Mongolian-Chinese Unsupervised Neural Machine Translation with Lexical Feature

Ziyu Wu[1], Hongxu Hou[*], Ziyue Guo[2], Xuejiao Wang[3] and Shuo Sun[4].

[1.2.3.4.*] Department of Computer Science, Inner Mongolia University, China
cshhx@imu.edu.cn

Abstract. Machine translation has achieved impressive performance with the advances in deep learning and rely on large scale parallel corpora. There have been a large number of attempts to extend these successes to low-resource language, yet requiring large parallel sentences. In this study, we build the Mongolian-Chinese neural machine translation model based on unsupervised methods. Cross-lingual word embedding training plays a crucial role in unsupervised machine translation which generative adversarial networks (GANs) training methods only perform well between two closely-related languages, yet the self-learning method can learn high-quality bilingual embedding mappings without any parallel corpora in low-source language. In this work, apply the self-learning method is better than using GANs to improve the BLEU score of 1.0. On this basis, we analyze the Mongolian word lexical features and use stem-affixes segmentation in Mongolian to replace the Bytes-Pair-Encoding (BPE) operation, so that the cross-lingual word embedding training is more accurate, and obtain higher quality bilingual words embedding to enhance translation performance. We reporting BLEU score of 15.2 on the CWMT2017 Mongolian-Chinese dataset, without using any parallel corpora during training.

Keywords: Mongolian-Chinese, neural machine translation, unsupervised method, Stem-Affix Segmentation.

1 Introduction

With the progress of deep learning (Sutskever I et al., [1]; Bahdanau D et al., [2]) and the availability of large parallel corpora, neural machine translation (NMT) have achieved excellent performance on some language pairs (Wu Y et al., [3]). However, these models can only perform well when they have large parallel corpora. Unfortunately, build parallel corpora is expensive because they require specialized expertise. In contrast, the monolingual corpora are easier to obtain than parallel corpora. Unsupervised NMT models that aim to train a model without using any labeled data have performed well in recent machine translation researches. Recent work has attempted to learn cross-lingual word embedding without parallel data by mapping monolingual embedding to the shared space using adversarial learning methods (Lample G et al., [4]; Artetxe M et al., [5]). However, these methods based on generative adversarial networks (GANs) is only applicable to bilingual dictionaries trained between two closely-related languages, but apply this method to build bilingual vocabulary between Mongolian and Chinese is poor. Previous works have shown that self-learning can learn high-quality bilingual embedding mappings without any parallel corpora in low-resource language. In this work, we use a fully unsupervised initialization based on self-learning methods to improve the performance of cross-lingual word embedding training. Different preprocessing methods in language corpora will also affect the training effect of cross-lingual word embedding. The BPE algorithm is mainstream methods, but the BPE based on the number of co-occurrences, this segmentation method can't consider the semantic features of Mongolian. It leads to the decline of cross-lingual word embedding training effects in low-resource language pairs. Therefore, we perform word stem-affixes segmentation operations on Mongolian in the corpora preprocessing stage.

In summary, this paper makes the following main works:

a. Constructing a Mongolian-Chinese NMT model based on unsupervised. The unsupervised method is used to first realize the translation of Mongolian-Chinese word-by-word, and then through the large-scale language model and back-translation to guide the optimization of model parameters until the model converge.
b. In the Mongolian language, we use the stem-affix segmentation instead of the BPE to preserve the semantic information of Mongolian as much as possible while ensuring the granularity of the segmentation.
c. Use a cross-lingual training method based on self-learning combined with stem-affix segmentation for the original unsupervised translation model to improve the accuracy of the bilingual dictionary.

2 Related work

Machine translation task is divided into supervised machine translation, unsupervised machine translation and semi-supervised machine translation, which is depending on whether supervision is performed.

2.1 Supervised Methods

There is a rich body of supervised methods for Mongolian-Chinese machine translation based on a large number of Mongolian-Chinese bilingual parallel corpora. Wu et al., [6] introduced the NMT model into the Mongolian Chinese machine translation task, and the machine translation model of cyclic neural network based on the attention mechanism is realized. Fan W et al., [7] proposes the method of using similar words instead of low-frequency words. Li J et al., [8] proposes a method of introducing a dictionary improves the translation effect of low-frequency words; Wang H et al., [9] proposes the method to control the segmentation granularity to improve the translation effect.

