ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

Key Research of Pre-processing on Mongolian-Chinese Neural Machine Translation

Jian Du, Hongxu Hou*, Jing Wu, Zhipeng Shen, Jinting Li, and Hongbin Wang

College of Computer Science, Inner Mongolia University, Hohhot, China

*Corresponding author

***Abstract*—Neural machine translation has recently achieved promising results with the big scale corpus. But there is little research on the small scale corpus, such as Mongolian. Mongolian belongs to the agglutinative language while Chinese is a pictograph. It is necessary to do some pre-processing for both Mongolian and Chinese before training the machine translation. In this paper, we successfully build an attention-based neural machine translation to do the CWMT2009 Mongolian to Chinese translation task. We also use four different approaches, respectively, to do the pre-processing for both Mongolian and Chinese, including segmenting Chinese into character, separating the Mongolian stem from the suffixes, addressing the case suffix and converting Mongolian into Latin. We carry out a lot of experiments to evaluate the approaches. We achieve the best BLEU with the score of 29.56. It is 1.82 points in BLEU score higher than the baseline which is trained with the original Mongolian and the general word segmentation of Chinese.**

Keywords-Mongolian-Chinese translation; neural machine translation; pre-processing; attention-based

I. INTRODUCTION

Neural networks have been successfully applied to the machine translation since Mikolov (2013) converted the words into vectors [1][2], and recently achieved promising results. It is the neural machine translation (NMT) that is usually applied to the language of big scale corpus like English and French. However, There is little research on the small scale corpus especially Mongolian.

The research on machine translation of Mongolian began in the 1980s. And now it has basically achieved the statistical machine translation. The research on the NMT of Mongolian is just starting up. Mongolian is an agglutinative language. Depending on the different suffixes and positions of a sentence, Mongolian words have a wealth change of inflections and meanings. Thus how to address Mongolian and make it easy to train the NMT models is still a pending problem. Otherwise, Chinese belongs to the pictograph, which is also a problem for the machine translation. Moreover, some pre-processing has successfully increased the performance of the statistical machine translation (SMT), we wonder if it is still useful to the NMT.

In this paper, we successfully build an attention-based NMT to do the CWMT2009 Mongolian to Chinese translation task. Different pre-processing has been done to analyze the influence of Mongolian-Chinese attention-based NMT. There is a brief introduction of the pre-processing as follows:

- We segment Chinese into character. By this way, it is effective to reduce the vocabulary size of the NMT and solve the problem of wrong segmentation. It also improves the alignment of the words.
- Separating the Mongolian stem from the suffixes is a basic method to do the pre-processing as Mongolian words are extremely rich in forms with different suffixes. We combine the dictionary, rule and statistics-based approaches to segment the stem.
- How to address the case suffix is a significant approach to preprocess Mongolian. In Mongolian, the case suffix is a special suffix that carries the grammatical information improving the fluency of a sentence. We cut down the case suffix in this paper to train the attention-based NMT.
- We use Latin as a conversion of Mongolian to train the models. The Latin character is easier to represent and it costs less memory than Mongolian. It can alleviate the mistaken spelling problem of Mongolian as well.

We also carry out a lot of experiments to evaluate these methods whether they are effective or not. At the end of the paper, it gives some analysis of each pre-processing approach. We have achieved the best BLEU score of 29.56 on average with the approaches of converting Mongolian into Latin and segmenting Chinese into character. It is 1.82 points in BLEU score higher than the baseline, which is without any pre-processing of attention-based NMT.

The rest of the paper is organized as follows. Section II will show the details of neural machine translation. The introduction of pre-processing is in section III. How we do our experiments will be described at section IV. Analysis and conclusions appear in section V and section VI.

