The Study of Graininess for Tibetan Named Entity Recognition

Fei-Fei Liu[1,2,a], Zhi-Juan Wang[1,2,b*]

[1]School of Information Engineering, Minzu University of China, Beijing, China
[2]. National Language Resource Monitoring & Research Center of Minority Languages, Beijing, China
[1]Liufeifei_muc@163.com, [2]wangzj_muc@126.com

Abstract: Tibetan named entity recognition (NER), which is a fundamental part in Tibetan natural language processing, is the important subtask of Information extraction. In this paper, we surveyed the methods, effect and problems of Tibetan NER. And we discussed which kind of tokens that should be taken as the graininess for Tibetan NER task. The paper used two kinds of different graininess in a comparative experiment for Tibetan person names, location names and organization names, based on syllables, or based on words. From the result, we know that the person names based on syllable have better result than that based on words. Location names have small difference while species differ. But the organization names are more suitable based on words.

1. Introduction

Named entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, locations, organizations, expressions of times, quantities, monetary values, percentages, etc. Tibetan NER is the fundamental and key subtask for Tibetan information extraction and text mining in Tibetan natural language processing.

Nowadays, named entity recognition had achieved good results in various languages, such as English. State-of-the-art NER systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%. However, Tibetan NER started late. It has yielded a great number of positive results， but is still a new study field in which there are series of problems.

This paper will summarize the present situation of Tibetan NER, introduce the existing research methods, the results obtained and the existing problems. By summarizing the results of studies, we find out that there are two kinds of graininess for Tibetan NER, including syllables, and words. Syllables are the basic units in Tibetan, and words created by syllables based on word segmentation.

Then the paper discusses the granularity problem in Tibetan NER by a comparative experiment. In addition to the graininess, syllables or words, all attributes are the same for Tibetan person names, location names and organization names. By comparing the accuracy, recall, F-score, we will get the conclusion that which graininess is more suitable.

The rest of the paper is organized as follows: In section 2, we briefly describe the background of Tibetan NER, including current situation, difficulties, methods, problems and improvement direction. In section 3, we introduce the algorithm of Tibetan named entity recognition based on conditional random field(CRF), this method is used in our comparative experiments about graininess. We report in section 4 our experimental results and give our conclusions on this work in section 5.

2. Tibetan Named Entity Recognition

The basic work of Tibetan natural language processing (NLP) includes Tibetan word segmentation, Tibetan POS tagging and named entity recognition. Nowadays, there are many practical methods in Tibetan word segmentation, such as automatic Tibetan word segmentation scheme based on lattice auxiliary word and continuation feature [3], Tibetan language word segmentation system [4], SegT [5], Yangjin [6], TIP-LAS [7], but for Tibetan NER, the research conclusions are immature. As the fundamental part in Tibetan NLP, there are many aspects of the need for further study and improve in Tibetan NER.

2.1 Introduction to Tibetan

Tibetan (བོད་ཡིག) refers to the use of Tibetan language Tibetan. The glyph structure is a letter as the core, the rest of the letters are based on this before and after the additional and overlapping from top to bottom, combined into a complete word table structure. Writing habits from left to right. The font is divided into "head" and "headless" two categories.

Tibetan is a phonetic alphabet, with 30 consonants and 4 vowels. One Tibetan syllable can have 1 to 7 basic characters, if you consider Sanskrit, characters may be more. The seven basic characters have a base character and a vowel, the other characters were added to the base word, the up, down, front, back, and then back[1][2].

There are fewer types of punctuation in Tibetan .Tibetan various syllables separate with a small point, this point named the syllable node (་).In addition to the syllable node, the most common punctuation is a single vertical line (།), as a full stop, colon and other situations. And the paragraph ends with a double vertical line (།།).

2.2 Methods of Tibetan NER

The methods of Tibetan NER can be divided into rule-based methods and based on supervised machine learning methods.

Rule-based methods

In the early days, the study of Tibetan NER was based on a rule-based approach. Yu et al. used a rule-based model based on case-auxiliary word and lexicon, and also adapt boundary information list static from large corpus to improve recognition. And experiments shows that recall rate and precise rate are respectively 90.13% and 94.02% in the newspaper corpus, 85.67% and 88.20% in the website text. Sun et al. used the internal features of names, contextual features and boundary features of names, and establishes the dictionary and feature base of Tibetan names. The results prove the algorithm is effective with 0.8391 F-score. Dou et al. used the Statistical Method of Mutual Information to, combining the rules of lattice auxiliary and the dictionary of person names, F value in the test can be up to 93.55%.

