Research on the key technologies of corpus preprocessing in Mongolian-Chinese SMT

Li Jin-ting[1,a], Hou Hong-xu[1,a], Wu Jing[1,a], Wang Hong-bin[1,a], Fan Wen-ting[1,a]

Department of computer, Inner Mongolia University, Hohhot, 010021
[a] E-mail: cshhx@imu.edu.cn

Keywords: Mongolian-Chinese SMT; corpus preprocessing; Mongolian morphological analysis; Chinese word segmentation.

Abstract. The traditional preprocessing method in morphology analysis uses Mongolian suffix segmentation and stemming. But there exists many cases in Mongolian. If the case is not processed, the corpus will suffer from data sparse problem and lead to poor translation performance. Therefore, we summarize and research the existing corpus preprocessing method, and focus on the effect of case processing, in order to improving the performance of Mongolian-Chinese SMT by analyzing Mongolian morphological. Experiments show improvements of about 3.22 relative in the BLEU score of SMT over baseline system 1 by optimizing the preprocessing method.

Introduction

Corpus preprocessing is one of the key technologies of Mongolian-Chinese SMT. Mongolian is an adhesive language, whose formation and configuration is connected by root and stem of different endings to complete [1].

Mongolian word segmentation methods are mainly based on dictionary, rules and statistics. A language model based on statistical methods includes the Mongolian word segmentation method is based on the Skip-N language model [2] and hierarchical Mongolian statistical language model [3]. The method of stemming mainly includes automatic segmentation of root, stem and suffix of Mongolian [1], the Mongolian word segmentation based on statistical language model [4], the Mongolian word segmentation based on conditional random fields [5], Mongolian word segmentation based on dictionary, rules and statistics [6]. At present, we use the Mongolian word segmentation based on dictionary, rules and statistics. For many cases existed in Mongolian, which are separated by space in Moses. Therefore the case forms a single word, which causes data sparse problem. We mainly study the key technology of case processing in the Mongolian to alleviate data sparse problem of corpus effectively, which improves the performance of SMT.

At present, Chinese corpus preprocessing uses the ICTCLAS Chinese word segmentation system [7]. But word-based segmentation exists the following problems: the recognition performance of word-based segmentation is poor [8]; the method of word-based segmentation is easy to be ambiguous and interferes with SMT [9]; word-based segmentation can also result in word segmentation errors, which causes alignment errors in word alignment [10]; coarser-grained word alignment suffers from more serious data sparse problem than finer-grained word alignment in the small scale parallel corpus [11]. We consider that fine-grained character segmentation should be adopted in the Chinese corpus preprocessing.

© 2016. The authors - Published by Atlantis Press

The phenomenon of different shapes with the same pronunciation exists in Mongolian, which is hard to avoid spelling errors in manual collection and collation of corpus. We use Latinization to reduce the error. Besides, it can be easy storage and processing Mongolian for computer [12].

Stemming

Mongolian is an adhesive language. Its formation and configuration is connected by root and stem of different endings. The composition of words and the representation of grammatical meaning depends on different suffixes, therefore the correct segmentation of root, stem and suffix can reveal the relationship between grammatical attribute and lexical category [1]. Figure 1 describes the Chinese "oil" by connecting different suffixes to express different semantic in Mongolian.

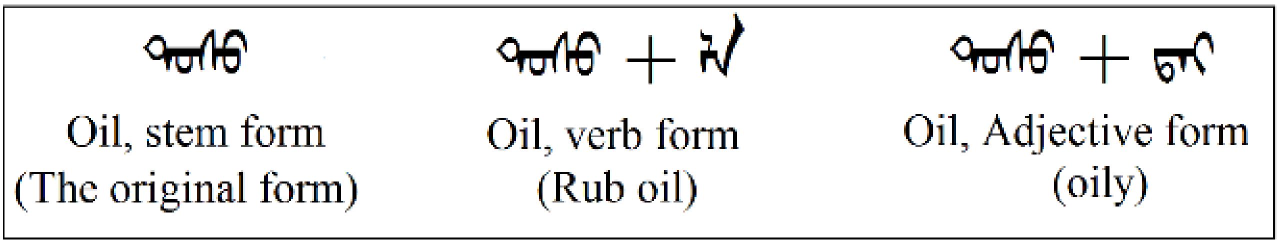


Fig. 1. Mongolian word stem and suffix

At present, we use the Mongolian word segmentation based on dictionary, rules and statistics. Figure 2 shows the extraction processing of Mongolian sentence stem.

