himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

Himalayan Linguistics

A Chinese to Tibetan machine translation system with multiple translating strategies

Huidan Liu
Chinese Academy of Sciences

Weina Zhao
Chinese Academy of Sciences; Qinghai Normal University

Minghua Nuo
Chinese Academy of Sciences; Inner Mongolia University

Jinling Hong
Chinese Academy of Sciences

Xin Yu
Chinese Academy of Sciences

Jian Wu
Chinese Academy of Sciences

ABSTRACT

This paper proposes a Chinese to Tibetan machine translation system with multiple translating strategies. The key corpora and technologies are explained in detail. Experiments show the sub-systems output the translation of each phrase in the same order as they are in the Chinese sentence rather than in a Tibetan sentence, which leads to unsatisfactory translation quality. Consequently, an order adjusting model is essential to Chinese to Tibetan translation system. The phrase recall of the SMT (MaxProb) sub-system makes an improvement of 9.71% over the popular off-the-shelf language neutral statistical machine translation programme Moses. Our translation system achieves a speed of about 0.175s per sentence on a PowerEdge R710 server with 2 4-core Intel E5620 (2.40 GHz) CPUs and 16 GB memories, which meets the requirement of a computer aided translation system.

KEYWORDS

Tibetan, NLP, machine translation

This is a contribution from *Himalayan Linguistics, Vol. 15(1): 149–166.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

A Chinese to Tibetan machine translation system with multiple translating strategies

Huidan Liu

Chinese Academy of Sciences

Weina Zhao

Chinese Academy of Sciences; Qinghai Normal University

Minghua Nuo

Chinese Academy of Sciences; Inner Mongolia University

Jinling Hong

Chinese Academy of Sciences

Xin Yu

Chinese Academy of Sciences

Jian Wu

Chinese Academy of Sciences

1 Introduction

The Tibetan language is the main carrier of Tibetan culture; it plays a very important role in Tibetan people's daily communication (Chen 2003). According to a previous investigation, there are about 100 organizations which have Chinese-Tibetan translation business in Tibet and about 1,000 translators who make translation their profession. But the annual amount of Chinese text to be translated into Tibetan is more than 50 million Chinese characters every year (Luo et al. 2010). Thus, there is a big gap between the translating ability and the market requirement. Machine translation technology is an urgent requirement, which contributes to the improvement of translation speed as well as its quality.

In the paper, we propose a Chinese to Tibetan machine translation system with multiple translation strategies. The paper is organized as follows: In Section 2 we recall related work on machine translation in general and Tibetan related machine translation in particular. In Section 3, we introduce the architecture of our machine translation system. The core corpora and technologies are explained in Section 4. Then, in Section 5, we experimentally compare the performances of the system with some other methods. Section 6 offers some conclusions.

2 Related Work

There are mainly three types of machine translation methods, namely rule-based machine translation (RBMT), example-based machine translation (EBMT), and statistical machine translation (SMT). Research of Chinese-Tibetan machine translation focused on rule-based methods before 2010 due to a lack of parallel corpora and other basic Tibetan language resources (Degai 2001; Cai 2005; Zhaluo 2005; Kanzhuo et al. 2006). In recent years, statistical Chinese-Tibetan machine translation methods are playing a more and more important role.

Generally speaking, a large rule set is needed in a RBMT system. However, it takes huge human resources and requires a high level of language knowledge. An EBMT system takes less human resources and requires a lower level of language knowledge. But if we can't find a very similar instance for the input text, the translation can't be satisfactory. A parallel corpus is essential to a SMT system. As long phrase takes more memory and will confront the problem of data sparseness, the well-known Moses statistical translation programme sets the limit of phrase length to 7 words or less (Koehn 2007).

Hou (2007) presents an example based Chinese-Mongolian machine translation method. The method consists of several parts, including example searching, segment splitting, matching and recombining. The method is based on word alignment. It uses word alignment information for segment matching, and computing the similarities by the number of matching words and length， and selects the best example. Kang et al. (2007) proposed a hybrid method which combines a statistical method with linguistic rules to extract Chinese multi-word chunks for translation purposes.

In Chinese-Tibetan machine translation, there are several RBMT systems reported. Cai (2005) built a Chinese-Tibetan machine translation system for government documents, which is based on dictionary and grammar rule templates. Degai et al. (2001) also built a machine translation system based on knowledge rules. Zhaluo (2006) proposed a method to build translation rules for complex sentences. Kanzhuo (2006) discussed the classification of verbs, different forms of verbs and the changeable regular pattern of verbs when they are in different tenses, and suggested several methods to improve the quality of translation.

There are two crucial problems to solve in Chinese-Tibetan machine translation. First, at the sentence level, all Chinese sentences have SVO structure, but Tibetan has SOV structure. So, long distant order adjusting is essential in Chinese-Tibetan machine translation. Second, verbs in Chinese do not inflect for tense, but in Tibetan a verb may have many variants. For instance, the forms of the verb "complete" are listed in Table 1.

	present tense	future tense	past tense	imperative mood
inactive voice	འགྲུབ་	འགྲུབ་	བགྲུབས་	འགྲུབས་
active voice	འགྲུབ་	འགྲུབ་	གྲུབ་	

Table 1. Forms of the Tibetan verb 'complete'

In this paper, we focus on building a Chinese to Tibetan machine translation system with multiple translating strategies to improve the translation quality subject to a poor language resource. It is expected that the system can find the translation if a sentence itself or a very similar sentence is in the parallel corpus, while it uses phrase-based machine translation decoding to generate a

translation in the many cases in which no similar sentences can be found, and a dictionary based decoding makes the greatest effort to assure that at least one translation is found for every word in the sentence.

3 System Structure

The proposed system is a combination of three different types of machine translation models, namely EBMT model, SMT model and RBMT model, as shown in Figure 1.

In the training phase of the system, Chinese articles and their translations are collected from several government translation organizations. These articles are processed into bilingual sentence pairs to form the bilingual sentence level parallel corpus. The sentence pairs are segmented into words by ICTCLAS (Zhang 2003) using Chinese and Tibetan word segmentation tools (Liu 2011; Liu 2012), and indexed to form the EBMT model. The segmented sentence pairs are word aligned and used to train the SMT model. Bilingual phrase pairs are extracted from the segmented sentence pairs. They form the RBMT model with word pairs from bilingual dictionaries (Zhang 1993; The ethnic publishing house 2002).

