Introduction of Traditional Mongolian-Chinese Machine Translation

J. Wu, H.X. Hou, Monghjaya, F.L. Bao, C.J. Xie
College of Computer Science, Inner Mongolia University,
Hohhot, China

***Abstract*—With the rapid development of Mongolian information technology and the closer communication within different nationalities, machine translation between Mongolian language and Chinese becomes more and more important. Traditional Mongolian-Chinese machine translation has achieved fruitful results in recent years but also confronts with many challenges. This paper reviews the recent development of Traditional Mongolian-Chinese machine translation, including the translation model, corpus collection and proofreading, traditional Mongolian code conversion, lemmatization, named entity recognition, parsing and system combination, etc. We also introduce our group's research and achievement in Traditional Mongolian-Chinese machine translation.**

Keywords-machine translation; Traditional Mongolian; codeconversion; word lemmatization; recognition of name entity; parsing

I. INTRODUCTION

Machine Translation (MT) is to translate by computer. The methods of MT include Rule-Based Machine Translation (RBMT), Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT) [1], and Template-Based Machine Translation (TBMT), which derivatives out from EBMT has better generalization ability [2]. Phrase-based statistical machine translation (PBSMT) is currently the mainstream method for machine translation. Recently, extending PBSMT with syntactic knowledge, mainly hierarchical phrase-based machine translation (HPBMT)model [3]and tree-to-string alignment template(TAT) model[4], has achieved very good results and become a new research hotspot. Besides, system combination of different machine translation systems [5] has been widely concerned and achieved good performance in the recent years NIST[6] and CWMT[7]machine translation evaluation.

The research of Traditional Mongolian-Chinese MT begins in the late 80s of the 20th century and gains more and more attention. Mongolian language has been used by millions of people in different countries and regions in the world. There are three kinds of writing Mongolian, which are traditional Mongolian, todo Mongolian and Cyrillic Mongolian. Traditional Mongolian is the official language of Inner Mongolia Autonomous Region and also the treasure of Chinese culture. Above all, the translation between Traditional Mongolian-Chinese is important and necessary. The Mongolian in this paper especially refers to the traditional Mongolian.

Research of Mongolian-Chinese MT starts late, but develops fast. There are several mature translation systems in both Mongolian-Chinese and Chinese-Mongolian translation. Besides, some important research outputs, such as corpus collection and proofreading, Mongolian code conversion and Mongolian word lemmatization, etc comes out as well. However, comparing with the international and domestic advanced MT technology, research of Mongolian-Chinese MT is lagging behind and some areas are even in the blank.

II. REPROCESSING OF TRADITIONAL MONGOLIAN–CHINESE MT

Preprocessing plays a crucial role in Machine Translation. A properpreprocessing can avoid some kinds of mistakes in MT, and enhance the translation accuracy. Screening and preprocessing of parallel corpora, code conversion and code correction, spelling correction, Mongolian word lemmatization and named entity recognition are all important parts in preprocessing of Traditional Mongolian-Chinese MT.

A. Code Conversionand Spelling Correction of Mongolian

The standard of Mongolian coding rules follows the national standard GB13000 and the international standardized coding standards150/IEC10646.However, currently there are many different code systems like Menkeli Mongolian coding, Oyuta Mongolian coding and Saiyin Mongolian coding systems are widely used, which makes information resources of the different coding systems cannot be shared and Mongolian corpus extremely difficult to obtain.

In addition, there are many mistakes in Mongolian international standard code corpus. For example, the 4, 5 vowels and 6, 7 vowels are easy to be mistaken when input. So although and looks like the same word "4", the computer process them as different words because the vowel input errors.

Researchers put forward code conversion and spelling correction to salve the above two problems. Gong Zheng raises a conversion method between some un international coding and Mongolian international standardized coding[8]. S·Loglo develops a kind of Mongolian code converting algorithm based on the grapheme [9]. These two methods are both based on dictionary and rules and cannot deal with homograph and unknown word. Hua Shaba implements a rule-based spelling check system, which can determine four kinds of errors which are root errors, affix errors, dative errors and vowels errors, but do not correct the errors[10]. Su Chuanjie et al. propose an automatic spelling correction method for Mongolian based on SMT framework and enhances the percentage of correct words from 48.21% to 97.55%[11].

