Research on Tibetan Culture Domain Entity Recognition

Yinghui Feng[1, a], Zhijuan Wang[1,2, b]

[1]College of Information Engineering, Minzu University of China, Beijing 100081, China;

[2] Minority Languages Branch, National Language Resource Monitoring & Research Center, Beijing 100081, China.

[a]fengyinghui_muc@163.com, [b]Wangzj.muc@gmail.com

Keywords: Named Entity Recognition, Tibetan Culture Domain, Bootstrapping, Maximum Entropy.

Abstract. Named Entity Recognition (NER) is the premise of other tasks in Information Extraction. At present, most NER studies are focus on person names, place names and organization names. However, domain entity recognition is still a challenging task. Tibetan culture domain entity recognition has important significance for studying Tibetan culture. This article extracts domain keywords based on improved TextRank algorithm. Then domain words bank is structured using domain keywords, and word segmentation is conducted. On the basis, Tibetan culture domain entities are recognized based on the improved Bootstrapping. The method in this article has better extracting performance and good generalization.

Introduction

With the popularity of computers and the rapid development of Internet, huge volume of information floods in the form of electronic files in front of people. In order to meet the challenge brought by information explosion, there is an urgent need for automatic tools to find the exact information among the magnanimous information sources. Information Extraction (IE) is generated based on this background. Information Extraction refers to extracting specific information from non-structured data and forming structured data pool for people to search and use. The mainly tasks of Information extraction are: named entity recognition, entity relation extraction and event extraction. Information extraction technology can be used to extract all kinds of information from texts to databases. Then data mining technology can be used to discover knowledge from a database. This is a viable approach of knowledge mining from texts. Moreover, the application of information extraction can also improve the precision and recall of text retrieval.

Information extraction has been developed and applied well in many fields such as military, economy, medicine and sports field. However, there is little research on Tibetan culture domain. Tibetan culture is colorful. Tibetan culture domain-oriented information extraction is very important for studying Tibetan culture.

Named Entity Recognition (NER) is the premise of other information extraction tasks. The recognition of proper names, such as person name, place name and organization name and numeral phrases are the mainly tasks of named entity recognition. This paper mainly research on Tibetan culture domain entity recognition.

Related researches on domain entity recognition

The initial entity recognition mainly focuses on person name [1, 2], place name [3] and organization name [4]. With continuous development of information extraction, information extraction technology is gradually applied into different fields, such as military, economy, medicine and sports field. For example, sport field information extraction system can extract interesting sporting events results from news texts, including event name, place, competing teams, performance and so on; Military field information extraction system can extract details of terrorist incidents from military field news, including time, place, criminals, victims, targets, weapons and so on; Economy field information extraction system can extract the personnel changes in the company from economy

news, including company name, job title, successor, job leavers and so on; Medicine field information extraction system can extract symptoms, diagnostic records, test results, prescription and so on. The extracted information is presented in a structured manner and stored into databases for diversified applications such as question answering and so on [5].

As the complexity of domain entity types and the lack of training corpora, common methods cannot be used in domain entity extraction completely.

Named Entity Recognition. Named entity recognition can be divided into two methods: named entity recognition based on semi-structured data and named entity recognition based on nature language. Named entity recognition based on semi-structured data roughly be divided into method based on wrapper and method based on HTML; Named entity recognition based on nature language mainly including two methods: Rule-based method and statistics-based method; Hybrid approach can combine the advantages of two methods mentioned above [6].

1) Named entity recognition based on semi-structured data

Semi-structured data is between structured data (database) and nature language data. HTML document is a kind of semi-structured data. Web document is a typical semi-structured data. This paper researches on Tibetan culture domain named entity recognition by getting relevant topics information. Named entity recognition based on semi-structured data is roughly divided into method based on wrapper and method based on HTML.

The method based on wrapper is first proposed by Nicholas Kushmerick in 1997[7]. The method automatically analyzes structural features of information to be extracted in web pages. The main idea is getting extraction rules using inductive learning approach. Method based on HTML locates information according to the structure of web pages. Parse web documents to syntax trees using resolver before information extraction. Then extract entities from syntax trees. The typical systems using this information extraction technology are LIXTO [8], XWRAR [9] and so on.