II. NEURAL MACHINE TRANSLATION

Neural networks used in machine translation have recently achieved promising results. Sutskever, Vinyals, and Le (2014) [3] and Bahdanau, Cho, and Bengio (2014) [4] directly built neural networks to perform end-to-end translation, named neural machine translation (NMT). A basic form of NMT contains two components: an encoder that converts a source sentence $s_1,s_2,\ldots,s_n$ into a fixed vector c, and a decoder which uses the fixed vector c to generate one target word t_j at a time. The NMT system is aimed to predict the max conditional

probability $p(t|s)$ of translating a source sentence to a target sentence using the following formula:

$$\log p(t|s) = \sum_{j=1}^{m} \log p\left(y_j | y_{j<m}, c\right) \quad (1)$$

A. RNN Encoder-Decoder

The most commonly used NMT work is recurrent neural network (RNN) which is also one of the basic NMT architectures. Other NMT architectures, different from the RNN, differ in terms of which architectures are used for the decoder and how encoder computes representation c of the source sentence. Kalchbrenner and Blunsom (2013) [5] used a standard RNN hidden unit for the decoder and a convolutional neural network for encoding the source sentence representation. However, at both the encoder and the decoder, Sutskever et al. (2014) [3] and Luong et al. (2015) [6] used multiple layers of an RNN with a Long Short-Term Memory (LSTM) hidden units. Cho et al. (2014) [7], Bahdanau et al. (2015) [8], and Jean et al. (2015) [9] all adopted a different version of the RNN with an LSTM-inspired hidden unit, the gated recur-rent unit (GRU), for both components. Bahdanau et al. (2015) [8] successfully applied the attention mechanism into the NMT and proposed attention-based NMT to change the fixed vector *c*.

We can see the general process in the FIGURE I. The left of the dotted line is encoder which converts the source inputs $s_1, s_2, \ldots, s_n$ into a fixed vector *c*. The other side is decoder that generates one target word t_j at each time with the fixed vector *c*.

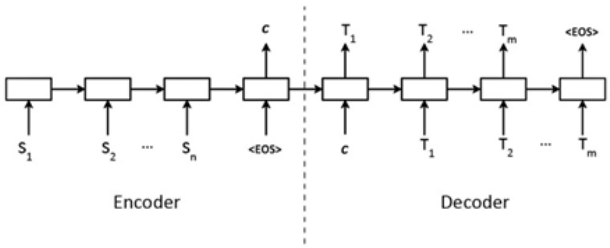

FIGURE I. A BRIEF INTRODUCTION OF RNN. THE LEFT OF THE DOTTED LINE IS ENCODER AND THE OTHER SIDE IS DECODER. THE EOS IS THE SYMBOL OF AN END OF A SENTENCE

B. Attention-based NMT

The attention mechanism has gained popularity recently in neural network. It allows models to learn alignments from different modalities. But it was usually used in the task of image objects, speech frames and visual features of a picture until Bahdanau et al. (2015) [8] successfully applied the attention mechanism to jointly translate and align words. Minh-Thang Luong et al. (2015) [10] proposed effective approach that can improve the translation effects of attention-based NMT. Attention-based NMT encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation. See FIGURE II for illustration of a single step of decoding in attention-based neural machine translations. We put Mongolian as source sentence into the encoder then compute the vector *c* and soft alignment α_{ij}. At the decoder, we predict the word "雨 ($y\check{u}$)" with the vector *c* and alignment α_{ij}.

In this paper, we use the attention-based neural machine translation to do our research. We build the attention-based NMT with soft alignment, bi-directional hidden units at encoder and general units at decoder.

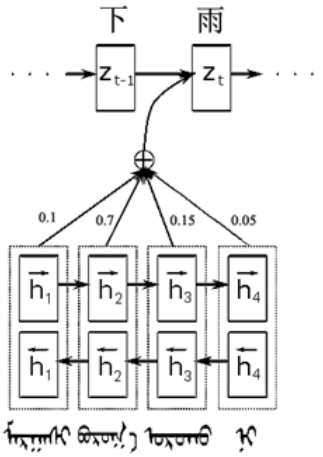

FIGURE II. AN ILLUSTRATION OF A SINGLE STEP OF DECODING IN ATTENTION-BASED NMT.

III. PREPARE YOUR PAPER BEFORE STYLING

Chinese, as the synthetic language, belongs to the Sino Tibetan language family. Because Chinese is a pictograph, we need to do some pre-processing while training the attention-based NMT, such as segmentation. However, Mongolian, as an agglutinative language, it is extremely rich in changing and meanings with different suffixes and different positions of a sentence. We need more pre-processing that will be introduced in the following passage of Mongolian to achieve the Mongolian-Chinese attention-based NMT.

