Study on Tibetan Word Segmentation as Syllable Tagging

Yachao Li and Hongzhi Yu

Key Lab of Chinese National Linguistic Information Technology,
Northwest University for Nationalities, Lanzhou, China 730030
harry_lyc@foxmail.com

Abstract. Tibetan word segmentation (TWS) is the basic problem for Tibetan natural language processing. The paper reformulates the segmentation as a syllable tagging problem, and studies the performance of TWS with different sequence labeling models. Experimental results show that, the TWS system with conditional random field achieves the best performance in the condition of current 4-tag set, at the same time, the other models achieve good results too. All the above show that, the segmentation as a syllable tagging problem that is an efficient approach to deal with TWS.

Keywords: Tibetan, word segmentation, sequence label.

1 Introduction

Tibetan is alphabetic writing that contains 30 vowels and 4 consonants and spoken by about 6 million people in China. There is no space delimiter between adjacent Tibetan words, therefore, tokenization itself, is challenging task in Tibetan information processing.

Tibetan word segmentation dates back to the work by Zhaxiciren in 1999 [1], and has many important researches. Chen [2] proposed a TWS scheme based on case auxiliary words and continuous features, which could detect and eliminate segmentation ambiguities and deal with unknown words. The scheme has much more practical and achieves better performance. Qi [3] proposed a three level method to segment Tibetan text, this approach is based on the research of Tibetan form logic case, semantic logic case, phonological tendency studies. Caizhijie [4] introduced a TWS system using reinstallation rules to identify abbreviated words (AW) for the first time. Liu [5-6] proposed a method to identify Tibetan numbers based on classification of number components, and presented a novel approach for TWS using the conditional random fields. The approach combines the TWS and abbreviated word recognition in a unified tag set, is one of the most major results of TWS.

Most of the methods above are based on dictionary matching (maximum matching) or linguistic rules, and use some simple statistical information as auxiliary method, such as word frequency, entropy and so on. The TWS with machine learning has received less attention, because of there is lack of human-annotated corpus in Tibetan. Liu reformulates the segmentation as a syllable tagging problem, one of the latest research results, this approach uses statistical machine learning model and achieves the best performance [6].

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 363–369, 2013.

TWS as a syllable tagging problem dates back to the fundamental work by Xue [7], published in the first SIGHAN in 2002, this approach reformulates word segmentation as a sequence tagging problem, namely identify the position information of a character. In recent years, many experimental results show that, the method with character tagging effetely, becoming the mainstream, and implement in the TWS successfully.

The paper studies the TWS as syllable tagging, and implements experiment with maximum margin markov networks (M^3N), maximum entropy (ME), conditional random fields (CRF) respectively. The plan of the paper is as follows. In Section 2 we introduce the Tibetan syllable tagging problem, and compare the performance of different sequence labeling models. Section 3 introduces the sequence labeling models used in the paper. Section 4 gives the experimental results, and Section 5 concludes the paper.

2 The TWS as Syllable Tagging

TWS based on syllable tagging dates back to the work by Xue [7], which reformulates the segmentation problem as a character tagging problem. The approach has become the mainstream in Chinese word segmentation.

Tibetan is alphabetic writing and Tibetan word constituted by syllables. Many Tibetan syllables can occur in different position within different words. We can get segmentation results according to the position of syllable. Therefore, it is an effective method that, reformulates the segmentation as a syllable tagging problem, and then use machine learning model to label syllables automatically.

Table 1. The Tibetan syllable can occur in many word-internal positions

Position	Example	Meaning	Tag
Single	ཡོན་	Reword	S
Begin	ཡོན་ཏན་	Knowledge	B
Middle	ཤེས་ཡོན་ཅན་	Intellectual	M
End	འདོད་ཡོན་	Desire	E

We can use “BMES” to denote the four tag-set of Tibetan syllables. It is tagged B if it occurs at the begin of a word. It is tagged M if it occurs in the middle of a word. It is tagged E if it occurs at the end of a word. It is tagged S if it forms a word by itself. In the light of the 4-tag set of Tibetan syllables, Liu introduced a method, it add two tag SS and ES besides “BMES”, SS denotes a monosyllabic word contains abbreviated word, ES denotes a multi-syllable word contains abbreviated word. Li [8] proposed an approach called TagSet-2 in the next section, which is the prophase study of this paper. Experiments show that the TWS system adopted TagSet-2 achieved a better performance. So in this paper, we adopt the TagSet-2 as our syllable tag set. The different between Liu and TagSet-2 showed in table 2.

Table 2. Examples of TagSets

type of word	Example	TagSet-2	Liu
1 syllable+AW	ངས་ (ngs)	S-S	SS
2 syllable+AW	གནས་པའི་(bnas pvi)	B-E-S	B-ES
3 syllable+AW	མ་བྱས་པའམ་(ma byas pavm)	B-M-E-S	B-M-ES

Syllables segmented by "་" (tsheg) in the ancient Tibetan, however, there are no "tsheg" between some case auxiliary words and its prior syllable, these words called abbreviated words. For example "འདས་/པ/འི་/ལོ་/ལྔ་/།" (In the past five years，vdas pai lo lnga), there is no "tsheg" between in the third segmentation unit and the second segmentation unit. In Tibetan word segmentation, we should properly handle six abbreviated words, namely "ས་" (sa), "ར་" (ra), "འི་" (vi), "འོ་" (vo), "འང་" (vang), "འམ་" (vam). Abbreviated words recognition has great influence on Tibetan syllable recognition, which is an important problem that we must face in the Tibetan word segmentation.

