Fusion of Word Clustering Features for Tibetan Part of Speech Tagging Based On Maximum Entropy Model

Ning Ma*, Yachao Li, Xiangzhen He

Key Laboratory of National Language Intelligent Processing
Northwest University for Nationalities, Lanzhou, Gansu, 730030, China

***Abstract* - Tibetan Part of Speech (POS) tagging, the foundation of Tibetan natural language processing, judges word classification according to contextual information of words. Based on the framework of the maximum entropy model, the paper studied the fusion of morphological features for Tibetan part of speech with maximum entropy model with the integration of word clustering features. Experimental results show that Tibetan POS based on maximum entropy achieves much better results and word cluster features can increase the performance of Tibetan POS significantly. Additionally, the accuracy rate of Tibetan POS based on maximum entropy is 0.81% higher than that of baseline system.**

Keywords - Tibetan; Part of speech; Maximum entropy; Clustering features

I. INTRODUCTION

Part of Speech (POS) tagging, the foundation in dealing with natural language, is a process of judging word classification according to contextual information of words. It is widely applied in the fields involving machinery translation, voice recognition, information retrieval and so on. Studies about Part of Speech (POS) tagging in Chinese and English are abundant and they have strict theoretical research and sufficient data source of annotation, therefore the Part of Speech (POS) tagging in these two languages has come to a practical level.

Tibet language is an ancient language. However, the study about its word classification appears later. The previous research mainly focuses on the word-formation and morphological change of Tibet. QU Aitang analyses something about Tibet, covering the characteristics of word –formation, affix, and the function of additional affix in verb present tense, future tense, past tense and imperative form. SONG Jinlan undertakes similar study about the change and differentiation of constructing homologous derivative in Chinese and Tibet. LONG Congjun builds Tibet adjective morpheme database of facing information processing according to the features that most adjectives are mainly based on syllables; disyllables are usually constructed by root morphemes and suffixes; root morphemes decide word meaning; affix only decides grammatical meaning.

In the development of technique machinery translation system between Chinese and Tibet, CHEN Yuzhong divides Tibet words into different classifications, starting the research of Tibet Part of Speech (POS) tagging facing information processing. CAI Rangjia and ZHA Luo analyze the grammar of Tibetan parts of speech and the description of semantic information in detail. Additionally, attribute description and rules of the Tibetan Part of Speech (POS) tagging in corpus are conducted and the case-auxiliary, junctions and auxiliary words in Tibet are classified and tagged. Tibetan Part of Speech (POS) tagging and classification method based on corpus are proposed and referential mark and method are provided according to the requirement of building Tibetan corpus. The above studies decide the basic research work of Tibetan Part of Speech (POS), including classification criteria, labeling and so on.

Since 2005, experts have started to do experiments on Tibetan Part of Speech (POS) tagging and relevant tagging system has been built. Deep analysis about Tibetan word order, morphological change and expression in Sakya Pandita document translation system of Chinese – Tibet by CAI Zangtai is widely applied in Chinese-Tibet machinery translation of official document. SU Junfeng studied Tibetan Part of Speech (POS) tagging method based on HMM. The accuracy of this system reaches 88%-90% in close tests. YANG Maozhuome adopts HMM model to establish a Tibetan Part of Speech (POS) tagging system whose accuracy reaches 89.56% in the Part of Speech (POS) tagging of open corpus. Three methods about Tibetan Part of Speech (POS) tagging, involving rule, statistics and the combination of rule and statistics, are compared by YANG Maozhuome and ZHA Xijia. Contrastive analysis among descriptive approach, tagging efficiency and tagging accuracy is studied when these three method are used in Tibetan Part of Speech (POS) tagging. ZHA Xijia and GAO Dingguo build knowledge database of functional word and functional word common border database according to the syntactic function and semantic function of functional word in Tibet. Integrated processing model about Tibetan word segmentation and Part of Speech (POS) tagging are constructed with the combination of tagged corpus. YU Hongzhi and LI Yachao studied the Part of Speech (POS)

tagging based on maximum entropy model according to the characteristics of Tibetan word formation and the basic framework of maximum entropy model. Experiment results show that syllabic features can dramatically improve the accuracy of Tibetan Part of Speech (POS) tagging.