2) Named entity recognition based on nature language

Named entity recognition based on nature language mainly including two methods: Rule-based method and statistics-based method. Rule-based method always uses rule templates made by specialists in linguistics. The features include statistics, punctuations, keywords, indicators, locality, the position of the words (eg, the end of the word), and the center words. The mainly methods are pattern matching and string matching. Rule-based method is the earliest used method. The representative foreign systems are ANNIE system in GATE program, FACILE system in MUC evaluation and so on [10]. In China, Wang Ning[4] recognized financial field company names based on rules. The system has a strong dependence on knowledge base. And the method has some limitations in the closed test and the open test.

Because rule-base systems have restrictions on extracting ability, people try to investigate the new approaches to improve performance of recognizing named entity. Besides, the advent of large-scale tagged corpus makes it possible for processing language information using tagged corpus. Statistics-based method is widely used in nature language information processing. The fundamentals of the machine learning based on the statistical learning theory are briefly introduced following. First, part of tagged corpus is selected as training corpus. Second, correlated features are extracted according to a certain strategy. Thirdly, the target model can be got by learning an algorithm. Finally, corpus is predicted using the model. The common statistical methods for named entity recognition are: N-Gram, HMM, Maximum Entropy, CRF, SVM, A decision tree and so on.

3) Hybrid method

Named entity recognition based on semi-structured data and named entity recognition based on nature language can be combined. Rule-based method and statistics-based method can be combined. Many researchers combine several models.

Domain entity extraction. More detailed and complex rules are required for extracting domain entity. And statistics-based method is combined usually. HMM, Maximum Entropy, CRF, SVM are the common statistical models for domain entity extraction. Domain words and domain knowledge base are required for some domain. Tagging corpus, especially automatic tagging corpus, is also very important.

Introduction of Tibetan culture domain entity

The Tibetan culture is extensive and profound, and it has a long history. Tibetan culture includes literature, art, region, calendar, Tibetan medicine and other different kinds of specific culture forms. And the deeper cultural awareness, such as ethics, mental, aesthetic, is included. Besides, Tibetan culture also includes ontology, epistemology, practice and other deeper thought contents. All of these things develop and change with Tibetan society developing and changing. It is with the development and change, Tibetan culture displays great vitality and capabilities of cultural integration in different historical stages.

Tibetan culture domain entity. Domain entity recognition is a necessary step of researching domain entity relation. There are many other domain entities besides person name, place name and organization name, as shown in Table 1.

Table 1 Tibetan culture domain entity class and examples

Class	Example	Meaning
Religious sect	Nyingma sect	A important sect in Tibetan buddhism.
Title	Renboqie	A title of respect for a master of Tibetan Buddhism
Living Buddha system	Lama	A title given to a spiritual leader in Tibetan Buddhism
Temples	Dazhaosi	A Tibetan Buddhism temple
Festival	Sour Milk Drinking Festival	A traditional festival in Tibet
Culture	Tibetan Medicine	A traditional medicine in Tibet

The difficulties in recognizing Tibetan culture domain entity. Word segmentation is a must before recognizing entity, but the most of Tibetan culture domain entities are not included in common dictionaries. The accuracy of word segmentation directly affects domain entity recognition. Tibetan culture domain corpus is less, the kinds of domain entities are various and tagging is complex.

Tibetan Culture Domain Entity Recognition

Fig.1 shows the framework of Tibetan culture domain entity recognition. It is divided into the two modules: keywords extraction module and entities recognition module. Tibetan culture domain keywords can be extracted using the method based on improved TextRank algorithm. A new Tibetan culture domain words can be gotten by collecting keywords. These domain words can be used for word segmentation. Then part of speech tagging is made again. Finally, automatic learning field words based on bootstrapping is used for extracting Tibetan culture domain entities.

