http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (22): 6543-6547, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.6543.6547

Algorithms of Generating and Recognizing the Abbreviation of Chinese Organization Names

[1]Zhou Faguo, [1]Zhao Meijiao, [2,3]Sun Zhen, [2]Zhao Jie, [2]Sun Tai and [2]Li Shengfei
[1]School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, 100083, China
[2]National Administration for Code Allocation to Organizations, Beijing, 100029, China
[3]Department of Information Management, Peking University, Beijing, 100087, China

Abstract: Named Entity Recognition (NER) is a meaningful part of the research of natural language document understanding and it is an important branch of natural language processing. Organization name recognition is an important task in NER of Information Extraction (IE), which intends to recognize and extract organization name from unstructured text documents. In this study, on the basis of the research status quo of the researchers and experts both in China and abroad, some abbreviation name recognition algorithms of organization are proposed. The first algorithm is generating the abbreviation names of organizations, the second is recognizing the abbreviation names of organizations and the third is to match the abbreviation name and its full name.

Key words: Named entity recognition, organization name, information extraction, abbreviation name of organization

INTRODUCTION

With the development of information society, how to get structural information from the text, how to quickly read the new information has become a serious problem. To solve this problem, information extraction came into being, recognizing and extracting lot of information of name entities, their relations and even entire events.

In this study, the abbreviation name of Chinese organizations is the main research topic. The algorithms of abbreviation names generation, recognition and matching the abbreviation with its full name are discussed and put forward at last.

LITERATUE REVIEW

In foreign countries, the English named entity recognition carried out earlier than in China. The first study is extracting company names from text (Rau, 1991). In this study, to recognize and extract company names is proposed firstly. From mid 1990s, some evaluating indicators of named entity were put forward.

Background: From 1996, named entity recognition and extraction as a subtask of information extraction was led into the International Conferences such as MUC-6 (Grishman and Sundheim, 1996), MET-2 of MUC-7 (Chinchor, 1998), IEER-99, LREC and so on. Named entity recognition and extraction turned into part of information extraction inseparable.

Chinese named entity recognition is developed from mid 1990s. The research topic of Chinese named entity is mainly focused on person names, location names and organization names. In some studys, person names is recognized and extracted through computing the probability of surname and second name appeared in context (Sun *et al*., 1995); organization names is recognized and extracted according to rules organized by experts (Zhang and Wang, 1997). The main methods of named entity recognition have three types, rules-based, statistics-based and their combination.

As same as the named entity recognition, name of Chinese organization recognition has the following three main algorithms, such as rule-based (Wang *et al*., 2002), statistics-based (Cucerzan, 2007; Kazama and Torisawa, 2007; Zhang *et al*., 2008) and their combination, also called based statistics and rules.

The rule-based method is easy to materialize. To recognize the specific field on a small scale could achieve a relatively high accuracy.

The statistics-based method of organization name identification is mostly widely studied method currently and also is the advanced method of named entity recognition.

Corresponding Author: Zhou Faguo, School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, 100083, China

The statistics-based machine learning algorithm has three major forms which are the supervised machine learning algorithm, semi-supervised machine learning algorithm and unsupervised machine learning algorithm.

Since description of natural language need to meet certain grammatical rules and the problem is not entirely a random process when solving a problem in the method of natural language processing, using statistical methods will expand the state space in solving the problem. So, we need to prune the rule base and knowledge base while analyzing and solving problems.

It results in that in natural language processing, we often put the rule-based method and the statistic-based method together instead of using merely one of these methods. Not only using the rule-based method or statistic-based method (Lin *et al*., 2004; Yu *et al*., 2006) means using a hybrid method combining with rules and statistics.

Difficulties: Up to now, it have made wonderful progress in research on Chinese named entity recognition, especially in recognizing person names and location names with high accuracy rate and recall rate is high comparatively. But, there is little research in Chinese organization name recognition, because of the big difficulties of the recognition of Chinese organization names.

The quantity of Chinese organization name is tremendous, there existing probably approximately 40,000,000 organizations.

