The Analysis of Named Entity Recognition Methods

Wei Rao [a] and Ziyi Wang [b]

School of computer and software engineering, Xihua University, Chengdu 610039, China

[a]649650334@qq.com, [b]344410323@qq.com

Abstract

Named entity recognition is the key technology of information processing, and its recognition effect directly affects the follow-up work, such as information extraction, information retrieval and so on. After nearly two decades of research, named entity recognition technology has also made some progress. This paper reviews the development process of named entity recognition, and analyzes the research results of named entity recognition, and discusses the development trend of named entity recognition.

Keywords

Naming entity recognition, Information extraction, Machine learning.

1. Introduction

With the growing size of the Internet and the popularity of computer applications, massive data information to people for effective information access has brought severe challenges, named entity recognition, as a common basis, is attracting more and more attention. Named Entity (NE), as an atomic element of information, is the basis for correct understanding of the text. The main task of NER is to identify the proper name (e.g. Person name, location name and others) , the meaningful phrase and to categorize them[1]. Named entity recognition is the first step in understanding the content of the text, and the results of its research will directly affect the study of the automatic processing of text information. Wang etc. [2] identify the name entity based on natural language, and achieve the organization name, job name and other important information extraction in the Web recruitment information. Mao etc.[3] presented a novel unified named entity recognition model based on the conditional random field, and applied it to the tourist attractions hotel Q & A system. In summary, named entity recognition as the basis of other information processing technology, it's a core position in NLP.

2. The Development and Current Situation of Named Entity Recognition

2.1 The Development of English naming entity recognition

The study of English named entity recognition began relatively early. Lisa F.Rau described a system that could extract company names at the Seventh Artificial Intelligence and Applications IEEE Conference, which used heuristic and man-made rules in 1991[4]. In 1996, named entities recognition as sub-tasks of information extraction, and their evaluation results were first introduced into MUC-6, which accelerated the development of named entity recognition technology. In MUC-7, ACE-8, CoNLL-2002, CoNLL-2003, named entity recognition is also a designated task. Bike etc.[5] use the hidden Markov model for English named entity recognition, the English name, place names and organization name recognition accuracy rate of 95%, 97% and 94%. Lee etc.[6] proposed a two-stage biomedical named entity recognition method, he divides the recognition task into two stage: entity boundary detection stage and semantic classification, which improves the system performance and reduce the training time.

2.2 The Development of Chinese naming entity recognition

The Chinese name entity recognition started late, the development speed is relatively slow. Sun etc.[7] proposed the Chinese name automatic identification algorithm, and the recall rate reached 99.77% in news corpus. Relative to the English named entity recognition, the Chinese name entity recognition

has the following main difficulties: First, the Chinese text don't have an obvious identifier, such as a space in English, so the first step in the Chinese naming entity recognition is to determine the boundary of the word, that is the word segmentation. Second, the quality of Chinese word segmentation directly affects the result of name entity recognition. Third, the complexity and disagreement of the Chinese language has effect the quality of named entity recognition. Finally, the different named entity has different internal characteristics, and it's difficult to establish a unified model that it covers all the characteristics.

3. The Method of naming entity recognition

At present, the research on Chinese named entity recognition has made some progress. Technically, it can be divided into the following four methods: Ⅰ)Dictionary-based method; Ⅱ)Rule-based (Linguistics) method; Ⅲ)Statistical machine learning-based method; Ⅳ) mixing method. The five methods are described as follows:

(1) Dictionary-based method

Dictionary-based approach is the easiest and simplest method, it's a simple match between words by querying dictionary. Li etc.[8] used the dictionary-based method to extract those pre-defined entities in the document to be processed, and proposed an effective entity filtering algorithm. Le etc.[9] use the method of geography code and Chinese word segmentation to identify the space named entity.

Because the dictionary-based approach is a kind of simple matching recognition mode, it is very simple and efficient in method operation. However, the method of named entity recognition based on dictionary is depend on the size of the dictionary and the overall quality of the dictionary itself, the named entity in Chinese don't follow the uniform format,, which led to recognition performance is not high, accuracy and recall rate is very low.

(2) Rule-based (Linguistics) method

The method of name entity recognition based on rules (linguistics) is based on the "named entity as a separate language unit, and its language structure should also be relatively fixed" conditions, through artificial or semi-automatic way, according to the internal and External features, the development of a series of relevant rules (such as grammar rules, specific affixes or other forms of rules, etc.)[10]. However, the rule-based (linguistic) method is highly targeted, and it is necessary to constantly spend considerable manpower and time to update the rules to ensure the best performance.