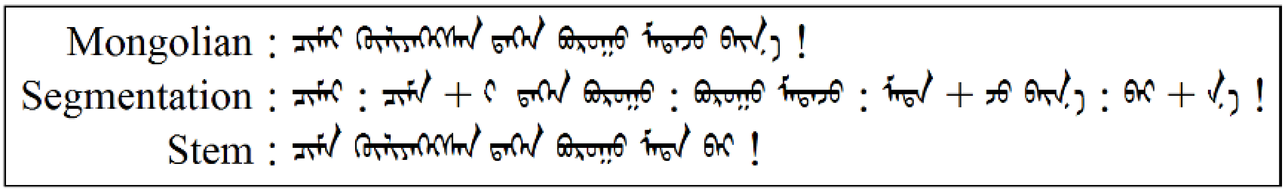


Fig. 2. Stemming

For many cases existed in Mongolian, which are separated by space in Moses. Therefore the case forms a single word, which causes data sparse problem. We carry out frequency analysis in the corpus. Latinization can correct the spelling mistakes of corpus, Therefore in this paper, the Mongolian corpus frequency analysis method is based on Latinization.

Table 1. Corpus frequency distribution

	Word	Stem
Total Tokens	42781	34335
Unique Tokens	23734	20818
Secondary Tokens	6216	4812
thrice Tokens	2840	1988
Repeatedly Tokens	9991	6715
Percentage of unique tokens (%)	**55.5**	**66.6**

Table 1 shows the token distribution of Mongolian words and Mongolian stems on corpus. We can see that the unique tokens in stem-based corpus increase almost 11.1% than those in word based corpus. Data shows that due to the case processing approach is not perfect, data sparse problem existed in corpus. Although stemming can make the alignment effect in process of Mongolian-Chinese SMT promotion, the corpus scale of Mongolian-Chinese SMT is small, so stemming brought serious data sparse problem. It was hypothesized that the stemming in the Mongolian-Chinese SMT brought more serious data sparse problem and affect translation performance.

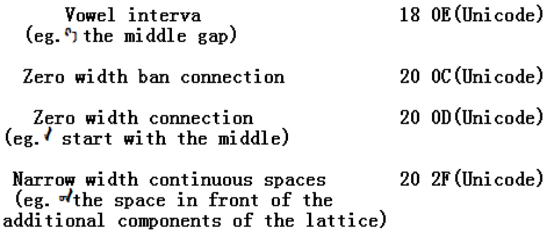
Case processing

Case processing in corpus is the focus of this paper. We research the case in the stem corpus. Case is also another kind of suffix, which has no semantic meaning in expressing, and only has the grammatical meaning. But in the process of SMT, cases are recognized as spaces, which may lead to a Mongolian word separated, and become two words or more. This situation will cause data sparse problem, and affect the performance of word alignment as well as translation. Table 2 shows 571075 Mongolian words in our Mongolian corpus. Four kinds of cases appear for 50518 times, which show that cases have great influence on the experiment.

Table 2. The frequency statistics of cases

Classification of cases	Frequency
Vowel interval	592
Zero width ban connection	5
Zero width connection	0
Narrow width continuous spaces	49921
Total	50518

Figure 3 shows the four kinds of case, and we identified their codes.



Vowel interva (eg. the middle gap)	18 0E(Unicode)
Zero width ban connection	20 0C(Unicode)
Zero width connection (eg. start with the middle)	20 0D(Unicode)
Narrow width continuous spaces (eg. the space in front of the additional components of the lattice)	20 2F(Unicode)

Fig. 3. The code of cases

We use two kinds of methods for case processing: the first method is removing four kinds of case, and combining stem with the additional components form a new word; the second method is removing the case and its additional components. Table 3 shows the frequency distribution of cases: in terms of the total number of words, because the first method takes the additional components add to the stem and makes the same stem become different words, the total number of words have increased. The second method is to remove the case directly. Stemming is more refined, and more same stem appears, so the total number of words have decreased. In the frequency of unique tokens the corpus, the first method stemming and removing the case, and the unique tokens reduce 2.62% than those in after stemming the Mongolian corpus. The stemming corpus uses the second method to remove the case, the unique tokens reduce 8.8% than those in after stemming the Mongolian corpus. Data prove that two methods both can alleviate data sparse problem. Obviously, the second method is more efficient. We selects to remove the case and its additional components in the paper.

Table 3. Two kinds of case processing corpus word frequency distribution

	Stem	The first method	The second method
Total Tokens	34335	37065	27517
1	20818	23717	15906
2	4812	4882	3917
3	1988	1984	1662
4+	6715	6481	6031
Percentage of unique tokens (%)	66.6	63.98 (-**2.62**)	57.8 (**-8.8**)

Figure 4 shows the Mongolian corpus of stemming; Figure 5 shows the Mongolian corpus of stemming and processing cases. Case processing can alleviate data sparse problem.

