Tibetan Base Noun Phrase Identification Framework Based on Chinese-Tibetan Sentence Aligned Corpus

NUO Ming Hua[1,2] LIU Hui Dan[1,2] ZHAO Wei Na[1,3] MA Long Long[1] WU Jian[1] DING Zhi Ming[1]
(1) Institute of Software, Chinese Academy of Sciences, Beijing, China
(2) Graduate University of the Chinese Academy of Sciences, Beijing, China
(3) Qinghai Normal University, Xining, China
Minghua,huidan,weina,longlong,wujian,zhiming@iscas.ac.cn

ABSTRACT

This paper presents an identification framework for extracting Tibetan base noun phrase (NP). The framework includes two phases. In the first phase, Chinese base NPs are extracted from all Chinese sentences in the sentence aligned Chinese-Tibetan corpus using Stanford Chinese parser. In the second phase, the Tibetan translations of those Chinese NPs are identified using four different methods, that is, word alignment, iterative re-evaluation, dictionary and word alignment, and sequence intersection method. We implemented and tested these methods on Chinese-Tibetan sentence aligned unlabelled corpus without Tibetan POS tagger and Treebank. The experimental results demonstrate these methods can get satisfactory results, and the best performance with 0.5283 precision is got using sequence intersection identification method. The identification framework can also be extended to extract Tibetan verb phrase.

Title and abstract in Chinese

基于汉藏句子对齐语料的藏文BaseNP识别框架

本文提出藏文BaseNP识别框架，它分两步完成。先通过句法分析得到汉语BaseNP。再为这些汉语BaseNP从汉藏句子对齐语料中识别出藏文对应短语。我们应用四种方法识别藏文BaseNP，分别是词对齐、迭代重估算法、词典和词对齐相结合的方法以及基于序列相交的方法。评价实验表明，没有藏文词性标注和树库的前提下，基于序列相交的方法性能最好。本文提出的框架可以用于藏文动词短语识别任务中。

KEYWORDS : Tibetan information processing; base noun phrase; head-phrase;
CHINESE KEYWORDS :藏文信息处理,基本名词短语,中心语块

Proceedings of COLING 2012: Technical Papers, pages 2141–2158,
COLING 2012, Mumbai, December 2012.

1 Introduction

Shallow parsing identifies the non-recursive cores of various phrase types in text, possibly as a precursor to full parsing or information extraction (Abney, 1991). The paradigmatic shallow parsing problem is NP chunking, which finds the non-recursive cores of noun phrases called BaseNPs. It can help to solve many natural language processing tasks, such as information extraction, named entity extraction, machine translation, and text summarization and so on.

In general, researchers consider chunking as a kind of tagging problem or a sequence labelling task. Machine learning techniques are often applied to chunking. In Tibetan information processing, the shortage of Tibetan language resource leads to the fact that most of the techniques related text processing are still developing. Since 2003, research on Tibetan corpus, Tibetan word segmentation are reported. Although there is no public available Tibetan annotated corpus and Tibetan Treebank, we intend to extract Tibetan BaseNP using machine translation techniques based on our Chinese-Tibetan sentence aligned corpus. The research on Tibetan BaseNP is still in the initial stage. So far, there is no related report. In this paper, we will propose several methods for automatic identification of Tibetan base noun phrases.

The concept of BaseNP is initially put forward by (Church, 1998). In English, BaseNP is simple and non-recursive noun phrase which does not contain other noun phrase descendants. It cannot meet the needs in Tibetan information processing. Presently, different definitions of Chinese BaseNP are used on the basis of research field. According to our Chinese-Tibetan corpus, restrictive attribute phrase is in the scope of our BaseNP extraction. Observing that the Tibetan BaseNP is different from English, BaseNP in Tibetan can be recursively defined as follows, which is in accordance with the definition of Chinese BaseNP in (Zhao and Huang, 1998).

Definition 1: Tibetan base noun phrase (abbreviated as *BaseNP*)
BaseNP ::= *BaseNP* + *BaseNP*
BaseNP ::= *BaseNP* + *Noun*
BaseNP ::= *Determinative modifier* + *BaseNP*
BaseNP ::= *Determinative modifier* + *Noun*
Determinative modifier ::= *Adjective* | *Distinctive Adjectives (DA)* |*Nominalized Verb*| *Noun* |*Location* |*Numeral* + *Quantifier*

The Determinative modifiers have agglutinative relation with the heads.

The rest of this paper is organized as follows. Section 2 introduces related work of BaseNP chunking. Section 3 describes the outline of our framework. In Section 4, we propose four methods to automatically identify the Tibetan BaseNP which are very convenient to manual proofreading. In Section 5, we make an experiment to evaluate the four methods by three metrics, namely coverage, quasi-precision, and precision, then concludes this paper.

2 Related work

2.1 English BaseNP chunking

In 1991, Abney proposed to approach parsing by starting with finding correlated chunks of words (Abney, 1991). The pioneering work of Ramshaw and Marcus (1995) introduced NP chunking as a machine-learning problem, with standard datasets and evaluation metrics. Their work has inspired many others to study the application of learning methods to noun phrase chunking. Other chunk types have not received the same attention as NP chunks. At home and abroad, many statistical and machine learning methods are applied to the English BaseNP identification, and have achieved good recognition performance.

