Research Advances in Minority Language Machine Translation at the Institute of Computing Technology, Chinese Academy of Sciences (Postprint)

Authors: Lü Yajuan, Liu Qun, Jiang Wenbin

Date: 2017-03-09T00:00:00+00:00

Abstract

This paper analyzes the background, current research status, and development trends of machine translation research for minority languages, and introduces the research advances of the Institute of Computing Technology, Chinese Academy of Sciences in minority language processing and machine translation, including fundamental language processing technologies for Uyghur, Mongolian, and Tibetan, analysis and translation modeling for morphologically rich languages, knowledge acquisition and translation techniques for low-resource languages, as well as the organization of evaluation tasks for minority language machine translation at the China Workshop on Machine Translation.

Full Text

Research Progress on Minority Language Machine Translation at the Institute of Computing Technology, Chinese Academy of Sciences

Authors: Lü Yajuan, Liu Qun, Jiang Wenbin
Institute of Computing Technology, Chinese Academy of Sciences

Abstract

This paper analyzes the background, current research status, and development trends of minority language machine translation research in China. It introduces the research progress on minority language processing and machine translation at the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS), covering fundamental language processing technologies for Uyghur, Mongolian, and Tibetan; analysis and translation modeling for morphologically

rich languages; knowledge acquisition and translation techniques for resource-scarce languages; and the organization of minority language machine translation evaluation tasks at the China Workshop on Machine Translation (CWMT).

Keywords: Minority Languages; Machine Translation; Morphologically Rich Languages; Machine Translation Evaluation

1. Introduction

China is a unified multi-ethnic nation comprising 56 ethnic groups, yet significant language barriers persist among them. The minority population totals 106 million, accounting for 8.41% of the national population. Among China' s 55 minority groups (excluding the Han), 53 have their own languages spoken by over 60 million people, while 22 ethnic groups use 28 distinct writing systems employed by approximately 30 million people. Despite increasing inter-ethnic exchanges driven by economic and social development, the language divide in minority regions remains severe. According to Ministry of Education surveys, 70% of farmers and herders in Xinjiang and Tibet cannot speak Mandarin Chinese, while 70-80% of the population in Guizhou and Yunnan cannot communicate in Mandarin. Additionally, over 30% of county and township officials in Xinjiang' s minority areas cannot speak Mandarin, and 50-80% of Chinese language teachers in southern Xinjiang have Mandarin proficiency below Level 3A. Conversely, the vast majority of Han Chinese do not understand minority languages.

These language barriers not only hinder external communication and economic development in minority regions but also seriously affect ethnic unity and social stability. On one hand, communication obstacles prevent minority areas from accessing timely knowledge and information in industry, agriculture, trade, and technology, while also impeding the promotion of their unique ethnic cultures, thereby fundamentally constraining regional development. On the other hand, language barriers create difficulties in implementing national policies, laws, and regulations in minority regions, making minority populations vulnerable to manipulation by separatist forces. Since the beginning of the new century, external "Tibetan independence" and "Xinjiang independence" forces have consistently targeted minority regions, intensifying their separatist propaganda and posing serious threats to national security and border stability. Therefore, bridging the language gap between minority languages and Chinese is crucial for promoting harmonious and rapid development in minority regions and safeguarding national unity.

Machine translation, which automatically translates one language into another using computers, represents one of the most powerful tools for overcoming these language barriers. In recent years, machine translation technology—particularly statistical machine translation—has made tremendous progress, with practical applications already widely deployed for some language pairs. However, research on translation between Chinese minority languages and Chinese has progressed slowly, with no practical systems yet available. Compared to the relatively

mature English-Chinese machine translation, Chinese minority language translation faces numerous challenges:

