Tibetan Syllable-based Functional Chunk Boundary Identification

Shumin Shi[1, 2*], Yujian Liu[1], Tianhang Wang[1], Congjun Long[3], Heyan Huang[1, 2]

[1] School of Computer Science & Technology Beijing Institute of Technology, Beijing 100081, China
[2] Beijing Engineering Research Center of High Volume Language Information Processing & Cloud Computing Applications, Beijing 100081, China
[3] Institute of Ethnology & Anthropology Chinese Academy of Social Sciences, Beijing 100081, China
{bjssm,yjliu,hbcdwth,hy63}@bit.edu.cn,
longcj@cass.org.cn

Abstract. Tibetan syntactic functional chunk parsing is aimed at identifying syntactic constituents of Tibetan sentences. In this paper, based on the Tibetan syntactic functional chunk description system, we propose a method which puts syllables in groups instead of word segmentation and tagging and use the Conditional Random Fields (CRFs) to identify the functional chunk boundary of a sentence. According to the actual characteristics of the Tibetan language, we firstly identify and extract the syntactic markers as identification characteristics of syntactic functional chunk boundary in the text preprocessing stage, while the syntactic markers are composed of the sticky written form and the non-sticky written form. Afterwards we identify the syntactic functional chunk boundary using CRF. Experiments have been performed on a Tibetan language corpus containing 46783 syllables and the precision, recall rate and F value respectively achieves 75.70%, 82.54% and 79.12%. The experiment results show that the proposed method is effective when applied to a small-scale unlabeled corpus and can provide foundational support for many natural language processing applications such as machine translation.

Keywords: Tibetan Syntactic Functional Chunk, Chunk Boundary Recognition, Syllable, Syntactic Marker, CRF.

1 Introduction

Syntactic chunk parsing plays an important role in chunk parsing. It aims to annotate the essential syntactic constituents of a sentence through top-down splitting. Then it can obtain the basic structure information units of the sentence. In Machine Translation, syntactic parsing is able to reduce the difficulty of recombining the sentence of over segmentation and reduce the disorder in the target-language generation. And also

[*] corresponding author

the recognition of syntactic functional chunks on the basis of the sentence preprocessing plays a very important role in rule based Machine Translation.

There are rich research results in English and Chinese that can provide good references for the syntactic functional chunks parsing of Tibetan Language. However, in the past, the research in this area must be based on the word segmentation and part of speech (POS) tagging. On the one hand, the errors of word segmentation and POS tagging will directly affect the correctness of syntactic functional chunk parsing. On the other hand, it increases the time and space cost. From the point of view of the current study on Tibetan, the accuracy of word segmentation and tagging remains to be improved and there is hardly any related software as well. According to the characteristic that there are abundant formal syntax markers in Tibetan, we hope to explore a new way to resolve this problem without word segmentation and POS tagging.

In this paper, we try to use the CRFs model to identify the boundary by syllables without word segmentation and POS tagging based on the Tibetan syntactic functional chunk description system. In the experiment, we add the syntax markers as characters to the identification model. Through it, we have achieved satisfactory result on a small-scale unlabeled corpus.

2 Related Work

There are rich research results in English and Chinese about the chunk boundary identification and some related study. For example, English noun phrases identification using HMM (Hidden Markov Model) [1], tagging part of speech and chunk boundaries using a mixed model combined by finite state automaton and 2-gram model [2]. For the researches in Chinese, many scholars have tried to take advantage of the characteristics of Chinese and use methods based on rules including the identification of verb-object construction using rules [3], the identification of noun phrases consisting of noun phrases based on rules [4], the longest noun phrase identification using the method of tagging and phrase boundary co-occurrence probability [5]. Also, many scholars have applied machine learning methods such as Support Vector Machine (SVM), Hidden Markov Model (HMM), Maximum Entropy (ME) and Conditional Random Fields (CRFs) into this area. Huang and others carry out Chinese chunking parsing using CRFs and then use the error-driven learning method to correct the identification results [6]. Dai and others carry out the identification of longest noun phrases using CRFs then correct the results by the internal structure information [7]. In the aspect of functional chunk analysis, Zhou and others use a top-down approach to define the functional chunks of Chinese in order to describe the basic structure of a sentence [8].Subsequently, they construct the Chinese chunk library ChunkBank on the basis of functional chunk system [9], and furthermore they consider the functional chunking as a process including the segmentation of a sentence and labeling chunks with different functional tags [10].