As the corpus size of Tibetan Part of Speech (POS) tagging is limited and Tibetan language is complex, the accuracy of existing and open Tibetan Part of Speech (POS) tagging is about 89% which is based on the test of private corpus. The public experiment results show that the tagging effect of Tibet is much lower than that of Chinese and English. Therefore, the Tibetan Part of Speech (POS) tagging shoulders heavy responsibilities.

II. THE TIBETAN POS TAGGING METHOD BASED ON MAXIMUM ENTROPY

A. The Maximum Entropy Model

Making a great achievement in the Part of Speech (POS) tagging study of Chinese and English, the maximum entropy model, with a feature of word fusion, is successfully applied in Tibetan Part of Speech (POS) tagging. It is firstly proposed in 1950 by E.T. Jaynes and Della Pietra applies it in the processing of natural languages. The basic idea of the maximum entropy illustrates that the chosen probability distribution which meets all the known facts is accordant with the given training samples. Under the condition where there is no more limitation and hypothesis, uniform probability distribution will be added to the uncertain part. Entropy stands for the uncertainty of random variable. The higher the uncertainty is, the bigger the entropy is and the more uniform the distribution is. The maximum entropy model shows below:

$$P^* = argmax \frac{\max}{p \in C} H(p) \qquad (1)$$

H(P) is the entropy of model P, C is the model set that satisfying the constraints. P* is need to seek, the form of P* is as follows:

$$P^*(y \mid x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x, y)) \qquad (2)$$

Z(x) is the normalized constant, which is expressed as follows:

$$Z(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y)) \qquad (3)$$

*Z(x)*is the weight parameter of feature.

This paper chooses the maximum entropy model as sequence annotation tool. For one thing, the maximum entropy shows a great performance in the speed and accuracy of Tibetan word segmentation; for another thing, the previous study shows that compared with other sequence annotation model, the maximum entropy model makes a better achievement in Tibetan Part of Speech (POS) tagging.

B. Feature Selection

B1. context features:

The part of speech of a target word is mainly decided by its context. Therefore, we can judge the part of speech of the target word according to the n words before or after the target word. Feature template is shown in table.1:

TABLE 1. THE TEMPLET OF CONTEXT FEATURES

Templet	Instructions
$C_n(n=-2,-1,0,1,2)$	The n^{th} word before or after the target word
$C_nC_{n+1}(n=-2,-1,0,1)$	Two consecutive words
$C_{-1}C_1$	Two words which are before and after the target word

B2. The word syllable features:

Tibet, belonging to alphabetic writing, is a language with plentiful forms. Its verb present tense, future tense, past tense and imperative mood are expressed by affix and additional affix. Generally speaking, Tibetan verb forms can be divided into same-rooted type and different-rooted type. In order to make a comparison with reference 11, the definition of characteristic function within words is presented below:

$$f(x_i, y_i) = \begin{cases} 1 & f(\text{prefix}(x_i = \text{"rko"} \text{and } y_i = v)) \\ 0 & \text{Otherwise} \end{cases}$$

The characteristic function definition of the last syllable of words is shown below:

$$f(x_i, y_i) = \begin{cases} 1 & f(\text{suffix}(x_i = \text{"cag"} \text{and } y_i = v)) \\ 0 & \text{Otherwise} \end{cases}$$

If the characteristics of the first and the last syllable of target word, words before or after target word and the last syllable of precursor word and the first syllable of subsequent word are mixed together, the defined syllable features are shown in table.2:

TABLE 2. THE TEMPLET OF WORD SYLLABLE FEATURES

Syllables	Instruction
w_0(prefix(w)), w_0(suffix(w))	The first and the last syllables of the target word
w_{-1}(suffix(w))	The last syllable of precursor word
w_1(prefix(w_1))	the first syllable of subsequent word

B3. The word clustering features:

As the corpus of Tibetan Part of Speech (POS) tagging is limited, methods based on statistics model will be influenced by data sparsity. For example, སྒྲ་སྒྲག (microphone)" and "ཅན་ཟིང་འཕྲུལ་ཆས། (headset) share similar

semantic. In practical corpus, training corpus can not obtain all the features because of corpus covering.