2.2 Unsupervised Methods

NMT for scarce resources has become a hotspot in recent years' research, and unsupervised method is one of them. The unsupervised method enables the NMT task to train well-behaved machine translation models without bilingual parallel corpora (Artetxe M et al., [10]). There are three key steps in an unsupervised machine translation system, including translation model initialization, language models, and back-translation. Facebook proposed an unsupervised machine translation method (Lample G et al., [4]), which achieved good translation results in English-French machine translation. Based on this model, this paper proposes an unsupervised neural machine translation model from Mongolian to Chinese.

2.3 Semi-Supervised Methods

Di He et al., [11] proposes a semi-supervised neural network model based on dual-learning, which can translate low-resource languages using some monolingual corpora and small parallel corpora. The result shows that semi-supervised neural machine translation can achieve reasonable results with parallel corpora which are insufficient to train a common neural model.

3 Unsupervised Mongolian-Chinese Neural Machine Translation

The Mongolian-Chinese machine translation task is limited by the lack of bilingual parallel corpora. The translation model often cannot be fully trained, which leads to the translation performance not being improved. Recent researches have shown that unsupervised methods enable machine translation tasks to train well-performing machine translation models without bilingual parallel corpora. In the unsupervised machine translation system, the three key steps are translation model initialization, training of language models and back translation. The overall architecture of the system is shown (see Fig. 1).

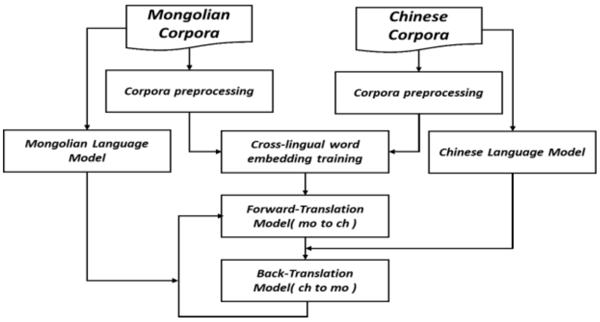

Fig. 1. The overall architecture of the unsupervised Mongolian−Chinese neural machine translation system.

3.1 Word Stem-Affix Segmentation

In recent years, the BPE Segmentation technology has been used to corpora preprocessing in machine translation models and obtain good performance. However, this method only relies on word frequency merging and does not take into account the semantic characteristics of any language itself. This makes the effect of Mongolian BPE operation less than that of stem-affix segmentation. In this paper, we use the discriminant stem-affix segmentation based on a directed graph morphology analyzer. The method uses the idea of discriminant classification to model the stem affixation of words into the labeling problem of the letters in the word. This method owns generalization ability and can deal with the problem that the word contains unregistered stems.

In the Mongolian sentence $S = W_1 W_2 \cdots W_r$ ，$W_i (1 \leq i \leq r)$, for the Mongolian word $W_i = C_{i\_1} C_{i\_2} \cdots C_{i\_n}$. $C_{i\_j} (1 \leq j \leq n)$ is the jth char of W_i . n represents the length of word. The problem of stem-affix segmentation becomes the division of the alphabetic sequence: $C_{i\_1} C_{i\_2} \cdots C_{i\_n} \rightarrow C_{i\_1:e_1} C_{i\_e_1+1:e_2} \cdots C_{i\_e_{m-1}+1:e_m}$ ， $e_m = i\_n$ .The letter sequence $C_{i\_1:n}$ is divided into m subsequences, the first subsequence is the stem, and others are affixes. Jiang W et al., [13] based on the Chinese word segmentation, divided each Mongolian letter C_{i_j} into four categories: b is the letter of the beginning of the stem or the affix; m in the middle of the stem or affixes is the end of the stem or the affix; s indicates that the word is a single stem or affix. The marked corpora are trained by using the maximum entropy toolkit to obtain the final segmentation result. As shown (see Fig. 2).