In Tibetan, the researchers carried out a great deal of basic research on Tibetan morphology and grammar. For instance, scholars have tried to deal with the stick-from writing in Tibetan [11], sort out the functions and meanings of synaptic markers in

Tibetan [12] and then the Tibetan basic chunk description system was proposed[13]. To identify the chunk boundary, Wang tried to identify the Tibetan functional chunk boundary using an error-driven method [14]. Furthermore, Wang and his co-workers tried to take advantage of word segmentation and part of speech tagging using CRFs to identify the chunk boundary [15].

3 Tibetan Syntactic Markers

Tibetan uses an alphabetic writing system which has 30 consonants and 4 vowels. Through different collocation, we can get different syllables and then syllables make the word. In Tibetan, the punctuation mark "་" (tsheg) acting as delimiter between two syllables, similar to the informational value of a character in the orthography of Chinese. Syntactic markers in Tibetan especially in modern Tibetan are abundant. Generally speaking, syntactic markers here refer to form markers in sentence such as case marking and auxiliary marking, which can be used to divide the sentence into different functional chunks. For instance, there might be locative case marking after adverbial of place, agent case marking after subject and patient case marking after object. However, due to the habit of text writing, some case and auxiliary markings contract to one syllable which is called the sticky writing form [16]. In order to make full use of the case and auxiliary markings, not only do we need to take advantage of the case and the auxiliary makings that form to be independent syllable, but also to identify the sticky writing form and then separate them correctly.

3.1 Tibetan Abbreviated Syllable Mark

Tibetan has the following abbreviated syllables: (1) syllable+ས (-s) (agentive/instrumental markers), (2) syllable +འི་ (-vi) (genitive markers), (3) syllable +ར (-r) (dative and locative markers), (4) syllable +འང/འམ (vang/vam) (conjunction), (5) syllable +འོ (-vo) (sentence end markers). But for the purpose of syntactic functional chunk identification, different sticky writing forms are not of equal importance. Crudely put, (1) and (2) have a relatively high frequency and they play the most important role in chunking identification while (4) and (5) have a low frequency and they contribute a little to chunking even though they occur as the boundary of chunks at times. There are three approaches to identify the sticky writing form. The first is to implement the sticky writing identification simultaneously with the word segmentation [17]. The other one is to implement these two ways consecutively. Among them, one way is only recognize whether it is abbreviated syllables without distinguishing the type. The other one is to recognize both the form and the type. The other one is to identify both the form and type. This method is able to facilitate the identification of the chunk boundary and type. We adopt the last strategy in this paper. The labels we design for the experiment are as follows:

Table 1. Types of abbreviated syllables.

Type	Example	Label
Syllable +ས(-s)(agentive/instrumental markers	ངས་ ”I do”	S
Syllable +འི་(-vi)(genitive markers)	ངའི་”Mine”	V
Syllable +ར(-r)(dative and locative markers)	ངར་”For me”	D
Syllable +འང/འམ(vang/vam)(conjunction)	ངའང་“And me”	C

3.2 Tibetan Non-abbreviated Syllable Form Mark

Non-abbreviated syllable form in Tibetan language mainly refers to the case and the auxiliary markings which are independent syllables. The case makings include agentive marker, causative marker and instrumental marker such as གིས, གྱིས, ཡིས, ཀྱིས, time marker, locative marker, target marker, genitive marker, allative marker སུ, རུ, ལ, དུ, ཏུ, ན (variants of la don), comparative marker ལས, comitative marker དང་ and so on. The auxiliary word markers include analogical auxiliary word, pause auxiliary word, enumerative auxiliary word, manner auxiliary word, result auxiliary word, purpose auxiliary word markers and so on. Moreover, we can design different labels to annotate them according to their function and identify the chunk type at the same time. But in this paper, we only want to get the chunk boundary, so we annotate them with the same label "M".