In our research, we generally use spelling correct, fuzzy matching, Mongolian word lemmatization and manual proofreading, etc to minimize the impact on the translation caused by coding problems. However, Mongolian and Chinese language parallel corpus still need manual proofreading after auto-preprocessing.

B. Traditional Mongolian Word Lemmatization

Mongolian language belongs to mount type language. Etyma and morphological affix both play important roles in Mongolian word and contain lots of part of speech and grammar information. The word lemmatization of Mongolian is complicated but significant to translation. We use the translation results of Moses[12] system to verify the impact of Mongolian word lemmatization. The size of training corpus is 60,000. We set for 1000 to develop and 1000 to test. The BLEU[13] results after processing are as table 1 show:

TABLE I. EXPERIMENT RESULTS OF BLEU.

	Development Set	Test Set
Word level	0.1864	0.1692
Stemming level	0.1943	0.1785

Besides, when we need to query the Mongolian dictionary, Mongolian word lemmatization can improve the matching rate significantly. We set 1000 Mongolian sentences which contains 731 words as test to query the dictionary and get the matching rate as Table 2 shows:

TABLE II. MATCHING RATE OF TRADITIONAL MONGOLIAN DICTIONARY.

	Matching Rate
No Lemmatization	28.61%
After Lemmatization	42.92%

There are many achievements in the research of Mongolian word lemmatization. Nasanurtu et al. proposes a Mongolian word lemmatization system based on dictionary and rules[14]. In order to solve the problem that dictionary and rules based Mongolian word lemmatization method cannot cover all lemmatization cases and deal with ambiguity, HouHongxu et al. raises a Mongolian word lemmatization based on statistical language model[15]; Zhao Wei proposes a conditional random fields based Mongolian word lemmatization method[16]; MingYu combines three different methods: dictionary based, rule based and statistical language model based methods to build a Mongolian word lemmatization system[17].

C. Named Entity Recognition

The task of named entity recognition is to identify three major categories: the entity class, time class and digital class [18]. Currently, research work and mature tools of named entity recognition in Mongolian are both very few.

In our research, we recognize the named entity of time class and digital class, which are easy to process based on dictionary, rules and templates. The three processing methods are as follows. First method is to build a dictionary which contains time and digital words, which most of are not in the Mongolian and Chinese dictionary. Second method is to write rules to process time and digital phrases. Thirdly method is to build a sentence template collection to collect time and digital sentence template.

III. MONGOLIAN-CHINESE TRANSLATION MODEL

Mongolian-Chinese Machine Translation experienced a rule-based machine translation stage. Current methods are mainly phrase-based SMT methods. Hierarchical phrase-based MT model becomes the new trend.

A. Phrase-Based Mongolian-Chinese Translation SMT

Development of SMT model has experienced word-based model, phrase-based model and syntax based model[19]. Currently phrase-based model is most mature of Mongolian translation. Especially after Mongolian-Chinese machine translation joining in CWMT, more machine translation institutes start to focus and research on Mongolian-Chinese translation.

In the CWMT'2013[20], there are 6 institutes attend the Mongolian-Chinese task and the best BLEU score is 0.1964, which is higher than 0.1851 in CWMT'2011[7]. Most system use hierarchical phrase-based SMT and some systems combine different SMT system together. However, the BLEU score of Mongolian-Chinese task is the lowest in five translation tasks which Chinese is target language as figure 1 shows. What we can get from the figure is firstly, the translation result of the best Mongolian-Chinese MT system is not that good; secondly, as the BLEU scores of English-Chinese news task and Mongolian-Chinese task in figure 1 use the same hierarchical phrase-based SMT model named Chiero [21], we can infer that in addition to the MT model, methods and preprocessing which are suited for Mongolian language characteristics are also important to Mongolian-Chinese MT system.