Supervised machine learning methods

After 2014, supervised machine learning methods are increasingly applied to Tibetan NER. Jia et al. came up with Maximum entropy (ME) and conditional random field(CRF), and the F-score of the recognition of names can be 92.08%. Hua et al. proposed a syllable features with Perceptron training model to identify Tibetan name entity with detail analysis NE structure rule and word segmentation ambiguity. The F-score of NE identification is 86.03% for the testset. Kang et al. defined a feature tag set to fit in with the characters of Tibetan names, used CRF as tagging model to train and test corpus data. The highest F-score obtained in the experiment can reach 94.31%. Zhu et al. studied Tibetan name recognition technology using conditional random fields (CRF) principle， focuses on analysis of the internal structure of the Tibetan names, contextual features, feature selection and data preprocessing, etc. and evaluated the effectiveness of different features through experiments. The recognition rate of Tibetan names can reach 80% of F-score.

2.3 Difficulties in Tibetan NER

Tibetan belongs to the Sino-Tibetan language family. In theory, the natural language processing methods used in Chinese can be used in Tibetan information processing. But in practice, it must be considered in the specific problems. The main difficulties in Tibetan NER are as follows:

Tibetan is a complex system of phonetic logic. The basic unit of the sentence is syllable. Syllables are separated by syllable node. One syllable or more syllables constitute words. There is no obvious mark between the word and next word. The boundaries of named entities are difficult to determine. And too few punctuation types, just single vertical line (།) and double vertical line (།།), will make the too long analysis object length, increasing the difficulty of recognition algorithm.

There is no morphological difference between named entities and unnamed entities in Tibetan. Unlike English, the person names, location names and organization names in English with the capitalized first letter, are easy to extract. And compared to Chinese person name, most of the Tibetans do not have the family name and the length of the name which can be from single syllable to twenty-six syllables.

The name dictionary, the labeled corpus and other related resources is insufficient. Nowadays, the main method of Tibetan Named Entity Recognition is supervised learning algorithms which require large-scale of labeled corpus. But Tibetan resource is not easy to obtain.

2.4 Summary of Tibetan NER

All kinds of methods are proven the feasibility and accuracy in the Tibetan NER. However, due to the different corpus of different methods, we cannot judge which method is better based on the experimental results alone. Today, the problems in Tibetan NER: the conflicts between Tibetan names and ordinary words, the misinterpretation of translations, and the difficulties in identifying Tibetan NE boundaries.

In the future, building a large amount of manually annotated training data is urgent for improving effect. It is necessary to improve the recognition accuracy of Tibetan named entities by using syntactic information, extending boundary information, making full use of boundary information, expanding the translation of the thesaurus, and testing other possible recognition models such as Support Vector Machine (SVM).

3. Tibetan NER based on CRF

3.1 Description

CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text.

Specifically, CRFs find applications in shallow parsing, named entity recognition.

Lafferty, McCallum and Pereira define a CRF on observations X and random variables Y as follows:

Let G = (V , E) be a graph such that Y=(Yv)v∈V, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field when the random variables Yv , conditioned on X, obey the Markov property with respect to the graph: p(Yv |X, Yw ,w≠v)=p(Yv |X, Yw ,w~v), where w~v means that w and v are neighbors in G.

What this means is that a CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets X and Y, the observed and output variables, respectively; the conditional distribution p(Y|X) is then modeled.

By now, CRF has become a widely used technique which is applied in named entity recognition on low resource language [17], such as Hindi, Bengali, Tamil, and Telugu [18].

3.2 Tibetan NER based on CRF

Tibetan NER can be defined as a sequence labeling problem for determining whether a observations belongs to a labeled set of markers. Suppose that a given marker sequence y= (y_1, y_2…, y_n) is labeled, n is the length of the sequence. The sequence of Tibetan NE is represented as w= (w_1, w_2,⋯,w_m),m is the length of the NE. The model of CRF is defined as follows:

$$p(y|w) = \frac{1}{Z(w)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_i, y_{i-1}, w)\right)$$

Z (w) is normalization factor, determined by the observation sequence.

$$Z(w) = \sum_y \exp\left[\sum_k \lambda_k f_k(y_i, y_{i-1}, w)\right]$$

λ_k is the weight of the k-th function, f_k (y_i,$y_{(i-1)}$,w) is a characteristic function.