4 Tibetan Functional Chunks

4.1 Tibetan Functional Chunk System

In this paper, we use the Tibetan functional chunk system which is defined following the descriptive definition in [18], and it includes subject chunking, predicate chunking, object chunking, adverbial chunking, completive chunking and syntactic markers chunking.

4.2 Tibetan Functional Chunk Annotation

We regard the identification of the boundary of each chunk as a problem of sequence labeling. We use the BIE tag set to mark chunks. That is, we tag the syllable with “B” when it is the start of a chunk, tag the syllable with “I” when it is inside of a chunk and tag it with “E” when it is the end of a chunk. Furthermore, we tag punctuation with “B”.

Take the example of the sentence ཁོང་གིས་ང་ལ་དེབ་དེ་འཁྱེར་འདུག (They bring me a parcel, khong tshos nga la bskur ma zhig bskur shag.). Firstly, we identify the syntactic tag of it, we can get the intermediate results as follows: [ཁོང་ཚོས་] [ང་ལ་] [བསྐུར་མ་ཞིག་] [བསྐུར་ཤག]. Furthermore, we can get the final label results as Fig. 1.

Eg1: ཁོང་/Bཚོ/Mས་E ང་/Bལ་/E བསྐུར་/Bམ་/M ཞིག་/E བསྐུར་/Bཤག/E

Latin: khong/B tsho/M s/E nga/B la/E bskur/B ma/M zhig/E bskur/B shag/E.

En: They bring me a parcel.

Fig. 1. Tibetan functional chunk boundary identification mark examples.

5 Tibetan Functional Chunk Boundary Identification Based on CRFs

5.1 Conditional Random Fields Model

CRF model is a sequence labeling and disaggregated model which was put forward by John Lafferty in 2001. In this paper, we only give a simple introduction to CRF model, for details please see the reference. It is a conditional distribution model based on undirected graph. Given a certain observed sequence needs to be annotated, it calculates the joint probability of the whole sequence to find the optimal result of the labeling. CRF is able to express long distance dependence and overlapping features, which is conducive to the resolution of the problem of labeling (classification) bias, so as to get the optimal result. For the given observed sequence $x=x_1x_2\ldots x_n$, with x_i denotes a word in the sequence. We define $y=y_1y_2\ldots y_n$ is the sequence to output, which is the tag of each word. For a CRF which is given the parameter $\Lambda=\lambda_1\lambda_2\ldots\lambda_k$, we will get the probability of the Y with the input of the sequence:

$$P_\Lambda(y|x) = \frac{1}{Z(x)} exp\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, t)\right)$$

Z(x) is the normalized functions and $f_k(y_{i-1}, y_i, x, t)$ denotes a feature function. The symbol f_k denotes the weight parameter which is relevant to λ_k. We will obtain it through training. And then the most possible labeling sequence $Y^* = arg_Y \, max P_\Lambda(Y|X)$ is the output.

In this paper, we use the CRF++[1] which is developed by Taku Kudo as the CRFs model to accomplish the task of functional chunk parsing.

5.2 Text Preprocessing

Before identifying the chunk boundaries, we firstly do the text preprocessing to identify whether the Tibetan syntactic markers are sticky writing or not. We use the CRFs to solve the problem using the current syllable and the context feature as the template.

[1] http://taku910.github.io/crfpp/

Table 2. Atomic Feature Template

ID	Template	ID	Template
1	CurSyllable	4	Syllable+1
2	Syllable-2	5	Syllable+2
3	Syllable-1		

In Table 2, the size of the context window is 5 and the +/- indicate the syllables after/before the current syllable. Then we combine the atomic feature template to get the complex feature template as Table 3

Table 3. Complex Feature Template.