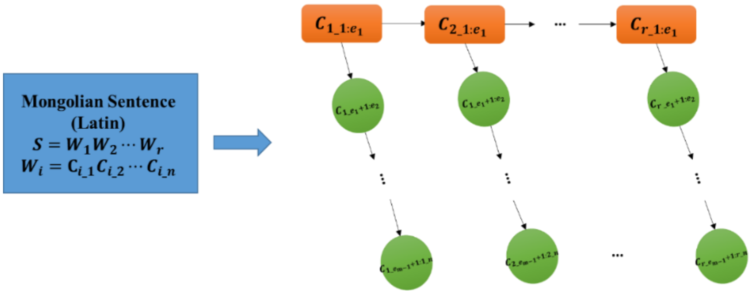

Fig. 2. The process of word stem-affix segmentation.

Orange squares indicate stem, and the number of the stem is one. A green circle indicates an affix, and the affix can be 0 or more. In this paper, the experiment of the main system uses the stem-affix segmentation for the Mongolian corpora. For Mongolian, the number of affixes is limited, and the data of the training corpora can easily cover all affixes. The situation of stems is much more complicated, and new words will continue to emerge with social development. When the stem of the word to be analyzed does not exist in the training material, the simple enumeration method cannot find the correct candidate for the analysis result. However, the discriminative stem-affix strategy may have a good generalization ability, just like the situation in Chinese word segmentation.

The segmentation results based on the BPE algorithm and the stem-affix segmentation method are shown in Table 1.

Table 1. Mongolian Sentences with Different Segmentation Methods

Method	Example
Source	[illegible] ··
BPE(35000)	[illegible] ··
Stem-Affix	[illegible] | [illegible] | [illegible] ··

Table 1 illustrated that the granularity of sentences after BPE segmentation is similar to that after Stem-Affix segmentation, such as "[illegible]" is divided into "[illegible]" and "[illegible]" in both methods. While the stem-affix segmentation method contains more semantic information, like "[illegible]", according to the result of BPE segmentation, the word attribute is changed into a verb, and the meaning of the stem-affix segmentation is consistent with the meaning of the BPE, and the part of speech still no change. So, this method more helpful to our model training.

3.2 Cross-Lingual word embedding

In the early training of the unsupervised machine translation, we need to construct a mapping relationship between Mongolian and Chinese, this mapping is called cross-lingual word embedding. Facebook proposed in their unsupervised machine translation system to use GANs (Gouws et al., [14]) training cross-lingual word embedding. However, their evaluation has focused on closed-related languages, while in cross-lingual learning from Mongolian to Chinese, they are often failing.

There are many ways to calculate the distance between the source language word embedding and the target language word embedding, including maximum mean difference, cosine similarity, and Cross-domain Similarity Local Scaling (CSLS) method (Lample G et al., [11]). In this work, we adopt the CSLS as a criterion for training word embedding pairs in cross-lingual word embedding training. It’s used to represent the average cosine similar measure that word embedding from source X to target Y. This part is an important part of the unsupervised Mongolian-Chinese machine translation model. The quality of cross-language word vector training will directly affect the quality of the Mongolian Chinese bilingual dictionary. The training methods can be based on GANs (Yang Z et al., [18]; Carone et al., [19]), self-learning method (Artetxe M et al., [20]) and so on. However, experiments show that the method based on GANs is not suitable for translation tasks between two languages with low similarity, but the self-learning method is more suitable for tasks like Mongolian Chinese machine translation, so we adopt the self-learning method.

The process of cross-lingual word embedding training by self-learning is shown (see Fig. 3).

