Tibetan Word Segmentation System Based on Conditional Random Fields

Tao Jiang, Hongzhi Yu

Key Laboratory of China's National Linguistic Information
Technology
Northwest University for Nationalities
Lanzhou, China
xinxiyuanjt@126.com

Yangkyi Jam

Key Laboratory of China's National Linguistic Information
Technology
Northwest University for Nationalities
Lanzhou, China
jamyangkyi2007@yahoo.com.cn

Abstract—Unlike English and other western languages, there are no delimiters to mark word boundaries in both Chinese and Tibetan. Therefore, word segmentation is the first step for Chinese and Tibetan natural language processing such as machine translation and information retrieval. However, Chinese word segmentation has been studied for many years and the technology is relatively mature. In contrast, Tibetan word segmentation is less concerned by researchers. In this paper, we learn from Chinese word segmentation approach and analysis the characteristic of Tibetan language, designs a Tibetan word segmentation system based on conditional random fields. The experiment shows that the algorithm is effective and can be preliminary applied.

Keywords- Tibetan word segmentation; conditional random fields; Natual language processing

I. INTRODUCTION

Tibetan language belongs to the Tibetan-Burman branch of the Sino-Tibetan language family which is spoken by about 6 million people in China (Tibet, Qinghai, Gansu, Sichuan and Yunnan), India, Bhutan, Sikkim, Ladakh and Nepal. It is alphabetic language which consists of thirty consonants, four vowels, five inverted letters (for the renting of foreign words) and the punctuations. Unlike English and other western languages, there are no delimiters to mark word boundaries in Tibetan, where sentences are represented as strings of Tibetan syllables. Therefore, word segmentation is first step for most Tibetan natural language processing (NLP) tasks such as machine translation and information retrieval. As its errors would be accumulated into the following phases, word segmentation is crucial for Tibetan NLP. In recently years, Tibetan word segmentation has been gradually concerned by researchers, but the progress is much slower than Chinese word segmentation. Chen Yuzhong [1] proposed a Tibetan word segmentation method based on case-auxiliary word and continuous feature. In this work, they used the grammar roles and dictionary to segment Tibetan text. Huidan Liu [2] studied the characteristics of Tibetan numbers and proposed a method to identify Tibetan numbers based on classification of number components. Cai zhijie [3] gave a new method of restoration to identify the abbreviated word for Tibetan word segmentation, which is to re-establish the abbreviated Tibetan word to its original form by the reinstallation rules. Qi kunyu [4] proposed a three level method to segment Tibetan text. In his work, the

case division was used on sentence structure level, and the bordering signal was used on phrase division, then the model matching was used on words block division.

At present, researchers mainly use the rules method based on dictionary in Tibetan word segmentation. However these approaches have been limited by the impossibility of creating a lexicon that includes all possible Tibetan words and by the lack of robust statistical inference in the rules. Statistical machine learning approaches such as CRFs are more desirable and have been successfully applied in Chinese word segmentation. In order to use the CRFs model in Tibetan word segmentation, we have built mount of manually segmented corpus as the training set. Then we implemented a Tibetan segmentation system used CRF++ package version 0.54.1. In our system, a Tibetan syllable is labeled by a tag which indicates for its position in the related Tibetan word.

The rest of this paper is organized as follows: In section 2, we briefly describe the conditional random fields. In section 3, we discuss the tag set selection and the feature templates. We report in section 4 our experimental results and give our conclusions on this work in section 5.

II. CONDITIONAL RANDOM FIELDS

Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting structured data [5], such as sequences, trees and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models. CRFs outperform both MEMMs and HMMs on a number of real-world tasks in many fields, including bioinformatics, computational linguistics and speech recognition.

¹ <http://crfpp.sourceforge.net/>

CRFs defines a conditional probability for a label sequence $y = y_1 \cdots y_T$ given an input sequence $x = x_1 \cdots x_T$ to be

$$P(y|x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^T \sum_k \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

Where Z_x is the per-input normalization that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, x, t)$ is a feature function which is often binary-valued, but can be real-valued, and λ_k is a learned weight associated with feature f_k . The feature functions can measure any aspect of a state transition [6], $y_{t-1} \rightarrow y_t$, and the entire observation sequence, x , centered at the current time step, t . The most probable label sequence y for an input x is

$$y^* = \arg \max_y P(y|x) \quad (2)$$

III. TAG SET SELECTION AND FEATURE TEMPLATES

A. Tag Set Selection

Character based tagging method for Chinese word segmentation, either based on maximum Entropy or CRFs, views Chinese word segmentation as a label tagging problem. Tibetan word is composed of syllables separated with syllable dots, such as ཁང་གི (house), there are two Tibetan syllables, ཁང་ and གི . Similar to the Chinese tagging method [7], we use the syllable as the basic unit in our system.

In Chinese tagging method, many different tag sets have been adopted such as 6-tag set, 4-tag set, 3-tag set and 2-tag set. In order to select an effective tag set, we analysis 100 thousands Tibetan words which are mainly from the Tibetan-Chinese Dictionary edited by Zhang Yisu and Tibetan Dictionary by Chos-kyi-grags-pa, then calculate that the average syllables of the Tibetan words is 2.76. The details are list in Table I.