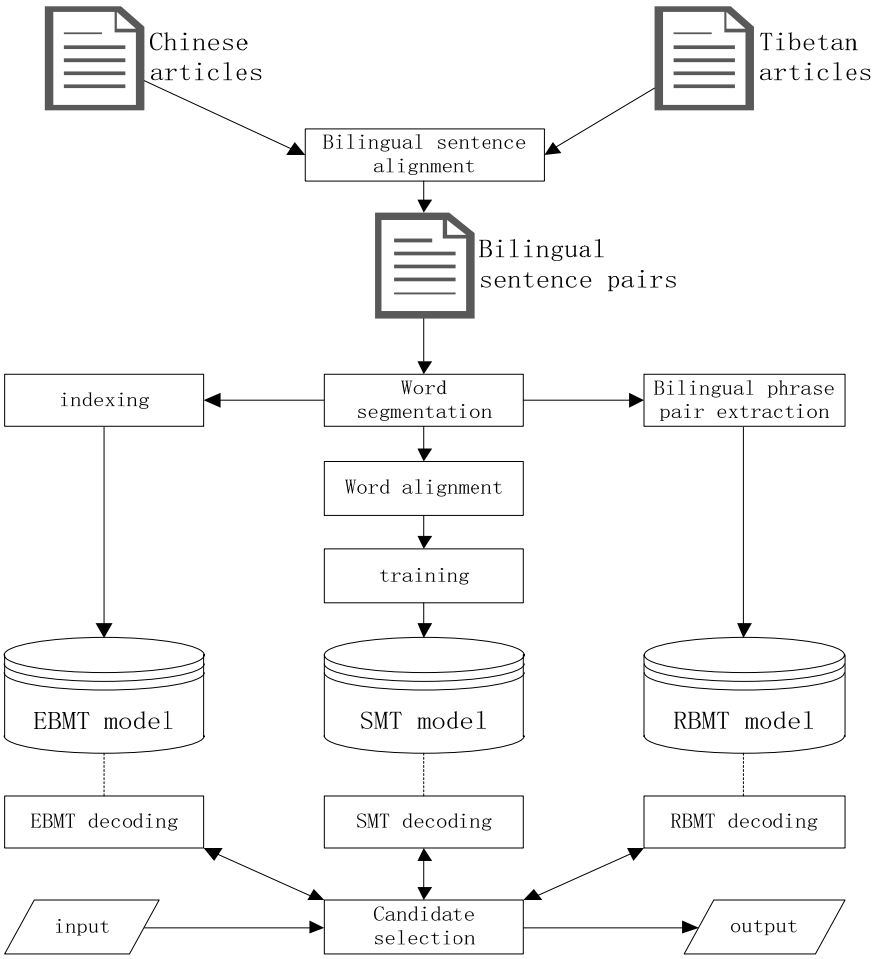

Figure 1. Data flow of the proposed system.

In the application phase of the system, as in the input of the system, a Chinese sentence is sent to the three sub-systems to be translated. The EBMT decoder segments the sentence into words, and trys to find similar sentences in the indexed bilingual sentence pairs by Levenshtein distance (Hirschberg 1975). If the EBMT decoder finds a Chinese sentence similar to the input one, and the similarity is larger than a predefined threshhold, its Tibetan counterpart is extracted and taken as the output of the EBMT sub-system. Otherwise, the input sentence will be sent to the SMT decoder. The SMT decoder splits the input sentence into words and phrases, and computes the probability for every potential candidate. The probability determines how good each candidate is. The best candidate is selected as the output of the SMT sub-system. If the former two decoders cannot offer a good enough translation, the Chinese sentence is sent to the RBMT decoder, word to word translation is performed with a maximum matching algorithm with a trie structure (Fredkin 1960).

In the following section, we dicuss the key corpora and technologies in both of the training phase and the application phase in detail.

4 Key corpora and technologies

4.1 Tibetan word segmentation

4.1.1 Rule based method

We design and implement a Tibetan word segmentation system named "SegT". It identifies critical words with a fast algorithm while segmenting each Tibetan sentence to chunks with case-auxiliary words, such as གི, གྱི, གྱི, ཡི, གིས, གྱིས, གྱིས, and ཡིས. Each chunk is segmented into words by both forward maximum matching and backward maximum matching with a Trie tree structure1. It detects ambiguities by bidirectional segmentation, and disambiguates making use of pre-determined word frequencies, such that the segmentation yielding the more frequent words is selected.

In the procedure, the structure of each syllable is analyzed to identify abbreviated syllables while segmenting each block into words. In Tibetan text, some syllables, including འི, ས,ར, འང, འམ, འོ (We call them abbreviation marker (AM) in this paper), can adhere to the previous word without a syllable delimiter "tsheg". The combination of these abbreviation markers with their preceding syllables produce 'abbreviated syllables'. For example, when the genitive case word འི follows the word རྒྱལ་པོ (king), no "tsheg" appears between them and they are fused to form རྒྱལ་པོའི (king[+genitive], king's), in which པོའི is an abbreviated syllable. When the ergative case word ས follows the word ངཱ་ཚོ (we), it forms ངཱ་ཚོས (we[+ergative]), in which ཚོས is an abbreviated syllable. Table 2 shows more examples.

[1] http://www.chasen.org/~taku/software/darts/

word	AM	result	explanation
ང་	སྟ་	ངའི་	Tsheg is omitted.
གཡ་ཆེ་	འི་	གཡ་ཆེའི་	Tsheg is omitted.
གོ་	འང་	གོའང་	Tsheg is omitted.
རྣ་བ་	འམ་	རྣ་བའམ་	Tsheg is omitted.
སྣ་	འམ་	སྣའམ་	Tsheg is omitted.
དབང་པོ་ལྔ་	འོ་	དབང་པོ་ལྔའོ་	Tsheg is omitted.
རྒྱ་མཚའ་	འི་	རྒྱ་མཚའི་	འ(/a/) and tsheg are omitted.
ནམ་མཁའ་	འི་	ནམ་མཁའི་	འ(/a/) and tsheg are omitted.
བཤད་པ་	ར་	བཤད་པར་	Tsheg is omitted.
རྒྱལ་པོ་	འི་	རྒྱལ་པོའི་	Tsheg is omitted.

Table 2. Examples of Tibetan abbreviated syllables.

In the rule based word segmentation system, Tibetan words are collected from several dictionaries, namely "Tibetan Chinese General Dictionary" (Zhang 1993), "The Antitheses Chinese-Tibetan Dictionary" (The ethnic publishing house 2002), and "The Antitheses Chinese-Tibetan Oral Dictionary" (Yu 1983), as well as some digital dictionaries. About 220 thousand words are included in the segmenting dictionary. Word frequencies are collected by segmenting a Tibetan text corpus with 230 thousand sentences and 2 million words in total, which includes newspaper articles, law text, political papers and books. In a previous study, experiments show that the precision of the system reaches 96.98% (Liu 2012).