Through unsupervised clustering method, words in corpus are divided into specialized categories. In this essay, open source word2vec of Google is adopted to achieve words automatic clustering. The feature function of clustering is shown below:

$$f(x_i, y_i) = \begin{cases} 1 & f(\text{suffix}(x_i = \text{"headphone"} \text{and } y_i = c1)) \\ 0 & \text{Otherwise} \end{cases}$$

Word clustering feature template is shown is shown in table 3.

TABLE 3. THE WORD CLUSTERING FEATURES

Templet	Instructions
$C(w_0)$	The clustering of target word
$C(w_n)$ $(n = -2,-1,0,1,2)$	The n^{th} word clustering before or after the target syllable

III. EXPERIMENT AND ANALYSIS

A. Experiment Preparation

There are not the unified standard Tibetan part of speech tagging set, the University of Tibet, Qinghai Normal University, Northwest university for nationalities have their own Tagging specification. The Tibetan part of speech tagging set of Professor Qi Kunyu who is our colleague was adopted in the paper. In the basis of "information processing based on Modern Chinese part of speech tagging norms ", according to the features of Tibetan grammar, increased part of the category, the total of categories are 21 and 61 child subclasses, because of the limitation of corpus size, there only do the experiment of categories.

TABLE 4.THE TIBETAN POS TAGGING SET

No.	Category	Tagging	Explaination
1	noun	n	Common nouns, names, etc.
2	verb	v	Substantive verb, judgment verb,etc.
3	adjective	a	Adjectives of nature, state
4	adverb	d	Degree adverbs, time adverbs, etc.
5	pronoun	r	Personal pronouns, demonstrative pronouns, etc.
6	prepostion	p	Excluding prepositions, object prepositions, etc.
7	auxiliary word	u	Modal auxiliary, question auxiliary,etc.
8	numeral	m	The cardinal number, ordinal number.
9	quantifiers	q	Noun-quantifiers,verb quantifiers,etc.
10	end word	e	end words
11	Conjunction	c	Progressive conjunction, transition conjunction, etc.
12	time word	t	Time words
13	onomatopoeia	o	onomatopoeia
14	interjection	y	interjection
15	idiom	i	idiom
16	Idioms	I	Idioms
17	abbreviations	j	abbreviations
18	noun of locality	f	noun of locality
19	morpheme	g	morpheme
20	non-morpheme	k	non-morpheme
21	punctuation character	w	Punctuation character of Tibetan Chinese and English

Primary school textbooks tagging corpus of Tibetan language are applied to the paper, the statistics of tagging corpus are as shown in table 5:

TABLE 5. THE STATISTICS OF CORPUS

	Training set	Testing set
The number of sentence	2,000	126
The total number of words	102,418	7,108
The proportion of multi-category words	16.9%	11.6%
The proportion of unknown words	-	5.6%

As the corpus of Tibetan Part of Speech (POS) tagging is rare and the number of presented corpus of Part of Speech (POS) tagging with poor coverage is comparatively small, the selection of corpus may have an influence on experiment results. So the testing corpus in this paper were randomly selected from the whole corpus. The corpus for the training of word clustering features from the 200,000 words corpus of Tibetan middle school textbooks, high school textbooks.

TABLE 6. CORPUS STATISTICS

Category	Tagging	Training (%)	Test(%)
noun	n	23.9	24.3
verb	v	14.9	14.7
adjective	a	3.9	3.5
adverb	d	2.7	2.6
pronoun	r	5.0	4.0
prepostion	p	17.6	18.0
auxiliary word	u	7.1	6.8
numeral	m	3.2	3.6
quantifiers	q	0.5	1.1
end word	e	0.3	0.4
Conjunction	c	5.7	5.9
time word	t	0.9	1.0
punctuation character	w	11.8	11.1

The detail statistics information of training and testing corpus can be seen from Table VI, the part of speech distribution of training set and test set is basically same,

which indicating that the training corpus and the test corpus represent better the category distribution feature of Tibetan words. Among them, nouns, verbs, prepositions, punctuation character and auxiliary word have the larger proportion which appearing corpus.

