Named Entity Recognition for Tibetan Texts Using Case-auxiliary Grammars

Hongzhi Yu, Tao Jiang and Ning Ma

Abstract—**Tibetan named entity recognition, which is an important part in multilingual processing, is one of the challenging tasks of text mining. In this paper, we present a *recognition* system of Tibetan person name using a rule-based model based on case-auxiliary word and lexicon, and also adapt boundary information list static from large corpus to improve recognition. In addition, our 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.**

Index Terms—**named entity recognition, case-auxiliary words, Tibetan texts, name lexicon**

I. Introduction

The goal of named entity recognition (NER) is to recognize phrases in a document that indicate the names of persons, organizations, locations, times or quantities. As an important subtask of information extraction and text mining, NER has been attracting more and more attention in the NLP community. English NER has been well developed and a wide variety of machine learning methods have been applied, including Hidden Markov Models (HMMs)[1][2], Maximum Entropy Models (ME)[3], and Support Vector Machines (SVMs)[4]. Recently, a number of methods have been reported for Chinese NER. Sun et al. proposed a class-based language model approach to Chinese NER [5]. In their work, they used different models to identity different types of NEs in Chinese text, including a character-based trigram for person, a word-based model for location and a more complicated model for organization. Zhang et al. put forward a stochastic role model to recognize Chinese NEs[6]. In this work, they defined a set of roles about component tokens within a Chinese NE and the relevant contexts. Their experiment showed that the role-based model was effective for different NEs. Guo et al. presented a robust risk minimization (RPM) classification method to Chinese NER, which was able to incorporate the advantages of character-based and word-based models [7].

However, person name recognition in the Tibetan text presents a special difficulty in NER, and there is no good performance has been achieved yet in this field. In addition to the challenging difficulties existing in the counterpart problem in English and Chinese, the problem also exhibits the following more difficulties: (1) In a Tibetan text, unlike English, there is no space to mark the word boundary and no standard definition of words. 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. (2) The statistical NER systems typically require a large amount of manually annotated training data, as there is not amount of labeled training data available in Tibetan at the moment.

This paper will focus on person name recognition in Tibetan texts, which is a very important problem in multilingual text processing. Based on the case-auxiliary words and lexicon, we propose a simple rule-based approach for person name recognition in Tibetan texts, which includes Tibetan person name recognition and translated person name recognition. The grammar of Tibetan language is strict, and case-auxiliary word is a very important part of the grammar. Case-auxiliary is regarded as useful cues in text processing, and it always appear in the left or right of the name, therefore we can use it to recognize person name.

The rest of the paper is organized as follows: In section 2, we briefly describe the background of the Tibetan person name. In section 3, we show the features which are used to identify person name. In section 4, we discuss the person name lexicon and present a rule-based algorithm for person name recognition in Tibetan texts. We report in section 5 our experimental results and give our conclusions on this work in section 6.

II. Background of Person Name

Tibetan language belongs to the Tibetan-Burman branch of the Sino-Tibetan language family which is widely used in all areas inhabited by Tibetans. It is alphabetic language which consists of thirty consonants, four vowels, five inverted letters (for the renting of foreign words) and the punctuations.

Different from Chinese person names, most of the Tibetan person names don't have surname. Only the upper-class gets it, which includes nobility、Living Buddha and officials in the old days and so forth. A large majority of the person names are consists of four syllables, and each of them has two syllables, few have three syllables which appear in the Kang and Amdo regional dialect. The upper-class of religious figure often has a long name, which is up to twenty-six

Manuscript received December 10, 2009. This work was supported in part by the National Nature Science Fund under Grant No. 60970071.

Hongzhi Yu is with the Key Laboratory of National Languages Information Technology, Northwest University for Nationalities, lanzhou, gansu 730030, China (e-mail: yhz@xbmu.cn).

Tao Jiang is with the Key Laboratory of National Languages Information Technology, Northwest University for Nationalities, lanzhou, gansu 730030, China (e-mail: xinxiyuanjt@126.com).

Ning Ma is with the Key Laboratory of National Languages Information Technology, Northwest University for Nationalities, lanzhou, gansu 730030, China (e-mail: maning@xbmu.edu.cn).

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

syllables. For example, རྗེ་བཙུན་འཇམ་དཔལ་ངག་དབང་བློ་བཟང་ཡེ་ཤེས་བསྟན་འཛིན་རྒྱ་མཚོ་སྲིད་གསུམ་དབང་བསྒྱུར་མཆུ་ངས་པ་མེད་སྤྲོ་དཔལ་བཟང་པོ (14th Dalai Lama). What's more, there is a habit of choosing the first and third syllable for short in the four syllables person names, which is common in speaking and written forms. For example, བསོད་ནམས་དར་རྒྱས (Sonamthargye), བསོད་དར (Sodar) for short[8].