4.1.2 Statistical based method

As Statistical based methods detect out of vocabulary items much more effectively than rule based methods, we also developed a Tibetan word segmentation tool base on the Conditional Random Fields (CRFs) machine learning model. We reformulate Tibetan word segmentation as a syllable tagging problem, and propose an approach using conditional random fields (CRFs) for Tibetan word segmentation.

We convert the segmented words in the corpus into a tagged sequence of Tibetan syllables (or sub-syllables). We tag each syllable with one of the four tags, B (Begin), M (Middle), E (End) and S (Single) depending on its position within a word. Two additional tags, namely ES (End and Single) and SS (Single and Single) are used when we take the syllable rather than the sub-syllable as the tagging unit (Liu 2011; Liu 2015). For each syllable:

(1) It is tagged B if it is the left boundary of a word.
(2) It is tagged M if it is at middle of a word.
(3) It is tagged E if it is the right boundary of a word.
(4) It is tagged S if it is a word by itself.
(5) It is tagged ES if it comes from a multiple-syllable word and an AM.
(6) It is tagged SS if it comes from a single-syllable word and an AM.

Then, for the Tibetan sentence in (a), which means (b), it's segmented into (c) manually. Consequently, it's converted into (d) or (e) by applying the aforementioned tags to form a word

segmentation corpus to be used as the training set for the CRFs. Earlier research shows that this method achieves precisions higher than 94.43% (Liu 2015).

(a) ང་ཚོས་སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་སྤྱི་ལ་དབང་བའི་ལས་ལུགས་དང་རྩོལ་བསྟུན་ཐོག་སྐྱོང་ཀྱི་རྒྱ་དེན་མཐའ་འཛོམས་བས་ཡོད།

(b) We have always followed the principles of socialist public ownership and distribution according to work.

(c) ང་ཚོ/ ས་/ སྤྱི་ཚོགས་རིང་ལུགས་/ ཀྱི་/ སྤྱི་ལ་དབང་བའི་ལས་ལུགས་/ དང་/ རྩོལ་/ བསྟུན་/ ཐོག་/ སྐྱོང་/ ཀྱི་/ རྒྱ་དོན་/ མཐའ་འཛོམས་/ བས་/ ཡོད་/ །/

(d) ང་/B ཚོས་/ES སྤྱི་/B ཚོགས་/M རིང་/M ལུགས་/E ཀྱི་/S སྤྱི་/B ལ་/M དབང་/M བའི་/M ལས་/M ལུགས་/E དང་/S རྩོལ་/S བསྟུན་/S ཐོག་/S སྐྱོང་/S ཀྱི་/S རྒ་/B དོན་/E མཐའ་/B འཛོམས་/E བས་/S ཡོད་/S །/S

(e) ང་/B ཚོ/E ས་/S སྤྱི་/B ཚོགས་/M རིང་/M ལུགས་/E ཀྱི་/S སྤྱི་/B ལ་/M དབང་/M བ/M འི་/M ལས་/M ལུགས་/E དང་/S རྩོལ་/S བསྟུན་/S ཐོག་/S སྐྱོང་/S ཀྱི་/S རྒ་/B དོན་/E མཐའ་/B འཛོམས་/E བ/B ས་/E ཡོད་/S །/S

4.2 Bilingual sentence level parallel corpus building

Constructing corpora is a basic necessity for Natural Language Processing. For Chinese-Tibetan machine translation, large scale bilingual parallel corpora are still basic resources which are urgently needed. We collected Tibetan text and Chinese text from several translating organizations. A bilingual sentence level parallel corpus was built as part of our project, which includes 571 thousand bilingual sentence pairs. As shown in Table 3 and Table 4, two versions are included. Set A is a long sentence version and Set B is a short sentence version. Each sentence pair in Set A has a complete Chinese sentence which ends with a period, while each sentence pair in Set B has a shorter Chinese sentence which may end with comma. Both of the sets are used in the machine translation system.

Long version	Chinese	西藏自治区面积 122 万平方公里, 平均海拔在 4000 米以上, 有着独特的自然生态和地理环境。
	Tibetan	བོད་རང་སྐྱོང་ལྗོངས་ཀྱི་ས་ཁྱོན་ལ་སྦྱི་ལེ་ཤུ་བཞི་མ་ཁྲི་122ཡོད་ཅ་དང་། རྒྱ་མཚོའི་ངོས་ལས་མཐོ་ཚད་ཆ་སྙོམས་ལུ་སྦྱི་4000ཡན་ཟེན་ཅ་དང་། རང་བྱུང་གི་སྐྱེ་གནས་དང་ས་གཤིས་ཀྱི་ཡུལ་གནས་དམིགས་བསལ་ཡིན།
Short version	Chinese	西藏自治区面积 122 万平方公里,
	Tibetan	བོད་རང་སྐྱོང་ལྗོངས་ཀྱི་ས་ཁྱོན་ལ་སྦྱི་ལེ་ཤུ་བཞི་མ་ཁྲི་122ཡོད་ཅ་དང་།
	Chinese	平均海拔在 4000 米以上,
	Tibetan	རྒྱ་མཚོའི་ངོས་ལས་མཐོ་ཚད་ཆ་སྙོམས་ལུ་སྦྱི་4000ཡན་ཟེན་ཅ་དང་།
	Chinese	有着独特的自然生态和地理环境。
	Tibetan	རང་བྱུང་གི་སྐྱེ་གནས་དང་ས་གཤིས་ཀྱི་ཡུལ་གནས་དམིགས་བསལ་ཡིན།

Table 3. Long sentence pair and the corresponding short sentence pairs.

Domain	#Sentence (Set A)	#Sentence (Set B)
Law text	115,299	68,535
Leader's book	53,292	96,181
News	26,613	4,270
Government reports	72,849	102,795
Dictionary	31,234	
Total	299,287	271,781

Table 4. Bilingual sentence level parallel corpus.

The Chinese part of the corpus is processed by some rules and open source tools. Some rules are used to segment Chinese text into sentences. A Chinese word segmentation tool named "ICTCLAS"[2] is used to segment Chinese sentence into words. Tibetan word segmentation methods described in the previous subsection are used to segment Tibetan sentence into words. There are also two problems to solve. The first is how to find the boundary of a Tibetan sentence. The second is how to align Tibetan sentences with Chinese sentences.

4.2.1 Tibetan sentence boundary detection

Tibetan sentence boundary detection is a nontrivial problem because punctuation marks in Tibetan are not used exclusively to mark sentence breaks, in other words the existence of a punctuation marker does not necessarily suggest the boundary of a sentence. In particular, "། (SHAD)" is one of the most common and significant punctuation marks in Tibetan, which functions like a period, a caesura sign, and a comma.