The task was extended to additional phrase types for the CoNLL-2000 shared task (Sang and Buchholz, 2000), which is the standard evaluation task for shallow parsing now. In this conference, many systems used the Machine learning methods, and among them, the most representative and effective one is Support Vector Machine (SVM) based method (Kudo and Matsumoto, 2000). Recently, some new statistical techniques, such as CRF (Lafferty et al. 2001), Winnow algorithm (Zhang, 2001) and structural learning methods (Ando and Zhang, 2005) have been applied to the BaseNP chunking task. Sha and Pereira (2003) considered chunking as a sequence labelling task and achieved good performance by an improved training method of CRF. Ando and Zhang (2005) presented a novel semi-supervised learning method on chunking and produced higher performance than the previous best results.

2.2 Chinese BaseNP chunking

Researchers apply similar methods of English BaseNP chunking to Chinese. Zhao and Huang (1998) made a strict definition of Chinese BaseNP in terms of combination of determinative modifier and head noun and put forward a quasi-dependency model to analyse the structure of Chinese BaseNP. There are some other methods to deal with Chinese phrase (not only BaseNP) chunking, such as HMM (Li et al., 2003), Maximum Entropy (Zhou et al., 2003), Memory-Based Learning (Zhang and Zhou, 2002) etc. Xu et al. (2006) propose a hybrid error-driven combination approach to chunking Chinese BaseNP, which combines TBL (Transformation-based Learning) model and CRF. In order to analyse the results respectively from the two (TBL-based and CRF-based) classifiers and improve the performance of the BaseNP chunker, an error-driven SVM based classifier is trained from the classification errors of the two classifiers. The hybrid method outperforms the previous works.

In general, the flexible structure of Chinese noun phrase often results in the ambiguities during the recognition procedure. Compared with English, internal grammatical structure of phrases is not rigorous; long noun phrase in Chinese is richer. Usage of Chinese word may serve with multi POS (Part-of-Speech) tags. Therefore, the chunker is puzzled by those multi-used words. Furthermore, there are no standard datasets and evaluation systems for Chinese BaseNP chunking as the CoNLL-2000 shared task, which makes it difficult to compare and evaluate different Chinese BaseNP chunking systems.

2.3 Tibetan chunking

In Tibetan information processing, the shortage of Tibetan language resource leads to the fact that most of the techniques related text processing are still developing. Recently, the focus of Tibetan information processing is gradually transferred from word processing to text processing. The Tibetan text processing started in the early 1990s, mainly analyse statically at the beginning. Since 2003, research on Tibetan syntactic chunks is reported.

Jiang (2003a) describes the basic types of syntactic chunks and their formal markers in modern Tibetan, and propose a scheme of automatic word-segmentation based on chunks according to the features of Tibetan syntactic structures. This paper finds the left and right boundaries of each chunk on the basis of pre-processing, setting up small tables of formal markers of each chunk, the verbal paradigm, special tables of homographs, etc, and goes on segmenting words with a dictionary and tagging within chunk. In (Jiang, 2003b), they discuss the automatic recognition strategies of nominalization markers in the modern Tibetan language. The purpose of identifying the nominalization markers is to make automatic word-segmentation within non-finite VP chunks. Due to the complexity of the formal features as well as distribution of the markers, the paper proposes the major recognition approach of the nominalization markers which distinguish between nominal markers and their homographic words. Huang et al. (2005) defines the nominal chunks of Tibetan according to the structures and syntax. Their identification strategy of nominal chunks depends mainly on the markers on the right boundary of the chunks. These previous works define the basic types of syntactic chunks and their formal markers. Identification of chunk is rule-based, including word order rule and syntactic rules of chunk. These papers just illustrate chunking result of several example sentences without experimental data.

The research on Tibetan BaseNP Chunking is, however, still at its initial stage. There is no public available Tibetan Treebank, even a POS tagger at present. In addition, there is no annotated Tibetan corpus available which contain specific information about dividing sentences into chunks of words of arbitrary types. Since we have large-scale Chinese-Tibetan sentence aligned corpus, public available Chinese parser, and word segmentation software etc., we can identify Tibetan BaseNP using these existing resources. Therefore, a Tibetan BaseNP identification framework based on Chinese-Tibetan sentence aligned corpus is proposed in the following.

3 Brief description of Tibetan BaseNP identification framework

The proposed Tibetan BaseNP identification framework consists of three main steps: pre-processing step, Chinese BaseNP extraction step, and Tibetan BaseNP identification step, which are in boldface in FIGURE 1(A). Chinese BaseNP extraction step and the Tibetan BaseNP identification step are the core of the identification framework.

In pre-processing step, sentence aligned Chinese and Tibetan corpus are word segmented and stored separately to the two documents, one sentence per line. Then words in sentence pairs are aligned using Giza++ toolbox (Och and Ney, 2003). FIGURE 1(B) shows the data flowchart of pre-processing.