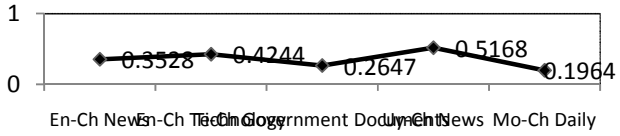

FIGURE I. CWMT'2013 EVALUATION RESULTS IN FIVE TASKS.

B. Traditional Mongolian Parsing

Phrase-based SMT achieved great success, but it is also flaws. On the one hand, phrase-based SMT depends on the training corpus and its generalization ability is poor, for those phrases which are not in training corpus, the model is helpless. Especially, Mongolian Chinese parallel corpus scale is very small, which greatly limits the effects of phrase-based Mongolia-Chinese SMT system. On the other hand, performance of phrase-based SMT model is poor in the long distance reordering. Phrase-based SMT can carry out simple reordering within the phrase, but can do nothing for reordering between the phrase and long distance reordering [22]. We find out that there are many word order errors in the Mongolian-Chinese phrase-based SMT translation result which reduces the translation accuracy and readability.

As a solution for above problems of phrase-based SMT, hierarchical phrase-based SMT model and tree-to-string alignment template model, which are both syntax-based SMT model, achieve good results in international and domestic Machine Translation Evaluation. Syntax-based SMT, reordering and template-based MT model all need parsing.

English and Chinese syntactic parsing is very mature, but there is almost no work in Mongolian parsing in addition to S·Loglotried to design and implement a Mongolian automatic syntactic Parsing system based on dependency grammar in 2002[23].

On the current condition, our general solution to get syntactic information of Mongolian in parallel corpus is to get the Chinese parsing result using Chinese parser, and aligned to Mongolian by dictionary or Giza++. Dictionary-based alignment has higher accuracy rate, but lower recall rate due to limited dictionary size. While the small scale of Mongolian and Chinese parallel corpus also limits the accuracy of Giza++.

C. System Combination

More and more machine translation researchers put their focus on system combination and try to produce a new translation that has better quality than all of them. According to the different levels of combining the outputs come from different machine translation systems, approaches to system combination can be classified into three types: sentence-level combination, phrase-level combination, and word-level combination[24]. There are many researches on system combination and the performance of multi-engine machine translation system is generally able to get better translation results, but at present, scarcely any work has been down on system combination of Mongolian and Chinese.

We build a multi-engine machine translation system to combine examples-based, template based and hierarchical phrase-based MT system in sentence-level. The system firstly try to get translation by examples-based method, and output the translation result if match, if not, try to match in template-based system, if still mot match, output the translation result using hierarchical phrase-based MT, if match in sentence template base, compare the translation result produced by template-based system and hierarchical phrase-based translation system on the bases of N-gram language modelthen output the best translation.

IV. CONCLUSION AND PROSPECT

Mongolian-Chinese MT has achieved great strides in recent years. Firstly, advanced MT methods have been applied to Mongolian-Chinese MT and many translation systems based on different models have been built. Besides, research on corpus reprocessing, Mongolian code conversion and word lemmatization get remarkable achievements. Furthermore, researchers make useful exploration in named entity recognition, parsing and system combination, etc.

However, due to the reasons that short research time of Traditional Mongolian-Chinese MT, lack of researchers and resources, and the complexity and particularity of Mongolian language grammar and writing mode, there is a large gap between Mongolian-Chinese MT and advanced MT research. This defect is reflected in two aspects: one is in the basic research, the other one is in the advanced research. In the basic research: Mongolian and Chinese parallel corpus is in a small scale and not good quality; basic but necessary research like Mongolian named entity recognition and parsing is almost in blank. In the advanced research: research on Mongolian and Chinese translation technology does not keep pace with the advanced level; A number of international and domestic advanced translation models and methods have not been applied to Mongolian and Chinese MT yet, or applied but not achieved very good results.

Therefore, we must speed up the basic research of Mongolian-Chinese translation, such as access to quality bilingual corpus, named entity recognition and parsing tools. What is more, research for the techniques and methods which are suited for Mongolian language characteristics; furthermore, keep up with the pace of international and domestic advanced research of MT and apply those sophisticated techniques to Mongolian and Chinese MT.

