Phrase-Based Tibetan-Chinese Statistical Machine Translation

YONG Cuo[1,2]，Xiaodong SHI [1*]，NYIMA Tashi[2*]，Yidong CHEN[1]
School of Information Science and Technology，Xiamen University，Xiamen, P.R.China[1]
School of Information Science and Technology，Tibet University，Lhasa, P.R.China[2]
yongtso@163.com; mandel@xmu.edu.cn; nmzx@utibet.edu.cn; ydchen@xmu.edu.cn

***Abstract*—Statistical machine translation has made great progress in recent years, and Tibetan-Chinese machine translation has many needs. A phrase-based translation model is suitable for machine translation between Tibetan and Chinese, which have similar morphological changes. This paper studies the key technologies of phrase-based Tibetan-Chinese statistical machine translation, including phrase-translation models and reordering models, and proposes a phrase-based Tibetan-Chinese statistical machine translation prototype system. The method proposed in this paper has better accuracy than Moses, the current mainstream model, in the CWMT 2013 development set, and shows great performance improvement.**

Keywords-Machine translation;Statistics;Phrase;Tibetan-Chinese

I. INTRODUCTION

Machine translation studies the use of computers to automatically convert between natural languages. It is one of the important research directions in the field of artificial intelligence and natural language processing [Koehn et al. 2003]. Machine translation is a key technology that breaks through the language barrier when transmitting information between countries. It is particularly important for strengthening cultural exchange and promoting foreign trade under the national strategy of the One Belt and One Road Initiative.

People have gradually realized that artificial translation rules have become a bottleneck in machine-translation research. Since the 1980s, the rise and rapid development of corpus linguistics have been refreshing. At the MT Summit IV conference held in Japan in 1993, the British scholar John Hutchins pointed out in a guest lecture that machine-translation research had entered its third generation. A major indicator is the introduction of corpus methods, including statistical and instance-based methods. Different from other natural language-processing research, the corpus used for machine-translation research is generally a parallel corpus, i.e., it contains translations between multiple languages. A common one is a bilingual parallel corpus consisting of two languages. Because the bilingual corpus contains contrasting translation information between two different languages, it has high research value and practical value in the field of natural language processing. Statistical machine translation has made great progress in recent years, with many significant achievements in international evaluation.

In China, the study of machine translation began in the 1950s. Several institutes and universities developed machine-translation systems such as Russian–Chinese, English–Chinese, Chinese–English, Japanese–Chinese, and Chinese–Japanese, and performed much research on the development of natural-language understanding of Chinese.

Tibetology is studied worldwide. Therefore, research on machine translation between Tibetan and other languages is of high concern. With the implementation of the One Belt and One Road Initiative and the advancement of informationization in Tibetan areas in China, there is wide demand for Tibetan-Chinese machine-translation technology in journalism, education, academic research, publishing, information security management, and cultural exchange. Hence, it is particularly important to study Tibetan-Chinese machine translation, which will not only enrich machine-translation theory, but also promote the development of Tibetan information technology, laying a solid foundation for the ultimate development of a Tibetan-Chinese two-way machine-translation system.

II. RELATED WORK

The study of machine translation in Tibetan and other languages started late, in the 1990s. Based on the research and development practice of the “863” project “BanZhida Chinese-Tibetan Official Document Machine Translation System”, Cai [2009] and Cai and Hua [2005] discussed the principle of combining lexical information with grammatical rules, and proposed a dichotomy of syntactic analysis centering on verbs. Combining their study with Tibetan language corpora segmentation specifications, they also discussed the establishment and design of a Tibetan grammar information dictionary for the segmentation and labeling of Tibetan corpora, and focused on the content construction, annotation of grammar information, index structure, and search algorithms of the dictionary. Chen and Yu [2003] introduced and evaluated related research work on the development of Tibetan-language information processing in China—the Tibetan language operating system, Tibetan information technology standards, Tibetan information processing, and comprehensive applications. They specifically mentioned their two "863" projects—a Chinese-Tibet science and technology machine translation system and a practical Chinese-Tibetan machine translation system, which adopted a conversion-translation model and a rule-based machine translation method. Based on the actual needs of constructing Tibetan corpora, Cai and Ji [2005] proposed a corpus-based Tibetan word-class annotation and classification method. Suo et al. [2004] proposed a British–Tibetan translation system based on rules and corpora, and outlined the design ideas and principles of the system. Cao and Suo [2009] proposed a language model and structural design of a British–Tibetan machine-translation system based on rules and corpora. He et al. [2015] analyzed Tibetan case-auxiliary words, added semantic information of the