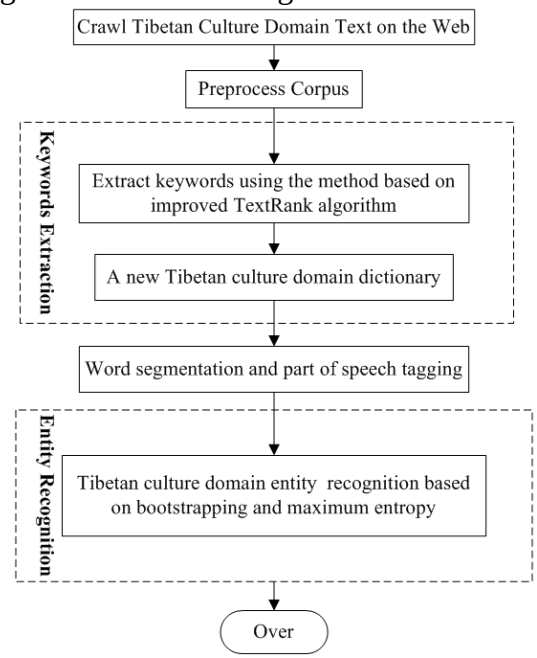

Fig. 1 The framework of Tibetan culture domain entity recognition

Keywords extraction. As most of Tibetan culture domain words are not included in common dictionaries, the precision of word segmentation is not high. However, the precision of entity recognition builds on the foundation of the precision of word segmentation. There is a need to build a Tibetan culture domain dictionary. This paper extracts Tibetan culture domain keywords using the method based on improved TextRank. Then Tibetan culture domain dictionary is gotten by proofreading.

Tibetan culture domain entity recognition based on the improved bootstrapping. Supervised learning algorithms need better tagged corpus and the large-scale tagged corpus costs much human and material resources. The common tagged corpuses are static corpus tagged several years ago. The corpuses cannot reflect the current language characters, and they are difficult to be updated. However, it is convenient to get untagged corpus, as the corpus scale is very large in the Internet. Semi-supervised learning methods are more suitable. The typical algorithms are bootstrapping [11, 12] and maximum entropy [13]. Bootstrapping [14] is a machine learning technology which has been widely applied in knowledge acquisition. Ellen Riloff [15] uses bootstrapping to build knowledge base for information extraction. David Yarowsky [16] uses bootstrapping to research ambiguity elimination and so on. In general, seeds are collected manually. Then new seeds are learned from corpus by self-learning model.

This paper uses a semi-supervised learning algorithm. The scale of tagged corpus is very small and seeds are the only tagged corpus. Then entities extractor recognizes new entities for each iteration. This paper makes entity extractor based on maximum entropy.

1) Automatic learning module based on Bootstrapping

Firstly, a certain number of domain words are selected for seed words. Then, the input to the algorithm is a handful of seed words and unannotated training texts. The basic idea is the co-concurrence frequency of seed words and the domain words is much higher. The steps are following: Firstly, parts of domain entity words are selected as seeds. Secondly, features are selected. Thirdly, maximum entropy is used for extractor to extract candidate words. Fourthly, evaluate candidate words and some of words are selected as new seeds appended to seed words. Finally, start iteration the progress till no new domain entity is extracted. The flow chart is shown in Fig.2.

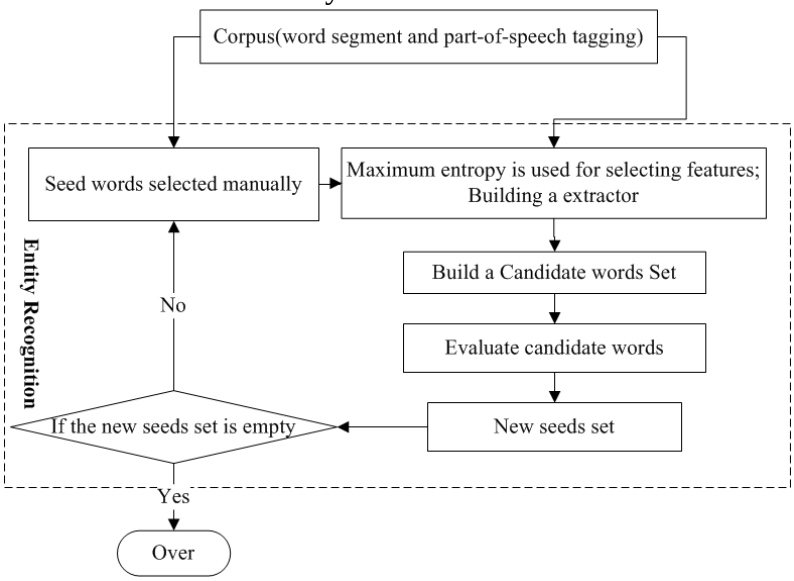

Fig. 2 The flow chart of Tibetan culture domain entity recognition algorithm

2) Maximum Entropy features

Maximum Entropy Model, (MEM) is a probability evaluation method which is widely used for nature language process. It comprehensively observes relevant and irrelevant probability knowledge. It has strong ability to express knowledge. Good results are obtained in text classification, data mining, part-of-speech tagging and so on. Maximum Entropy model keep to the maximum entropy principle, which means select the statistic model that has the maximum entropy and satisfy all the constrains. For the training, there are N training samples: (a_1,b_1), (a_2,b_2), (a_3,b_3), (a_4,b_4),…,(a_n,b_n).