In the light of abbreviated words recognition problem, Li [8] proposed an AW recognition method with sequence labeling, which reformulates segmentation problem as a binary classification problem, and then adopts sequence labeling model to recognition syllables. Tibetan word segmentation system, using AW recognition method listed above, need two steps, first recognizes the syllable sequence; second syllable labeling. The system is time consuming. To solve the problem, we consider the six abbreviated words as a unit, in order to alleviate the balance between of the precision of abbreviated words recognition and the system efficiency. The approach is the results of this paper's preliminary study, and has a best performance [8].

Example1:"ཚོང་ཟོག་རྫུན་མ་བཟོ་འཚོང་བྱེད།" (Manufacturing and selling inferior products, tsong sog rdzun ma bso vtsong byed). Example1 is listed to illustrate the Tibetan word segmentation based on syllables tagging. First, sequence labeling model was used to label syllables, the result is "ཚོང་/Bཟོག་/Eརྫུན་/Bམ་/Eབཟོ་/Sའཚོང་/Sབྱེད་/S།/S"; on the basis of labeling result, we can reduction preliminary segmentation result "ཚོང་ཟོག་/རྫུན་མ་/བཟོ་/འཚོང་/བྱེད་/།/"; in the next, the processing of digital, time and date; finally, output the segmentation result.

3 Sequence Labeling Model

Statistical machine learning models are widely used in word segmentation tasks, these models can be classified into two categories: one is based on maximum margin learning, such as support vector machine (SVM) [9], which is widely used in classification; the other is based on the criterion of maximum likelihood estimation, such as maximum entropy (ME) [10], conditional random field (CRF) [11], these models are successfully used in Chinese word segmentation.

Maximum entropy and conditional random field are widely used in Chinese word segmentation. Maximum entropy was used in the early study of Chinese word segmentation; conditional random field has been proved to be an excellent sequence tagging model; at the same time, maximum margin markov networks (M^3N) was used in segmentation also. Conditional random field is successfully used in Tibetan word segmentation too, but no references have given about the effect of others model. We hope in this article that, implement these three models in the Tibetan word segmentation and compare these models in the same condition. The feature templates used in this paper showed in table 3.

Table 3. Feature templates

Features	Meaning
$C_n(n = -2,-1,0,1,2)$	The *ith* syllable to the left/right of the current syllable
$C_nC_{n+1}(n = -2,-1,0,1)$	Adjacent two syllables
$C_{-1}C_1$	Two syllables before and after the current

3.1 Maximum Entropy

Maximum entropy was proposed by E.T.Jaynes in 1950s [10], and was used in natural language processing by Della Pietra. The basic principle of ME is that using the given samples, and select a probability distribution conform to the training sample, it must be satisfy all the facts that known. Without additional assumptions and constraints conditions, for those uncertain samples, ME will give a uniform probability distribution. Entropy is measure of the uncertain, the greater the uncertain, the greater the entropy, and the more uniform distribution. Maximum entropy model:

$$P^* = \arg\max_{p\in C} H(P) \tag{1}$$

$\mathrm{H(P)}$ is the entropy of model P,C is the collection of model that satisfy constraints, the following need to seek P^*, P^* represented as follows:

$$P^*(y|x) = \frac{1}{Z(\mathrm{x})}\exp\left(\sum_i \lambda_i f_i\,(x,y)\right) \tag{2}$$

$\mathrm{Z(x)}$ is the normalization constant, represented as follows:

$$Z(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right) \tag{3}$$

λ_i is the weight parameters of features.

3.2 Conditional Random Field

Conditional random field, proposed by Lafferty [11], is a statistically sequence labeling model, for more information in the reference 12.

We reformulate the segmentation problem as a syllable tagging problem, and generates a linear-chain CRF based on a undirected graph G = (V,E). V is a set of random variables Y, Y = {Yil1 ≤ i ≤ n}, for the n units needed to label in the input sentence, E = {(Yi-1, Yi) |1 ≤i ≤ n} is the linear-chain composed of n-1 edges. For each sentence x, define two non-negative factors:

$$\text{For each edge: } \exp\left(\sum_{k=1}^{k} \lambda_k f_k(y_{i-1}, y_i, x)\right) \tag{4}$$

$$\text{For each node: } \exp\left(\sum_{k=1}^{\acute{K}} \acute{\lambda}_k \acute{f}_k(y_i, x)\right) \tag{5}$$

f_k is a binary feature function, K and K' is the number of features defined in each edges and each nodes respectively.

Given a sequence x need to label, conditional probability corresponding tag sequence y is:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{(i,k)} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{(i,k)} \acute{\lambda}_k \acute{f}_k(y_i, x)\right) \tag{6}$$

Z(x) is normalized function, given the sequence x, the corresponding tag sequence y is given by $Argmax_{y'}P(y'|x)$.