ID	Template
6	CurSyllable,Syllable-1
7	CurSyllable,Syllable+1
8	Syllable-1,Syllable+1

5.3 Tibetan Functional Chunk Boundary Identification Based on CRFs

After the identifying of syntactic markers, we label each syllable by the rule shown in 3.2. We define the atomic feature template in the process of functional chunk boundary identification, they are listed in Table 4.

Table 4. Atomic Feature Template

ID	Template	ID	Template
1	CurSyllable	6	CurCase
2	Syllable-2	7	Case-2
3	Syllable-1	8	Case-1
4	Syllable+1	9	Case+1
5	Syllable+5	10	Case+2

In Table 4, the symbol “Syllable” indicates the syllable. The symbol “case” indicates the result of text preprocessing. When the characteristic function takes a specific value, the template is instantiated.

Eg3: [ཁོང་/][རྒྱལ་ནང་/ལ་/][ཕྱིར་ལོག་བྱས་/པ་/རེད་/]།

Latin: khong rgyal nang la phyir log byas ba red.

En3: He returns his home country.

Fig. 2. Atomic Template Feature Selection Sample.

In eg3, we select the syllable “གསར” as the CurSyllable, then we can get the features using templates in Table 4. For example, the Syllable + 2 will indicate the syllable “བཤད་”and the Case + 1 will indicate "Non-abbreviated form".
We build on combinations of the atomic feature templates to get the complex feature template as Table 5.

Table 5. Complex Feature Template

ID	Template
11	CurSyllable,Syllable-1
12	CurSyllable,Syllable+1
13	Syllable-1,Syllable+1
14	CurCase,Case-1
15	CurCase,Case+1
16	Case-1,Case+1

6 Experiment and Analysis

6.1 Syntactic Marker Identification Result

In the experiment of identifying syntactic marker, we use the F-value as the evaluation criterion. The results are shown in Table 6.

Table 6. Syntactic Marker Identification Result

	S	D	V	N	C	M	total
F	0.95	0.85	0.93	1.00	0	0.93	0.98

From Table 6, we find that the overall effect is satisfying but for the identification effect of R is not very ideal. Through the analysis of the training corpus, the cause we find is the syllable ར (-r) can be suffixed to different syllables (some are abbreviated, but others not) and its frequency is very high. For example, “ངར”, it is possibly a syllable, meaning “strength”, and possibly an abbreviated form, meaning “ང+possessive ར”. So we cannot train an effective model for it. It also occurs on abbreviated syllables, but the frequency is much lower than D type.

6.2 Syntactic Functional Chunk Identification Results

In order to verify the effect that syntactic markers have in the identification of Tibetan functional chunk boundary, we have conducted two experiments with treating the experiment 1 as the baseline. In experiment 1, we do not carry out the text preprocessing and identify the chunk boundary directly based on the “་” between different syllables. In experiment 2, we firstly identify the syntactic marker, then we split the abbreviated syllables and finally we identify the chunk boundary. The results are shown as Figure 3.

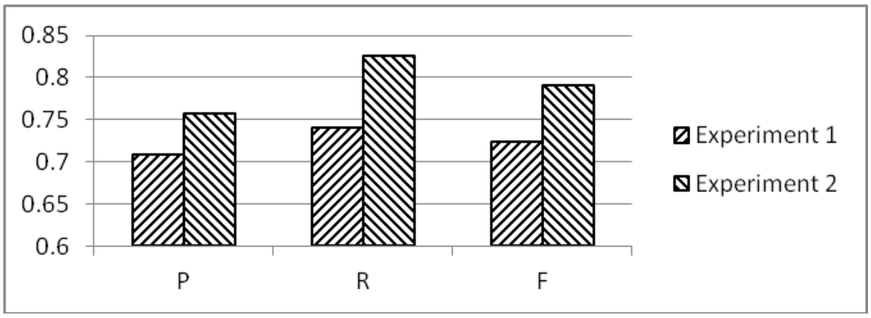

Fig. 3. Syntactic Functional Chunk Identification Result.