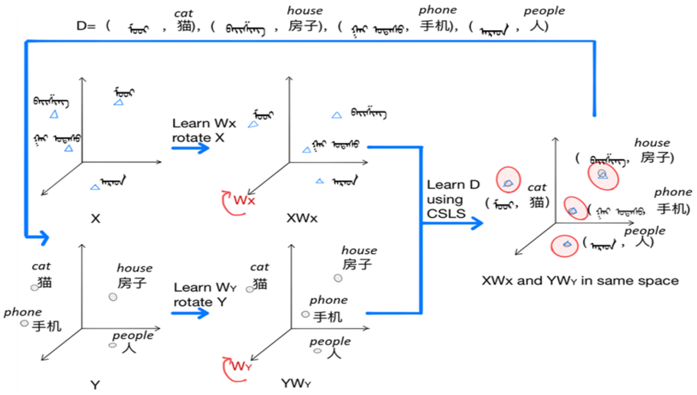

Fig. 3. A process sketch of bilingual dictionary generation by cross-lingual word embedding. The bilingual dictionary is constructed by learning the mapping matrix between $X$and Y.

Let X and Y be the word embedding matrices in Mongolian and Chinese, their ith row X_{i*} and Y_{i*}denote the embedding of the ith word in their respective vocabularies. Our goal is to learn the linear transformation matrices W_X so the mapped embedding XW_X and YW_Y are in the same cross-lingual space. We build a dictionary between Mongolian and Chinese, encoded as a matrix D where $D_{ij} = 1$ if the jth word in Chinese is a translation of the ith word in Mongolian. It is divided into three parts: word embedding normalization, dictionary initialization and self-learning, and symmetric reweighting.

Word embedding normalization. The implementation of this part requires two steps: the first normalize according to the length of word embedding, then average the center of each dimension and normalize again according to the length. The advantage of this operation is that we can guarantee that the final embedding has a unit length. In other words, for any two word embedding, their dot product is their cosine similar distances.

Dictionary Initialization. The difficulty of initializing a bilingual dictionary in this paper is that the Mongolian word embedding X and the Chinese word embedding Y are not aligned (no matter which dimension is not aligned). Therefore, we construct two aligned the Mongolian word embedding matrices X_1 and the Chinese word embedding matrices Y_1 as the initial dictionary. There are many methods for initializing a dictionary, including random dictionary induction, word frequency-based lexical cut-off, nearest neighbor search, and Cross-Lingual Similarity Local Scaling (CSLS). We adopt the CSLS method for dictionary initialization.

Given two map embedding matrices X_1 and Y_1, respectively calculate$r_T(x)$ and $r_S(y)$, $r_T(x)$is expressed as the average cosine similarity of the k nearest neighbors of the Mongolian word embedding x in the Chinese word embedding matrix Y_1, $r_S(y)$ is expressed as the average cosine similarity of the k nearest neighbors of the Chinese word embedding y in the Mongolian word embedding matrix X_1. The calculation method of CSLS is shown in formula (1)

$$CSLS(x,y) = 2\cos(x,y) - r_T(x) - r_S(y) \tag{1}$$

The process uses a self-learning method. After calculating the initial dictionary, X_1 and Y_1 are discarded, and the remaining self-learning iterations are performed on the original X and Y.

Self-learning Iterative Improvement. Corresponding rotation matrixW_X and W_Y are obtained by singular value decomposition. As shown in Equation (2)-(4):

$$\mathrm{U}SV^T = X^T DY \tag{2}$$

$$W_X = \mathrm{U} \tag{3}$$

$$W_Y = VS \tag{4}$$

The Mongolian-Chinese machine translation model based on neural under unsupervised method mainly includes the following four parts: the Mongolian-Chinese bilingual dictionary training; the Mongolian language model and the Chinese language model training; the translation model initialization from Mongolian to Chinese; back-translation. Next, we will introduce in detail.

Mongolian-Chinese bilingual dictionary. In the Mongolian-Chinese NMT with the supervised method, the bilingual dictionary consists of word pairs in the parallel corpora, but in the unsupervised method, the Mongolian corpora and the Chinese corpora are not aligned, so it's impossible to find the one-to-one correspondence of Mongolian and Chinese through the traditional method of supervising machine translation. So before building the bilingual dictionary, we use the fasttext to train the word embedding in Mongolian and Chinese monolingual corpora, then build a Mongolian-Chinese bilingual dictionary by aligning monolingual word embedding spaces in an unsupervised way, which is CSLS method.