- **Large Language Type Span:** China' s minority languages exhibit rich typological diversity. From a genealogical perspective, while Chinese and Tibetan both belong to the Sino-Tibetan family, they belong to different branches: Chinese to the Sinitic branch and Tibetan to the Tibeto-Burman branch. Uyghur and Kazakh belong to the Turkic branch of the Altaic family, while Mongolian belongs to the Mongolic branch of Altaic. From a morphological perspective, Uyghur, Mongolian, Kazakh, and Korean are highly inflectional agglutinative languages, whereas Chinese, Tibetan, Yi, Zhuang, and Miao are isolating languages with virtually no morphological changes. These substantial differences in linguistic properties mean that simply applying conventional research approaches yields poor results for such diverse language pairs.
- **Scarce Language Resources:** Current mainstream statistical machine translation methods require large-scale language resources. Insufficient parallel corpora significantly impact translation quality. Due to relatively backward economic and cultural development in minority regions, available language resources (dictionaries and bilingual parallel corpora) are far scarcer than for Chinese. Under these conditions, pure statistical methods struggle to achieve satisfactory results, necessitating the integration of multiple translation strategies to maximize utilization of various linguistic knowledge and resources.
- **Weak Basic Language Processing Technologies:** Compared to Chinese, some minority language processing technologies remain immature. Fundamental issues such as encoding conversion, word segmentation, stemming, part-of-speech tagging, and named entity recognition have not been adequately resolved, while deeper problems like syntactic parsing are still in their infancy and far from practical application in machine translation, requiring further intensive research.

These challenges render existing mature machine translation methods unsuitable for translation between Chinese minority languages and Chinese. In fact, automatic translation between Chinese minority languages and Chinese involves numerous complex scientific problems, such as machine translation for morphologically rich languages and resource-scarce languages, which constitute important research directions in current statistical machine translation.

2. Research Status and Development Trends of Minority Language Machine Translation

Since the 1990s, the rapid development of statistical machine translation technology has fundamentally transformed machine translation research and applications. Statistical machine translation builds probabilistic models for the translation process, estimates model parameters through statistical analysis of

large-scale parallel texts, and then performs translation using these parameters. Beginning with IBM' s first statistical machine translation model in 1993, the field has evolved through several major stages: word-based models, phrase-based models, and syntax-based models. In recent years, incorporating semantic analysis techniques has also yielded progress. This technological advancement has driven machine translation applications, with domestic companies like Baidu and NetEase Youdao launching online translation services and products following Google and Microsoft. Machine translation has become ubiquitous in daily life. Statistical machine translation has become the mainstream technology in both academia and industry, as it overcomes the difficulties of manual knowledge engineering in traditional rule-based systems and can be easily adapted to new domains and languages.

Relative to the rapid international development of statistical machine translation, domestic research on minority language machine translation has progressed slowly, focusing primarily on a few languages such as Uyghur, Mongolian, and Tibetan.

For Uyghur, research on Chinese-Uyghur computer-assisted translation began in the mid-1990s. In 1995, Wang Shijie and colleagues at Xinjiang University' s National Natural Science Foundation project "Basic Research on Xinjiang Minority Language-Chinese Machine Translation Systems" conducted preliminary experiments. In 2004, Halimurat and colleagues' Xinjiang Autonomous Region Science and Technology Department project "Computer Chinese-Uyghur Assisted Translation Software" built a prototype system. In 2006, another NSFC project at Xinjiang University, "Research on Bilingual Aligned Corpus and Phrase Bank Construction Technology for Chinese-Uyghur Machine Translation," prepared resources and technologies for bidirectional translation. In recent years, Xinjiang University and Xinjiang Normal University have undertaken large-scale Chinese-Uyghur bilingual corpus construction and preliminary statistical translation research.

For Mongolian, machine translation has explored various methods. Domestic scholars have conducted rule-based and example-based research on Chinese-Mongolian translation with some success. Recently, statistical Chinese-Mongolian translation has been explored, along with preliminary studies on English-Mongolian, Japanese-Mongolian, and Mongolian-Chinese translation. Currently, more research focuses on Mongolian as the target language than as the source language.