3.3 Maximum Margin Markov Networks

M^{3}N is the extension of multi-class support vector machine (SVM) in the condition of structure prediction [12], the goal of M^{3}N is constructing a mapping function $h: X \to Y$ from observation example set $S = \{(x^i, y^i = t(x^i))\}(i = 1, \ldots, m)$. $h: X \to Y$ is defined by weight coefficient vectors w$_i$(i=1,…,n), each weight coefficient vector corresponds a feature function $f(x, y_i, y_j)$, the function denoted by $f(x,y)$, the goal of classifier is the solution of function h_w:

$$h_w(x) = \arg\max \sum_{i=1}^{n} w_i f_i(x, y) = \arg\max w^T f(x, y) \tag{7}$$

4 Experiments

In our experiments, the training set and the test set are come from CWMT2011. Total corpora divided into two parts according to the ratio of 3:7. The training set contains 856647 words, and the test set contains 300000 words, and out of vocabulary rate is 1.7%.

We can evaluate the approach by Precision (P), Recall (R) and F-score (F). P, R and F calculated as follows:

$$\text{P} = \frac{\textit{the number of correctly segmented words}}{\textit{the total number of words in the segmented corpus}} \times 100\% \tag{8}$$

$$\text{R} = \frac{\textit{the number of correctly segmented words}}{\textit{the total number of words in the gold standard}} \times 100\% \tag{9}$$

$$\text{P} = \frac{2 \times R \times P}{R+P} \times 100\% \tag{10}$$

We implemented three Tibetan word segmentation systems based on CRF, ME, M^{3}N respectively. CRF model, ME model and M^{3}N mode are implemented by

CRF++[1], maximum entropy toolkit[2] and pocket_m3n[3] respectively. There are illegal tags in ME and M3N, but is rarely in M3N. Therefore, we add feature tags and use dynamic programming to post-processing in the system with ME, called ME+D below. Experimental results showed in Table 4.

Table 4. Experimental results

TWS	R(%)	P(%)	F(%)
ME	94.11	93.02	93.56
ME+D	95.05	93.89	94.47
M^3N	94.38	94.34	94.36
CRF	95.35	95.32	95.33

Table 4 shows that, F-score of all the system reach 93%. This suggested that, reformulates the segmentation as a syllable tagging problem gets a good treatment of Tibetan word segmentation problem.

In the four TWS systems, F-score of the system based on CRF achieves 95%, gets the best result, which shows that, at the condition of four-tag set, CRF model gets a better treatment of Tibetan word segmentation problem compared to the other models.

5 Conclusion

This paper adopts the method based on syllable tagging, and compares the performance of different sequence labeling models. Experiments on our training set and test set show that the TWS system based on CRF outperforms all the system and achieves the best F-score. According to the experimental results Tibetan word segmentation based on syllable tagging can achieves a good performance.

Acknowledgments. The research work has been funded by the Natural Science Foundation of China under Grant No.61032008 and 61262054.

References

1. Bai, G.: Research on the Segmentation Unit of Tibetan Word for Information Processing. Journal of Chinese Information Processing 24(3), 124–128 (2009)
2. Chen, Y., Li, B., Yu, S.: A Tibetan Segmentation Scheme Based on Case-auxiliary Word and Continuous Features. Journal of Chinese Information Processing 17(3), 15–20 (2003)
3. Kun-Yu, Q.: On Tibetan Automatic Participate Research with the Aid of Information Treatment. Journal of Northwest University for Nationalities (Philosophy and Social Science) (4), 92–97 (2006)

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[2] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[3] http://sourceforge.net/projects/pocket-crf-1/

4. Zhi-Jie, C.: Identification of Abbreviated Word in Tibetan Word Segmentation. Journal of Chinese Information Processing 23(1), 35–37 (2009)
5. Liu, H., Zhao, W., Nuo, M., Jiang, L., Wu, J., He, Y.: Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation. In: Proceedings of the 23rd International Conference on Computational Linguistics (Posters Volume) (Coling 2010), pp. 719–724 (2010)
6. Liu, H., Nuo, M., Ma, L., Wu, J., He, Y.: 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), pp. 168–177 (2011)
7. Xue, N., Converse, S.P.: Combining classifiers for Chinese word segmentation. In: Proceedings of the First SIGHAN Workshop on Chinese Language Processing, Taipei, Taiwan, pp. 63–70 (2002)
8. Yachao, L., Yangkyi, J., Chengqing, Z., Hongzhi, Y.: Research and Implementation of Tibetan Automatic Word Segmentation with Conditional Random Field. Journal of Chinese Information Processing 4(27), 52–58 (2013)
9. Cortes, C., Vapnik, V.: Support- vector networks. Machine Learning 20(3), 273–297 (1995)
10. Berger, A.L., Pietra, S.A.D., Pietra, V.J.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics (22), 39–71 (1996)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of ICML 2001, pp. 282–289 (2001)
12. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In: Processing Syst., Vancouver (2003)