Here are some rules when people choose a name:

- Use words which mean good wishes.
  - Tashi (auspicious, prosperity) བཀྲ་ཤིས
  - Tsering (longevity) ཚེ་རིང་
- Use words which mean glorification.
  - Loden (wisdom) བློ་ལྡན
  - Phakpa (preeminence) འཕགས་པ
- Use Buddhist terminology.
  - Dorjee (vajra) རྡོ་རྗེ
  - Jamyang (Manjushri) འཇམ་དཔྱངས
- Use things form nature as person name.
  - Gyatso (ocean) རྒྱ་མཚོ
  - Dawa (moon) ཟླ་བ
- Use the date of birth.
  - Nima (birth on Sunday) ཉི་མ
  - Pasang (birth on Friday) པ་སངས

III. Features Description

The key of person name recognition is to identify left and right boundary words of the name. Some features of such words are identified by analyzing the 30 megabyte corpus, which consist most of the Tibet Daily. The feature includes case-auxiliary words, title words, punctuation, affix, etc. All of that can be summarized as follow:

A. Case-auxiliary words

Tibetans express themselves grammatically with the help of function words and different order of the words. Certain auxiliary words originating from verbs are playing an important role similar to that of the function words. Auxiliary verb can be added at the end of the verbs to indicate tense; it can be added between words or phrases to indicate different relationship between components of the sentence; it can also be added at the end of the sentence to indicate the mood. Generally speaking, there are three kinds of case-auxiliary words, which always appear in both of the left and right of the person name. A kind of case-auxiliary word appears between the subject and the transitive verbs, which include five forms, that is གྱིས, ཀྱིས, གིས, ཡིས, ～ས. Another kind is used to modify nouns and denote dependency relation, which consist of གྱི, ཀྱི, གི, འི, ཡི.

Table1 shows the occurrence frequency of the case-auxiliary word in the boundaries, which based on the statistic of our corpus. The statistical data shows that the case-auxiliary is very common and important boundary information.

The key of person name extraction is to identify left and right boundary words of the name. Some features of such words are identified by analyzing the 30 megabyte corpus, which consist most of the Tibet Daily. The feature includes case-auxiliary words, title words, punctuation, affix, etc. All of that can be summarized as follow:

Table1: Frequency of the case-auxiliary word

Location	Words	Times	Frequency%
Left boundary	གྱིས་ཀྱིས་གིས་ཡིས་ས	100	0.358
Left boundary	སུ་ར་རྣ་དུ་ན་ལ་ཏུ	50	0.179
Right boundary	གྱིས་ཀྱིས་གིས་ཡིས་ས	11900	42.652
Right boundary	གྱི་ཀྱི་གི་འི་ཡི	900	3.226
Right boundary	སུ་ར་རྣ་དུ་ན་ལ་ཏུ	200	0.717

B. Title Words

A person's title and appositive phrases are the mainly heuristic information for name identification, which helps us identify the candidates of person name.

- Post title
  - Chairman ཀྲུའུ་ཞི
  - Minister དུའུ་ཀྲང
- Professional title
  - Teacher དགེ་རྒན
  - Educator སློབ་གསོ་བ
- Specialty title
  - Father ཨ་པ
  - Mother ཨ་མ

C. Honorific Affix

Some honorific words are frequently appeared in front of or behind a person's name. རྒན, which is used in front of a person name, meaning a teacher or an elder. For example, རྒན་སྒང་པ, and སྒང་པ is the name. While ལགས is used behind of a person name to expresses reverence. For example, མིག་དམར་ཆོ་ཉིང་ལགས, མིག་དམར་ཆོ་ཉིང is the

name.

IV. Rules for Person Name Identification

Considering the huge number of Tibetan person name and translated person name, it is not realistic to only maintain a fixed list of person name. Then combining the name lexicon and the feature of the boundary words, we propose a rule-based language model for person name recognition in Tibetan texts.

A. Name Lexicon

There are three lexicons used in our system, that is, common Tibetan person name lexicon, title lexicon and translated Chinese surname lexicon. The first lexicon contains 3530 the most commonly used Tibetan person name, which are taken from a Tibetan person name dictionary[9]. The second lexicon contains different titles which are attached in front of or behind a name. And the third one contains major surnames of Han Nationalities, which are translated into Tibetan.