A. Segmentation of Chinese

Chinese, as a standard synthetic language, has much semantic information only by one word. It is a pictograph, so it does not need spaces to separate the words in a sentence. However, before training the attention-based NMT, we need to separate the words from each other. Thus, we do the segmentation first.

segment into word 今天 天气 很 好。
It is a nice day, today.

segment into character 今 天 天 气 很 好。
It is a nice day, today.

FIGURE III. A SCHEMATIC DIAGRAM OF TWO TYPES OF SEGMENTATION

We try two types of segmentation. First we use the general segmentation, which separates the sentence through the word

one by one. But it has some drawbacks such as it easily introduces error through the ambiguity of one word in semantics [11][12]. It also creates a large vocabulary that is bad for NMT to reduce model complexity [13]. So we use the second type of segmentation that separates the sentence into characters. It can solve the problem of separation error, and extremely reduces the vocabulary size for the character of Chinese is less than 10K. We can see the differences between the two types of segmentation through FIGURE III.

B. Pre-processing of Mongolian

Different suffixes linked after one word root or stem have different grammatical meanings, so that Mongolian words present an extremely variety of transformation, since Mongolian is an agglutinative language. There are 30 Mongolian characters. Each character has at least three shapes according to the position it appears in a word [14]. The Mongolian words are constituted with the root or stem linking with different suffixes. Thus, it has different semantic meanings as well as grammatical information. There are two types suffixes named derivational suffix, which is also called word-building suffix, and inflection suffix [15]. One or more derivational suffixes linked to the root make the stem, and inflection suffix makes the sentence fluent. There is a special inflection suffix named case suffix, which is important to the machine translation. According to the character of Mongolian, we need to do some pre-processing to train the model of machine translation.

1) Stem segmentation

A Mongolian word shows a variety of transformations with different suffixes linked with it. So it is significant to segment the Mongolian stem and suffixes correctly. By this way, it can reveal the true attribute of a word and its grammatical relation. We put FIGURE IV for illustration to introduce the change. From the FIGURE IV we can see the original form of the noun word *oil* is ᠰᠡᠪᠰᠡ. When the word ᠰᠡᠪᠰᠡ linking with the suffix ᠬᠦ, it becomes verb form with the meaning rub oil. However, if we put the suffix ᠲᠠᠢ after the word ᠰᠡᠪᠰᠡ, it has anther meaning oily with the adjective form.

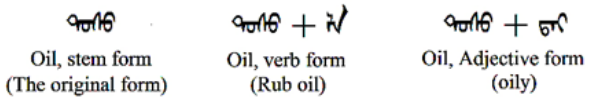

FIGURE IV. AN EXAMPLE OF OIL LINKING DIFFERENT SUFFIXES SHOWING DIFFERENT

There are three main approaches to separate the stem from the suffixes, called dictionary-based approach, rule-based approach and statistics-based approach. In 2011, Yu Ming et al. firstly combined the three methods to segment the Mongolian word [16]. In this paper, we follow the Yu Ming's work to separate the stem from suffixes combining the three main approaches. We can see the result of the stem segmentation according the FIGURE V.

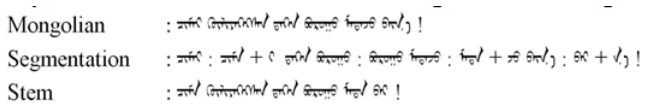

FIGURE V. THE RESULTS OF OUR METHOD THAT SEPARATE THE STEM FROM SUFFIXES THROUGH A MONGOLIAN SENTENCE.

2) Addressing the case suffix

There is a special suffix named the case suffix in Mongolian. As a special suffix, it does not have the actual semantic meanings but the grammatical meanings. The case suffix, as an additional element adding to the Mongolian word, is helpful to increase the fluency of Mongolian sentences. We can summarize the case suffix as the following four types that will be showed in the TABLE I with their own different coding.

TABLE I. THE CODING OF DIFFERENT CASE SUFFIX SYMBOL

Symbol of Case Suffix	Unicode	UTF-8
Vowel Connect Symbol	18 0E	E1 A0 8E
Zero Width Forbidden Symbol	20 0C	E2 80 8C
Zero Width Connect Symbol	20 0D	E2 80 8D
Narrow Width Continuous Space Symbol	20 2F	E2 80 AF

In our Mongolian corpus, there are totally 571075 Mongolian words. However, the four different case suffixes have appeared 50518 times. It is important to address the case suffix. There are two ways to address the case suffix. One is to cut the case suffix and then combine the stem with the additional component of the case suffix as a new word. This method could make the same stem become different words. So we choose the other method, which is to remove the case suffix and keep the stem only, to address the case suffix.