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

Tibetan ontology based on the Tibetan phrase tree library, and proposed a Tibetan-Chinese machine-translation method that integrates Tibetan semantic information. Wang [2016] proceeded from the linguistic characteristics of the Tibetan language and carried out a study of identifying functional blocks of Tibetan sentences for an practical Tibetan-Chinese machine translation system.

In summary, Tibetan-Chinese machine-translation technology is mainly based on rules and linguistic materials, and does not involve phrase-based statistical machine translation. In particular, the Tibetan-Chinese statistical machine-translation research foundation is weak in the research on reordering models.Therefore, the key technology research and achievements of Tibetan-Chinese machine translation have strong academic research value and a wide range of application prospects, for both the actual needs of the society and the development of Tibetan natural-language processing research.

III. PHRASE-BASED TIBETAN-CHINESE STATISTICAL MACHINE

A relatively mature statistical translation method is a phrase-based translation method, which improves on IBM's word-based translation model. Word- and phrase-based machine translation use words and phrases, respectively, as the basic unit of translation. Taking the phrase as a basic unit does not need to consider the grammatical information between the words in a phrase, and therefore the obtained translation result is more accurate and reasonable than word-based translation. A phrase-based translation model is suitable for machine translation between Tibetan and Chinese, which share similar morphological changes.

A. Phrase Translation Model

A phrase translation model reflects the probability of translating a source-language phrase to a target-language phrase. Through translation model training, namely, word alignment and phrase extraction, Tibetan-Chinese phrase-translation probabilities are obtained from Tibetan-Chinese parallel training corpora.

1) Word Alignment

Word-alignment training obtains word-alignment models from sentence-aligned corpora. Using GIZA++ to perform bidirectional training between Tibetan and Chinese after pre-processing Tibetan-Chinese training corpora, bidirectional alignment results are obtained.

2) Phrase Extraction

Phrases are extracted from Tibetan-Chinese word-alignment results, translation probabilities are calculated, and the Moses statistical machine-translation system is used to obtain a phrase translation model. A Perl script optimizes bidirectional alignment results to obtain an alignment matrix. Phrases are extracted from the alignment matrix to obtain a phrase table, the probability of phrase-translation for each phrase pair is calculated using maximum likelihood estimation, and an IBM Model 1 is used to obtain the lexical translation probability and a Tibetan-Chinese phrase-translation probability table.

3) Decoding

Using the currently most popular and stable Moses decoder, a beam-search algorithm with dynamic programming is used to implement the search function. The decoder reads the configuration file and test file and outputs a file of translation results. The translation outcome is improved by adjusting the feature weights of the model.

4) Translation Rules

To guarantee the quality of translation, pruning and clustering methods are used to reduce the number of translation rules. The goal is to reduce the number of translation rules to about 1/4 so that the translation speed can be increased to more than 100 words per second, with only a slight decrease in quality. Specific strategies include:

- To introduce part-of-speech rules for abstraction, rather than the current phrase-system translation rules that has only a specific vocabulary. As long as the parts of speech are the same, translation rules can be matched.
- To remove translation rules with too small a probabilities.
- To remove reordering rules with too small probabilities.

B. Language Model

This paper uses the traditional n-gram language model with SRILM [Stolcke 2002] to train the five-gram language model.

C. Reordering Model

The reordering model adopts a new reordering algorithm, which can better solve the syntactic differences between Tibetan and Chinese due to the difference in verbal order.