Tibetan is an ergative language whose structure is SOV (subjective, objective, and verb), so a predicate almost always found at the end of a sentence. Therefore, we choose the predicate to help us disambiguate the end-of-sentence patterns. The diverse constituents of predicates can be divided into four types:

(1) Verb. For example: ང་དགེ་རྒན་ཡིན།
(2) Verb with Auxiliary. For example: ངས་རང་བོད་ཡིག་འཇི་ཤེས།
(3) Verb with Aspect-Evidentially Mark. For example: ངས་ཁྱེད་རང་གི་ཐུག་ལས་ལ་གཅོད་བསྒལ་མ་ཡོང་།
(4) Verb with Modal Particle. For example: ཁྱེད་རང་ཟླ་ས་བ་ཡིན་ནམ།

We observed from the corpus that in a majority of cases the last word of a Tibetan sentence is an auxiliary. Specifically, 78.79% of sentences end with auxiliaries. Considering the importance of auxiliaries, we built an Auxiliary List. A Tibetan verb lexicon is also used in our algorithm. They are used to find the sentence boundaries in a Tibetan article. A previous paper shows that the method is robust and efficient, and its accuracy reaches 99.26% (Zhao 2010).

[2] http://ictclas.nlpir.org/newsdownloads?DocId=389

#	positive	negative	#	positive	negative
1	ཐུབ།	མི་ཐུབ།	16	ཚོད།	མི་ཚོད།
2	སྲིད།	མི་སྲིད།	17	འདོད།	མི་འདོད།
3	ཤེས།	མི་ཤེས།	18	པར་འདོད།	པར་མི་འདོད།
4	མཁྱེན།	མི་མཁྱེན།	19	བར་འདོད།	བར་མི་འདོད།
5	འོས།	མི་འོས།	20	དད།	མི་དད།
6	བར་འོས།	མི་བ་འོས།	21	པར་དད།	པར་མི་དད།
7	པར་འོས།	མི་པ་འོས།	22	བར་དད།	བར་མི་དད།
8	རུང་།	མི་རུང་།	23	བཏུབ།	མི་བཏུབ།
9	རུང་དང་མ་རུང་།		24	སུ་བཏུབ།	སུ་མི་བཏུབ།
10	ཀྱང་རུང་།		25	ཏུ་བཏུབ།	ཏུ་མི་བཏུབ།
11	ཆོག	མི་ཆོག	26	ད་བཏུབ།	ད་མི་བཏུབ།
12	ཉན་།	མི་ཉན།	27	རུ་བཏུབ།	རུ་མི་བཏུབ།
13	ལ།	མི་ལ།	28	ར་བཏུབ།	ར་མི་བཏུབ།
14	དགོས།	མི་དགོས།	29	ལ་བཏུབ།	ལ་མི་བཏུབ།
15	ནུས།	མི་ནུས།	30	དཀའ།	དཀའ་མིན།

Table 5. A part of the Auxiliary List

4.2.2 Bilingual sentence alignment

A Chinese-Tibetan dictionary with 137,873 items was collected by combining several published dictionaries (Liu 2011; Liu 2012; Liu 2015). Bilingual articles are respectively segmented into monolingual sentences. They are further segmented into words. As the correspondences of some words in Tibetan sentence to their Chinese translations in Chinese sentence exist, a dynamic programming algorithm is applied to find the correspondence of the sentences in each pair of Bilingual articles. A previous study shows that the aligning precision of this approach is 84.8% (Yu 2010). We implemented a tool for further proofreading to correct alignment errors (Yu 2012).

4.3 Bilingual word alignment

As Bilingual word pairs are core resources in statistical machine translation, each bilingual sentence pair is segmented into word sequence pairs and aligned. Word alignment finds their correspondences in the target language for the words in the source language. Figure 2 shows a word alignment example. In the system, the alignment is denoted by α={1→14;2→1;3→2;4→3;5→4;6→6,7;7→10;8→11;9→16}

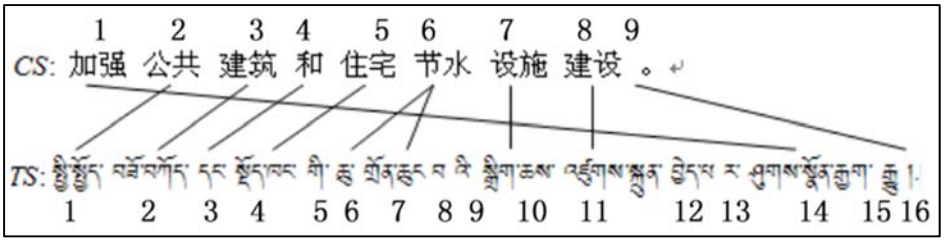

Figure 2. Word alignment result of an example sentence pair.

In the training procedure of statistical machine translation, the word pairs are extracted and the translation probabilities from each word in the source language to each of its correspondences in the target language are fitted, which results in the translation model that will be used in the translation decoding procedure.

We also use the Giza++ toolkit[3] (Och 2003), which is widely used in many machine translation systems, to enact the word alignment.

4.4 Bilingual phrase pair extraction

As mentioned above, we have a Chinese-Tibetan dictionary with 137,873 items. This scale is too small for many NLP purposes. In particular, we need many more word or phrase pairs to build the dictionary based machine translation sub system. As we have a bilingual sentence level parallel corpus, we can extract phrase pairs from it. Such a procedure has two stages.

4.4.1 Stage 1: extracting Chinese phrases

A phrase with more than one word is also called Multi Word Expression (MWE). A collocation measure is used to find the left and right boundaries of a Chinese phrase. For an adjacent pair of words (w1 ,w2), the collocation is defined by the following formula (Nuo 2011):

$$Collocation(w1, w2) = \frac{VMI(w1, w2)}{H(w1) + H(w2)}$$

where w1 and w2 represent the occurrence of two words, H(w1) is the entropy of the word w1, and *VMI(w1,w2)* is the average mutual information of the two words (Nuo 2011), defined as follows:

$$VMI(w1, w2) = P(w1, w2) log \frac{P(w1, w2)}{P(w1) * P(w2)} + P(\overrightarrow{w1}, \overrightarrow{w2}) log \frac{P(\overrightarrow{w1}, \overrightarrow{w2})}{P(\overrightarrow{w1}) * P(\overrightarrow{w2})} + P(\overrightarrow{w1}, w2) log \frac{P(\overrightarrow{w1}, w2)}{P(\overrightarrow{w1}) * P(w2)} + P(w1, \overrightarrow{w2}) log \frac{P(w1, \overrightarrow{w2})}{P(w1) * P(\overrightarrow{w2})}$$

where P(w1) is the occurrence probability of the word w1, P(w1,w2) is the occurrence probability of the adjacent pair (w1,w2), and $P(w1, \overline{w2})$ is the occurrence probability of adjacent pairs starting with w1 but followed by any word other than w2.