3.3 Language model

In this work, language modeling is accomplished via de-noising auto-encoding (Lample G et al., [4]), it's loss function as formula (5), our goal is minimizing L^{lm} :

$$L^{lm} = E_{x\sim S}\left[-\log P_{s\to s}\left(x \middle| N(x)\right)\right] + E_{y\sim T}\left[-\log P_{t\to t}\left(y \middle| N(y)\right)\right] \tag{5}$$

where $N(\cdot)$ is a noise function with some words dropped in Lample G et al., [4]. $P_{s\to s}$ and $P_{t\to t}$ are combinations of encoder and decoder operating on the Mongolian side and the Chinese side, respectively.

3.4 Translation model initialization

According to the already trained Mongolian-Chinese bilingual dictionary and two language models, through the word-by-word method to initialize the translation model, get the initial translation results from Mongolian to Chinese. The Chinese language model is used to adjust the sequence of translation results. At the same time get the first translation model.

3.5 Back translation

To train the new system in a real translation environment without violating the limitations of using only monolingual corpora, we introduce the back translation method proposed by Sennrich R et al. [16]. Specifically, this method is an input sentence for a given language, and the system uses greedy decoding to translate it into another language in an inferred mode (using a shared encoder and a decoder of another language). Using this method, we can get pseudo-parallel corpora and then train the system to predict the original text based on the translation.

The translation results in those previous Mongolian-Chinese translation model is translated into the Chinese-Mongolian translation model (still through word-by-word). The Mongolian language model is used to correct the translated results and the source Mongolian after back translation. Repeat the previous two translation processes until the model converges. The loss of the back translation model is shown as formula (6)

$$L^{back} = E_{y\sim T}\left[-\log P_{s\to t}\left(y|u^*(y)\right)\right] + E_{x\sim S}\left[-\log P_{t\to s}\left(x|v^*(x)\right)\right] \quad (6)$$

$u^*(y) = argmaxP_{t\to s}(u|y)$ is the back translation result from Chinese to Mongolian, $v^*(x) = argmaxP_{s\to t}(v|x)$ is the back translation result from Mongolian to Chinese, and $(u^*(y), \mathrm{y})$, $(\mathrm{x}, v^*(x))$ are pseudo-parallel sentences.

In the process of translation, the final objective function is the weighting of the loss of language models and the loss of the back translation. As shown in formula (7):

$$L = \alpha L^{lm} + （1 - \alpha）L^{back} \quad (7)$$

4 Experiment

4.1 Dataset

Monolingual Data. All the methods being evaluated in all tasks (except for supervised translation systems) take monolingual word embedding in each language as the input data. Use CWMT2017 Mongolian-Chinese parallel corpora for 0.26M as a training set. Randomly disrupted the corpora sentences to ensure the model run in unsupervised. We use BPE to segment Mongolian corpora according to the number of word combinations. In both baseline systems we choose the BPE method. In the main system performs the technology of stem-affix segmentation in Mongolian. Remove the noise sentences and keep sentences from 1 to 100 in length. Our unsupervised Mongolian-Chinese machine translation tasks based on the transformer. Experiments in NVIDIA TITAN X.

Bilingual Data. Use the 1001 sentence pair test set of the CWMT2017 Mongolian-Chinese daily language translation as the test set for our experiments. The corresponding dataset statistics are summarized in Table2.

Table 2. Corresponding dataset statistics

Method	language	vocab.size
WBW	mo/ch	69413/5288
Unsupervised	mo/ch	31738/20754
Ours	mo/ch	41909/5288

Election of various parameters. The number of BPE codes is 35000 in Mongolian. The number of encoder layers is 4 and the number decoder layers is 4. The number of share encoder and decoder layers both are 3. The dimensionality of the word embedding is 100. The hidden units are 100, dropout is 0.1, blank is 0.2, the learning rate is 0.0001, the batch size is 32, the epoch size is 500000. We take α is 0.5 to train the model in turn. At the decoding time, we generate greedily.

4.2 Baselines

We used two baseline systems as a comparison of the experiments.