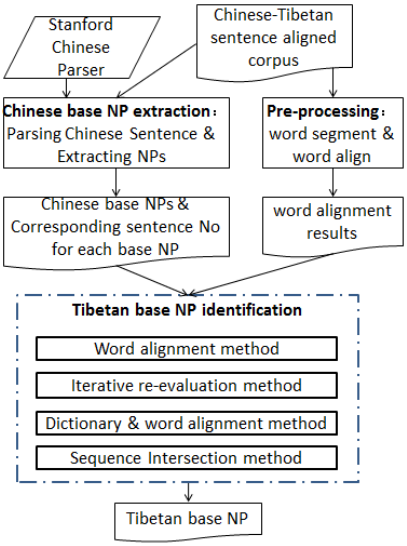

FIGURE 1 (A) – Flow chart of Tibetan BaseNP identification

(B) – Flow chart of pre-processing

In Chinese BaseNP extraction step, we use Stanford Chinese parser to parse all sentences in Chinese corpus from step 1; Extract all NP phrases from parsing results and note the sentence number in which there is a Chinese BaseNP.

The final step is Tibetan BaseNP identification. Aligned sentence pairs, their word alignment results and Chinese BaseNP extracted in step 2 is the input of Tibetan BaseNP identification. To determine the alignment for each Chinese BaseNP from step 2, different approaches within the dotted line in FIGURE 1(A) are proposed.

In the condition that there are no annotated Tibetan corpus, Treebank and Tibetan POS tagger, the framework regards the Tibetan BaseNP identification as translation problem, which confirms correct correspondence for those extracted Chinese BaseNP based on Chinese-Tibetan sentence aligned corpus. We assume that when a phrase in a source language is a BaseNP, its translation in target language is BaseNP too. The term "correspondence" is used here to signify a mapping between words in two aligned sentences. Specify that symbol '↔' is used to represent alignment. Any sentence-pair is symbolized by *SP*, and notated as $SP = CS \leftrightarrow TS$, where *CS* and *TS* stand for Chinese and Tibetan sentence respectively. A word sequence in *CS* is defined here as the correspondence of another sequence in *TS* if the words of one sequence are considered to represent the words in the other. In other words, Chinese-Tibetan BaseNP correspondence is an alignment at phrase-level.

Definition 2: Chinese-Tibetan BaseNP correspondence

$< C_{mi}, C_{mi+1}, \cdots, C_{mi+p} > \leftrightarrow < T_{nj}, T_{nj+1}, \cdots, T_{nj+q} >.$

On the left of alignment symbol is a Chinese BaseNP. Definition of Chinese BaseNP is the same to definition 1. On the right is its translation in Tibetan sentence of aligned sentence pair where Chinese BaseNP located, it conforms to definition 1 too. In this paper, Tibetan BaseNP consist of two or more words are taken into consideration, because of the task objectives.

In next section, we describe in detail how to identify Tibetan BaseNP correspondence to the extracted Chinese BaseNP. Different methods are evaluated, and we will select the method with best performance to generate referable Tibetan BaseNP, which is fully or partial correct, for further manual proofreading.

4 Tibetan BaseNP identification methods

Four different methods, that is, word alignment, iterative re-evaluation, dictionary and word alignment, and sequence intersection method are proposed to determine Tibetan correspondences for Chinese BaseNP. In the following, we elaborate the four methods.

4.1 Word alignment method

This subsection presents word alignment (WA, hereafter) method. In this method, phrase-level alignments are obtained on the basis of the optimal two-way word alignment results from Giza++. Outline of WA method is given below.

Step 1: Run Giza++ to get word alignment results of aligned sentence pair in Chinese-Tibetan corpus.

Step 2: For each Chinese BaseNP, get the word alignment results of located sentence pair and corresponding Tibetan sentence based on the sentence number it located.

Step 3: For each word in Chinese BaseNP, obtain the aligned Tibetan word according to word alignments using a heuristic. These Tibetan words constitute a Tibetan BaseNP for current Chinese BaseNP.

A number of different word alignment heuristics are implemented; in the end, grow-diag-final heuristic shows better performance than others on our aligned corpus. Consequently WA method with grow-diag-final heuristic is regarded as a baseline method.

4.2 Iterative re-evaluation method

This subsection describes iterative re-evaluation (IRE, hereafter) method, which is based on correlations of Chinese phrase and Tibetan words, to complete the identification of Tibetan BaseNP. It is an instance of a general approach to statistical estimation, represented by the EM algorithm (Dempster et al., 1977). Iterative re-evaluation algorithm assumes that a Chinese phrase has a corresponding probability with every Tibetan word; we call it relevancy (R). First, we assign an initial value to R, and then iteratively update the value of R based on correlations between Tibetan word

and Chinese phrase. If the absolute value of the difference between the latest updated two R value, the iterative process stops. Eventually we obtain satisfactory correlations of Chinese phrase and Tibetan words. We will describe the details of iterative re-evaluation algorithm in this section.

