quality of the neural network is poor. Secondly, the existence of the unknown words increases the ambiguity of the source sentence, affects the alignment accuracy of attention network and the quality of translation result.

2 Framework of Our Method

This paper proposes an approach of unknown words processing with WordNet. In translating process, we firstly mark the input sentence with the part of speech tag (POS-tag), then get the candidate in-vocabulary words set of the unknown words in WordNet. Then, we calculate the semantic similarity between the unknown words and the words in the candidate in-vocabulary words set by integrating the language model and the semantic concept network of WordNet to attain the word with highest similarity as a replacement word. After replacement, we use the trained NMT model to translate the replaced input sentences. During translating, the alignment probabilities of each target words are obtained by the attention network. Finally, the translation of replaced words will be restored by the alignment probabilities and a bilingual dictionary.

The framework of our method is shown in Figure 1.

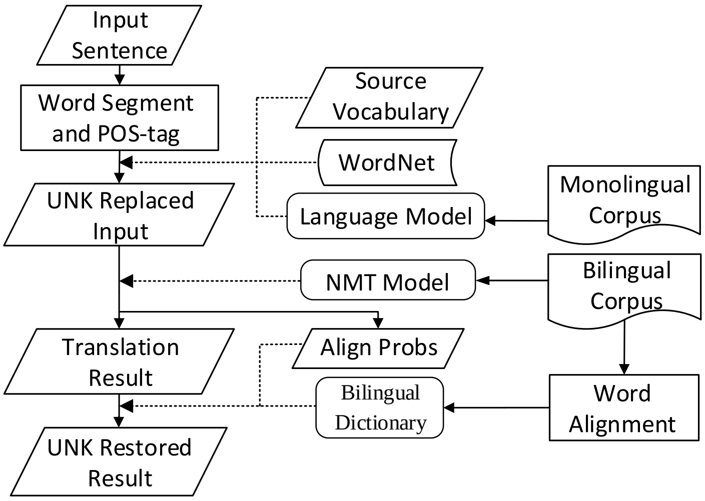

Fig. 1. Framework of our unknown words processing method

2.1 WordNet

WordNet[5] is an English semantic dictionary. WordNet3.0 contains 117659 concept nodes, with 82115 noun nodes, 13767 verb nodes, 18156 adjective nodes and 3621 adverb nodes. These concepts are separately organized into four networks, each semantic concept corresponds to a synonym set. The semantic concepts are connected by various relationships. The most commonly used relationships are Synonymy and Hypernymy/Hyponymy, which are the semantic relations used in this paper.

Synonymy: Synonymous relationship is the most basic semantic relation in WordNet. The synonym set of a semantic concept is derived from the synonymy relationship.

Hypernymy/Hyponymy: Hypernymy/Hyponymy relationships are the most important semantic relations in WordNet, there is an "*is a*" relationship between the hypo-concept and the hyper-concept, which is transitive.

In this paper, WordNet is adopted to obtain the candidate in-vocabulary words of unknown words. Compared with the corpus-based word vector similarity approach, it has the following advantages: Firstly, it is feasible to find various concepts of polysemy in WordNet, and ensure that the candidate in-vocabulary words set contains replacement words of various semantics. Secondly, as long as the corresponding semantic concept of the unknown word can be found in WordNet, the difference between rare words and common words can be eliminated effectively.

2.2 Replacement of Unknown Words

This paper proposes a set of near-synonym which contains all the similar in-vocabulary words of an unknown word based on semantic concepts. The near-synonym set includes the synonyms of all concepts of the unknown word, the synonyms of the hyper-concepts, and the synonyms of the hypo-concepts. The maximum depth of the semantic concept of the hyper-concept is defined as m. The maximum depth of the hypo-concepts is defined as n. And m and n are adjusted by the preliminary experiment. The strategies for selecting the near-synonyms of the unknown words are as follows:

1. Give the unknown word w and its POS-tag w_{pos}, we find all the semantic concepts of w whose POS-tag is w_{pos}, as a semantic concept set *Synsets_w*.
2. Get the synonyms of all semantic concepts in *Synsets_w*, add all the in-vocabulary words in them to the near-synonym set of w and mark their distance to w as 0.
3. If the near-synonym set of w is empty, then find all the hyper-concepts of concepts in *Synsets_w* whose POS-tag is w_{pos}, within the depth of m, denoted as a semantic concept set *Hyper_synsets_w*.
4. Get the synonyms of all semantic concepts in *Hyper_synsets_w*, add all the in-vocabulary words in these synonyms to the near-synonym set of w and mark their distance to w as the depth of their corresponding hyper-concepts.
5. If the near-synonym set of w is empty, then find all the hypo-concepts of concepts in *Synsets_w* whose POS-tag is w_{pos}, within the depth of n, and form a semantic concept set *Hypo_synsets_w*.