In this paper, the tagging accuracy is used to evaluate the tagging results, and the tagging accuracy is defined as follows:

$$P = \frac{\text{The number of correct tagging words}}{\text{the total words}} \times 100\% \quad (4)$$

B. The Experimental Setup and Results Analysis

The maximum entropy toolkit1 is used to implement the maximum entropy model. According the analysis in the second part before, different features are applied to do experiments as follows, before and after dependency information of words in Table 2 are used to the following experiments, the difference is syllable features, the setting of experiments and the results of experiments are shown in table 7. The word dependency features in table 1 are represented by T1, as our baseline system, Table 2 shows the contextual word clustering features in table 2 are represented by C1.

TABLE 7. THE RESULTS OF EXPERIMENTS

Experiments	The setting of experiments	Accuracy (%)
Experiment 1	T1	90.32
Experiment 2	T1, C1	90.73
Experiment 3	T1, C1, w_0(prefix(w)), w_0(suffix(w)),	90.90
Experiment 4	T1, C1, w_0(prefix(w)), w_0(suffix(w)), w_{-1}(suffix(w_1))	90.78
Experiment 5	T1, C1, w_0(prefix(w)), w_0(suffix(w)), w_1(prefix(w_1))	91.13
Experiment 6	T1, C1, w_0(prefix(w)), w_0(suffix(w)), w_{-1}(suffix(w_1)), w_1(prefix(w_1))	90.61

Experiment 1 as the paper's datum system uses the traditional word dependence features and the accuracy is 90.32%. The accuracy of 90.73% was obtained after adding clustering features to the experiment 2. Comparing with Experiment 1, the accuracy increase by 0.41%, which proved that the effect of Tibetan part of speech tagging based on maximum entropy helped a lot.

The clustering features were added in before and after of the word syllables and other mixed features in Experiment 3, 4, 5. In which Experiment 4 achieved the best experimental results, the results of the Tibetan POS tagging based on the syllable features can be improved, after clustering features were added in before and after of the current word syllables, in the first syllables of the succeeding word.

The clustering feature of words can improve the accuracy of the Tibetan POS tagging in a certain extent and can improve the tagging accuracy of some unknown words, such as nouns, named entity by the experiments. All of this illustrates that the clustering features of the word proposed in this paper is more effective and can achieve the expected results.

C. Detail Experimental Results of the Method

Table 8 is the experimental results of the method proposed in the paper.

TABLE 8. CONTRAST EXPERIMENT OF POS TAGGING

Category	Tagging	CRF (%)	ME (%)
the total	-	89.80	91.13
noun	n	94.40	95.00
verb	v	94.00	92.00
adjective	a	69.60	72.60
adverb	d	60.50	73.20
pronoun	r	92.12	95.00
prepostion	p	98.00	96.4
auxiliary word	u	87.00	87.50
numeral	m	72.21	73.0
quantifiers	q	49.38	65.00
end word	e	89.71	100.0
conjunction	c	86.34	88.20
time word	t	42.21	46.00
punctuation character	w	97.89	98.80

D. Error Analysis

By analyzing the tagging results of the errors, the results are shown in Table 9, which shows that the errors proportion of category.

TABLE 9. TAGGING ERRORS OF ME MODEL

Tagging	Category of word	Proportion
a	adjective	9%
c	conjunction	9%
d	adverb	6%
f	noun of locality	5%
m	numeral	7%
n	noun	19%
p	preposition	6%
q	quantifier	5%
t	time word	8%
u	auxiliary word	9%
v	verb	11%

Table 9 is detail errors analysis of the Tibetan part of speech tagging based on maximum entropy, and tagging errors of verbs, nouns, adjectives words and numerals accounted for a very large proportion. Main error

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

including adjectives tagged as the nouns, numerals, verbs; numeral tagged as the nouns and auxiliary words; verbs tagged as nouns and auxiliary words; nouns tagged as adjectives frequently, and nouns can be tagged as conjunctions, adverbs, ending words, noun of locality.