Word-by-word translation (WBW) (Lample G et al., [11]): The first baseline system is that it performs word-by-word translation using an initialized Mongolian-Chinese bilingual dictionary. Simultaneously, this model is also our initial translation model.

Unsupervised training (Artetxe M et al., [5]): Unsupervised Mongolian Chinese neural machine translation model, in which the corpora preprocessing part uses the BPE segmentation technique to segments. The parameters of this model are also the same as the parameters of our experiment. The training time is one week.

4.3 Experimental results and analysis

Through several comparative experiments, we made the following analysis. As shown in Table 3-Table 5.

Table 3. Comparison of three unsupervised methods

Models	BLEU
WBW(Word-by-word)	5.4
Unsupervised(BPE & GAN)	13.5
Ours_model1(BPE & Self-learning)	14.5
Ours_model2(Stem-Affix Segmentation & GAN)	14.3
Ours_model3(Stem-Affix Segmentation & Self-learning)	15.2

Table 3 shows the BLEU scores in different unsupervised Mongolian-Chinese neural machine translation models. Compare to WBW (baseline system 1), the BLEU score increase 9.8. Compare to Unsupervised (baseline system 2), the BLEU score increase 1.7. We analyze the reason for this situation is the affixation of the Mongolian corpora can preserve as much as possible reducing the size of the dictionary, which can further effectively reduce the out of the vocabulary problem and the unknown word problem. According to the second and the third line, we verify the effectiveness of cross-lingual word embedding training based on self-learning. To compare the second and fourth line, we verify the advantage of the stem-affix segmentation method.

Table 4. Translation results (Short Sentences) for different models

	Example
Source	[illegible] ··
WBW	但是以为说着。
Unsupervised	但是的话蒙古。
Ours_model3	但听过蒙古的。
Reference	但是听说过很多关于蒙古地区的传说。
English	But I have heard a lot about the legends of Mongolia.

Table 5. Translation results (Long Sentences) for different models

	Example
Source	[illegible] 457 ·· 3 [illegible] ··
WBW	乌力吉、呼锦德政权、残疾人发展475，托拉、康复、补助、西塔格玛、工作措施。
Unsupervised	呼和残疾人事业发展资助资金4个。下拨3万元, 残疾人主要用于多取、康复、托养、西塔高加、工作补贴措施，为残疾人正镜，推动残疾人发展事业。
Ours_model3	呼和浩特共有475.3万元残疾人事业资金。主要用于残疾人技术、康复、托养。为残疾人正镜，推动残疾人发展事业。
Reference	近日，呼和浩特市财政下达残疾人事业发展补助资金457.3万元，主要用于落实残疾人在技能培训、康复救助、托养服务补贴、燃油补贴、机构补贴、工作补贴等方面的保障措施，提升服务机构对残疾人的服务水平，促进残疾人事业的全面发展。
English	Recently, Hohhot issued a subsidy of 4.573 million yuan for the development of the disabled, mainly for the implementation of safeguards for skills training, rehabilitation assistance, care support subsidies, fuel subsidies, institutional subsidies, work subsidies, etc. The level of service provided by the institution to the disabled and the overall development of the cause of the disabled.

Table 4 shows the translation result of the three models in short sentences (sentence length less than 20), where WBW uses simple word-to-word translation to obtain a correct translation result. The second baseline system received the correct translation of two words ("But" and "Mongolia"), but the location of "Mongolia" in the target language was incorrect. Through analysis we find is due to the BPE segmentation causes the Mongolian words part of speech changed, so it's position in the decoding process change. The results of the third line confirm our analysis. Our model does not fully translate the correct results, but the result is best in these unsupervised models.

Table 5 explains the translation effects of the three models in long sentences (sentence length of more than 50). The results are similar to those obtained in Table 4. In WBW, two target words ("disabled" and "rehabilitation") were translated; five words were translated on the Unsupervised model(baseline system 2), but there are still cases where the translation results do not match the target end position; better translation performance is still achieved in our model than baseline systems. Regarding the phenomenon of lack of translation, we analyze the reason that the training corpus is still small, resulting in insufficient training of the model.