For Tibetan, research on Chinese-Tibetan machine translation began in the 1990s with Professor Li Yanfu at Qinghai Normal University, who completed two National "863" projects: "Chinese-Tibetan Scientific and Technical Machine Translation System" and "Chinese-Tibetan Official Document Machine Translation Technology," implementing prototype systems. Qinghai Normal University also conducted research on practical Chinese-Tibetan systems. Since 2003, theoretical research and technical preparations have included verb processing, syntactic analysis, named entity recognition, and Tibetan-Chinese parallel cor-

pus construction. Northwest Minzu University, the China Tibetology Research Center, and the Institute of Ethnology at the Chinese Academy of Social Sciences have also made progress in Tibetan corpus construction and information processing research, laying foundations for further translation technology development.

In recent years, translation research between minority languages and Chinese has gained increasing attention. With support from the NSFC and Ministry of Science and Technology, the Hefei Institute of Intelligent Machines at CAS has researched statistical machine translation models for morphologically rich languages, achieving good results in Chinese-Mongolian translation. Beijing Institute of Technology has conducted multi-strategy minority language-Chinese machine translation research based on ontologies, while Inner Mongolia Normal University has studied Chinese-Mongolian statistical machine translation incorporating linguistic knowledge.

The 7th China Workshop on Machine Translation (CWMT 2011) evaluation, held in 2011, introduced the first minority language-to-Chinese translation evaluation tasks, including five languages: Uyghur, Mongolian, Tibetan, Kazakh, and Kyrgyz. Ten research institutions and universities participated. Notably, all 24 systems submitted by the 10 participating units, including those from minority universities, employed statistical translation technology, demonstrating its widespread adoption in minority language machine translation research. However, direct application of existing statistical techniques faces many problems. First, the large typological differences between minority languages and Chinese mean that a single model cannot address all language pairs. Mainstream statistical models treat all languages equally, which fails to adequately capture differences between morphologically divergent languages (e.g., agglutinative languages like Uyghur and Mongolian versus Chinese), resulting in poor translation quality. Second, the severe resource scarcity means pure statistical methods cannot achieve good results. In fact, rule-based and statistical methods each have advantages and disadvantages. Rule-based methods can more effectively utilize expert knowledge and achieve high-precision translation for regular phenomena such as temporal and numeral expressions. Under resource-scarce conditions, multiple translation strategies should be integrated. Additionally, basic minority language processing technologies remain weak, lacking high-performance morphological analysis and named entity recognition tools, while syntactic parsing research is still immature. These foundational technologies significantly impact translation model selection and training.

Based on this analysis, we believe that minority language machine translation research should, while drawing on existing statistical machine translation methods and experience, pay greater attention to language-specific characteristics. On one hand, basic minority language processing technologies must be strengthened to better support machine translation applications and system development. On the other hand, translation models and methods suitable for minority languages and Chinese must be studied, such as methods for morphologically

rich languages and resource-scarce languages. In-depth research on these issues is significant for solving machine translation problems for many low-resource languages and for advancing machine translation research more broadly.

3. Research Progress on Minority Language Processing and Machine Translation at ICT-CAS

The Natural Language Processing Group at ICT-CAS has focused on machine translation research for over 20 years, developing rule-based, example-based, and statistical machine translation systems. In the past decade, the group has made significant progress in statistical machine translation research and applications, proposing a series of source-language syntax-based statistical translation models. The group has published over 50 papers in leading international journals (*Computational Linguistics*) and conferences (ACL, EMNLP, COLING), applied for 18 technology patents, and received widespread attention from domestic and international peers. The group' s machine translation systems have achieved excellent results in prestigious international evaluations such as NIST and IWSLT, and the technology has been applied to patent translation, mobile translation, and news translation.