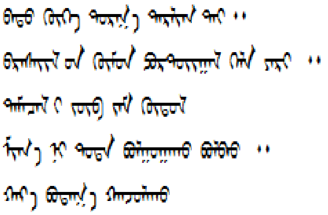


Fig. 4. Stemming the Mongolian corpus

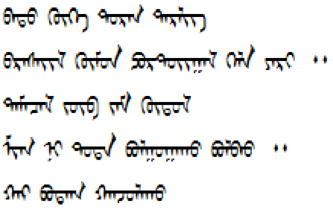


Fig. 5. Stemming and case processing

Table 4 shows that in the frequency of unique tokens, the corpus removes the case and its additional components. We see that the unique tokens reduce 2% than those in the original Mongolian corpus. Data shows that case processing can alleviate data sparse problem. We assume that case processing can improve the performance of SMT.

Table 4. The frequency distribution of corpus

	Mongolian corpus	Case Processing corpus
Total Tokens	42781	30502
1	23734	16301
2	6216	4368
3	2840	2089
4+	9991	7742
Percentage of unique tokens (%)	55.5	53.4(-**2**)

Experiments

Introduction of the experimental environment and data

We implemented Moses as our basic SMT system and built it as follows: alignment was performed by GIZA++ [13]. A phrase-based MT decoder similar to the work of Koehn et al [14] was used with the decoding weights optimized by MERT [15]. We used a 3-gram language model. Chinese language corpus was processed by the ICTCLAS Chinese word segmentation system [8].

CWMT’2009 was used for the experiments. As a small language, public Mongolian and Chinese parallel corpus is limited. We select within 50 of the length of the 65752 sentences of the bilingual corpus as training set. In order to adjust the appropriate training parameters, the Mongolian and

Chinese SMT's development set was established by selecting 1000 sentences from the training corpus, and being removed from the training set. For testing the final translation results, we built a test set of 1000 sentences. The test set and train set can't be repeated but that should belong to the same field. Table 5 shows the data set in detail. Mo is the abbreviation of Mongolian and Ch is the abbreviation of Chinese.

Table 5. Data set

	Train	Dev	Test
Bilingual sentence pairs	64752	1000	1000
Ch Scale	2.72MB	40KB	40.3KB
Mo Scale	7.32MB	107KB	109KB
Total Mo words/stems	571075	7371	7555
Total Ch words	591521	8670	8792
Total Ch characters	722099	10484	10556

SMT automatic evaluation standard is the necessary for SMT discriminant training, and it is also an important indicator to quickly measure the results of SMT [16]. This experiment uses the BLEU evaluation standard [17].

Comparison of experimental results and analysis

We have combined group experiment to SMT by using different preprocessing methods of Mongolian-Chinese SMT bilingual corpus. Two groups of baseline were selected: the baseline 1 uses Mongolian original corpus and Chinese corpus with word segmentation. The baseline 2 uses Mongolian original corpus and Chinese corpus with character segmentation.

Table 6. The experimental results

Experiment system			BLEU
	Mo	Ch	
baseline 1	word	word	29.48
baseline 2	word	character	30.66
system 1	stem	word	29.08 (**-0.40)**
system 2	stem	character	29.49 (**-1.17**)
system 3	case processing	word	30.26 (+**0.78)**
system 4	case processing	character	31.04 (+**0.38)**
system 5	stem + case processing	word	29.50
system 6	stem + case processing	character	30.73
system 7	case processing + Latinization	word	31.98 (+**2.50)**
system 8	case processing + Latinization	character	32.70 (+**1.64)**

As the experimental results are shown in Table 6, we can see that the BLEU of baseline 2 is higher than baseline 1, which exemplifies that the method of character segmentation can effectively alleviate the problem of word segmentation, and can improve the translation performance.

System 1 uses the stem of Mongolian corpus and Chinese corpus with word segmentation, and system 2 uses the stem of Mongolian corpus and Chinese corpus with character segmentation. The BLEU score of system 1 reduces 0.4, compared with baseline 1. The BLEU score of system 2 reduces 1.17, compared with baseline 2. The results exemplify that stemming cause data sparse problem. We reserve suffixes to the Mongolian corpus and count the frequency of words, in order to verify that stemming can cause data sparse problem. Table 7 shows the token distribution of Mongolian stems in corpus of retaining the suffix. The frequency of unique tokens in Mongolian corpus of retaining a suffix reduces 9.44% than the stem of Mongolian corpus. The frequency of unique tokens in Mongolian corpus of retaining two suffixes reduces 10.22% than the stem of Mongolian corpus. The results exemplify that stemming caused data sparse problem.