TABLE I. SITUATION OF TIBETAN SYLLABLE IN THE TIBETAN WORDS

Syllable type	Ratio of total words(100 thousands)	Example
1-syllable	5.8%	བླ་ (son)
2-syllable	48.0%	ཁང་གི (house)
3-syllable	22.4%	རང་བྱུང་ཁམས་ (nature world)
4-syllable	17.7%	རྒྱ་བུ་མའི་ཁྱུ་ (cube)

Multi-syllable	3.1%	ཨ་ཇ་ཀའི་དྲི་ག་ (the name of a flower)
----------------	------	---

Table I shows that most Tibetan words contains no more than 4 syllables, therefore we chose the 4-tag set, and tag each syllable with one of the four tags, B, E, M and S depending on its position in a word. It is tagged B if it occurs on the begin position of a word, and forms a word with the syllables on its right. It is tagged E if it occurs on the end position of a word, and forms a word with the syllables on its left. It is tagged M if it occurs in the middle of a word. It is tagged S if it forms a word by itself. For example, the manually segmented sentence (a) will be tagged as (b):

གནའ་བྱས་ སྤྱི་ མིང་ ལ་ སི་མ་ཀོང་ ཟེར་བ
འི་ ཁྱིམ་ ཞིག་ ཡོད།

(a)

གནའ་/B སྤྱི་/E སྤྱི་/S མིང་/S ལ་/S
འི་/B མ་/M ཀོང་/E ཟེར་/B བ/E
འི་/S ཁྱིམ་/B སྤྱི་/E ཞིག་/S ཡོད་/S
/S

(b)

B. Feature templates

A feature template is a pattern to extract features. It may correspond to several features, while a pair of context and candidate word may correspond to several features. One of the advantages of CRFs is that they are flexible enough to capture many correlated features, including overlapping and non-independent features. In our system, we use the basic feature template and the additional feature template, which are shown in Table II and Table III. In the feature templates, C refers to a Tibetan syllable, and C_0 denotes the current syllable.

TABLE II. BASIC FEATURE TEMPLATE

Feature	Meaning
$C_n (n = -2, -1, 0, 1, 2)$	The syllable n positions to the right(left) of the current character
$C_n C_{n+1} (n = -2, -1, 0, 1)$	The second previous (previous, current, next) syllable and the previous (current, next, second next) syllable
$C_{-1} C_1$	The previous syllable and the next syllable

TABLE III. ADDITIONAL FEATURE TEMPLATE

Feature	Meaning	Example
$Au(C_0)$	The current syllable is case-auxiliary word.	འི་ ཁྱིམ་ ན་ ཏུ་
$Pu(C_0)$	The current syllable is a punctuation symbol.	$ \quad \quad \text{།}$

IV. EXPERIMENT

At present, there is no Tibetan word segmentation specification and standard, so comparing Tibetan word segmentation accuracy can be difficult because research papers use different data sets and different ground-rules. To evaluate our system, we manually segmented 2500 sentences for training and 225 sentences for testing. The training data is from Tibetan Daily and the testing data is from the Tibetan web sites. Then a Tibetan segmentation system based on the reverse directional maximum matching method (RMM) is developed to be a comparative system. Final results of the RMM and CRFs are summarized in Table IV.

TABLE IV. COMPARISON BETWEEN RMM AND CRFS RESULTS

System	Precision	Recall	F1
RMM	0.847	0.834	0.84
CRFs	0.933	0.935	0.934

As is shown in Table IV, the CRFs method is better than the RMM method, and all evaluation indexes are significantly improved. In future work, we will use external dictionary and more Tibetan grammar information to improve the performance of our system.

V. CONCLUSIONS

Tibetan word segmentation is a key precursor for Tibetan language processing tasks. Till now, there is no Tibetan word segmentation specification and standard, and researchers mainly use the rules method based on dictionary in Tibetan word segmentation. Statistical machine learning approaches such as CRF are more desirable and have been success in Chinese word segmentation. In this paper, we take the Tibetan

syllable based on their word-internal positions and implemented a Tibetan segmentation system used CRF++ package. In the testing result, the segmentation accuracy is 93.3%. We will use more Tibetan grammar information to improve the performance in future.

ACKNOWLEDGMENT

This work was supported by the Open Program of the National Laboratory of Pattern Recognition, the Open Program of the Key Laboratory of China's National Linguistic Information Technology, and the Natural Science Foundation of China under Grant No.60970071.

REFERENCES

- [1] Chen Yuzhong, Li Baodi, Yu shiwen, Lan cuoji. "An Automatic Tibetan Segmentation Scheme Based on Case Auxiliary Words and Continuous Features," *Applied Linguistics*, 2003(01),pp. 75-82.
- [2] HuiDan Liu, Weina Zhao, and Minghua Nuo. "Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation," *In Proc. of COLING*, poster volume, 2010, pp. 719-724.
- [3] Cai Zhijie, "Identification of Abbreviated Word in Tibetan Word Segmentation," *Journal of Chinese Information Processing*, 2009, pp. 107-112.
- [4] Qi Kunyu, "On Tibetan Automatic Participate Research with the Aid of Information Treatment," *Journal of Northwest University for Nationalities*, 2006, pp. 92-97.
- [5] J.Lafferty, A.McCallum, and F.Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *ICML-01*, 2001, pp. 282-289.
- [6] Fuchun Peng, Fangfang Feng, and Andrew McCallum, "Chinese Segmentation and New Word Detection Using Conditional Random Fields," *In Proc. of COLING*, 2004.
- [7] Nianwen Xue, "Chinese Word Segmentation as Character Tagging," *Computational Linguistics and Chinese Language Processing*, 2003, pp. 29-48.