In recent years, our group has conducted research on machine translation for minority languages and neighboring countries' languages. For minority languages, we focus on Uyghur, Mongolian, and Tibetan—the most widely spoken minority languages in China. We have established close collaborations with Xinjiang University, Inner Mongolia University, and Qinghai Normal University. Through several years of effort, we have made substantial progress in processing these languages and developing machine translation systems between them and Chinese. We have collected and processed large-scale parallel corpora and translation dictionaries for Uyghur-Chinese, Mongolian-Chinese, and Tibetan-Chinese, developed a series of practical minority language processing tools (language identification and encoding conversion, Uyghur morphological analysis, Mongolian morphological analysis, Tibetan sentence segmentation/word segmentation, named entity recognition and translation), researched translation models for morphologically rich languages and methods for resource-scarce languages, and built statistical machine translation systems for Uyghur-Chinese, Mongolian-Chinese, and Tibetan-Chinese. Our minority language translation systems have been deployed in relevant national departments and received positive user feedback. Our group has also organized the minority language machine translation evaluation at CWMT, contributing to the development of domestic minority language machine translation.

This section introduces our main progress in minority language processing and translation research, while the next section describes our organization of the CWMT minority language machine translation evaluation.

3.1 Basic Uyghur, Mongolian, and Tibetan Language Processing Technologies

Language processing is fundamental to machine translation. Basic language processing technologies are indispensable for both machine translation itself and corpus processing. For Uyghur, Mongolian, and Tibetan, we have focused on solving essential processing technologies required for machine translation, including encoding, morphological analysis, word segmentation, and named entity recognition.

- **Encoding Identification and Conversion:** Many minority languages, including Uyghur, Mongolian, and Tibetan, have multiple encoding schemes. For example, besides Unicode, Tibetan commonly uses Baima, Huaguang, Tongyuan, and Sangbozha encodings. Processing these languages requires encoding identification and conversion. Since language identification and encoding identification can be unified as encoding problems in computer representation, we address them simultaneously using a unified model. We propose a general statistical language model-based method for language and encoding identification. The method first coarsely identifies encodings into three character encoding families, then combines three granularity-level language models to simultaneously identify language and encoding. This approach does not rely on language-specific rules, facilitating extension to new languages and encodings. The three granularity-level language models are byte-based, character-based, and word-based models, which distinguish languages and encodings from three perspectives to improve identification performance. The overall processing flow is shown in [Figure 1: see original paper].

Based on this method, we have implemented a language and encoding identification tool currently supporting 11 languages and their mainstream encodings: Chinese, English, Tibetan, Uyghur, Mongolian, Arabic, Turkish, Russian, Kazakh, Kyrgyz, and Japanese, with average identification accuracy exceeding 95%. Encoding conversion tools for mainstream Tibetan, Uyghur, and Mongolian encodings have also been developed.

- **Uyghur and Mongolian Morphological Analysis:** Uyghur and Mongolian are morphologically rich agglutinative languages where words typically consist of stems and multiple affixes. Morphological analysis parses the stem-affix structure and identifies their categories. For agglutinative languages, we designed a graph-based discriminative model that represents morphological analysis results as graph structures. Through feature engineering, edges in the graph describe associations between morphological components within words and across adjacent words. Specifically, in the graph model, edges exist between stems and affixes within each word (capturing local features) and between stems and affixes across adjacent words (capturing long-distance features). These two feature types respectively describe linguistic association constraints within and between words.

Compared to traditional linear models, this approach better captures linguistic associations among morphological components.

Experiments show that graph-based morphological analysis achieves significant performance improvements over linear modeling and substantially surpasses previous work. Based on this model, we have implemented practical morphological analysis tools for Uyghur, Mongolian, and Korean. Another article in this special issue details the latest progress, and related work can be found in our published papers [33, 34].

- **Tibetan Word Segmentation:** Tibetan word formation is far more complex than Chinese. We constructed a word segmentation model suitable for Tibetan by organically combining rule-based and statistical methods. First, sentences are segmented into minimal granularity sequences called unit sequences based on Tibetan-specific word formation rules. Then, using weights from a perceptron model, coarse segmentation generates a directed graph, with dictionary lookup assigning different weights to graph edges. Finally, dynamic programming finds the shortest path in the weighted directed graph to produce the final segmentation result. [Figure 2: see original paper] illustrates the Tibetan word segmentation system workflow.