Relevancy of Chinese phrase and Tibetan words are used to identify the potential Tibetan BaseNP translation for each Chinese phrase. Tibetan words with higher relevancy should be within the translation of Chinese phrase. If we denote the relevancy of a Chinese phrase *c* and a Tibetan word *t* by $R(c,t)$, then we can calculate it with the following formula.

$$R(c,t) = \frac{W(c,t)}{\sum_{q \in V_F} W(c,q)} \qquad \text{satisfied} \quad \sum_{q \in V_F} R(c,q) = 1 \tag{4.1}$$

In formula(4.1), $W(c,t)$ represents weighted frequency (product of co-occurrence frequency and relevancy) of *c* and *t*. Every Chinese phrase *c* and every Tibetan word *t* has a weighted frequency. V_F indicates words set of Tibetan corpus.

Weighted frequency of Chinese phrase and Tibetan word is the key to the calculation of relevancy. Weighted frequency of *c* and *t* is defined as follow:

$$W(c,t) = \sum_{i=1}^{N} F(i,c,t) R(c,t) \tag{4.2}$$

In which *N* represents the number of sentence pairs in Chinese-Tibetan corpus. $F(i,c,t)$ indicates the number of simultaneous occurrence of *c* and *t* in i^{th} sentence pair. Equation (4.3) assumes that each Chinese BaseNP is initially equally likely to correspond to each Tibetan BaseNP. The weights $W_0(c,t)$ can be interpreted as the mean number of times that *c* corresponds to *t* given the corpus and the initial assumption of equivalent correspondences.

$$W_0(c,t) = \sum_{i=1}^{N} F(i,c,t) \frac{1}{\phi(i)} \tag{4.3}$$

Where $\phi(i)$ indicates the number of Tibetan words in i^{th} sentence pair.

Let *r* be the number of iterations, $P(c,t)$ and $W(c,t)$ for iteration *r* is formulated as in formula (4.4) and (4.5).

$$R_r(c,t) = \frac{W_{r-1}(c,t)}{\sum_{q \in V_F} W_{r-1}(c,q)} \tag{4.4}$$

$$W_r(c,t) = \sum_{i=1}^{N} F(i,c,t) R_{r-1}(c,t) \tag{4.5}$$

The procedure is then iterated using Equations (4.4) and (4.5) to obtain successively refined, convergent estimates of the probability that *c* corresponds to *t*. Experiment results shows that it works well when the iterative threshold is 0.001. It means that we can find Tibetan word *t* with highest corresponding probability to each Chinese phrase,

after several iterations. If $|R_r(c,t)-R_{r-1}(c,t)|<0.001$ then stop iterating.

IRE method can globally calculate the corresponding information, while Mutual Information is used to measure the relevance between Chinese phrase and Tibetan word in isolation without the information of other Tibetan words. For instance, if a few Tibetan words are correct translation for a certain Chinese phrase in a sentence pair, and its located sentence is long, it will lead to lower initial relevancy. However, after iteration, the weighted frequency will be increased due to the high co-occurrence frequency; and the proportion become greater in the sum of weighted frequency of all words, which makes the relevancy increased. Meanwhile, the proportion of weighted frequency of other words will decrease. In other words, relevancy of the error Tibetan correspondence will decrease. After each iteration, the difference between the correct and error Tibetan correspondence words gets bigger and bigger. This is the reason why we use IRE method.

4.3 Dictionary and word alignment method

To increase the overall performance of identification of Tibetan BaseNP, Dictionary and word alignment based (D&WA, hereafter) method are presented. Zhang et al. (2006) proposed a phrase alignment method. Their work obtains the translation head-phrase according to dictionary-based word alignment，and statistical translation boundary is determined based on the translation extending confidence.

Inspired by this research, we use the basic idea of head-phrase extension. The key problem is how to get the head-phrase and how to extend it to a correct Tibetan BaseNP. In other words, we decompose the identification of Tibetan BaseNP into head-phrase extraction and head-phrase extension steps. D&WA method use different way from (Zhang et al., 2006) in both steps.

4.3.1 Tibetan head-phrase extraction

Bilingual dictionary with 135,000 word pairs is used as an additional resource. Dictionary provides reliable alignments, but its coverage rate is low. Hence, we have modified the head-phrase extraction step in (Zhang et al., 2006). Our head-phrase extraction step is no longer solely dependent on Chinese-Tibetan dictionary. When there are no corresponding entries in bilingual dictionary for a Chinese word, we will use the word alignment result from intersect heuristic for current word, because intersect heuristic can achieve higher precision without any interference. Description of modified head-phrase extraction is as follows.

Firstly, for each word in a Chinese BaseNP, we search the bilingual dictionary and get the translation words list (TWL).

- If TWL is not null, judge whether one or more member of TWL occur in the corresponding Tibetan sentence of Chinese BaseNP, and mark the positions in Tibetan sentence.
- If TWL is null, directly use the word alignment correspondence from intersect heuristic, and mark the position in Tibetan sentence.

Then, continuous words between the most left and most right position from previous step constitute the head-phrase.