3) Converting Mongolian into Latin

There are many phenomena that the Mongolian words have the same font with different pronunciations. It is hard to avoid the mistaken spelling when people manually create and collate the corpus. To reduce these mistakes, we choose Latin as an intermediate language that can be easily converted. Another main reason we convert Mongolian into Latin is that using the Latin character to represent Mongolian costs less memory when training the models of NMT. We take advantage of the method that proposed by the computer institute of Inner Mongolia University to convert Mongolian into Latin character.

IV. EXPERIMENTS

We have run our experiments with the corpus of CWMT2009 translation task on Mongolian to Chinese. It contains about 65K sentences in both Mongolian and Chinese training datum. 1K sentences of the training data are chosen as a development set. We also use 1K new sentences to test the translation performance with BLEU. The baseline is the attention-based NMT, which is trained with the original Mongolian and the Chinese words.

A. Training Details

In our training corpus, there are 65752 sentences, whose length are limited to 50 words, in parallel with 571075

Mongolian words and 591521 Chinese words that contain 722099 Chinese characters in total without removing the duplicate words or characters. We choose both development corpus and test corpus with 1K sentences. TABLE II shows the details of corpus. It is worth mentioning that all the words or characters in the table we listed are the total number without removing the duplicates. We limit the vocabulary to the top 20K words in Mongolian and top 10K words in Chinese. Words that are not in the vocabulary are replaced by a universal symbol <UNK>. Our attention-based models contain bi-directional hidden units at encoder and general units at decoder. The hidden units of both encoder and decoder have 1000 cells, respectively. Other parameters of attention-based NMT that we choose are through the experiments. We run both the training and decoding on a single GPU device NVIDIA Tesla K80.

TABLE II. THE DETAILED INFORMATION OF THE CORPUS

Corpus	Mo-Ch sentence pairs	Ch Scale (size)	Mo Scale (size)	Total Mo words	Total Ch words	Total Ch characters
Train	65752	2.79MB	8.23MB	571075	591521	722099
Dev	1000	40KB	107KB	7371	8670	10484
Test	1000	40.3KB	109KB	7555	8792	10556

1) Get the best parameters

There are three main parameters used in the experiments, which are learning rate, word embedding dimension and shuffle, aimed to achieve a better translation performance. For each parameter, we have done several experiments to analyze which one can get the higher BLEU and cost less time.

From the FIGURE VI, we can see the influence of learning rate on BLEU. When we reduce the initial learning rate, the BLEU basically keeps increasing until the learning rate is 0.0001. From these experiments, we choose 0.0001 as the initial learning rate. If we still reduce the learning rate, it can not get the better translation but the more time is wasted. We train the attention-based NMT models combined the stochastic gradient descent(SGD) and Adadelta. Each mini-batch contains 32 sentence pairs and Adadelta is used to adapt the learning rate of parameters (ε=10^{-6} and ρ=0 .95).

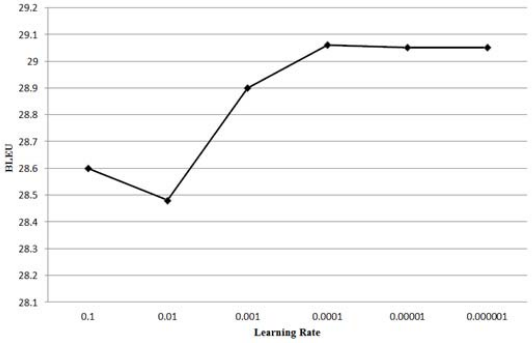

FIGURE VI. THE INFLUENCE OF LEARNING RATE ON BLEU.

FIGURE VII shows the effects of the word embedding dimension. We conduct 5 groups of experiments on the dimension of 100, 200, 500, 1000 and 2000. Through the experiments we can conclude that the higher word embedding dimension is, the better translation result can be obtained. But if we choose the dimension 2000, it costs double memory and time with only a little increase. Thus, we choose 1000 as word embedding dimension to achieve a relatively good performance.