Without specific and standard defining forms, Chinese organization names are hard to determine rules. Moreover, the name can be nested.

Chinese organization name is changing all the time, because some organizations often change their names.

Chinese organizations have no any standard form and symbol.

Words in Chinese organization name are unrestrained. The word construction is quite arbitrary and decentralized. Some words could be part of location names, could also be part of organizations. Some person names can be part of organization. What is more, some could either be the initial word or the end of word of organization.

Some Chinese organization names are very long, some are short. Without any limitation, their length is different.

Consequently the recognition of Chinese organization name is difficult, also failed to achieve an ideal result and it is far from the needs of practical application, become the bottleneck of Chinese Entity Recognition.

The recognition and extraction of organization names is one field of NER. The Entity recognition algorithm on chapter three could also be used in the recognition and extraction of organization name which will not give unnecessary details. We study questions about the recognition and the generation in abbreviation of Chinese organization name in this chapter.

GENERATING ALGORITHM OF CHINESE ORGANIZATION ABBREVIATION

Organization abbreviations often appear in many real-life official document exchanges, retained documents, user requirements and on the internet. And there are some conditions of naming barbarism many times.

Chinese organization abbreviation: Organization name and abbreviations have appeared in more and more locations, even in many localities and large institutions, there are many notices and announces distributing about the name and standardized abbreviations of organizations.

The organization name identification is the subtask of named entity recognition. Meanwhile, it is the most difficult part in the recognition of named entity. In the application system about named entity recognition on the market, the accuracy rate and recall rate is around 70% generally and few of them could reach 80%. It is harder in abbreviation recognizing which makes that there is no ripe product on the market.

Composition characteristic of chinese organization abbreviation: We particularly analyze the organization name and abbreviation in the part of organization code information which is offered by the national organization code management center. And we find that there are following relationship between its abbreviation and full name:

- The abbreviation is composed of the first word of each word in its full name
- If a proper noun is appeared in the full name of the organization, it is the abbreviation of this organization
- If the full name of an organization is begin with a location, the abbreviation consists of initials of location name and the other words
- Using the phrase in the words which are appeared in the organization full names other than the postfix of location and mechanism makes up the abbreviation
- Using initials in the words which are appeared in the organization full names other than the postfix of location and mechanism makes up the abbreviation

- Using initials in the other words in organization full names except the postfix of the organization makes up the abbreviation.

The relation above is based on the existing name in order to run smoothly, therefore, in this chapter the recognition of the abbreviation is on the basis of full name recognition.

Generating algorithm: Since there isn't any generating method and the organization abbreviation is as follows:

- **Rule 1:** The initial Chinese character of every vocabulary (in this situation, the abbreviation is not more than three Chinese character)
- **Rule 2:** For proper nouns only (in this situation, the proper name is appeared in organization name, especially the pronunciation.)
- **Rule 3:** Taking all phrases in addition to the common suffix of the address and organization name in organization name to make up it (two or three terms generally)
- **Rule 4:** Taking initial Chinese character of all phrases in addition to the common suffix of the address and organization name in organization name to make up it (two or three Chinese character generally)
- **Rule 5:** According to the abbreviation of the bank, taking the initial Chinese character of words in front of the word "bank" in addition to the address name
- **Rule 6:** According to the case of organization name nested, it should apply the above rules twice

According to the above the generation rules of the abbreviation, the algorithm of generating the Chinese organization abbreviation could be achieved as follows:

- **Algorithm 1:** The algorithm to generate Chinese organization abbreviation
- **Input:** The full name of a Chinese organization
- **Output:** All the possible abbreviations of the Chinese organization
- **Step 1:** Collating generation rules of the abbreviation
- **Step 2:** Making Fine-grained segmentation to the organization full name entered
- **Step 3:** Generating all possible abbreviations of the organization according to abbreviation generated rules

Organization abbreviation has been applied in information retrieval weightily. In the retrieval, users often input the abbreviation, while this abbreviation used to be nonstandard. Therefore when a column is indexed, we usually generate the abbreviation for organizations which are appeared in the document and build index to all the possible abbreviations in order to improve the recall ratio and precision ratio.