Comparison of experimental result in Figure 3 shows that the identification of syntactic markers can improve the result significantly with the F value reaching 79.12% (6.71% higher than the baseline). From the result we can see that it does have a positive effect on the boundary identification through semantic information implied by syntactic markers.

6.3 Error Analysis

It is difficult to carry on the identification of syntactic markers on the original corpus or the corpus after preliminary processing. From our experimental results, although we have achieved certain effect, there are plenty of errors which can roughly sum up to several types as follows:

Boundary Identification Error of Non-predicate Verb Structure. Non-predicate verb structure is a syntactic chunk in a sentence, which is consist of phrases and clauses with nominal tag. They are typically long distance chunks, difficult to identify. For instance, ཤོག་ཕེ་རྩེ་རྒྱུ་ནི་སྤྲོ་སྣང་སྐྱེས་པའི་བྱ་བ་ཞིག་རེད ། (shog phe rtse rgyu ni spro snang skyes bavi bya ba zhig red. Playing cards is a great pleasure.) The result of boundary identification is [ཤོག་ཕེ་རྩེ་] [རྒྱུ་] [ནི་] [སྤྲོ་སྣང་སྐྱེས་པའི་བྱ་བ་ཞིག་] [རེད།] while the correct result should be [ཤོག་ཕེ་རྩེ་རྒྱུ་] [ནི་] [སྤྲོ་སྣང་སྐྱེས་པའི་བྱ་བ་ཞིག་] [རེད།]. The nominal tag [རྒྱུ་] should be part of the previous chunk. This kind of errors contribute the most of all. It is the focus of following research.

Boundary Identification Error of Continuous Predicate Structure. For instance, ཁོ་མོ་མར་ལྷུང་སྟེ་མས། (kho mo mar lhung ster mas, she fell and hurt herself.) The result of boundary identification is [ཁོ་མོ་] [མར་ལྷུང་] [སྟེ་] [མས], while the correct result should be [ཁོ་མོ་] [མར་ལྷུང་སྟེ་མས].

Boundary Identification Error of Appositive Structure and Structure of Modification Being Lack of Markers. The appositive structure and structure of modification consist of many syllables are short of dominant makers hence there are much errors. For instance, ཁྱོད་ཚང་ལ་མི་དུ་ཡོད། (khyod tshang la mi du yod, How many people are there in your family?) The word ཁྱོད་ཚང་ is lack of attributive marker so that the word was spilt into two parts in the experiment.

Boundary Identification Error of Core Predicate Chunk Lack of Tense Marker. In Tibetan language, the verbs and some adjectives serve as the predicate and appear at the end of the sentence. In a sentence directly ending with a verb or an adjective, there are few linguistic features after verbs, reducing the confidence of the training model and causing the identification errors. For instance, གཞུང་དྲང་པའི་མིས་གཏན་ནས་རྫུན་མི་བཤད། (gzhung drang pavi mis gtan nas rdzun mi bshad. Good people don’t tell lies at all.) In this sentence, རྫུན་མི་བཤད is the predicate block errors. In spite of the errors above, we think if the identification features are further refined and the identification strategy is optimized, the results can be improved effectively.

7 Conclusions

Syntactic functional chunks represent different functional components of a sentence. Through the recognizing of syntactic functional chunks we can simplified the analysis of the structure in a sentence. In this paper, we propose a method that identifies the syntactic chunk boundary without word segmentation and POS tagging based on the Tibetan functional chunk system. Through the analysis of the Tibetan language and the experiment results, we add the Tibetan syntactic marker into the experiment and the precision, recall rate and F value respectively achieves75.70%, 82.54% and 79.12%. In the next step, on one hand, we are ready to select better features to improve the identification effect of syntactic markers; on the other hand, we will expand the quantity of the training corpus and then provide basic support for the use of other nature language processing applications such as machine translation and so on.