(3) Statistical machine learning-based method

At present, the statistical machine learning-based method is relatively flexible, the recognition effect is more accurate, therefore, it's the most commonly used naming entity recognition technology. Nowadays, two types of methods based on classifiers and Markov-based models are the most common, and quite a number of models, such as support vector machines (SVM), decision trees (DT), Hidden Markov Models (HMM), Maximum Entropy (ME), Conditional Random Fields (CRFs) and so on.

However, statistical machine learning-based method need marking the artificial corpus in the training set is rather cumbersome, and the corpus can be used, especially the large-scalethe corpus is extremely rare, which brings great inconvenience to the research work.

(4) mixing method

Hybrid method is combine two or more methods to avoid the weaknesses of the above methods, is a better solution to the problems of name entity recognition. Ju etc.[11] proposed the method of combining the random field with the rule, with the rich knowledge as the trigger condition, using the CRF model to mark the conditions of the text fragments for place names and organization name recognition. Cai etc.[12] proposed a novel framework to combine semantic entities with CRFs and SVM advantages, and combine context, language and statistical features in the algorithm. Wu etc.[13] proposed a conditional random field combined with dynamic names table mixed model for the Chinese named entity recognition.

4. Summary and Outlook

Named entity recognition is the basis for the development of natural language processing. From the perspective of linguistic language, because of the inconsistencies in the Chinese language, language ambiguity and so on, so the Chinese named entity can be put grammar, semantic information into the naming entity recognition. From the named entity field, named entity recognition is not only for formal text (news, articles, etc.), more and more to the professional field, irregular text expansion, such as the financial industry, automotive, medical terms, micro-blog mobile short This, especially for biomedical naming entity identification, now has a lot of research results. From method technology, the hybrid approach is an ideal method for naming entity recognition. The hybrid approach integrates the advantages and disadvantages of various methods, and improve the effect of recognition , so it's becoming the development trend of named entity recognition.

References

[1] Chinchor N: In Proceedings of the 7th Message Understanding Conference (1998).
[2] X.F.Wang, X.R.Zhang et al. Research on Named Entity Recognition in Web Recruitment Information Extraction, Computer and Digital Engineering, (2012)No.5, p.34-37.
[3] C.L. Mao, Z.T. Yu, X.L. Lei et al: Research on Question and Answer System of Tourist Attractions Hotel Based on Short Message, Technical Committee on Control Theory, Chinese Association of Automation(2011).
[4] Ran L E : Extracting company names from text, In Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications(1991), p.29-32.
[5] Bikel D.M, Schw R: An algorithm that learns what's in name, Machine Learning Journal Special Issue on Natural Language Learning(1999).
[6] Lee K.J, Rim H.C: Biomedical named entity recognition using two-phase model based on SVM, Journal of Biomedical Informatics,(2004)No.37, p.436-447.
[7] M.S. Sun, C.N. Huang: Automatic identification of Chinese names, Journal of Chinese Information Processing,Vol.9(1995)No.2, p:16-27.
[8] G.L.Li, D.Deng, J.H. Feng: Efficient Filtering Algorithms for Approximate Dictionary-based Entity Extraction, Proceedings of the 2011 ACM SIGMOD international conference on management of data(2011), p.529-540.
[9] X.H. Le, C.J. Yang, D.L. Liu: Spatial Named Entity Recognition in LBS, Computer Engineering, Vol.31(1995)No.20, p.49-53.
[10] Alfred R, Leong L C, On C K, et al. : Malay Named Entity Recognition Based on Rule-Based Approach, International Journal of Machine Learning & Compution, Vol.4(2014)No.3, p.300-306.
[11] J.P. Ju, W.W. Zhang, J.J. Ning: Geospatial Named Entities Recognition Using Combination of CRF and Rules, Computer Engineering, Vol.37(2011)No.7, p.210-212.
[12] P.Cai, H.Z.Luo, A.Y.Zhou: Semantic Entity detection by Interating CRFS and SVM, Web-age information management lecture notes in computer science, (2010)No.6184, p.483-494.
[13] Wu X, Wu Z, Jia J:Adaptive Named Entity Recognition based on Conditional Random Fields with Automatic updated Dynamic Gazetteers, 8th In ternational Symposium on.IEEE (2012), p.363-367