4.3.2 Tibetan head-phrase extension

The next step is the determination process of statistical translation boundary called head-phrase extension. Unlike (Zhang et al., 2006), D&WA method use commonly used Mutual information (MI) and t-value to determine left and right boundary of Tibetan BaseNP in the extension step. The formula is as follows:

$$MI(c,t) = \log \frac{P_r(c,t)}{P_r(c) \times P_r(t)} \tag{4.6}$$

$$t(c,t) \approx \frac{P_r(c,t) - P_r(c) \times P_r(t)}{\sqrt{\frac{1}{N} P_r(c,t)}} \tag{4.7}$$

Where *N* indicates the total number of sentences in bilingual corpus; *c* indicates Chinese phrase, *t* indicates Tibetan word; $P_r(c,t)$ denotes the co-occurrence probability of *c* and *t*. $P_r(c)$ and $P_r(t)$ denotes the occurrence probability of *c* and *t* respectively. For each Chinese BaseNP, we calculate MI and t-value between Tibetan words in corresponding sentence and reserve these values.

Suppose T is full or partial correspondence of Chinese BaseNP, it can symbolized as formula(4.8).

$$T = w_1 w_2 \cdots w_i \cdots w_n \tag{4.8}$$

Average Mutual Information and Average T-score between Chinese BaseNP c and T are based on formula(4.9) and formula(4.10).

$$AMI(c,T) = \frac{1}{n} \sum_{i=1}^{n} MI(c,W_i) \tag{4.9}$$

$$AT(c,T) = \frac{1}{n} \sum_{i=1}^{n} t(c,W_i) \tag{4.10}$$

Definition 4: head-phrase extension confidence.

For Chinese phrase Ph_C, the head-phrase of its translation in Tibetan sentence is denoted by $Ph_T(n)$, where *n* indicates the length of head-phrase; Extend to an adjacent Tibetan word of $Ph_T(n)$ and get $Ph_T(n+1)$, so the head-phrase extension confidence C_n defined as:

$$\begin{aligned} C_n = {} & \lambda_1 \left| AMI[Ph_C(n), Ph_T(n)] - AMI[Ph_C(n), Ph_T(n+1)] \right| \\ & + \lambda_2 \left| AT[Ph_C(n), Ph_T(n)] - AT[Ph_C(n), Ph_T(n+1)] \right| \end{aligned} \tag{4.11}$$

In which AMI and AT indicates the mean of *MI* and t-value in the scope of extended Tibetan BaseNP respectively.

In the extension step, word by word calculation of extension confidence in Tibetan sentence will be held, to both sides of head-phrase. For each extension-ready Tibetan word, note the head-phrase extension confidence C_n; if C_n is greater than the threshold, current Tibetan word is accepted as a member of Tibetan BaseNP, and extension continues; when C_n is less than the threshold extension stops. Statistical translation boundary for Chinese phrase Ph_C is obtained at the end of extension under head-phrase. FIGURE 2 shows the extension process in detail.

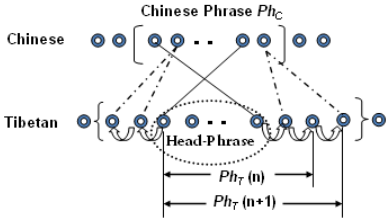

FIGURE 2 –Concept of head-phrase extension

In FIGURE 2, Chinese phrase Ph_C is in brackets; its final translation is in brace in Tibetan sentence. The extended $Ph_T(n+\omega), 0 \le \omega \le L-n$ is Tibetan BaseNP.

4.4 Sequence intersection identification method

Sequence intersection identification (SII, hereafter) method we proposed in this subsection is in accordance with the characteristics of Tibetan BaseNP. It uses intersection operation between Tibetan sentences and head-phrase extension strategy.

Analysis to the structure of Tibetan BaseNP indicates that, context based free translation style causes one Chinese phrase has different correct translation in Chinese-Tibetan sentence aligned corpus. Let's analyse the structure of following Example. "进出口货物收发货人" is "Shipper & Consignee as Declarant" in English. Three versions of Tibetan translation of "进出口货物收发货人" in Chinese-Tibetan sentence aligned corpus are given in TS1, TS2 and TS3.

TS1: ཕྱིར་གཏོང་ནང་འདེན་བྱ་རྒྱུའི་ཚོང་ཞོན་གཏོང་ལེན་བྱེད་མཁན

TS2: ཕྱིར་གཏོང་ནང་འདེན་བྱ་རྒྱུའི་ཚོང་ཞོན་ཐོབ་སྟོངས་ལེན་བྱེད་མཁན

TS3: ཕྱིར་གཏོང་ནང་འདེན་ཚོང་ཞོན་ཐོབ་སྟོངས་ལེན་མི་སྣ

During the intersection operation, "ཕྱིར་གཏོང་ནང་འདེན་" and "ཚོང་ཞོན་" is common string of TS1, TS2 and TS3. We can get "ཕྱིར་གཏོང་ནང་འདེན་བྱ་རྒྱུའི་ཚོང་ཞོན་" or "ཕྱིར་གཏོང་ནང་འདེན་ཚོང་ཞོན་" from different Tibetan sentences as head-phrase. Then we use head-phrase extension strategy to Tibetan sentence of sentence pair where "进出口货物收发货人" located and get TS1, TS2 or TS3. These different translations co-occur in our bilingual corpus. Moreover, the case is more common.