For the word sequence of w1w2w3, if we denote x=Collocation(w1,w2) and y= Collocation(w2,w3), then, the BindingDegree(x, y) is defined as follow (Nuo 2011):

$$BindingDegree(x, y) = \begin{cases} x/y, if\ y \geq x \\ y/x, if\ y < x \end{cases}$$

[3] http://www.statmt.org/moses/giza/GIZA++.html

So the Collocation(w1,w2) tells whether w1w2 forms a bi-word phrase, while the BindingDegree(x, y) tells whether w3 can be appended to the bi-word phrase w1w2 to form a tri-word phrase w1w2w3.

4.4.2 Stage 2: extracting the Tibetan correspondences for Chinese phrase

Generally, if a phrase occurs in the source language in every pair in a certain set of bilingual sentence pairs, its translation into the target language occurs also in the same set. When a Chinese phrase is extracted, we extract all the sentence pairs which it occurs in to form a candidate sentence pair set A. Then, words occurring in every Tibetan sentences in set A are extracted, which forms the Tibetan translation of the Chinese phrase.

4.5 Example based machine translation

When a Chinese sentence is to be translated, it is segmented into word sequences. If there is a similar sentence (or an identical sentence) in the bilingual sentence level parallel corpus, the Tibetan counterpart of the sentence pair can be taken as the translation of the Chinese sentence. This is the advantage of example based machine translation.

In the proposed system, for a Chinese sentence, all sentence pairs which have any word in common with in the Chinese sentence to be translated are extracted as the candidate set B. Then the Levenshtein Distances (Hirschberg 1975) (LD, word as the unit) between the Chinese sentence and all those in set B are computed. The similarity of two Chinese sentences CS_A and CS_B as word sequences is defined as follows:

$$\mathrm{Sim}(\mathrm{CS\_A}, \mathrm{CS\_B}) = 1 - \frac{\mathrm{LD}(\mathrm{CS\_A}, \mathrm{CS\_B})}{\max(\mathrm{len}(\mathrm{CS\_A}), \mathrm{len}(\mathrm{CS\_B}))}$$

The sentence with the maximum similarity to the input sentence is selected. If the similarity is larger than the predefined threshold (0.7 in our system, as determined by a test on some typical sentences), it is selected and the Tibetan counterpart in the sentence pair is taken as the translation of the input sentence.

4.6 Statistical machine translation

phrase	Candidates	Phrase	candidates
我们	ང་ཚོ་ས།	民族	མི་རིགས།
我们	ང་ཚོ་	民族	མི་རིགས།ལས།
我们 要	ང་ཚོ་ས།	民族 地区	མི་རིགས། ས།ཁུལ། གྱི།
我们 要	ང་ཚོ་ས། བྱ་དགོས།	民族 地区	མི་རིགས། ས།ཁུལ། དུ།
我们 要 不断	ང་ཚོ་ས། རྒྱུན་ཆད་མེད་པ་ར།	民族 地区 的	མི་རིགས། ས།ཁུལ། གྱི།
我们 要 不断	ང་ཚོ་ས། རྒྱུན་མེ་འཆད་པ་ར།	民族 地区 的	མི་རིགས། ས།གནས། ཀྱི།
要	བྱ་དགོས།	地区	ས།ཁུལ།
要	དགོས	地区	ས།ཁུལ།ལས། དང་
不断 加强	རྒྱད་རྒྱུ་ར། ཤུགས་སྤྲོན། ཟབ། མེ། ཚད་པ། བརྒྱབ། པ	干部	ལས།རྒྱད་པ།
不断 加强	ར། ཤུགས་སྤྲོན། ཟབ། མེ། ཚད་པ། བརྒྱབ། པ	干部	འགོ་ཁིད་ལས་རྒྱད་པ
加强	ཤུགས་སྤྲོན།	干部 队伍	ལས་རྒྱད་པ་ཞི། དཔུང་།ལས།
加强	རྒྱད་ཤུགས།	干部 队伍	ལས་རྒྱད་པ་ཞི། དཔུང་།སྡེ།
加强 民族	མི་རིགས་མཐུན་སྒྲིལ། ལ། ཤུགས་སྤྲོན།	干部 队伍 建设	ལས་རྒྱད་པ་ཞི། དཔུང་།སྡེ། འཛུགས་སྐྲུང་ རྒྱད་པ
加强 民族	མི་རིགས་མཐུན་སྒྲིལ། ལ། ཤུགས་སྤྲོན་རྒྱག། པ	干部 队伍 建设	ལས་རྒྱད་པ་ཞི། དཔུང་།ལས། ཞེས། སྐྲུང་

Table 6. Each phrase in the sentence has multiple translation candidates.

After word alignment, we can train a SMT model. The probability of any translation candidate for a phrase in the source language is stored in the model. The SMT decoder will make the decision to select the overall best candidate for each phrase in a sentence to be translated. We take an example to explain the procedure.

Sentence to be translated : 我们 要 不断 加强 民族 地区 的 干部 队伍 建设 。

English translation: We have to keep strengthening the construction of cadres in ethnic areas.

The decoder will find multiple candidates for nearly every phrase in the sentence. Table 6 shows some candidates for some phrases in the sentence to be translated.

Let's denote the Chinese sentence by f, and the Tibetan sentence by e, and denote the ith phrase in f by f_i, and the ith phrase in e by e_i. The probability of the Tibetan sentence is the translation of the Chinese sentence can be measured by the following formula:

$$P(\mathrm{e}|\mathrm{f}) = exp\left[\lambda_\phi \sum_{i=1}^{\mathrm{I}} log\phi(f_i|e_i) + \lambda_{LM} \sum_{i=1}^{|e|} \log p_{LM}(e_i \mid e_1 \cdots e_{i-1})\right]$$

Note the $\phi(f_i|e_i)$ is the translation probability to translate a Tibetan phrase to a Chinese phrase, and $P_{LM}(e_i \mid e_1 \cdots e_{i-1})$ is the probability of how likely the Tibetan phrases forms a Tibetan sentence, which is computed by the language modeling technology (Ponte 1998). $\lambda_\emptyset$ and λ_{LM} are the weights of the translation model and the language model respectively in the computation, and they are determined by training on the bilingual parallel corpus.