6. Find the synonyms of all semantic concepts in *Hypo_synsets_w*, add all the in-vocabulary words in these synonyms to the near-synonym set of *w* and mark their distance to *w* as the depth of their corresponding hypo-concepts.

According to the priority order of the same semantic concept, the hyper-concepts and the hypo-concepts, we generate the candidate in-vocabulary words set of *w*.

The replacement words should not only be close to the unknown words in semantics, but also should keep the semantics of the original sentence as much as possible. Therefore, this paper defines a semantic similarity calculation approach by integrating the source language model and the semantic concept network, to calculate the semantic similarity between the words in the candidate in-vocabulary words set and the target unknown words, and then selects the best replacement words.

Firstly, a n-gram language model is trained on a source language monolingual corpus. And scores all the candidate words in the context of source sentence. We trained the 3-gram language model in our experiment, so for an unknown word w_i and its candidate replacement word w_i', where *i* refers the position of *w* in source sentence, the score on the 3-gram language model is defined as formula 1:

$$Score_{3\text{-}gram}(w_i',w_i)=\frac{p(w_i'|w_{i-1},w_{i-2})+p(w_{i+1}|w_i',w_{i-1})+p(w_{i+2}|w_{i+1},w_i')}{3} \tag{1}$$

The similarity of word pair (w_i',w_i) in WordNet is defined as formula 2:

$$\mathrm{Sim}_{WordNet}(w_i',w_i)=\frac{1}{path(w_i',w_i)+1} \tag{2}$$

Where path(w_i',w_i) means the distance between the corresponding concepts of w_i' and w_i in WordNet, which is obtained in the selection strategies mentioned before.

The semantic similarity of the word pair (w_i',w_i) is finally defined as the formula 3:

$$Sim(w_i',w_i)=\sqrt{Score_{3\text{-}gram}(w_i',w_i)\cdot Sim_{WordNet}(w_i',w_i)} \tag{3}$$

Finally, the word in the candidate in-vocabulary words with highest similarity is chosen as the final replacement word W_{best}, as shown in formula 4:

$$W_{best}=\arg\max_{w'\in S} Sim(w',w) \tag{4}$$

2.3 Restore Translation for Unknown Words

NMT model is a sequence to the sequence model, we can only find the most likely alignment through the align probability of attention network. However, the performance of the attention network in NMT model is very unstable. In order to reduce the effect of alignment errors, a judged operation is added to the alignment: We align the words in the training corpus with GIZA++[1] to get a bilingual dictionary, which will contain all words in training corpus and their translations. For the word t_i in the output sentence, if t_i aligns to a replaced word s_j, the previously obtained bilingual dictionary will be used to determine the correctness of the alignment: If word pair(s_i, t_i) is in the bilingual dictionary, then the alignment is correct, then replace t_i with the translation of original source word. Otherwise t_i will be kept in the output sentence.

3 Experiments

Since WordNet is an English semantic dictionary, we verify our approach on the English to Chinese translation task.

3.1 Settings

The bilingual data used to train NMT model is selected from the CWMT2015 English-Chinese news corpus, including 1.6 million sentence pairs. The development set and test set are officially provided by CWMT2015, each with 1000 sentences. The word alignment is also carried out on the training set. The language model and the word vectors will be trained on the monolingual data, which contains 1.6 million source sentences in the training set and other 5 million source sentences selected from the CWMT2015 English-Chinese news corpus.

Use the BLEU score[6] to evaluate the translation results.

3.2 Training Details

The hyperparameters of our NMT system are described as follows: the vocabulary size of the source side is limited to 20k, and the target side, 30k. The number of hidden units is 512 for both encoder and decoder. The word embedding dimension of the source and target words is 512. The parameters are updated with Adadelt algorithm[7].The Dropout method[8] is used at the readout layer, and the dropout rate is set as 0.5.

[1] http://code.google.com/p/giza-pp/downloads/list

3.3 Preliminary Experiments

A preliminary experiment is required to determine the maximum depth of the hyper-concept (*m*) and that of the hypo-concept (*n*) in the replacement strategy mentioned in section 2.2. Here the unknown words coverage rate of replacement strategy on the development set and the translation quality after replacement on the development set are the basis of determining the maximum depth. Experiments show that when *m* is greater than 5 or *n* is greater than 1, the size of the near-synonym set is no longer increased. Therefore, the variation range of *m* is limited from 0 to 5, and the variation range of *n* is limited from 0 to 1. The experimental results are as table 1 and table 2:

Table 1. Unknown words coverage rate (%) of different maximum depth

n m	0	1	2	3	4	5
0	45.9	66.6	71.7	73.1	73.8	73.7
1	46.3	67.5	72.2	74	74.3	74.6

Table 2. BLEU scores (%) of different maximum depth

n m	0	1	2	3	4	5
0	26.02	26.44	26.74	26.81	26.79	26.78
1	26.04	26.46	26.77	**26.82**	26.8	26.79

It can be seen from Table 1 that when *m* is greater than 3 and *n* is greater than 1, the changes of coverage rate with depth is not obvious. Table 2 shows that when *m* equals 3 and *n* equals 1, the BLEU score is the highest. When *m* is greater than 3, the BLEU score appears to decrease. With the expansion of the search depth, some candidate words that differ greatly in semantics with target unknown words appear in the near-synonym set resulting in a decline in translation quality.