The extension step is the same as that we described in section4.3.2. The identification of Tibetan head-phrase in SII method is presented in this subsection.

4.4.1 Definition of sentence sequence intersection

Chinese-Tibetan bilingual corpus *CTBC* is composed of numerous aligned sentence pairs. Any sentence pair is notated as $SP = CS \leftrightarrow TS$, where *CS* and *TS* represent Chinese and Tibetan sentence respectively. Formula (4.12) and (4.13) give the expression of Chinese and Tibetan sentence as a word sequence.

$$CS = < C_1, C_2, \cdots, C_n > \tag{4.12}$$

$$TS = < T_1, T_2, \cdots, T_m > \tag{4.13}$$

Thus, SP can be expressed as words sequence form in formula (4.14) as below:

$$SP = CS \leftrightarrow TS = < C_1, C_2, \cdots, C_n > \leftrightarrow < T_1, T_2, \cdots, T_m > \tag{4.14}$$

Then, let's define sentence sequence intersection. Set $SP_r, SP_t \in CTBC$ are any two aligned sentence pairs. Representation of SP_r and SP_t in word sequence form is in formula (4.15) and (4.16).

$$SP_r = CS_r \leftrightarrow TS_r = < C_r, C_{r+1}, \cdots, C_{r+nr} > \leftrightarrow < T_r, T_{r+1}, \cdots, T_{r+mr} > \tag{4.15}$$

$$SP_t = CS_t \leftrightarrow TS_t = < C_t, C_{t+1}, \cdots, C_{t+nt} > \leftrightarrow < T_t, T_{t+1}, \cdots, T_{t+mt} > \tag{4.16}$$

Definition 3: Intersection of TS_r and TS_t

$$\begin{aligned} TS_r \cap TS_t = \{ & < T_{r+r1}, T_{r+r2}, \cdots, T_{r+rq} > = < T_{t+t1}, T_{t+t2}, \cdots, T_{t+tp} > \\ & | 0 \le r_1 < r_2 < \cdots < r_q \le n_r, 0 \le t_1 < t_2 < \cdots < t_p \le m_t \} \end{aligned} \tag{4.17}$$

In formula(4.17), the result of $TS_r \cap TS_t$ is a set of common substring of TS_r and TS_t. New subscripts $r_1, r_2 \cdots\cdots r_q$ and $t_1, t_2 \cdots\cdots t_p$ monotonously increase, they must be within the scope of original subscript n_r and m_t.

4.4.2 Identification of Tibetan BaseNP

SII method uses head-phrase extension like in the D&WA method. In head-phrase extraction step, SII method uses the idea of sentence sequence intersection, which is different from D&WA method.

Intuitively, if a Chinese BaseNP occurs in more than one sentence, denotes S_C, its Tibetan correspondence must occur in the aligned Tibetan sentences of S_C, denotes S_T; and sentences in S_T must have common substrings, denotes *TCS*, which is a set of multiword units including full or partial Tibetan correspondence. After sentence intersection, it is likely to get part of Tibetan BaseNP correspondence in terms of Tibetan tense, verb-endings, auxiliary word etc. Hence, one of the intersection parts must be regarded as head-phrase. It is to say, the preferred way to obtain translation of a Chinese BaseNP Q_i is searching for common substring of Tibetan sentences.

From above analysis, another form of sequence intersection for formula (4.17) is in formula(4.18).

$$TS_r \cap TS_t = T = \{T_1, T_2, \cdots, T_g\} \tag{4.18}$$

In(4.18), T is including T_j which is full or partial correspondence of Chinese BaseNP. Among these multiword units in T, we use selection function Ψ_j to determine the candidate. Commonly used mutual information (MI) and t-value statistical information are used to determine the candidate. Definition of Ψ_j for $T_j\ (1 \le j \le g)$ is given in formula(4.19).

$$\Psi_j = \lambda_1 \cdot AMI(Q_i, T_j) + \lambda_2 \cdot AT(Q_i, T_j) \tag{4.19}$$

Where *AMI* and *AT* are the means of *MI* and t-value between Chinese BaseNP Q_i and T_j respectively. $T_j\ (1 \le j \le g)$ with the highest Ψ_j is head-phrase of candidate Tibetan BaseNP. In the end, the Tibetan head-phrase is extended to BaseNP using the same extension process described in section4.3.2.

5 Experiments

5.1 Experimental corpus

The corpus used in this experiments is a domain-specific Chinese-Tibetan bilingual corpus in laws, regulations and official documents, which is the input of the Tibetan BaseNP identification. The original corpus is used to Chinese BaseNP extraction; we call it corpus1. It consists of 256,880 bilingual aligned sentence pairs including both long and short sentences. The size of Chinese corpus is 18,244 kilobytes; and that of Tibetan corpus is 58,650 kilobytes. We generate test corpus in TABLE 1 under random selection, to evaluate the proposed methods. TABLE 1 shows the basic information about the corpora. In TABLE 1, CS denotes Chinese sentence, TS denotes Tibetan sentence and SP denotes sentence pair.