B. ReferencesBoundary and Name Information list

Our rules are based on a set of word lists, cue-character lists getting from the corpus, as follows:

1) left_n_list: Left boundary words with title and some nouns.
2) left_u_list: Left boundary words with case-auxiliary.
3) left_p_list: Left boundary words with punctuation.
4) left_Ex_list: Exclusive words in the left boundary words.
5) right_n_list: Left boundary words with title and some nouns.
6) right_u_list: Left boundary words with case-auxiliary.
7) right_p_list: Left boundary words with punctuation.
8) right_Ex_list: Exclusive words in the right boundary words.
9) Inner_f_list: High frequency words of the name.
10) Inner_Ex_list: Exclusive words of the name.
11) Inner_h_list: Surnames of Han Nationality.
12) r_list: Personal pronoun list.

C. Algorithm

Step 1：Checking the left and right boundaries: for a word W, if $W \in$ left_n_list or $W \in$ left_u_list or $W \in$ left_p_list and $W \notin$ left_Ex_list, then W is recognized as a left boundary word. Likewise, for a word W, if $W \in$ right_n_list or $W \in$ right _u_list or $W \in$ right _p_list and $W \notin$ right _Ex_list, then W is recognized as a right boundary word.

Step 2: Finding the candidates: if a word is between the left boundary word and the right boundary word, then it is a person name candidates. Only if the word's left or right is the boundary word, we use the lexicon to check it.

Step 3: Determining the person name: if the candidate contains the Tibetan word, which belongs to Inner_f_list or Inner_h_list, then it is a person name. While the candidate contains the Tibetan word, which belongs to Inner_Ex_list or r_list, then it is not the person name.

V. Experiments

We have built a person name recognition system using a mix of rules, lexicons and some training strategies. Our system was implemented on Windows 2003 platform and was programmed using C# with the database support of Microsoft Access 2003.

In order to test the effectiveness of the methods proposed above, an experiment is conducted on large scale corpus. The corpora which built from text of Tibet Daily published during February 2007 and text of China Tibet Online website published during December 2008 is selected for experiment. We use Precision (P) and Recall (R) to evaluate the performance of our system, 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)

Using the recognition algorithm and test corpus mentioned above, the result of the experiment is in table 2:

Table2：The result of the experiment

	Corpus	P	R
Close	Daily 7.2.15-2.25	94.02%	90.13%
Open	Web 08.12.11	88.20%	85.67%

In [10], a rule-based name recognition for Xinjiang Minority Languages was proposed. Compared to the result of the paper, the precision is 75% and recall scores is 80%, our result enjoys a good advantage in close and open test. One of the reasons is that various types of information list and grammar information play a major role in person name recognition.

VI. Conclusion

In this paper, we proposed a rule-based method to recognize person name for Tibetan texts, which combines the lexicon and case-auxiliary words information. We created lots of boundary and name information lists to improve system performance. The result tested from large-scale corpus verified the validity of the method.

In the future, we plan to build a large amount of manually annotated training data. Then we can use more statistical information to improve system performance. We also plan to build system that handles personal names, location and organization names simultaneously, so that all the named entity can be recognized.

Acknowledgment

Authors would like to thank Yang-Kyi Jam for her help with labeling the corpus.

REFERENCES

[1] Bikel ,D.M., Schwartz, R., and Weischedel, R.M., "An algorithm that learns what's in a name", Machine Learning, 34, 1-3,1999, pp.211-231
[2] Zhou, G.D., and Su J.," Named entity recognition using an HMM-base chunk tagger", in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL-2002), Philadelphia, USA, 2002,pp.473-480.
[3] Borhwick, A., "A maximum entropy approach to named entity recognition", Ph.D. Thesis, New York University, 1999.
[4] Isozaki, H., and Kazawa, H., "Efficient support vector classifiers for named entity recognition", Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, 2002, pp.953-959.
[5] Sun, J., Gao, J., Zhang, L., Zhou, M., and Huang, C., "Chinese named entity identification using class-based language model", Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, 2002,pp.967-973.
[6] Zhang, H.-P., Liu, Q., Yu, H.-K, Cheng, Y.-Q, and Bai, S.," Chinese named entity recognition using role model", Computational Linguistics and Chinese Language Processing, 8, 2(2003), 2003,pp.29-60.
[7] Guo, H., Jiang, J., Hu, G., and Zhang, T.," Chinese named entity recognition based on multilevel linguistic features", Proceedings of the First International Joint Conference on Natural Language Processing, Sanya, Hainan Island, China, 2004,pp.294-301.
[8] Wang, G., "A Study of Tibetan Name, Publishing House of Minority Nationalities", Beijing,1991.
[9] Chen, G.S. and An Caidan, "Dictionary of Common Tibetan Personal and Place Names", Beijing Foreign Language Press, Beijing, 2004.
[10] Gulila A., "Rule-base Person Name Recognition for Xinjiang Minority Languages", Journal of Chinese Language and Computing, 15,4(2005), 2005,pp.219-226

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)