RECOGNITION ALGORITHM OF CHINESE ORGANIZATION ABBREVIATION

As for the company name or some units name in Chinese organizations name, we often use the abbreviation of these companies or enterprises. But people do not necessarily use the abbreviation because the ordinary abbreviation can't indicate this company or unit well. If using the abbreviation at first, people often do not understand what we say. So, in general, when we use an organization abbreviation, we always define what this organization name or company abbreviation name is. It is means that define the abbreviation first, people would know which specific organization or company represented by this abbreviation when it appears again.

The abbreviation of organization, company or unit is that we can't know which specific organization it stands for as soon as hearing the abbreviation generally, so it is hard to be accepted. And the specific meaning of this organization abbreviation could only be understood by some insiders. There is details pertain of the abbreviation in the document which contains the abbreviation like this except which uses or appears like this. That is pointing out which organization abbreviation it is and what the corresponding full name is before using it. The range of application of this abbreviation is sometimes limited to this document. Sometimes, some abbreviations with geographical regions make strangers understanding hardly what it means. The Second Breach of Jinan Iron and Steel Group is called Second Steel for short, for example.

In many cases, the organization abbreviation is always defined first when it appears for the first time. That is to say that point out the correspondence between the full name and the abbreviation in order to make everyone understand the specific meaning of this abbreviation and what the corresponding full name is.

Above all, we can generalize the generating rules of organization when the abbreviation exists. Thereby we can draw the recognition algorithm of abbreviation:

- The abbreviation of the organization name include the left and right borders in which the left boundary has a left bracket and the right boundary has a right bracket, that is "(",")"
- The condition that there is no "( )" in the left and right boundaries of organization abbreviation

There are mainly the following situations: A, called "B" for short, A is abbreviated as B, B is the abbreviation of A, B act as the abbreviation of A, as the abbreviation of A, B and so on:

- **Algorithm 2:** The algorithm of organization abbreviation
- Input: Text
- **Output:** The existent organization abbreviation
- **Step 1:** Segment the input text into words
- **Step 2:** Apply the algorithm mentioned in chapter three to identify organization full names
- **Step 3:** Analysis the matched abbreviation recognition rules for the context in which exists organization full names
- **Step 4:** Make the corresponding relationship between the identified organization abbreviation and their full name
- **Step 5:** Output the identified organization abbreviation

If the organization abbreviation is not defined, it can be identified with small possibility, so there is no exhaustive study in this study.

ALGORITHM TO MATCH THE ABBREVIATION AND ITS FULL NAME

After the recognition in addition to the organization abbreviation present in the document, it is ordinarily to find organization full name corresponding the abbreviation. While these full names are appear in the document generally, there is the relevant information in the information base of organization code. How to find the organization full name after getting the abbreviation?

This is the organization name matching problem between the abbreviation and full name. It is different from traditional string matching, because neither accurate string matching nor fuzzy matching could identify which organization the full name is.

In order to match the organization abbreviation with the full name better, the article has brought about an algorithm to match the organization abbreviation and full name combining the generating algorithm of organization abbreviation (algorithm 1):

- **Algorithm 3:** The algorithm to match the Chinese organization abbreviation and its full name
- **Input:** The organization abbreviation yet to be investigated, the organization full name database
- **Output:** The matched organization full name
- **Step 1:** Segment the organization full name into words to get rules about the corresponding full name
- **Step 2:** According the algorithm 1, all abbreviations of the organization name should be received
- **Step 3:** Using the string fuzzy matching to match the yet to be investigated abbreviation with the all abbreviation of the full name
- **Step 4:** Making the matching field between the yet to be investigated abbreviation and the full name which has through the fuzzy matching
- **Step 5:** If there are more than one matching results
- **Step 6:** The algorithm will make the post-processing for the matching result in the overall consideration of the number of vocabulary, the length of full name, the matching order
- **Step 7:** Take the matching results with highest score to be the full name of this abbreviation
- **Step 8:** Output the matching full name of the organization

EXPERIMENTS AND ANALYSIS OF ALGORITHMS

The algorithm 1 is the generating algorithm of organization abbreviation. It is mainly for information retrieval. The accuracy rate of the algorithm execute result is not taken into consideration in the practical application. It is the recall ratio and precision ratio via retrieval full name using the abbreviation to be considered primarily.