First of all, Chinese sentences in corpus1 are parsed by Stanford parser, and NPs are extracted. The number of Chinese BaseNP in corpus1 is 422,146 without any pre-processing. After duplicate removal, it decreases to 255,249 which is shown in the last column of TABLE 1, where CBNP denotes Chinese BaseNP. The size of Chinese BaseNP from corpus1 is too large. In order to quantify the result, Tibetan BaseNP identification is tested on the small size of test corpus, because we need the manual reference for those Chinese BaseNPs at present. The number of Chinese BaseNP in test corpus is 394 before pre-processing. We take no account of one word NP. After filtration of one word NP and deletion of error parsed NP, we get 212 Chinese BaseNP from the test corpus.

Corpora	CS(KB)	TS(KB)	Number of SP	Number of CBNP
corpus1	18,244	58,650	256,880	255,249
Test corpus	37	149	378	212

TABLE 1 – Information about corpora

5.2 Evaluation

We define the Coverage, quasi-precision and precision to evaluate the experimental results.

$$Coverage = \frac{N_1}{N} \times 100\% \tag{5.1}$$

$$Quasi - Precision = \frac{N_2 + N_3}{N_1} \times 100\% \tag{5.2}$$

$$Precision = \frac{N_3}{N_1} \times 100\% \tag{5.3}$$

Where, N denotes the number of Chinese BaseNP for test. N_1 denotes the total number of Chinese BaseNP for which we obtain its correspondence. N_2 denotes the number of Chinese BaseNP which obtained partial correct correspondence . N_3 denotes the number of Chinese BaseNP which obtained full correct correspondence . We ask Tibetan scholar to provide us reference for Chinese BaseNP in size 212 for test, then automatically judge N_1, N_2, and N_3.

5.2.1 Different word alignment heuristics for WA method

A number of different word alignment heuristics are used in WA method. The options are:

- intersect
- union
- grow-diag-final-and
- grow-diag-final
- grow-diag

Different heuristic may show better performance for a specific language pair or corpus, the experimental results of Tibetan BaseNPs are shown in TABLE 2.

heuristic	Coverage	Quasi-precision	Precision
intersect	0.9292	0.7716	0.4975
union	1.0	0.7972	0.5165
grow-diag-final-and	0.9811	0.7692	0.5096
grow-diag-final	1.0	0.7972	0.5212
grow-diag	1.0	0.7972	0.5142

TABLE 2 –Results of WA method using different word alignment heuristics

TABLE 2 shows that, from the overall perspective, grow-diag-final heuristic outperforms others, and intersect heuristic gets the lowest performance.

On analysis, the reason for lower precision is as follows.

- Some of determinative modifier in Chinese noun phrases is verb in test set. The target language is Tibetan, which is morphologically rich language with ample variations in terms of tense, verb-endings, auxiliary word etc. These lead to discontinuous Tibetan translation.
- Chinese and Tibetan word segmentation is in different granularity.
- The quasi-precision of five heuristic is more even, however, there are some boundary interference like stop words in word alignment result all but intersect heuristic.
- Giza++is a tool based on statistics; therefore, it does not work so well on low frequency phrases.

The finding by analysis is that the word alignment result from heuristics except intersect heuristic interfered by tense, verb-endings, auxiliary word or stop word at the boundary. Yet the results from intersect heuristic consist of one or more Tibetan words without any interference. Hence we select the word alignment results of intersect heuristic in D&WA method to supplement to bilingual dictionary. It is proved that modification to automatic Tibetan BaseNP candidate under partial correspondence turns out to be effective than pure manual translation. Consequently, the best overall performance will be regarded as baseline method in next subsection.

5.2.2 Different methods for Tibetan BaseNP identification

The motivation of this paper is produce referable Tibetan BaseNP with best overall performance. To compare the four proposed Tibetan BaseNP identification method, WA methods with grow-diag-final heuristic is selected as baseline. TABLE 3 shows the results of four methods which are proposed in this paper for Tibetan BaseNP identification.

Methods	Coverage	Quasi-precision	Precision
WA method(Baseline)	1.0	0.7972	0.2453
IRE method	0.9764	0.8261	0.4203
D&WA method	0.9670	0.8732	0.4976
SII method	1.0	0.8821	0.5283

TABLE 3 –Results of different methods

The precision of IRE method is higher than baseline, because IRE method is able to filter some interference. However, precision of IRE method is influenced by some of the high frequency words and different granularity of Chinese and Tibetan words segmentation. Its coverage is medium due to correct correspondences to the low

frequency phrases. In D&WA method, Chinese-Tibetan word dictionary as auxiliary resource significantly increases the precision. Combination of dictionary and intersect heuristic word alignment improves the coverage of D&WA method. In SII method, we use intersection of Tibetan sentences to improve the coverage and quasi-precision. During intersection, some interference on boundary like verb-endings, auxiliary word are filtered; meanwhile, the step for head-phrase identification does not rely on statistics, so it works even on low frequency phrases. The overall result of SII method outperforms other proposed methods. Obviously, our test corpus is in small size, we are working on the further verification of the framework on large-scale Tibetan BaseNP identification.