Experiments demonstrate that the discriminative model and word graph reranking-based Tibetan word segmentation model significantly advances previous best work [35]. The resulting Tibetan word segmentation tool has been effectively used in our Tibetan-Chinese machine translation system. Another article in this special issue provides detailed introduction to this work.

- **Named Entity Recognition and Translation:** Named entity recognition is crucial for subsequent processing stages such as syntactic analysis and machine translation. While Chinese and English named entity recognition technologies are relatively mature, the complex linguistic characteristics of Uyghur, Mongolian, and Tibetan prevent direct application of existing models. For instance, Uyghur named entities can take complex suffixes, making recognition inseparable from morphological analysis. For temporal and numeral expressions, which exhibit strong regularity across languages, we employ manual linguistic rules combined with bilingual dictionaries. For person names, locations, and organization names, we aim to build a general multilingual framework using language-independent statistical methods at the core, supplemented with rule-based processing for language-specific phenomena. We propose a framework combining rule knowledge and statistical modeling to leverage statistical models' stability, high precision, and language independence while accommodating language-specific lexical and syntactic patterns.

We have implemented numeral and temporal expression recognition and translation tools for Uyghur, Mongolian, and Tibetan, achieving high accuracy and improving translation quality. A discriminative statistical model-based named

entity recognition engine has also been developed, though further expansion of annotated named entity corpora and translation dictionaries for these languages is needed.

3.2 Analysis and Translation Modeling for Morphologically Rich Languages

Morphologically rich languages include agglutinative languages (e.g., Finnish, Japanese, Korean) and some highly inflectional fusional languages (e.g., German, French, Arabic, Russian). Many Chinese minority languages, such as Uyghur, Mongolian, Kazakh, and Korean, are morphologically rich. These languages can have hundreds or even thousands of word forms per word, whereas the most studied languages in machine translation—Chinese and English—have simple morphology. Chinese has virtually no morphological changes, while English verbs have only four to five forms. Mainstream machine translation methods generally ignore morphological variation, treating each word form as an independent word. For morphologically rich languages, this causes severe data sparsity, leading to numerous out-of-vocabulary words and seriously impacting performance. Moreover, many syntactic features in morphologically rich languages (tense, voice, person, number) are expressed through verb morphology, whereas in Chinese or English, these features are mostly expressed through specific words. This results in poor syntactic isomorphism between these language types, and existing machine translation models perform poorly on such structurally divergent language pairs.

To effectively achieve machine translation between morphologically rich and simple languages, mapping must be realized at deeper linguistic levels, enabling translation models to fully capture the characteristics of morphologically rich languages and their translational correspondences with Chinese.

Common word representation forms include: (1) treating complete words as independent units, which causes severe out-of-vocabulary problems in morphologically rich languages; and (2) segmenting words into stems plus multiple affixes, where each stem and affix is an independent unit. While this alleviates the out-of-vocabulary problem, it makes stems distant from each other, weakening statistical model effectiveness. Word representations can be flexible, such as the multi-granularity linear representation in [Figure 3: see original paper] or the graph-based representation in [Figure 4: see original paper].

Different word representations impact lexical analysis, word alignment, translation modeling, and algorithms. We investigate various representations to determine the most reasonable form for machine translation and corresponding alignment, translation models, and algorithms. Based on the graph-based representation in Figure 4: see original paper, we implemented a graph-structured morphological analysis model for morphologically rich languages, achieving excellent results for Uyghur, Mongolian, and Korean [33, 34]. For translation modeling, we improved translation quality from morphologically rich languages

to Chinese through multi-granularity representations and differentiated treatment of stems and affixes, achieving significant results for Uyghur, Kazakh, and Kyrgyz [36, 37]. Another article in this special issue details this work.