All available information for the two sets of phrases (phrases to be ordered and phrases for reference) needed for reordering, including the vocabulary itself, parts of speech, and even other lexical semantic information, is assembled. The two sets of phrases may be adjacent or non-adjacent. The reordering operation is expanded to include M, S, L, R, F, and I, which respectively represent monotonous, switch, left discontinuity, right discontinuity, end of sentence, and beginning of sentence. It is represented by the following equation (1):

$$p(o|e,f)=\prod_{i=1}^{k}p(o|<\overline{e_{t-n}},\overline{f_{t-n}}>,<\overline{e_{a_t}},\overline{f_t}>)$$

The difference between our reordering model and another model is shown in Figure 1.

Due to the sparseness of the data, we obtain more reliable reordering calculations through equivalence-like functions:

$$p(o|<\overline{e_{a_{t-n}}},\overline{f_{t-n}}>,<\overline{e_{a_t}},\overline{f_t}>)=\sum_{i=1}^{m}w_i p_i(o|<\varphi_i(\overline{e_{a_{t-n}}}),\varphi(\overline{f_{t-n}})>,<\varphi_i(\overline{e_a}),\varphi(\overline{f_t})>)$$

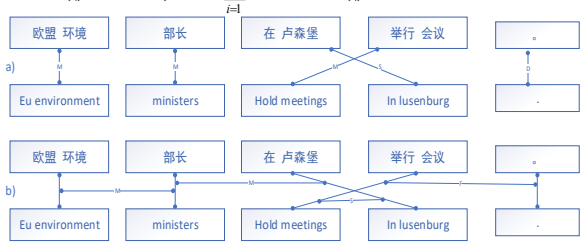

Figure 1. (a) Galley and Manning 2008 Model, (b) Our model.

The algorithm combines the reordering model with the linear decoding process from left to right. Our model differs from the previous model, as shown in Figure 2.

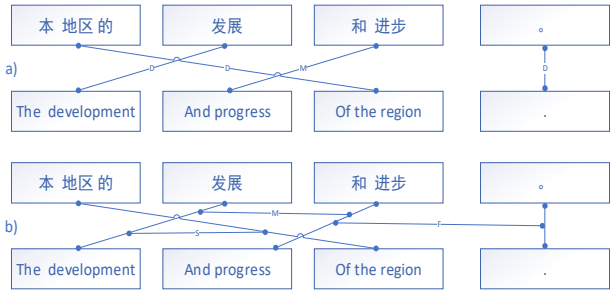

Figure 2. (a) Galley and Manning 2008 Model, (b) Our model.

Without considering the end of the sentence and sentence order, there are six reordering methods for adjacent phrases, as shown in Figure 3.

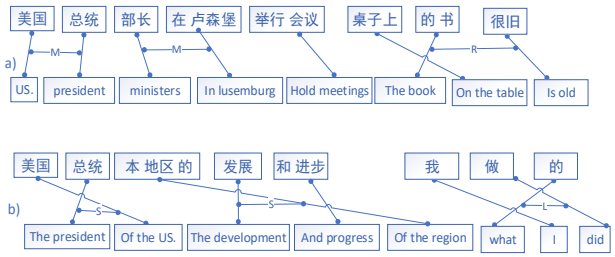

Figure 3. Six main ordering ways.

The reordering algorithm is as follows.

Algorithm 1 Computing Reordering Probabilities

find bilingual phrases using grow-diag-final-and method
foreach adjacent phrase pair <s_{i-1},t_{i-1},s_i,t_i> from left to right
  If t_{i-1},t_i are adjacent,
then If t_{i-1} is to the left of t_i,
  then o ← m
  else o ← s
else **If** t_{i-1} is to the left of t_i,
  then if there are phrases between t_{i-1} and t_i that align to the left of s_{i-1},
then o ← r
else o ← m
  else if there are phrases between t_{i-1} and t_i that align to the left of s_{i-1},
then o ← l
else o ← s
add <s_{i-1},t_{i-1},s_i,t_i,o,1> to the reordering frequency table
foreach equivalence function $\emptyset_i$
add <$\emptyset_i(s_{i-1})$, $\emptyset_i(t_{i-1})$, $\emptyset_i(s_i)$, $\emptyset_i(t_i)$,o,1> to the reordering freq table
Normalize the raw frequencies into probabilities

Experiments showed that our reordering algorithm had better accuracy than the current mainstream model (Moses). In the CWMT 2013 development set, there was a great performance improvement. Performance comparison is shown in Table 1.