For the algorithm 2, we organized 500 articles with the definition of including 654 organization abbreviations. On the basis of the algorithm 5.2, there are 643 abbreviations identified in which 632 are the errorless one. The accuracy rate achieved 98.3% and the recall rate achieved 96.6%.

For the algorithm 3, we matched the organization full name data base which includes the corresponding full name of these 654 abbreviations for using 654 abbreviations in the 500 documents above (4000 piece of test records). The number of organizations which have been matched is 619 in the results. 601 of them are errorless. The accuracy rate achieved 97.1% and the recall rate achieved 91.9%.

CONCLUSION

It introduces the chief contents, the research methods and the main problems of existence in the recognition of Chinese organization name in this study. Three algorithms are put forward in this study. And one algorithm is to generate the abbreviation names of rule-based, another is based on statistics and the other is based on statistics and rules. The analysis of the algorithms and experiments Chinese organization names, another is to recognize the abbreviation names of Chinese organizations and the other is to match the abbreviation name and its full name. The experiment results demonstrated that the algorithms have higher accuracy and recall rate.

ACKNOWLEDGEMENT

This study is partly supported by “the Fundamental Research Funds for the Central Universities”, National Key Technology Research and Development Program of the Ministry of Science and Technology of China with No. 2013BAK07B02, National High-tech R and D Program of China (863 Program) with No. 2013AA064303.

REFERENCES

Chinchor, N.A., 1998. Overview of MUC-7/MET-2. Proceedings of the 7th Message Understanding Conference, April 29-May 1, 1998, Fairfax, Virginia.

Cucerzan, S., 2007. Large-scale named entity disambiguation based on wikipedia data. Proceedings of Empirical Methods in Natural Language Processing, June 28-30, 2007, Prague, Czech Republic, pp: 708-716.

Grishnan, R. and B. Sundheim, 1996. Message understanding conference-6: A brief history. Proceedings of the 16th Conference on Computational Linguistics, August 5-9, 1996, Copenhagen, Denmark, pp: 466-471.

Kazama, J. and K. Torisawa, 2007. Exploiting wikipedia as external knowledge for named entity recognition. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, pp: 698-707.

Lin, Y.F., T.H. Tsai, W.C. Chou, K.P. Wu, T.Y. Sung and W.L. Hus, 2004. A maximum entropy approach to biomedical named entity recognition. Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics, August 22-25, 2004, Seattle, Washington, USA., pp: 56-61.

Rau, L.F., 1991. Extracting company names from text. Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications, February 24-28, 1991, Miami Beach, FL., pp: 29-32.

Sun, M., C. Huang, H. Gao and J. Fang, 1995. Identifying Chinese names in unrestricted text. J. Chinese Inform. Proc., 9: 16-27.

Wang, N., R. Ge, C. Yuan, J. Huang and W. Li, 2002. Company name identification in Chinese financial domain. J. Chinese Inform. Proc., 16: 1-6.

Yu, H.K., H.P. Zhang, Q. Liu, X.Q. Lv and S.C. Shi, 2006. Chinese named entity identification using cascaded hidden markov model. J. Commun., 27: 87-93.

Zhang, X. and L. Wang, 1997. Identification and analysis of Chinese organization and institution names. J. Chinese Inform. Proc., 11: 21-32.

Zhang, Z., F. Ren and J. Zhu, 2008. A comparative study of features on CRF-based Chinese named entity recognition. Proceedings of the 4th National Information Retrieval and Content Security, November 15-16, 2008, Beijing, China, pp: 111-117.