3.3 Knowledge Acquisition and Translation Techniques for Resource-Scarce Languages

Language resource scarcity is a major challenge in building minority language-to-Chinese translation systems. Current statistical machine translation methods rely on large-scale bilingual parallel corpora. Without such corpora, the advantages of statistical methods—low development cost and short cycles—disappear. Although minority language resource construction has received increasing attention in China, available bilingual parallel corpora remain limited. Moreover, minority language natural language processing foundations are relatively weak, lacking experts proficient in both minority languages and rule-based machine translation methods, making rule-based systems difficult to implement and maintain. For resource-scarce minority languages, neither pure statistical nor rule-based methods may achieve ideal results. Maximally utilizing various linguistic resources and human resources for rapid knowledge acquisition while effectively integrating multiple translation strategies to improve system performance constitutes an effective approach.

To address resource scarcity, we aim to identify knowledge gaps in existing systems through human-computer interaction, selectively introducing human expert knowledge to work closely with automatic learning processes, thereby accelerating learning and comprehensively leveraging human expertise and statistical learning capabilities. We have experimented with integrating human expert-written rules into statistical translation systems to solve long-distance reordering and sentence skeleton translation problems, achieving good results [38]. Additionally, we have researched multi-granularity fusion lexical alignment strategies [39] and bilingual-mapping-based unsupervised minority language syntactic analysis knowledge acquisition [40]. These efforts have partially alleviated resource scarcity difficulties with some success. Another form of resource scarcity is domain-specific resource shortage, which requires addressing domain adaptation. We have conducted some work in this area, though current results are not yet satisfactory.

Based on the above research, we have developed a series of practical minority language processing tools and built statistical machine translation systems for Uyghur-Chinese, Mongolian-Chinese, and Tibetan-Chinese. These systems have been deployed in relevant national departments and received positive feedback. Beyond minority languages, our group has also implemented prototype systems for Korean, Japanese, Thai, Russian, Arabic, and Vietnamese to Chinese, with progress in Korean morphological analysis, Japanese word segmentation, and Thai word segmentation.

4. Minority Language Machine Translation Evaluation

The China Workshop on Machine Translation (CWMT), initiated in 2005 by the Institute of Automation and Institute of Computing Technology of CAS and Xiamen University, aims to promote machine translation research development and domestic and international exchanges. Since 2007, CWMT has organized machine translation evaluation activities to facilitate substantive exchanges among research institutions and advance technology development. The Natural Language Processing Group at ICT-CAS has been responsible for organizing these evaluations. In 2011, under our group' s advocacy, CWMT 2011 introduced the first minority language-to-Chinese translation evaluation tasks, covering five languages: Uyghur, Mongolian, Tibetan, Kazakh, and Kyrgyz. CWMT 2013 continued with Uyghur-Chinese, Mongolian-Chinese, and Tibetan-Chinese evaluation tasks. and show the evaluation task settings, with minority language tasks highlighted in gray. The evaluation organizer, ICT-CAS, collaborated with minority universities to provide training corpora to participants. shows the training corpus scale for minority language tasks in these two evaluations and the corpus providers.

CWMT 2011 and CWMT 2013 minority language machine translation evaluations attracted 14 participating units, including ICT-CAS, Institute of Automation, Xinjiang Technical Institute of Physics and Chemistry, Harbin Institute of Technology, and Northeastern University. In CWMT 2013, we also collaborated with the Institute of Automation and Xiamen University to provide baseline systems (including full source code and tools for training and decoding) for Mongolian-Chinese, Uyghur-Chinese, and Tibetan-Chinese tasks. The minority language machine translation evaluation provides a cooperation and exchange platform for domestic machine translation research institutions and minority language information processing units, which will further promote research and application development. We look forward to more research teams participating in this evaluation.

5. Summary and Outlook

In recent years, the Natural Language Processing Group at ICT-CAS has conducted considerable work on minority language processing and machine translation, achieving certain progress. This paper provides an overview, with details available in other articles in this special issue. Future work will further expand minority language translation resources, research key technologies and methods suitable for minority language-Chinese machine translation, and hopefully significantly improve automatic translation quality between major minority languages (Uyghur, Mongolian, Tibetan) and Chinese, advancing minority language processing technology and practical machine translation systems. We will continue organizing minority language machine translation evaluations, hoping that minority language processing and machine translation research and applications will receive increasing attention.