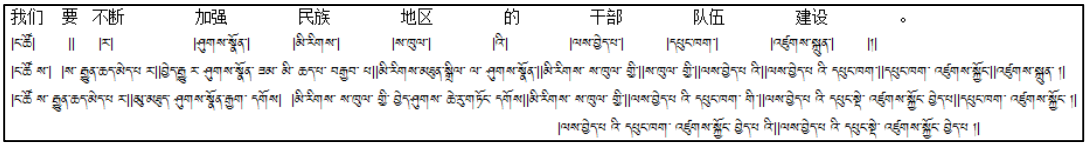

Figure 3. The candidates form a graph.

As each phrase has many translation candidates and the phases are overlapping, as shown in Table 6 and Figure 3. There are many candidate paths from the beginning to the ending of the decoding. In the proposed system, the decoder tries to select the best path which has the highest probability in terms of the translation model $\phi(f_i|e_i)$ and the language model $P_{LM}(e_i \mid e_1 \cdots e_{i-1})$. The corresponding Tibetan phrase sequence is taken as the output of the SMT subsystem.

For the example above, the SMT sub-system gives the translation: ང་ཚོ་ ས་ ཅེད་རྒྱུ་ ར་ ཤུགས་སྟོན་ ཟམ་མི་ཚད་ལྡན་ བཀྲལ་ ཨེ་རིགས་ ས་ཁུལ་ གི་ ལས་ཅེད་ཨེ་ དཔུང་ཁྲི་ འཕྲོགས་སྒོང་ ཅེད་ །

4.7 Rule based machine translation

In the rule (dictionary) based machine translation sub system, the bilingual dictionary is indexed by a structure call double array trie (Aoe 1989).

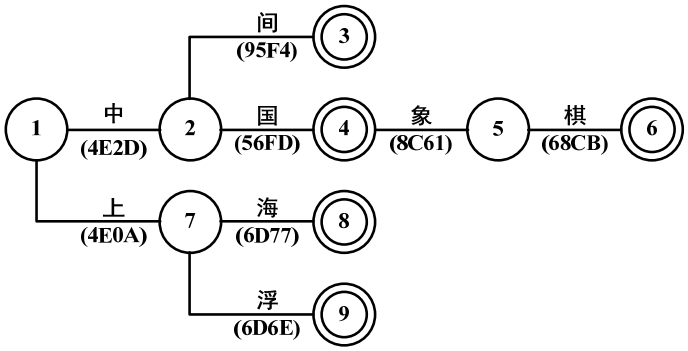

Figure 4. The trie of the key set *K*

We express the trie structure with an example. For a key set K={“中国”, “中国象棋”, “中间”, “上海”, “上浮”} in which the respective members mean “China”, “Chinese chess”, “middle”, “Shanghai” (a city), and “floating upward”, the trie is shown in

Figure 4. In the figure, a node with two circles indicating that it is a terminal node corresponding to an acceptance state, which means the end of a known word.

Previous research has found a method to store the trie in two arrays; this method makes better use of memory while keeping retrieval performance high. We can further decrease the memory to store the whole index of the bilingual dictionary by code splitting and code mapping.

Figure 5 shows the new trie of the key set K. Because some nodes are shared, just like the nodes 2 in

Figure 5 because “中”(4E 2D).and “上”(4E 0A) have a same high byte (4E), it is similar to node 13.

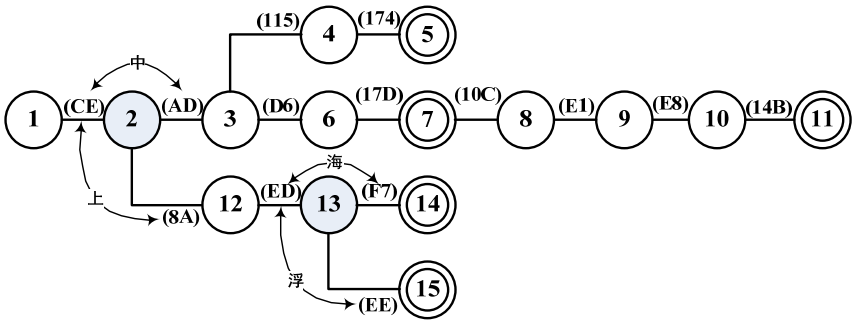

Figure 5. The new trie of key set *K*

In each terminal node, a pointer to the Tibetan translations of the Chinese word is indexed, which will be used to extract the corresponding translation. With this structure, the forward maximum matching method can be performed to translate a Chinese sentence to Tibetan. Backward maximum matching method can be performed similarly (Aoe 1989).

5 System Evaluation

In this section the performances of the different sub-systems are compared, and the well-known Moses[4] statistical machine translation system is also included in the comparison. 69,756 sentence pairs are included in the experiments, in which 429 pairs are randomly selected as the test set. The other pairs are included in the training set. The domain distributions of the training set and test set are listed in Table 7.

In the proposed system, two SMT sub systems are implemented, the first is maximum probability decoding model (MaxProb). The well-known Moses system is also taken as the second SMT sub system. Two RBMT sub systems are implemented too, which use Forward Maximum Matching (FMM) decoding and Backward Maximum Matching (BMM) decoding respectively.

The performances of different MT sub-systems are listed in Table 8. The performance data are collected on a PowerEdge R710 server with 2 4-core Intel E5620 (2.40 GHz) CPUs and 16 GB of memory. All sub systems (including Moses) are running on the same server. The two evaluation metrics BLEU4 and NIST are widely used in machine translation research (Papineni 2002; George 2002). BLEU4 simply calculates the geometric average of n-gram precisions for n=1 to 4 adding equal weight to each one, while NIST also calculates how informative a particular n-gram is. Obviously, Moses gives the best translations comparing with all sub-systems of the proposed system. However, it takes nearly 10 times more time. The EBMT sub-system gives the fewest correct translations. The other three sub systems have similar performance in both translation quality and time cost. However, 25,512 phrases are extracted by the SMT(MaxProb) sub system, in which 2258 long phrases are not extracted by Moses with default configuration. So at the phrase level, the recall of the SMT(MaxProb) sub system improves the results by 9.71% over Moses.

[4] http://www.statmt.org/moses/

	Domain	# of sentence pair
Training set	Law text	8,595
	Leader's book	25,112
	Government reports	34,610
Total of training set.		67,327
Test set	Law text	63
	Leader's book	151
	Government reports	215
Total of test set.		429
Total		69,756

Table 7. Domain distributions of the training set and test.

Table 9 shows the output of different sub-systems for the sentence mentioned in the former section. It shows that the EBMT sub system successfully find the sentence itself in the RBMT model because it's included in the training set. The other three sub-systems find translation candidates for every phrase in the sentence, but they are arranged in the same order as they are in the Chinese sentence rather than in a Tibetan sentence. That is why these sub-systems have a worse translation quality than Moses. So, an order adjusting model is essential to a Chinese to Tibetan translation system.