Conclusion and perspectives

We are in the initial stage of identification of Tibetan base noun phrase. At present, you know, Tibetan POS tagger, Tibetan Treebank or annotated corpus is not available. On the basis of the existing resources of our group, we take the BaseNP identification as a translation problem. Four methods, namely word alignment, iterative re-evaluation, dictionary and word alignment, and sequence intersection method are applied to identify Tibetan BaseNP. We define the Coverage, quasi-precision and precision as metrics to evaluate the experimental results. As a result, the best one (SII method) achieves 1.0, 0.8821 and 0.5283 respectively on the test corpus. Compared with English or Chinese BaseNP identification work, the proposed methods doesn't get the best score, but the approach is very novel to Tibetan BaseNP identification. Due to the lack of resources like POS tagger and previous technology, the result is acceptable.

In the future, on one hand, we will improve the coverage to identify more potential BaseNP. Methods proposed in this paper need further validation in large-scale corpus. On the other hand, we will make more research on the Tibetan BaseNP templates using grammatical rules to produce a high quality results. It means that Tibetan parts-of-speech tagging is one of our future direction too.

Acknowledgments

We are grateful to three anonymous reviewers for their insightful comments that helped us improve the quality of the paper. We are thankful for Guancho Dorjie providing reference of Tibetan BaseNP in experiment part. The research is partially supported by National Science and Technology Major Project (No.2010ZX01036-001-002, No.2010ZX01037-001-002), National Science Foundation (No.61202219, No.61202220), Major Science and Technology Projects in Press and Publishing (No.0610-1041BJNF2328/23, No.0610-1041BJNF2328/26), and CAS Action Plan for the Development of Western China (No.KGCX2-YW-512).

References

Abney, S.P. (1991). Parsing By Chunks. In (Robert Berwick, Steven Abney and Carol Tenny, 1991) *Principle-Based Parsing*, Kluwer Academic Publishers.

Ando, R.K. and Zhang, T. (2005). A High-Performance Semi-Supervised Learning Method for Text Chunking. Kevin Knight. In *Proceedings of the 43rd Annual Meeting of ACL*, pages 1-9. Ann Arbor, Michigan.

Church, K.W. (1998). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of 2nd Conference on Applied Natural Language* Processing (ANLC '88), pages 136-143.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39(1):1-38.

Fei, S. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In (*Eduard Hovy, 2003*) *Proceedings of HLT-NAACL*, pages 134-141. Edmonton, Alberta.

Huang, X., Sun, H.K., Jiang, D., Zhang, J.C., and Tang, L.M. 2005. The Types and Formal Markers of Nominal Chunks in Contemporary Tibetan. In *proceedings of the 8th Joint Conference on computational linguistics* (JSCL'2005).

Jiang, D. (2003a). On syntactic chunks and formal markers of Tibetan. *Minority Languages of China*,(3):30-39.

Jiang, D. and Long, C.J. (2003b). The Markers of Non-finite VP of Tibetan and its Automatic Recognizing Strategies. In *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages* (ICCPOL'2003).

Kudo, T. and Matsumoto, Y. (2001). Chunking with support vector machine. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*(NAACL '01),pages 1-8.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, pages 282-289.

Li, H., Webster, J.J., Kit, C.Y., and Yao, T.S. (2003). Transductive HMM based Chinese Text Chunking. *IEEE NLP-KE 2003*, pages 257-262. Beijing.

Liu, D.M. (2004). Chinese-English bilingual parallel corpus alignment method. Graduate dissertation. China, Shanxi University.

Och, F.J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1): 19-51.

Ramshaw, L.A. and Marcus, M.P. (1995). Text Chunking using Transformation-Based Learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge MA, USA.

Sang, E.F.T.K. and Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task: Chunking. In *Proc. CoNLL-2000*, pages 127-132.

Xu, F., Zong, C.Q., and Zhao, J. (2006). A Hybrid Approach to Chinese Base Noun Phrase Chunking. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 87–93, Sydney.

Zhang, C.X., Li, S., and Zhao, T.J. (2006). Phrase Alignment Based on Head-Phrase Extending. *Journal of Computer Research and Development*, 43(9):1658-1665.

Zhang, T., Damerau, F., and Johnson, D. (2001).Text Chunking using Regularized Winnow. In *Proceedings of the 39rd Annual Meeting of ACL* (ACL '01), pages 539-546. Morgan Kaufmann Publishers.

Zhang, Y.Q. and Zhou, Q. (2002). Chinese Base-Phrases Chunking. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, (SIGHAN '02), pages 1-5.

Zhao, J. and Huang, C.L. (1998). A Quasi-Dependency Model for Structural Analysis of Chinese BaseNPs. In *36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*, pages 1-7.

Zhao, J. and Huang, C.N. (1999), The model for Chinese baseNP structure analysis, *Chinese Journal of Computers*, 22(2):141-146.

Zhou, Y.Q., Guo, Y.K., Huang, X.L., and Wu, L.D. (2003). Chinese and English Base NP Recognition on a Maximum Entropy Model. *Journal of Computer Research and Development*. 140(13):440-446.