References

[1] *Study Guide for "Several Opinions on Further Prospering and Developing Minority Cultural Undertakings"* . 2009. Compiled and issued by the State Ethnic Affairs Commission. http://cpc.people.com.cn/GB/165240/167240/10085487.html
[2] "Language Barriers Affect Minority Students' Education and Employment." *China Youth Daily*. September 16, 2005. http://edu.people.com.cn/GB/1053/3672381.html
[3] Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.
[4] Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, Philadelphia, PA: 295-302
[5] Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*, Sapporo, Japan: 160-167
[6] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, Ann Arbor, Michigan: 263-270
[7] Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL 2001*, Toulouse, France: 523-530.
[8] Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING/ACL 2006*, Sydney, Australia: 609-616
[9] Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proc. of ACL/HLT 2008*: 559-567
[10] Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP 2007*: 61-72
[11] D. Xiong, M. Zhang, and H. Li. 2012. Modeling the translation of predicate-argument structure for SMT. In *Proc. of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. In *Proceedings of ACL 2012*: 902-911
[12] Xin Ke Jian Zi No. 215, Xinjiang Minority Language-Chinese Machine Translation Information System, Xinjiang Uyghur Autonomous Region Science and Technology Appraisal Certificate (98), November 1998.
[13] Halimurat, Turgun, Alifu. 2005. Development of computer Chinese-Uyghur assisted translation software. In *Proceedings of the 10th National Minority Language Information Processing Academic Symposium*: 112-115
[14] Aishan · Maoerniyazi, Tan Xun, Turgun · Yibulayin, Aishan · Wumaier. 2011. Research on word alignment technology for Uyghur-Kazakh-Kyrgyz bilingual corpus processing systems. *Computer Knowledge and Technology*, 7(28): 6895-6897
[15] Rexida · Taiyi, Turgun · Yibulayin. 2009. Research on sentence alignment methods based on dictionary translations in Chinese-Uyghur bilingual corpora. *Journal of Xinjiang University (Natural Science Edition)*, 2009(3): 359-363

[16] Xu Chun, Yang Yong, Dong Xinghua. 2011. Research on several issues in Chinese-Uyghur/Uyghur-Chinese statistical machine translation. *Computer Engineering and Applications*, 47(35): 150-154
[17] Ren Gaoju, Turgun · Yibulayin, Aishan · Wumaier. 2010. Research on Chinese-Uyghur phrase pair extraction in statistical machine translation. *Journal of Xinjiang University (Natural Science Edition)*, 2010(3): 349-352
[18] Nashunwuritu, Liu Qun, Badamafangdesier. 2001. Mongolian generation for machine translation. In *Natural Language Understanding and Machine Translation*, Tsinghua University Press: 285-291
[19] Hou Hongxu, Liu Qun, Nashunwuritu. 2007. Example-based Chinese-Mongolian machine translation. *Journal of Chinese Information Processing*, 2007(4): 65-72
[20] Hu Guanlong, Li Miao. 2007. Multi-engine Chinese-minority language machine translation system based on successive screening. In *Research on Minority Language Information Technology*, February 2007: 187-191
[21] Jirimutu. 2005. Research on template-based English-Mongolian machine translation systems. Master' s thesis, Inner Mongolia University: 1-55
[22] Baishun. 2008. Research on Japanese-Mongolian verb phrase machine translation based on derivational grammar. *Journal of Chinese Information Processing*, 22(2): 47-52
[23] Nabuqing. 2006. Research on statistical Mongolian-Chinese machine translation systems. *Journal of Inner Mongolia Agricultural University (Social Science Edition)*, 2006(2)
[24] Kanzhuo Caidan, Jin Weixun, Li Yanfu, Luozhihua, Pengmao Zhaxi. 2006. Research on verb processing in Chinese-Tibetan translation systems. *Terminology Standardization and Information Technology*, 2006(3): 28-32
[25] Cai Zangtai, Huaguanjia. 2005. Research on binary-based syntactic analysis methods in Baima Chinese-Tibetan official document translation systems. *Journal of Chinese Information Processing*, 19(6): 7-12
[26] Zhang Guoxi. 2004. Implementation of English-Tibetan named entities in machine translation systems. *Journal of Qinghai Normal University (Natural Science Edition)*, 2004(3): 26-28
[27] Cai Rangjia. 2011. Research on large-scale Chinese-Tibetan (Tibetan-Chinese) bilingual corpus construction technology for natural language processing. *Journal of Chinese Information Processing*, 25(6): 157-161
[28] Zhaxijia, Gao Dingguo. 2011. Discussion on TEI annotation standards for Tibetan corpora. *Journal of Chinese Information Processing*, 25(4): 66-70
[29] Duola, Zhaxijia, Ou Zhu, Daluosanglangjie. 2007. Standards for Tibetan word classes and tags for information processing. In *Proceedings of the 11th National Minority Language Information Processing Academic Symposium*: 441-452
[30] Yang Pan, Zhang Jian, Li Miao, Wudabala, Xue Yan. 2009. Research on morphological methods in Chinese-Mongolian statistical machine translation. *Journal of Chinese Information Processing*, 23(1): 50-57
[31] Zhao Hongmei, Lü Yajuan, Ben Guosheng, Huang Yun, Liu Qun. 2012. Summary of the 7th China Workshop on Machine Translation evaluation.