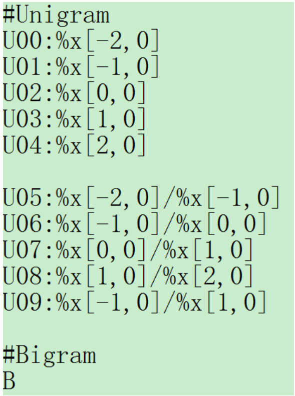
$$f_k(y_i, y_{i-1}, w) = \begin{cases} 1, \text{if } y_i = u, \text{and } y_{i-1} = v \\ 0, \text{otherwise} \end{cases}$$

3.3 Design of Feature Template

The use of the CRF model for a feature set is defined by a fixed pattern of feature templates, by a feature template to get features in a context (or window). The larger the window is, the better the relationship between the current word and the context can be observed, and the long distance dependency can be found in the text. But when the window is too large, the model will take a long time. This will lead to a decline in overall performance.

Considering the relationship between the Tibetan NE and its context correlation, the length of the window is set to 5, which can achieve a balance between training time and recognition effect.

The characteristics of the template used in the experiment shown in Figure 1.



```
#Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]

U05:%x[-2,0]/%x[-1,0]
U06:%x[-1,0]/%x[0,0]
U07:%x[0,0]/%x[1,0]
U08:%x[1,0]/%x[2,0]
U09:%x[-1,0]/%x[1,0]

#Bigram
B
```

Figure 1. Feature Template of CRF

4. Comparison Experiment for Graininess

4.1 Corpus Pretreatment

In the experiment, the corpora which built from text of the Tibetan version of People's Daily online website published. The train data is 5.97M and the test data is about 700K, which were included 2546 person names, 6469 location names, 4049 organization names.

In order to examine the effect of different graininess on Tibetan NER， We designed a comparative experiment used CRF on Tibetan person names, location names and organization names , based on syllables, or based on words. The syllable-based markers : B-PER(person names' first syllable),I-PER(syllable in person names' but not the first syllable),B-LOC(location names' first syllable),I-LOC(syllable in location names' but not the first syllable),B-ORG(organization names' first syllable),I-ORG(syllable in organization names' but not the first syllable),O(remaining syllables). The word-based markers: PER (person names), LOC (location names), ORG (organization names), O (remaining words).

We use Precision (P), Recall (R), F1 to evaluate the performance of each graininess, which are very common in NLP evaluation.

P= (number of correctly identified NE) / (number of identified NE)

R= (number of correctly identified NE) / (number of all NE)

F1= (2*P*R) / (P+R)

4.2 Result and Analysis

The result of the experiment is in table 1. We labeled “Words as the graininess” as group A, and” Syllables as the graininess” as group B.

Table1 Evaluation result for experiment

Tibetan NE	Words as the graininess(A)			Syllables as the graininess(B)		
	P	*R*	*F1*	*P*	*R*	*F1*
PER	87.72	34.25	49.26	76.95	77.74	77.34
LOC	97.57	74.31	84.36	88.38	82.84	85.52
ORG	92.11	71.92	80.77	88.66	58.90	70.78

As can be seen from the experimental results in Table 1, we get the following several conclusions.

The precision of group A is higher than that of group B. But its recall is short of group B, except in ORG. Compared F1, which considered as comprehensive effect, differs greatly in PER and ORG. For PER, the score of group B is 28.08% higher than that of group A. But for ORG, B is 9.99% less than A. It can be considered that the granularity of the syllable is small, more data can be obtained, which has a great influence on the recall rate. This also shows that the syllables-based approach because of the smaller particle size, can partially solve the problem of data sparseness.

For PER, although the Precision of group A is significantly higher than group B, but the recall is serious losses, resulting in poor F, while the results of group B were stable. This means that the identification of Tibetan person names should be based on syllables to achieve good results without Tibetan word segmentation.

For LOC, the Precision of group A is better than B, but the Recall is less. However, the F1 is slight difference between A and B. So, we can get conclusions that the identification of Tibetan location names can use both methods. But based on syllables, Tibetan word segmentation can be omitted.