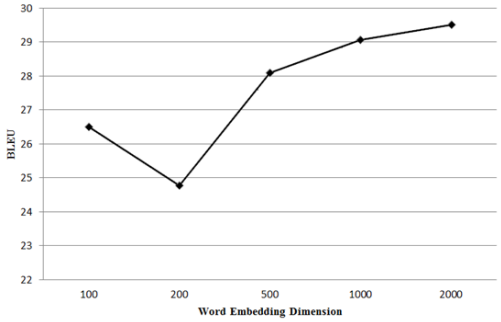

FIGURE VII. THE INFLUENCE OF WORD EMBEDDING DIMENSION ON BLEU.

We also start some experiments on the parameter of shuffle. From the FIGURE VIII we can deduce that the shuffle significantly improves the performance of the NMT translation. We do the shuffle after each epoch.

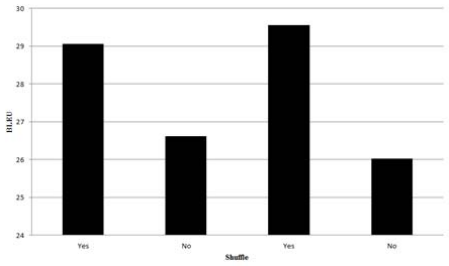

FIGURE VIII. THE INFLUENCE OF SHUFFLE ON BLEU.

B. Results

TABLE III lists the results on test set. In the table, the baseline is the attention-based NMT with the original Mongolian and Chinese words. *Stem* means we add with the stem segmentation we have mentioned to the baseline. We use *Case* to express the approach of addressing the case suffix in the table. And *Latin* represents that we convert Mongolian into Latin characters. *Word* and *Character* means the different methods on the Chinese word segmentation.

From the TABLE III, we can find the highest average BLEU score is 29.56 with the methods of Latin conversion and character segmentation on attention-based NMT, with the 1.82 points in BLEU score higher than the baseline. From each line of the table, we can deduce that the character segmentation of Chinese can improve the translation performance with the average 0.79 BLEU score increase. And the Latin conversion can increase the BLEU with average score of 0.6. But the approaches of stem segmentation and addressing the case suffix have a negative effect on the accuracy of translation. And in the next section we will discuss the influence of each method.

TABLE III. THE RESULTS OF TEST SET

System	Word	Character
Baseline	27.74	29.06
+Stem	26.66	27.11
+Case	26.00	26.72
+Stem+Case	26.57	26.58
+Latin	28.68	**29.56**
+Stem+Latin	25.93	27.35
+Case+Latin	26.59	27.04
+Stem+Case+Latin	27.53	28.58

V. ANALYSIS

It is easy to find that the methods of stem segmentation and addressing the case suffix cannot improve the translation results. There are two main reasons that cause this phenomenon.

- Mongolian belongs to the agglutinative language. One Mongolian root or stem with different suffixes has different semantic information. The approach of stem segmentation that keeps the root or stem only destroys the semantic information. As a result, it leads the word embedding get less effective information. It can also cause the problem of data sparseness which, likewise, decreases the translation performance of statistical machine translation (SMT).
- The case suffix, as a special suffix, does not have the actual semantic meanings but carries the grammatical meanings. The method of addressing the case suffix we have mentioned above is harmful to the grammatical information. When building the vector of words, it cannot take advantage of the grammatical information so that we get a negative result.

The strength of NMT lies in that the semantic and grammatical information can be learned by taking global context into consideration [17]. The methods of stem segmentation and addressing the case suffix destroy both semantic and structural information, which cause the decrease of BLEU score.

However, the method of Latin conversion has a nice effect. It can improve the translation results with 0.6 BLEU score increase, on average. The main reason of the improvement is that it can effectively reduce the error of original Mongolian spelling. It can not only reduce the length of the sentence but also the error rate of word alignment [18][19]. Moreover, converting Mongolian into Latin makes Mongolian easy to represent and be trained.

Likewise, segmenting Chinese into character also increases the BLEU score with 0.79 on average. Chinese, as the pictograph, makes the segmentation important to the machine translation. But the general segmentation segments Chinese into word which usually causes the segmentation faults. If we segment Chinese into character, it will successfully solve the problem. Also, segmenting Chinese into character is in favor of the word alignment and reducing the vocabulary size of NMT. Thus, the approach that segmenting Chinese into character is effective.