8 Acknowledgement

This work is supported by the National Natural Science Foundation of China (61671064, 61201352, and 61132009), the National Key Basic Research Program of China (2013CB329303) and the Fundamental Research Fund of Beijing Institute of Technology (20130742010).

References

1. Church, K. W.: A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the second Conference on Applied Natural Language Processing, pp. 136-143. Association for Computational Linguistics (1988)
2. Pla, F, Molina, A, Prieto, N.: An integrated statistical model for tagging and chunking unrestricted text. In: The Third International Workshop on Text, pp. 15-20. Speech and Dialogue, Brno, Czech Republic(2000)
3. Sun, H.L.: Induction of grammatical rules from an annotated corpus V+N sequence analysis. In: China National Conference on Computational Linguistics, pp. 157-163.Tsinghua University Press, Beijing(1997)
4. Liu, C.Z.: Research on binding of common noun sequences based on POS tagging corpus. In: Proceedings of the International Conference on Chinese information processing, Tsinghua University Press, Beijing (1998)
5. Li, W.J., Zhou, M., et al.: Automatic extraction of Chinese longest noun phrases based on corpus. In: Chen, L., Yuan, Q. Progress and Application of Computational Linguistics, pp. 119-124. Tsinghua University Press, Beijing (1995)
6. Huang, D., Yu, J.: The combination of distributed strategy and CRFs to identify Chinese chunk. Journal of Chinese Information Processing 23 (1):16-22(2009)
7. Dai, C., Zhou, Q.L., Cai, D.F., et al.: Automatic Identification of Chinese Maximum Noun Phrase Based on Statistics and Rules. Journal of Chinese Information Processing22(6):110-115(2008)
8. Drábek, E. F., Zhou, Q.: Experiments in learning models for functional chunking of Chinese text. In: IEEE International Conference on Systems, Man, and Cybernetics IEEE vol.2:859-864 (2001)
9. Chen, Y., Zhou, Q.: Analysis and Construction of Hierarchical Chinese Function Block Description Library. Journal of Chinese Information Processing 22(3):24-31(2008)
10. Zhou, Q., Zhao, Y.Z.: Automatic Parsing of Chinese Functional Chunks. Chinese Journal of information 21(5): 18-24(2007)
11. Jiang, D., Kang, C.J.: The methods of lemmatization of bound case makers in modern Tibetan. In: International Conference on Natural Language Processing and Knowledge Engineering, pp. 616-621(2003)
12. Jiang, D.: The method and process of block segmentation in modern Tibetan. Minority Language of China 2003 (4):30-39 (2003)
13. Long, C.J., Kang, C.J., Jiang, D.: The Comparative Research on the Segmentation Strategies of Tibetan Bounded-Variant Forms. In: International Conference on Asian Language Processing 2013(30), pp. 243-246. IEEE Computer Society (2013)
14. Wang, T.H., Shi, S.H., Long, C.J., et al.: Syntactic boundary block identification of Tibetan syntactic functions based on error driven learning strategy. Chinese Journal of information 28 (5):170-175(2014)
15. Wang, T.H.: Research on Tibetan functional block recognition for Machine Translation. Beijing Institute of Technology(2016)
16. Liu, H.D.: Research on Tibetan Word Segmentation and Text Resource Mining. University of the Chinese Academy of Sciences(2012)
17. Li, Y.C., Jia, Y.J., Zong, C.Q.: Research and Implementation of Tibetan Automatic Word Segmentation Based on Conditional Random Field. Journal of Chinese Information Processing 27(4):52-58(2013)
18. Li, L., Long, C.J., Jiang, D.: Tibetan Functional Chunks Boundary Detection. Journal of Chinese Information Processing 27(6):165-168(2013)