And a_i has k attributes.a_i={ a_{i1},a_{i2},…,a_{ik} }. b_i is a label. In the process of recognizing entities, a_i represents word, part-of-speech and other complex features. b_i represents an entity label, such as religious sect, person, title and so on. The problem-solving is labeling a text best when giving a new sample. The major characteristic of maximum entropy model is integrating information, such as string of words, part-of-speech, related words.

3) Construct seed words and select new seed words

Each entity type selects N words as initial seeds when making word segment and part-of-speech tagging for the second time. The seed words should cover many context features.

New seed words are selected from candidate words. Candidate words are gotten using an evaluation function, such as E(newseed$_i$). Top K words are appended to the training set.

There might also be some mistakes. Some adverse effects will be brought in the next iteration. So, extraction results need some collection.

Conclusions

Named entity recognition is often a challenging task. The research on Tibetan culture domain entity recognition is very little. But the research is very important for studying Tibetan culture domain information processing. This paper recognizes Tibetan culture domain entities based on bootstrapping and maximum entropy. The method based on improved TextRank is used for extracting keywords before entities building. The keywords can be used as the domain words and the domain words can be used for word segment at the second time. In the future research, Tibetan culture domain will be more specialized and multiple levels will be combined to extract Tibetan culture domain entity.

Acknowledgement

The research was sponsored by Key Program of National Natural Science Foundation of China (No. 61331013), Projects of The Chinese Language Committee (No. WT125-46 and WT125-11) and Graduate Student Project of Minority Languages Branch, National Language Resource Monitoring & Research Center (No.CML15A02).

References

[1] Zheng Jiahen, Li Xin, The Research of Chinese Names Recognition Method Based on Corpus. Journal of Chinese Information Processing, 2000. 14(1): pp.7-12.

[2] Liu Bingwei, et al., Statistical Chinese Person Names Indentification. Journal of Chinese Information Processing, 2000, 14(3): pp.16-24.

[3] Ye, T.H., ZHENG and LIU, Research on Method of Automatic Recognition of Chinese Place Name Based on Transformation. Journal of Software, 2001.

[4] Wang Ning, et al., Company Name Identification in Chinese Financial Domain. Journal of Chinese Information Processing, 2002, 16(2): pp.1-6.

[5] Gao Guoyang, Research on the Information Extraction System in Sports Domain, 2010, North China University of Electric Power (He Bei).

[6] Zhou Lei, Research of Complex Named Entity Extraction based on Hybrid Method, 2009, Shanghai Jiao Tong University.

[7] Kushmerick, N., Wrapper Induction for Information Extraction, Dissertation. In Intl. Joint Conference on Artificial Intelligence (IJCAI), 1997.

[8] Baumgartner, R., S. Flesca and G. Gottlob, Visual Web Information Extraction with Lixto. Proc Vldb, 2001, pp.119-128.

[9] Liu, L., C. Pu and W. Han, XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources. Icde, 2000, pp. 611.

[10] Sun Zhen, Wang Huilin, Overview on the Advance of the Research on Named Entity Recognition. New Technology of Library and Information Service, 2010,(6): pp.42-47.

[11] Niu, C., et al., A Bootstrapping Approach To Named Entity Classification Using Successive Learners. Proceedings of Annual Meeting of the Acl, 2003, 1: pp. 335-342.

[12] Becker, M., et al., Optimising selective sampling for bootstrapping named entity recognition. Proceedings of the Icml Workshop on Learning with Multiple Views, 2005, pp. 5-11.

[13] Quasthoff, U., C. Biemann and C. Wolff, Named Entity Learning and Verification. EM in large Corpora, Proceedings of CoNNL-2002, 2002.

[14] Blum, A. and T. Mitchell, Combining Labeled and Unlabeled Data with Co-Training. Colt Proceedings of the Workshop on Computational Learning Theory, 2000, pp. 92-100.

[15] Riloff, E. and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. in Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence. 1999.

[16] Yarowsky, D., Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of Annual Meeting of the Association for Computational Linguistics, 1995, pp. 189-196.