Based on the results of Table 1 and Table 2, in our comparative experiments, the maximum depth of the hyper-concept in the replacement strategy is set to 3, and the maximum depth of the hypo-concept is set to 1.

3.4 Comparative Experiments and Main Results

There are 7 different systems in our comparative experiments:

1. Moses[9]: An open-source phrase-based SMT system with default configuration.
2. RNNSearch: Our baseline NMT system with improved attention mechanism[10].
3. PosUnk: Add an approach proposed by Luong et al.[11] to the baseline NMT system in order to process unknown word problem.

4. w2v&lm: Based on our baseline NMT system, use the approach proposed by Li et al.[4] to replace the unknown words in source language based on source language word vector and the source language model. The word vector is trained by word2vec[12] toolkit, and the 3-gram language model with modified Kneser-Ney smoothing is trained by SRILM[13].
5. w2v&lm_restore&PosUnk: Based on system 4, the restore approach proposed in Section 2.3 is used to translate the replaced words, and the remaining unknown words are processed by the method that was used in system 3.
6. wn&lm： Based on the baseline NMT system, our method will use WordNet and the source language model to replace the unknown words in source language. The language model used is the same as the language model used in system 4.
7. wn&lm_restore&PosUnk: Based on the system 6, the replacing approach proposed in Section 2.3 is used to translate the replaced words, and the remaining unknown words are processed by the method that was used in system 3.

System 6 and 7 our approaches. System 6 is an NMT system that only adds the unknown words replacing module of source side, and besides that, system 7 also adds a replaced word restoring module of target side as well as the remaining unknown word processing module of target side, so it is our best system. The main experimental results are shown in Table 3.

Table 3. BLEU scores (%) of different systems

System	Development set	Test set	Average
moses	28.42	25.48	26.95
RNNSearch	25.71	23.22	24.47
PosUnk	27.58	24.89	26.24
w2v&lm	26.26	23.73	25
wn&lm	**26.82**	**24.91**	**25.87**
w2v&lm_restore&PosUnk	27.66	25.02	26.34
wn&lm_restore&PosUnk	**28.33**	**25.86**	**27.1**

The pre-processing NMT system based on our approach (wn&lm) outperforms the baseline system (RNNSearch) by 1.4 BLEU points on average; It also surpasses the pre-process NMT system based on traditional method (w2v&lm) by 0.87 BLEU points in the development set and 1.18 BLEU points in the test set. Our best system (wn&lm_restore&PosUnk) outperforms the baseline system (RNNSearch) by 2.63 BLEU on average; In addition, it surpasses the NMT system which add a simple unknown word processing module (PosUnk) by 0.86 BLEU points, it significantly improves the best NMT system of traditional approach (w2v&lm_restore& PosUnk) by 0.76 BLEU points.

3.5 Comparison of Translating Details

Here we compare the translating details of our system with other systems, we mainly analyze the translating process of unknown words. The main advantage of our approach is that the replacement words selected by our system are more appropriate.

Table 4. Translation instances table

Source	*This paragraph **restores(unk)** the old text to preserve a delicate balance.*
Reference	*这个 段落 恢复 旧 的 案文 以 保持 一 种 微妙 的 平衡 。*
RNNSearch result	*这 一 段 是 为了 保存 一 个 微妙 的 平衡 而 UNK 的 。*
PosUnk result	*这 一 段 是 为了 保存 一 个 微妙 的 平衡 而 恢复 的 。*
Replaced source(with w2v&lm)	*This paragraph **impair** the old text to preserve a delicate balance.*
w2v&lm result	*该 段 损害 了 保持 微妙 平衡 的 旧 案文 。*
w2v&lm_restore&PosUnk result	*该 段 恢复 了 保持 微妙 平衡 的 旧 案文 。*
Replaced source(with wn&lm)	*This paragraph **repair** the old text to preserve a delicate balance.*
wn&lm result	*该 段 修复 了 旧 的 案文 以 保持 微妙 的 平衡 。*
wn&lm_restore&PosUnk result	*该 段 恢复 了 旧 的 案文 以 保持 微妙 的 平衡 。*
Source	***Signifying(unk)** an exact **divisor(unk)** or factor of a quantity.*
Reference	*表示 某 数量 的 准确 除数 或者 因数 。*
RNNSearch result	*控制 数量 的 一 个 精确 的 UNK 。*
PosUnk result	*控制 数量 的 一 个 精确 的 除数 。*
Replaced source(with w2v&lm)	***Name** an exact **divisor(unk)** or factor of a quantity.*
w2v&lm result	*名称 一 个 数量 的 UNK 或 UNK 。*
w2v&lm_restore&PosUnk result	*名称 一 个 数量 的 确切 或 因素 。*
Replaced source(with wn&lm)	***Mean** an exact **factor** or factor of a quantity.*
wn&lm result	*指 数量 的 一 个 确切 因素 或 因素 。*
wn&lm_restore&PosUnk result	*表示 数量 的 一 个 确切 除数 或 因素 。*