MT (sub) system	BLEU4	NIST	Time(s)
EBMT	0.0410	3.1472	41
SMT(MaxProb)	0.2168	5.6882	75
SMT(Moses)	0.2771	6.2338	633
RBMT(FMM)	0.2040	5.6235	68
RBMT(BMM)	0.1986	5.5944	67

Table 8. Performances of different MT sub systems.

Source	我们(we) 要(have to) 不断(keep) 加强(strengthening) 民族(ethnic) 地区(area) 的('s) 干部(cadres) 队伍(group) 建设(construction) 。
English	We have to keep strengthening the construction of cadres group in ethnic areas.
Reference	ང་ཚོ་ས་ མི་རིགས་ ས་ཁུལ་ ཀྱི་ ལས་བྱེད་པ་ འི་ དཔུང་ཁག་ འཛུགས་སྐྱོང་བྱེད་ཉམས་ ཆེ་རུ་ ཟན་མི་ཚད་པ་ གཏོང་དགོས།
EBMT	ང་ཚོ་ས་ མི་རིགས་ ས་ཁུལ་ ཀྱི་ ལས་བྱེད་པ་ འི་ དཔུང་ཁག་ འཛུགས་སྐྱོང་བྱེད་ཉམས་ ཆེ་རུ་ ཟན་མི་ཚད་པ་ གཏོང་དགོས།
SMT	ང་ཚོ་ས་ བྱེད་རྒྱ་ ར་ ཉམས་སྐྱོན་ ཟན་མི་ཚད་པ་ བཏུབ་པ་ མི་རིགས་ ས་ཁུལ་ ཀྱི་ ལས་བྱེད་པ་ འི་ དཔུང་ཁྲེ་ འཛུགས་སྐྱོང་ བྱེད་པ།
RBMT(FMM)	ང་ཚོ་ས་ རྒྱན་ཚད་ཞེད་པ་ ར་ མི་རིགས་ ས་ཁུལ་ ཀྱི་ བྱེད་ཉམས་ ཆེ་རུ་གཏོང་ དགོས་ ལས་བྱེད་པ་ འི་ དཔུང་ཁྲེ་ ལེགས་ སྐྱོང་ བྱ་རྒྱ་ འི།
RBMT(BMM)	ང་ཚོ་ས་ རྒྱ་མཐུད་ ཉམས་སྐྱོན་རྒྱ་ དགོས་ མི་རིགས་ གྲངས་ཤུང་ ས་ཁུལ་ ཀྱི་ ལས་བྱེད་པ་ འི་ དཔུང་ལག་ ལེགས་ སྐྱོང་ །

Table 9. A comparison of sub systems' outputs.

6 Conclusion

There is a big gap between the capacity for Chinese to Tibetan professional human translation and the market requirement. Machine translation technology is an urgent requirement, which contributes to the improvement of translation speed as well as the quality. We made great effort to build a Chinese to Tibetan machine translation system and discuss the key corpora and technologies in the paper. Experiments show the sub systems output the translation of each phrase in the same order as they are in the Chinese sentence rather than in a Tibetan sentence, which leads to a low quality translation. So an order adjusting model is essential to a Chinese to Tibetan translation system. As the translation quality is not good enough, we will make efforts toward order adjusting in the future. A computer aided tool is another requirement to make full use of the output of the system, and generates a better translation by interacting with the translator.

Acknowledgements

We thank the reviewers for their critical and constructive comments and suggestions that helped us improve the quality of the paper. The research is partially supported by National Science Foundation (No.61202219, No.61202220, and No.61303165) and Informationization Project of the Chinese Academy of Sciences (No.XXH12504-1-10).

References

Aoe, Junichi. 1989. "An efficient digital search algorithm by using a double-array structure". *IEEE Transactions on Software Engineering* 15.9, 1066-1077.

Cai, Rangjia. 2011. "Tibetan corpus processing method". *Computer Engineering and Application* 47.6: 138-139, 146. (才让加.藏语语料库加工方法研究[J].计算机工程与应用,2011,47(6):138-139,146.)

Cai, Zangtai; and Hua, Guanjia. 2005. "Research of Banzhida Chinese-Tibetan document translation system based on the dichotomy of syntax analysis". *Journal of Chinese Information Processing* 19.6: 9-14. (才藏太,华关加. 班智达汉藏公文翻译系统中基于二分法的句法分析方法研究 [J]. 中文信息学报, 2005, 19 (6): 9-14.)

Chang, Pi-Chuan; Galley, Michel; and Manning, Chris. 2008. "Optimizing Chinese word segmentation for machine translation performance". Paper presented at ACL 3th Workshop on Statistical Machine Translation.

Chen, Yuzhong; and Yu, Shiwen. 2003. "The current status and future of Tibetan information processing technologies". *China Tibetology* 4: 97-107. (陈玉忠,俞士汶. 藏文信息处理技术的研究现状与展望[J]. 中国藏学, 2003, (4): 97-107.)

Degai, Cailang; Li, Yanfu; and Hangqing Chaojia, et al. 2001. "The design and implementation of a practical Chinese-Tibetan machine translation system". In: *Proceedings of the 863 high-tech project "Intelligent Computer Workshop"*, 405-411. (德盖才郎,李延福,项青朝加等.实用化汉藏机器翻译系统的设计与实现[C].//863 计划智能计算机主题学术会议论文集.2001:405-411.)

Doddington, George. 2002. "Automatic evaluation of machine translation quality using n-gram Co-occurrence Statistics". In: *Proceedings of the Second International Conference on Human Language Technology Research,* 138-145. San Francisco: Morgan Kaufmann Publishers Inc.

Fredkin, Edward. 1960. "Trie memory". *Communications of the ACM*, 3.9: 490-500.

Hirschberg, D. S. 1975. "A linear space algorithm for computing maximal common subsequences". *Communications of the ACM*. 18.6: 341–343.

Hou, Hongxu; Liu, Qun; and Nasun Urt. 2007. "Example based Chinese-Mongolian machine translation". *Journal of Chinese Information Processing* 21.4: 65-72. (侯宏旭,刘群,那顺乌日图等. 基于实例的汉蒙机器翻译[J].中文信息学报,2007,21(4):65-72.)

Kang, Byeong-kwu; Zhang, Qinlong; Chen, Yirong; and Chang, Bao-bao. 2007. "Chinese multi-word chunks extraction for computer aided translation". *Journal of Chinese Information Processing* 21.1:9-16. (姜柄圭,张秦龙,谌贻荣等.面向机器辅助翻译的汉语语块自动抽取研究[J].中文信息学报, 2007,21 (1):9-16.)