5 Conclusion

In this work, we build the Mongolian−Chinese NMT model based on the unsupervised method, which is greatly alleviated the problem of the sparse corpus. At the same time, we solve the solution that the previous cross-domain word embedding training performed poorly in low-resource language by self-learning and stem-affix segmentation. Laid a good foundation for the study of translation models between Mongolian-Chinese machine translation and another low-resource language. In future researches, we will consider using higher dimension word embedding size, deeper networks or some new unsupervised methods to improve the quality of translation.

References

1. Sutskever I, Vinyals O, Le Q. V.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems (NIPS). pp. 3104-3112.(2014)
2. Bahdanau D, Cho K, Bengio Y, et al.: Neural Machine Translation by Jointly Learning to Align and Translate. In: International Conference on Learning Representations (ICLR). arXiv preprint arXiv :1409(0473).(2015)
3. Wu Y, Schuster M, Chen Z, et al.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv: Computation and Language.(2016)
4. Lample G, Conneau A, Denoyer L, et al.: Unsupervised Machine Translation Using Monolingual Corpora Only. In: International Conference on Learning Representations (ICLR). arXiv preprint arXiv :1711(00043).(2018)
5. Artetxe M, Labaka G, Agirre E, et al.: Unsupervised Neural Machine Translation. In: International Conference on Learning Representations (ICLR). arXiv preprint arXiv:1710(11041).(2018)

6. Wu J, Hou H, Shen Z, et al.: Adapting Attention-Based Neural Network to Low-Resource Mongolian-Chinese Machine Translation. In: International conference on the computer processing of oriental languages (ICCPOL). pp. 470-480.(2016)
7. Fan W, Hou H, Wang H, et al.: Machine Translation Model of Mongolian-Chinese Neural Network Fusing Priori Information. In: Chinese Journal of Information Science, 32 (06). pp. 36-43.(2018)
8. Jinting L, Hongxu H, Jing W, et al.: Combining Discrete Lexicon Probabilities with NMT for Low-Resource Mongolian-Chinese Translation. In: Parallel and distributed computing: applications and technologies (PDCAT). pp. 104-111.(2017)
9. Wang H.: Multi-granularity Mongolian Chinese Neural Network Machine Translation Research. In: Inner Mongolia University. pp. 15-35.(2018)
10. Artetxe M, Labaka G, Agirre E, et al.: Learning bilingual word embeddings with (almost) no bilingual data. In: Meeting of the association for computational linguistics (MACL). pp. 451-462.(2017)
11. Lample G, Conneau A, Ranzato M, et al.: Word translation without parallel data. In: International conference on learning representations (ICLR). arXiv preprint arXiv : 1710(04087).(2018).
12. He D, Xia Y, Qin T, et al.: Dual Learning for Machine Translation. In: Neural information processing systems (NICS). pp. 820-828.(2016)
13. Jiang W, Wu J, Wu R, et al.: Discriminant stem affixation of Mongolian directed graph morphology analyzer. In: Journal of Chinese Information Processing, 25 (04). pp. 30-34.(2011)
14. Gouws, Stephan, Bengio Y, et al.: BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In: International conference on machine learning (ICML). pp.748-756.(2015)
15. Johnson M, Schuster M, Le Q V, et al.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. In: Transactions of the Association for Computational Linguistics (TACL) 5(1). pp. 339-351.(2017)
16. Sennrich R, Haddow B, Birch A, et al.: Improving Neural Machine Translation Models with Monolingual Data. In: Meeting of the association for computational linguistics (MACL). pp. 86-96.(2016)
17. Lample G, Ott M, Conneau A, et al.: Phrase-Based & Neural Unsupervised Machine Translation. In: Empirical methods in natural language processing (EMNLP). pp. 5039-5049.(2018)
18. Yang Z, Chen W, Wang F, et al.: Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets. In: North American chapter of the association for computational linguistics (NAACL). pp. 1346-1355.(2018)
19. Barone, Antonio.: Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In: Meeting of the association for computational linguistics (ACL). pp. 121-126.(2016)
20. Artetxe M, Labaka G, Agirre E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 789–798.(2018)