TABLE1. PERFORMANCE COMPARISON

	BLEU SBP	BLEU	LM
Moses phrase model	0.5562	0.5847	1016 M words, 5-gram
Our model	0.5819	0.6109	10 M words, trigram

D. System Architecture and Implementation

The Tibetan-Chinese machine translation system employs a hierarchical phrase model. A new semantic model is used to extract the hierarchical rules to solve the generalization problem of the model. Based on Tibetan sentence analysis, the syntax (boundary) restriction is used to further reduce the redundancy of rules.

The phrase-based Tibetan-Chinese statistical machine translation prototype system contains five major modules. Its overall architecture and translation engine are illustrated in Figures 4 and 5. In terms of word segmentation, the Chinese word-segmentation tool segtag developed by the Natural Language Processing Laboratory of Xiamen University is used for Chinese, and the Sunshine Tibetan-language word-segmentation tool developed by our laboratory is used for Tibetan.

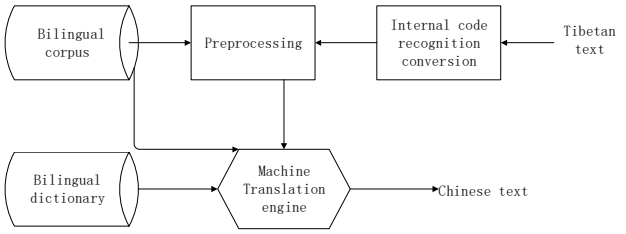

Figure 4. The system architecture

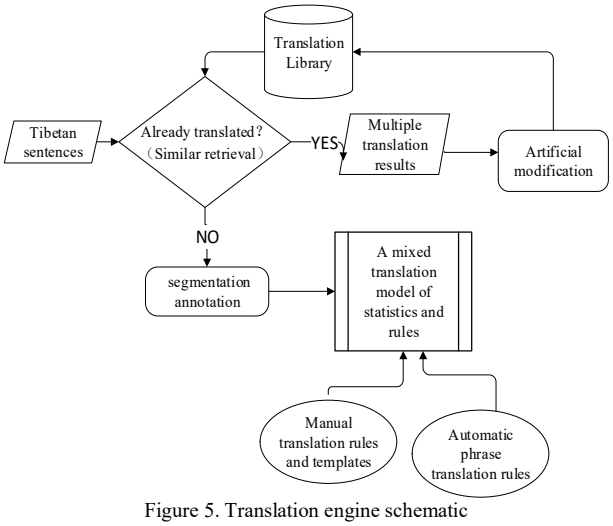

Figure 5. Translation engine schematic

Given a source-language sentence f, the statistical machine translation process based on a logarithmic display model searches for the target translation with the highest probability, e:

$$e^* = \arg\max \sum_{m=1}^{M} \lambda_m h_m(e, f)$$

where hm(e,f) is the characteristic function and λm is the feature weight. In the phrase-based model, the characteristic functions used are:

1) Phrase translation probability (two directions);
2) Word translation probability (two directions);
3) Language model;
4) Reordering based on source phrases;
5) Length penalty.

Thus, there are seven features. The decoding search strategy is a column-search algorithm, which ultimately generates a 1-Best result. At present, the weight of each characteristic function is set by experience.

Training is performed using the GIZA++ toolkit, and grow-diag-final heuristics are used for word-alignment expansion to generate phrases. When constructing the phrase table, the word translation probability of GIZA++ is fused to avoid the situation where the translation of certain words is not in the phrase table.

The phrase-based Tibetan-Chinese statistical machine translation system currently has about 75% readability of translation results in news and government documents.

IV. Conclusion and Future Work

Statistical machine translation has made great progress in recent years, and Tibetan-Chinese machine translation technology has many needs. Since a phrase-based translation model is suitable for machine translation between Tibetan and Chinese, which have similar morphological changes, this paper has studied the key technologies of phrase-based Tibetan-Chinese statistical machine translation, including phrase-translation and reordering models, and we have designed and implemented a phrase-based Tibetan-Chinese statistical machine translation prototype system.