Kanzhuo Caidan; Jin, Weixun; and Luo, Zhihua et al. 2006. "To search and solve problems about verbs in Chinese-Tibetan translation system". *Terminology Standardization and Information Technology* 3. (看卓才旦,金为勋,李延福等. 汉藏翻译系统中的动词处理研究 [J]. 术语标准化与信息技术, 2006, (3).)

Koehn, Philip; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Costantin, Alexandra; and Herbst, Evan. 2007. "Moses: Open source toolkit for statistical machine translation". In: *Proceedings of the ACL 2007* (demo and poster sessions) Prague, June 2007, 177–180.

Liu, Huidan; Long, Congjun; Nuo, Minghua; and Wu, Jian. 2015. "Tibetan word segmentation as sub-syllable tagging with syllable's Part-Of-Speech property". In: Maosong Sun; Zhiyuan Liu; Min Zhang; and Yang Liu (eds.), *Chinese computational linguistics and natural language processing based on naturally annotated big data* (LNAI 9427), 189-201.

Liu, Huidan; Nuo, Minghua; Ma, Longlong; Wu, Jian; and He, Yeping. 2011. "Tibetan word segmentation as syllable tagging using conditional random fields". In: *Proceedings of the 25th Pacific Asia conference on language, information and computation* (PACLIC-2011), 168-177.

Liu, Huidan, Minghua Nuo, Weina Zhao, Longlong Ma. 2012. "SegT: A pragmatic Tibetan word segmentation system". *Journal of Chinese Information Processing* 26.1: 97-103. (刘汇丹, 诺明花, 赵维纳, 吴健, 贺也平. SegT：一个实用的藏文分词系统[J]. 中文信息学报, 2012, 26(1): 97-103.)

Liu, Huidan; Zhao, Weina; Nuo, Minghua; Jiang, Li; Wu, Jian; and He, Yeping. 2010. "Tibetan number identification based on classification of number components in Tibetan word segmentation". Paper presented at International Conference on Computational Linguistics 2010, pp. 719-724.

Luo, Aijun; Gelang; Wujin Jiacan; and Ling, Yongxiao. 2010. "Investigation and analysis on the Chinese to Tibetan translator group in Tibet". *Tibet Technology* 5: 21-13. (罗爱军,格朗,伍金加参等. 西藏汉藏翻译队伍状况调查与分析 [J]. 西藏科技, 2010.)

Nuo, Minghua; Liu, Huidan; and Ma, Longlong et al. 2011. "Automatic acquisition of Chinese-Tibetan multi-word equivalent pair from bilingual corpora". In: *Proceedings of the International Conference on Asian Language Processing* (IALP2011), 177-180.

Nuo, Minghua; Wu, Jian; Liu, Huidan; and Ding, Zhiming. 2011. "Research on phrase translation extraction methods in the process of Chinese-Tibetan phrase pair extraction". *Journal of Chinese Information Processing* 25.3: 112-117. (诺明花, 吴健, 刘汇丹等. 汉藏短语对抽取中短语译文获取方法研究[J]. 中文信息学报,2011,25(3):112-117.)

Och, Franz Josef; and Ney, Hermann. 2003. "A systematic comparison of various statistical alignment models". *Computational Linguistics* 29.1: 19-51

Ponte, Jay M.; and Croft, W. Bruce. 1998. "A Language modeling approach to information retrieval". *SIGIR '98 Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 275-281. New York: ACM.

Papineni, Kishore; Roukos, Salim; Ward, Todd; and Zhu, Weijing. 2002. "BLEU: A method for automatic evaluation of machine translation". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 311-318.

The Ethnic Publishing House. 2002. *The antitheses Chinese-Tibetan dictionary*. Beijing: The ethnic publishing house. 2002.7. (民族出版社编, 汉藏对照词典. 民族出版社. 2002.7)

Yu, Daoquan. 1983. *The antitheses Chinese-Tibetan oral dictionary*. Beijing: The Ethnic Publishing House. (于道泉, 汉藏对照词典. 民族出版社. 1983.10)

Yu, Xin; Zhang, Liqiang; and Liu, Huidan. 2012. "Mechanism for Large-scale Chinese-Tibetan Bilingual Corpus Construction". In: *Proceedings of the 4th China National Youth Conference on Minority Language Information Processing*, 37-42.(于新, 张立强, 刘汇丹等. 大规模汉藏双语语料库构建机制[C].//第四届全国少数民族青年自然语言信息处理论文集.2012:37-42.)

Yu, Xin; Zhao, Weina; and Wu, Jian. 2010. "Dictionary-based Chinese-Tibetan sentence alignment". In: *Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Integrated Systems*.

Zhao, Weina; Liu, Huidan; Yu, Xin; Wu, Jian; and Zhang, Pu. 2010. "The construction of the Chinese-Tibetan bilingual parallel corpora for the Chinese-Tibetan bilingual computer-aided translation systems". In: *Proceedings of the 3rd China National Youth Conference on Minority Language Information Processing*, 43-46. (赵维纳,刘汇丹等. 面向汉藏辅助翻译系统的平行语料库建设. 第三届全国少数民族青年自然语言信息处理学术研讨会, 2010.)

Zhao, Weina; Yu, Xin; and Liu, Huidan. 2010. "Sentence boundary detection based on auxiliary verbs in Modern Tibetan". Paper presented at the Conference on Language Investigation and Information Processing (LIIP2010).

Zhang, Huaping; Liu, Qun; Cheng, Xueqi; Zhang, Hao; Yu, Hongkui. 2003. "Chinese lexical analysis using Hierarchical Hidden Markov Model". Paper presented at the Second SIGHAN Workshop Affiliated with 41st ACL; Sapporo Japan, July, pp. 63-70.

Zhaluo; Suonan Renqian. 2006. "Research on the translation rules for complex sentences in Chinese-Tibetan machine translation". In: *Chinese information processing Frontiers - Proceedings of the 25th Anniversary Conference of the Chinese Information Processing Society*. (扎洛,索南仁欠. 汉藏机器翻译中复句的翻译规则研究 [A]. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议[C]. 2006.)

Zhang, Yisun. 1993. *Tibetan Chinese general dictionary*. Beijing: The Ethnic Publishing House. 1993.12. (张怡荪编, 藏汉大辞典. 民族出版社.1993.12)

Huidan Liu
huidan@iscas.ac.cn

Weina Zhao
zhaoweina1999@yahoo.com.cn

Minghua Nuo
csnmh@imu.edu.cn

Jinling Hong
jinling@iscas.ac.cn

Xin Yu
yuxin08@iscas.ac.cn

Jian Wu
wujian@iscas.ac.cn