REFERENCES

[1] Fei Xia, Michael McCord, Improving a statistical MT system with automatically learned rewrite patterns.*Proceedings of the International Conference on Computational Linguistics*, Geneva, pp. 132-139, 2004.

[2] Kaji, Hiroyuki, Kida Telal, Learning translation templates from bilingual text. Proceedings of the 15th International Conference on Computation Linguistics, Nantes, pp. 672–678, 1992.

[3] Chiang, Hierarchical phrase–based translation. International Conference on Computational Linguistics, 33(2), pp. 201–228, 2007.

[4] Liu, Q. Liu, S. Lin, Tree–To–String Alignment Template for Statistical Machine Translation. International Conference on Computational Linguistics, Sydney, pp. 609–616, 2006.

[5] Rosti, S. Matsoukas, R. Schwartz, Improved word–level system combination for machine translation. Association for Computational Linguistics, Prague, pp. 312–319, 2007.

[6] NIST Open Machine Translation 2012 Evaluation Plan (OpenMT12), http://www.nist.gov/itl/iad/mig.

[7] Zhao Hongmei, LvYajuan, ben Guosheng, Summary on CWMT2011 MT translation evaluation.Journal of Chinese Information Processing, 26(1), pp. 22–30, 2012.

[8] Gong Zheng, Hao Li, Yang Xuhua, A method for conversion from nonstandard Mongolian character codes to the ISO IEC Mongolian codes. Journal of Inner Mongolia University, 39(2), pp. 216-219, 2008

[9] S.Loglo, Research on the General Algorithm of Mongolian Code Conversion. Journal of Inner Mongolia University, 42(2), pp. 133-136, 2009.

[10] Hua ShaBao, Modern Mongolian automatic calibration system. Journal of Inner Mongolia University, 04(2), pp. 49-53, 1997.

[11] Su Chuanjie, HouHongxu, Yang Ping, A spelling correction method for traditional Mongolian based on statistical translation framework. Journal of Chinese Information Processing, 27(6), pp. 175-179, 2013.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 2007.

[13] K.Papineni, S.Roukos, T.Ward, et al , BLEU: a method for automatic evaluation of machine translation evaluation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002.

[14] Nasanurtu. Mongolian automatic word lemmatization subsystem of root, stem and affix. Journal of Inner Mongolia University, 29(2), pp. 53–57,1997.

[15] HouHongXu, Liu Qun, Nasanurtu, et al, Mongolian word segmentation. based on statistical language model. Pattem Recognition and Aitificial Intelligence, 22(1), pp. 108-112, 2009.

[16] ZhaoWei,HouHongxu, Cong Wei et al, Research on conditional random fieldsbased Mongolian word segmentation. Journal of Chinese Information Processing, 24(5), 31-35+84. 2010.

[17] Ming Yu, HouHongxu, Researching of Mongolian word segmentation system based ondictionary, rules and Language model.2011.

[18] ZHAO Jun, National laboratory of pattern recognition. Journal of Chinese Information Processing, 23(2), pp. 3-17, 2009.

[19] Brown, P.F., J.Cocke. A statistical approach to machine translation. Workshop on Speech Hand Natural Language,Paris, pp. 146–151, 2012.

[20] Hongmei Zhao, Jun Xie, YajuanLü, et al. CWMT 2013 machine translation evaluation official report. 2013. http://nlp.ict.ac.cn/index_zh.php.

[21] Macherey, F. J. Och. . An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. Proceedings of EMNLP–CoNLL2007, Prague,2007: 986–995.

[22] LIU Qun. Recent developments in machine translation research, Contemporary Linguistics ,11(2), pp. 21-26,2009.

[23] SLoglo, Hua Shabao, Sarula. Mongolian dependency parsing model based on statistical methods. Journal of Chinese Information Processing, 23(4), pp. 27-32, 2012.

[24] Li Maoxi, ZongChengqing. A survey of system combination for machine translation. Journal of Chinese Information Processing, 24(4), pp. 74-84+118,2010.