At present, the performance of the system in translating news and government documents is acceptable. However, due to the limited scale of Tibetan-Chinese parallel corpora, the quality of translation in other areas must be improved. The system uses a relatively ancient hierarchical phrase model, with much room for improvement.

Research on neural networks has progressed greatly in recent years. In the future, a Tibetan-Chinese statistical machine translation method integrating hierarchical phrase models and neural network translation models [Bahdanau et al. 2014; Cho et al. 2014; Devlin et al. 2014; Kalchbrenner and Blunsom 2013; Wang et al. 2017] should be studied to further improve the quality of Tibetan-Chinese machine translation.

V. Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2017YFB1402200) and the National Team and Key Laboratory Construction Program of Computer and Tibetan Information Technology of the Education Department of Tibet Autonomous Region. Xiaodong SHI and NYIMA Tashi are the corresponding authors.

VI. Additional Authors

Additional authors: GAMA Trashi (Tibet University, email: 80957997@qq.com),Yang GUO (Tibet University, email: sdfg33445@qq.com).

References

[1] BAHDANAU, D., CHO, K. AND BENGIO, Y. 2014 Neural machine translation by jointly learning to align and translate, Comput. Res. Repos. https://arxiv.org/abs/1409.0473.

[2] CAI, R.J. AND JI, T.J. 2005 Researches of speech classification methods based on tibetan repertoire, Journal of Northwest University for Nationalities (Natural Science) 26, 2, 39-42.

[3] CAI, Z.T. 2009 Design of Tibetan segmentation dictionary and its algorithm study, Journal of Computer Applications 7, 2019-2021.

[4] CAI, Z.T. AND HUA, G.J. 2005 Research of banzhida Chinese-Tibetan document translation system based on the dichotomy of syntax analysis, J. Chin. Inf. Proc. 19, 6, 9-14.

[5] CAO, Y.L. AND SUO, N.D.Z. 2009 English-Tibetan machine translation system model and population structure design, Journal of Southwest University for Nationalities (Natrual Science Edition) 2, 365-370.

[6] CHEN, Y.Z. AND YU, S.W. 2003 Research status and prospect of Tibetan information processing technology, China Tibetology 4, 97-107.

[7] CHO, K., van MERRIENBOER, B.M., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. AND BENGIO, Y. 2014 Learning phrase representations using encoder-decoder for statistical machine translation, Comput. Res. Repos. https://arxiv.org/abs/1406.1078?context=cs.NE.

[8] DEVLIN, J., ZBIB, R. AND HUANG, Z. 2014 Fast and robust neural network joint models for statistical machine translation, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics, Baltimore, USA, 1370-1380.

[9] HE, X., WAN, F., YU, H. AND WU, X. 2015 Machine translation technology based on Tibetan semantic parsing, Comput. Eng. Appl. 15, 134-137,173.

[10] KALCHBRENNER, N. AND BLUNSOM, P. 2013 Recurrent continuous translation models, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, Seattle, USA, 18-21.

[11] KOEHN, P., OCH, F.J. AND MARCU, D. 2003 Statistical phrase-based translation, In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Edmonton, Edmonton, Canada, 48-54. doi:10.3115/1073445.1073462

[12] STOLCKE, A. 2002 SRILM-an extensible language modeling toolkit, In 7th International Conference on Spoken Language Processing ISCA, Denver, USA.

[13] SUO, N.D.Z., MA, N.C. AND CAO, Y.L. 2004 Regulation and corpus-based English to Tibetan machine translation system design model, Terminology standardization and Information Technology, 4, 37-42.

[14] WANG, M., LU, Z., ZHOU, J. AND LIU, Q. 2017 Deep neural machine translation with linear associative unit, Comput. Res. Repos. https://arxiv.org/abs/1705.00861.

[15] WANG, T.H. 2016 An MT-oriented research on recognition of Tibetan syntactic functional chunk, Master's thesis. Computer Science and Technology Beijing Institute of Technology, Beijing, China.