Journal of Chinese Information Processing, 26(1): 22-30
[32] Zhang Haibo, Liu Kai, Lü Yajuan, Huaque Cairen, Liu Qun. 2012. A general method for minority language and encoding identification. In *Proceedings of the 4th National Minority Youth Natural Language Information Processing Academic Symposium*: 24-29
[33] Jiang Wenbin, Wu Jinxing, Chang Qing, Nashunwuritu, Liu Qun, Zhao Lili. 2011. A directed graph model for Mongolian morphological analysis. *Journal of Chinese Information Processing*, 25(5): 94-100
[34] Mai Rehaba · Aili, Jiang Wenbin, Wang Zhiyang, Turgun · Yibulayin, Liu Qun. 2012. A directed graph model for Uyghur morphological analysis. *Journal of Software*, 23(12): 3115-3129
[35] Sun Meng, Huaque Cairen, Cai Zhijie, Jiang Wenbin, Lü Yajuan, Liu Qun. 2013. Tibetan word segmentation based on discriminative classification and reranking technology. *Journal of Chinese Information Processing*, accepted.
[36] Wang Zhiyang, Lü Yajuan, Liu Qun. 2011. Multi-granularity translation fusion for morphologically rich languages. *Journal of Chinese Information Processing*, 4: 75-81
[37] Zhiyang Wang, Yajuan Lü, Meng Sun, Qun Liu. 2013. Stem Translation with Affix-Based Rule Selection for Agglutinative Languages. In *Proceedings of ACL 2013*, Sofia, Bulgaria
[38] Fu Lei, Huang Jin, He Zhongjun, Lü Yajuan, Liu Qun. 2006. A translation method fusing sentence pattern templates and statistical machine translation technology. China, Invention Patent, Patent No.: ZL200610165532.6, Authorization Date: 2009-09-23
[39] Zhiyang Wang, Yajuan Lü, Qun Liu. 2011. Multi-granularity Word Alignment and Decoding for Agglutinative Language Translation. *Machine Translation Summit XIII (MT-summit XIII)*, Xiamen, China
[40] Kai Liu, Yajuan Lü, Wenbin Jiang and Qun Liu. 2013. Bilingually-Guided Monolingual Dependency Grammar Induction. In *Proceedings of ACL 2013*, Sofia, Bulgaria

Author Biographies:

Lü Yajuan: Associate Professor, Institute of Computing Technology, Chinese Academy of Sciences. lvyajuan@ict.ac.cn
Liu Qun: Professor, Institute of Computing Technology, Chinese Academy of Sciences.
Jiang Wenbin: Assistant Professor, Institute of Computing Technology, Chinese Academy of Sciences.

Note: Figure translations are in progress. See original paper for figures.