For ORG, all the evaluation results in a group A are higher than those in group B. It indicates that the better results can be achieved by Tibetan word segmentation for organization names. Based on the analysis, we think organization names are complex, nesting and more syllables. These lead to difficulties in boundary identification based on syllables. But the boundary is confirmed on the method based on words.

5. Conclusions

In this paper, to get which kind of tokens that should be taken as the graininess for Tibetan NER, we did a comparison experiment for graininess. The paper used two different graininess, syllables or words, for identification of three kinds of named entities, including Tibetan person names, location names and organization names. From the result, we know that the method which syllable as graininess for person names is better. But the organization names are more suitable based on words. And both are fit into location names. In other words, identification of person names and location names can achieve very good results without Tibetan segmentation. But it is inconformity for organization names.

Acknowledgment

The project was supported by Key Program of National Natural Science Foundation of China (Grant No. 61331013), National Language Committee of China (Grant No. WT125-46), Minzu University of China Tier-1 Universities and Tier-1 Subjects Graduate independent research projects (Grant No. 10301-0170040601-184).

References

[1] LIU Huidan Rui Jianwu, Wu Jian. Encoding Detection and Conversion of Tibetan Web Pages [C] // 25th Anniversary Conference of Chinese Information Society of China.

[2] Wu Jian, Rui Jianwu, Liu Huidan. Tibetan web page and its code identification method: CN, CN 101055593 A[P]. 2007.

[3] Chen YZ, Li BL, Yu SW. The Design and Implementation of a Tibetan Word Segmentation System[J]. Journal of Chinese Information Processing, 2003.

[4] Cai Z J, Cai R. Design of a Tibetan Word Segmentation System[J]. Computer Engineering & Science, 2011, 33(5):151-154.

[5] Liu H, Minghua N, Zhao W, et al. SegT:A Practical Tibetan Word Segmentation System[J]. Journal of Chinese Information Processing, 2012, 26(1):97-103.

[6] Shi XD, Lu YJ. A Tibetan Segmentation System—Yangjin[J]. Journal of Chinese Information Processing, 2011, 25(4):54-56.

[7] Li YC,Jiang J, Jia YJ, Yu HZ.TIP-LAS: TIP-LAS: An Open Source Toolkit for Tibetan Word Segmentation and POS Tagging [J]. Journal of Chinese Information Processing, 2015, 29(6):203-207.

[8] Li YC, Jia YJ, Zong CQ, Yu HZl. Research and Implementation of Tibetan Automatic Word Segmentation Based on Conditional Random Field[J]. Journal of Chinese Information Processing, 2013, 27(4):52-58.

[9] Sun Z, Wang. Overview on the Advance of the Research on Named Entity Recognition[J]. New Technology of Library & Information Service, 2010.

[10] Yu HZ, Jiang T, Ma N.Named Entity Recognition for Tibetan Texts[J].Lecture Notes in Engineering and Computer Science,2010,2180.

[11] Dou R, Jia YJ, Huang W.Automatic recognition of tibetan name with the combination of statistics and regular. Journal of Changchun Institute of Technology (Social Science Edition), 11 (2) 113-115.,2010.2:113-115.

[12] Sun Y, Yan X, Zhao X, et al. Research on automatic recognition of Tibetan personal names based on multi-features[C]// International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2010:1-5.

[13] Jia Y, Yachao L I, Zong C, et al. A Hybrid Approach to Tibetan Person Name Identification by Maximum Entropy Model and Conditional Random Fields[J]. Journal of Chinese Information Processing, 2014.

[14] Hua Q, Jiang W, Zhao H, et al. Tibetan name entity recognition with perceptron model[J]. Computer Engineering & Applications, 2014.

[15] Kang C, Long C, Jiang D. Tibetan names recognition research based on CRF[J]. Computer Engineering and Applications, 2015.

[16] Zhu J, Li T, Liu S. Research on Tibetan name recognition technology under CRF[J]. Journal of Nanjing University, 2016.

[17] Hänig C, Bordag S, Thomas S. Modular classifier ensemble architecture for named entity recognition on low resource systems[C]//Workshop Proceedings of the 12th Edition of the KONVENS Conference. 2014: 113-116.

[18] Das A, Garain U. CRF-based Named Entity Recognition @ICON 2013[J]. Computer Science, 2014