VI. CONCLUSION

In this paper, we build an attention-based NMT on a small scale corpus of CWMT2009 translation task on Mongolian to Chinese. We carry out a lot of experiments to evaluate our pre-processing on both Mongolian and Chinese and to choose the relatively better learning rate, word embedding dimension and the shuffle. We achieve the best translation with the pre-processing of Latin conversion on Mongolian and character segmentation on Chinese with the average BLEU score of 29.56, which is 1.82 points higher than the baseline. But the methods of the stem segmentation and addressing the case suffix do not have a good performance. Both of them cause the decrease of the BLEU score. At last, we analyze the reasons why each method of pre-processing is effective or invalid.

The case suffix, as a special suffix of Mongolian, is very important to the machine translation seeing that it carries the grammatical information. We can get a lot of effective information by analyzing the case suffix. But the method of addressing the case suffix now is useless to the NMT. In the future, we need to find a new way to address the case suffix to make it effective to the NMT. Moreover, we also need to consider how to do the post-processing to improve the translation performance.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of China (Grant No. 61362028).

REFERENCES

[1] Tomas Mikolov et al.: Efficient Estimation of Word Representations in Vector Space. In: Proceeding of Workshop at ICLR, 2013.

[2] Mikolov, Tomas et al.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT, 2013.

[3] Sutskever I., Vinyals O., and Le Q. V.: Sequence to sequence learning with neural net-works. In: NIPS, 2014.

[4] Cho K., Bahdanau D., Bengio Y. et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Em-pirical Methods in Natural Language Processing, EMNLP, 2014.

[5] N. Kalchbrennerand, P. Blunsom.: Recurrent continuous translation models. In: EMNLP, 2013.

[6] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba.: Addressing the rare word problem in neural machine translation. In: ACL, 2015.

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.: Learning phrase representations using RNN encoder-decoder for sta-tistical machine translation. In: EMNLP, 2014.

[8] D. Bahdanau, K. Cho, and Y. Bengio.: Neural machine translation by jointly learning to align and translate. In: ICLR, 2015.

[9] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio.: On using very large target vocabulary for neural machine translation. In: ACL, 2015.

[10] Minh-Thang Luong, Hieu Pham, Christopher D. Manning.: Effective Approaches to Atten-tion-based Neural Machine Translation. In: Proceedings of the 2015 Conference on Empiri-cal Methods in Natural Language Processing, EMNLP, 2015.

[11] Guohe Feng and Wei Zheng.: A summary of the research on Chinese automatic word seg-mentation in China. In: Library and information service, pp. 41-45, 2011.

[12] Xiao Chen, Guangjin Jin, Changning Huang.: Experimental research on word segmentation based on character. In: The research and application of Content Computing--Proceedings of the Ninth National Conference on Computational Linguistics, 2007.

[13] J. Chung, K. Cho, Y. Bengio.: A Character-Level Decoder without Explicit Segmentation for Neural Machine Translation. arXiv:1603.06147, 2016

[14] Qinggeertai.: Traditional Mongolian grammar. Inner Mongolian Press, Huhhot, China, 1992.

[15] Xiangdong Su, Guanglai Gao, Yupeng Jiang, Jing Wu and Feilong Bao.: Mongolian Inflec-tion Suffix Processing in NLP: A Case Study. In: the 4th Conference on Natural Language Processing and Chinese Computing, pp. 347-352 NLPCC, 2015.

[16] Y. Ming.: Researching of Mongolian Word Segmentation System Based On Dictionary, Rules and Language Model. In: Inner Mongolia University, 2011.

[17] Wei He, Zhongjun He, HuaWu, HaifengWang.: Improved Neural Machine Translation with SMT Features. In: the Thirtieth AAAI Conference on Artificial Intelligence, AAAI, 2016.

[18] Wang.siriguleng, Nasun-urtu, siqintu.: Description for Chinese-Mongolian Hybrid Machine Translation System of CWMT09. pp. 137-140. CWMT, 2009.

[19] Xiang Li, Linfeng Song, Fandong Meng et al.: The ICT Technique Report of All Language Tracks of CWMT 2013. CWMT, 2013.