IV. THE SUMMARY AND PROSPECT

This essay introduces the study work about Tibetan Part of Speech (POS) tagging of the maximum entropy which is fused with word clustering characteristic. According to the Tibetan morphological features, word clustering characteristic is added into the mixture morphological characteristics of the first and the last syllable of target word and the last syllable of precursor word and the first syllable of subsequent word. The Tibetan Part of Speech (POS) tagging is established. Word clustering characteristic can obviously raise the accuracy of Tibetan Part of Speech (POS) tagging. The accuracy of the experiment in this essay is 91.13% which is 0.81% higher than that of the benchmark system. As the corpus size in the experiments of this essay is limited, the entire performance of Part of Speech (POS) tagging needs further improvement. We hope that Tibetan language features will be further studied in the future and practical Tibetan Part of Speech (POS) tagging will be finally established.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGMENT

This work is supported by “the Fundamental Research Funds for the Central Universities of Northwest University for Nationalities” (No. zyp2015001).

REFERENCES

[1] ZHONG Chengqing, Statistics natural language processing. Tsinghua University Press, 2008.
[2] Qu Aitang. The structure and evolution of Tibetan Verbs of inflectional morphology [J]. Ethnic Chinese, 1985,1:1-15.
[3] Song Jinlan . Differentiation of morphological variants in Sino Tibetan Languages. Ethnic Chinese, 2002,1:29-33.
[4] Long Congjun. dragon army. Tibetan adjective morpheme research [J]. Journal of Chinese Language and Computing. 2006, 15 (4):193-201.
[5] Chen Yuzhong, Yu Shiwen. The research status and prospects of Tibetan information processing technology [J]. Chinese Tibet, 2003, 04:97-107.
[6] Cai Rangjia, Kyrgyzstan. Part of speech classification method of Tibetan corpus [J]. Journal of Qinghai Normal University (PHILOSOPHY AND SOCIAL SCIENCES EDITION), 2005,4:112-113.
[7] Zha Luo. Study on speech classification method of Tibetan corpus. Journal of Qinghai Normal University (PHILOSOPHY AND SOCIAL SCIENCES EDITION) 2005,4:112-114
[8] Cai Rangjia.. Words classification and tag set research of Tibetan corpus , [J]. Chinese Journal of information, 2009,4 (23) 107:112.
[9] Cai Rangjia. Computer engineering and corpus of Tibetan lexical category description method [J]. 2011, 47(4):146-148.
[10] Cai Zangtai. Hua Guanjia. Sakya Pandita Chinese Tibetan document translation system based on the dichotomy of syntax analysis method research [J]. Chinese Journal of information, 2005,6(19)7:12.
[11] Su Junfeng. Study on part of speech tagging for Tibetan texts based on HMM. [D] Northwest University For Nationalities, master's degree thesis, 2010
[12] Yang Maozhuome. Research and implementation of Tibetan POS tagging. [D] University of Tibet, master's degree thesis, 2012
[13] Yang Maozhuome, Zha Xijia. Comparison of automatic annotation of Tibetan speech [J]. Information communication, 2013,2:270-271.
[14] Zha Xijia,Gao Dingguo. Study on the integration of Tibetan text POS tagging. Journal of University of Tibet (NATURAL SCIENCE EDITION) 1, 2012 (27): 57-61.
[15] Yu Hongzhi, Li Yachao, Wang Kun, Leng bin Zahi. Study on maximum entropy fusion characteristics of the Tibetan part of speech syllable annotation[J]. Chinese Journal of information science, 2013,5 (27) 160:165.
[16] J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In Proceedings of ICML-2001, 2001:282-289.