On the one hand, our proposed method keeps the original meaning of source sentences better and provides less impact on subsequent translations. For example, in Example 1, the source word "*restores*" is an unknown word; An "*UNK*" symbol appears, and there is no proper translation of "*restores*"; even after post-processing, "*restores*" is translated to "*恢复*", other parts of the translation still need improving; The approach based on word vector and language model replaces "*restores*" with an in-vocabulary word "*impair*", but the semantic changes have occurred. It affects the subsequent of translation result; Our system replaces "*restores*" with an in-vocabulary word "*repair*", basically keeping the original meaning. The translation result and the reference are basically the same, and a post-processing module is followed to correct the translation of "*restores*", and ultimately get a better translation results.

On the other hand, appropriate replacement words make the attention better and solve the problem of over-translating and under-translating to some extent. As in Example 2, the source word "*signifying*", "*divisor*" are unknown words; An "*UNK*" symbol appears, there is no proper translations of "*divisor*" and “*factor*”, and the translation of unknown word “*signifying*” is not correct; Because it only generates one “UNK”, post-processing only get back the correct translation of “*divisor*” and lose others; The approach based on word vector and language model replaces "*signifying*" with an in-vocabulary word "*name*", but the meaning has been changed. In addition, it fails to replace the unknown word "*divisor*". The translation result contains two “*UNK*”, affects the alignment of attention network, and post-processing fails to get the correct translation of "*signifying*", parts of the translation are then retrieved, but the translation of “*divisor*” are still missed; Our system replaces "*signifying*" with an in-vocabulary word "*mean*", and replaces "*divisor*" with an in-vocabulary word "*factor*". There is no “*UNK*” in translation, and the alignment is quite good. Therefore, we successfully correct the translation of "*signifying*" and "*divisor*" by post-processing.

4 Conclusion and Future Work

This paper proposes an unknown words processing approach in NMT by integrating semantic concepts of the source language and source language model. This method not only improves the translation of the unknown words but also ensures that the semantics of the source language sentence is complete, and enhances the quality of the entire translation. Moreover, this approach provides a new idea for NMT to integrate external knowledge base. Experiments on English to Chinese translation show that our method not only achieves a significant improvement over the baseline, but also provides some advantages compared with the traditional unknown words processing methods.

Our future work mainly contains three aspects. Firstly, the replacement method proposed in this paper is limited to the replacement of word level. We are going to challenge the phrase level method; Secondly, our proposed method is difficult to deal with some items, for instance, named entities. These entities can be classified and marked as the corresponding symbols; Thirdly, in this paper, the external semantic dictionary is only used in the decoding part. We will focus on integrating the external semantic dictionary into the training process of NMT model.

Acknowledgments. The authors are supported by the National Nature Science Foundation of China (Contract 61370130 and 61473294), and the Fundamental Research Funds for the Central Universities (2015JBM033), and the International Science and Technology Cooperation Program of China under grant No. 2014DFA11350.

Reference

1. Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]// 2013.
2. Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in Neural Information Processing Systems, 2014, 4:3104-3112.
3. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.
4. 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.
5. Miller G A. WordNet: a lexical database for English[J]. Communications of the Acm, 1995, 38(11):39--41.
6. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for auto matice valuation of machine translation. InProceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania,USA,July2002.
7. Zeiler M D. ADADELTA: An Adaptive Learning Rate Method[J]. Computer Science, 2012.
8. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
9. Collins M, Koehn P. Clause restructuring for statistical machine translation[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005:531-540.
10. Meng F, Lu Z, Li H, et al. Interactive Attention for Neural Machine Translation[J]. 2016.
11. 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.
12. Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
13. Stolcke A. Srilm --- An Extensible Language Modeling Toolkit[C]// International Conference on Spoken Language Processing. 2002:901--904.