Table 7. The token distribution of Mongolian stems in corpus of retaining the suffix

	stem	an suffix	two suffixes
Total Tokens	34335	37248	38170
1	20818	21291	21521
2	4812	5264	5475
3	1988	2326	2459
4+	6715	8366	8714
Percentage of unique tokens (%)	66.6	57.16 (**-9.44**)	56.38 (**-10.22**)

While system 3 uses case processing of Mongolian corpus and Chinese corpus with word segmentation, system 4 uses case processing of Mongolian corpus and Chinese corpus with character segmentation. System 3 outperforms baseline 1, and system 2 outperforms baseline 2. While system 5 uses the stemming and case processing of Mongolian corpus as well as Chinese corpus with word segmentation, system 6 uses the stemming and case processing of Mongolian corpus as well as Chinese corpus with character segmentation. System 5 outperforms system 3, and system 6 outperforms system 4. The results of two different groups of corpus all prove the previous hypothesis: case processing can alleviate the data sparse problem effectively, as well as improve the performance of Mongolian-Chinese SMT.

In order to demonstrate the influence of Mongolian Latinization, system 7 uses case processing and Latinization of Mongolian corpus as well as Chinese corpus with word segmentation, and system 8 uses case processing and Latinization of Mongolian corpus as well as Chinese corpus with character segmentation. System 7 outperforms baseline 1, and system 8 outperforms baseline 2. The results exemplify that Latinization could avoid spelling errors in manual collection and collation of corpus, as well as improve the performance of Mongolian-Chinese SMT.

Summary

This paper summarizes the main methods of the corpus preprocessing of Mongolian-Chinese SMT, and puts forward the method of case processing. Group experiment is combined with corpus obtained from different processing methods. We analyze the influence of different corpus preprocessing methods on SMT and the result exemplifies that case processing in this paper can significantly improve the performance of the Mongolian-Chinese SMT. Corpus preprocessing method is proposed, and the BLEU score increase 3.22, compared with the baseline 1. Besides, we also analyze and summarize a series of problems, and prove the effectiveness of our methods.

We will continue to research on the key technologies of Mongolian Chinese bilingual corpus preprocessing. Especially, in the aspect of stemming, we will focus on how to reduce the data sparse problem. As for processing the case on stemming corpus, we will strive to get better Mongolian-Chinese SMT corpus preprocessing method to improve the performance of SMT.

Acknowledgments

We thank ICITIT 2016 reviewers. This work is supported by the National Natural Science Foundation of China (No. 61362028) and the Postgraduate Scientific Research Innovation Foundation of Inner Mongolia (No. 11200-12110201).

References

[1] Nanwushuritu: Automatic segmentation system of Mongolian root, stem and suffix. In: Journal of Inner Mongolia University, pp.02:53-57 (1993)

[2] H.X. Hou, Q. Liu, Z. Wang, G. Zhang: Skip-n Mongolian statistical language model. In: Journal of Inner Mongolia University, pp. 02:220-224 (2008)

[3] H.X. Hou, G. Zhang, Z. Wang: Hierarchical Mongolian statistical language model. In: Journal of Inner Mongolia University, pp. 03:336-340 (2009)

[4] H.X. Hou, Q. Liu, Nasanurtu: Mongolian word segmentation based on statistical language model. In: Pattern Recognition and Artificial Intelligence, 22(1): 108-112 (2009)

[5] W. Zhao, H.X. Hou, W. Cong, M. Song: Based on conditional random field of Mongolian word segmentation. In: Journal of Chinese information, pp. 05:31-35+84 (2010)

[6] Y. Ming, H.X. Hou: Researching of Mongolian word segmentation system based on dictionary, rules and language model. In: Inner Mongolian University

[7] P. Zhang, K. Yasuda, and Eiichiro Sumita.2008. Improved statistical machine translation by multiple Chinese word segmentation. In Proceedings of the Third Workshop on Statistical Machine Translation, pp. 216-223

[8] C. Huang, H. Zhao: A review of Chinese word segmentation in ten years. In: Chinese Journal of information and information, pp. 03:8-19 (2007)

[9] X. Chen, G. Jin, C. Huang: Experimental research on word segmentation based on character. In: The Ninth National Conference on Computational Linguistics (2007)

[10] G. Feng, W. Zheng: A summary of the research on Chinese automatic word segmentation in China. In: Library and information service, pp. 02:41-45 (2011)

[11] J. Wu, H.X. Hou, C.J. Xie: Realignment from Finer-grained Alignment to Coarser-grained Alignment to Enhance Mongolian-Chinese SMT (2015)

[12] X. Shen: Minority languages Latinization of the significance and method. In: Chinese Information Institute of national language committee information (2007)

[13] F.J. Och, H. Ney: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)

[14] P. Koehn, H. Hoang, A. Birch, et al.: Moses: Open source toolkit for statistical machine translation. ACL (2007)

[15] F.J. Och: Minimum error rate training in statistical machine translation. In: Proceedings of ACL, pp. 160–167 (2003)

[16] N. Yang: Machine Translation research based on neural network learning. In: University of Science and Technology of China (2014)

[17] P. Koehn, S. Roukos, T. Ward, et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 311-318 (2002)