The Overview of Chinese Information Extraction

ZHU Qian[*1] and CHENG Xian-yi[1,2] ,

[*1] *School of Computer Science and Telecommunications Engineering, Jiangsu University， Zhenjiang, Jiangsu, 212013,China*

[2] *College of Computer Science, Nantong University, Nantong, Jiangsu, 226019, China*

Summary
The purpose of Information Extraction (IE) is to analyze and process large amounts of data, to discover useful knowledge and to provide the answer to the question for the user. It is a kind of effective way to realize information standardization. In the recent ten years, IE has developed to be one kind of technology that is parallel to Information Retrieval (IR). After discussing the research status of Chinese IE, the paper presents Chinese IE system architecture, and point out the opportunities and challenges of Chinese IE: semantic annotation.

Key words:
information extraction, Chinese, semantic annotation

1. Introduction

1.1 The status of IE

IE is to extract appointed information from non-restricted text according to user's need, and then fills it into corresponding slot of template in accordance with the defined template formats. For example, we can extract the information from economic news about new products released by company, such as: company name, product name, release data, product performance etc. We can also extract information about terrorist activities from news, such as: data, perpetrators' names, victims' names, the number of victims etc.

The research of information extraction was started in the 1960's[1], and it really flourished since MUC (Message Understanding Conference) was held. The conference has been held for seven times from 1987 to 1998. It not only tested the information extraction system, but also gave definite define of extraction template and the slot filling rules, evaluation criteria, system assessment mission and other norms[2]. MUC has become one part of TIPSTER [3]since MUC-4, which is held in June 1992. TIPSTER was organized by the US Defense Advanced Research Projects Agency, and its goal is to improve text-processing technology, especially Document Detection, IE, and Summarization.

Generally speaking, the object of information extraction system is natural language text and unstructured text in particular. While broadly speaking, the system can not only deal with electronic text, but speech, image, video and other media types as well.

At present, the IE Technology in English and Japanese had made much progress (English especially), and has been put into practice in many fields. Many companies with the main produce is information extraction software have emerged. Some of the most famous information extraction system is listed in Table 1 as below.

Now, except for the intense semantic demand, the main power that promotes the development of information extraction is the evaluation conference of Automatic Content Extraction (ACE), which is organized by American National Institute of Standards and Technology (NIST) [6]. The conference has been held for three times:

Table 1 Some famous information extraction system

Name	Basic function
InfoXtract[4]	NE、TE、TR expansion, unrelated to the domain, customizable, mobile, support open field QA system
FASTUS[5]	Base on NLP， cascaded finite-state automata(Mainly use of template matching)
PALKA[6]	Base on rules, semi-automatic acquisition of knowledge
SIFT[7]	NE, integrated TE/TR extraction system, complete use of statistical methods, train sentence-level model

May 2000, February 2002 and September 2002. The purpose of the conference is to develop automatic content extraction technology to automatically process three sources of linguistic text: common text, automatic speech

Manuscript received September 5, 2010
Manuscript revised September 20, 2010

recognition text and optical character recognition text. The primary content of research is how to automatically extract useful information from news corpus, such as: entity, relation, event etc.

Be compare with MUC, ACE is not aiming at certain domain or certain scenario, it uses a set of evaluation system, which based on omissions and misstatements, and can valuate the system's capability to process cross-document. The evaluation conference will raise the research on information extraction technology to a new level.

1.2 The research status of Chinese IE

The research on Chinese IE starts relatively late, the efforts mainly focus on Chinese named entity recognition, but the design and realization of complete Chinese information extraction system is still in the exploration stage. National Taiwan University and Kent Ridge Digital Labs (Singapore) evaluated Chinese named entity recognition in MUC-7. Zhang YiMin et al at Intel China Research Center [8] demonstrated their information extraction system on ACL-2000, which can extract Chinese named entity and the relationship between them. The system used memory-based learning algorithm to acquire rules to extract named entity and their relationship. Che Wanxiang et al at Harbin Institute of Technology[9] evaluated named entity recognition in ACE 2004. With the training data provided by ACE, they respectively used two types of supervised machine learning algorithms which based on eigenvector (SVM and Winnow) to extract entity relation, these two algorithms all choose two words around the named entity as characteristic words, and both of the F value reach 73%. But they all referred to the existing English information extraction models and methods without take the character of Chinese itself into consideration, and didn't completely implement the system. While the following two systems meet the requirements above: Li Lei et al at Beijing University of Posts and Telecommunication realized the Chinese IE System which based on comprehensive information [10], Zheng Jiaheng et al at Shanxi University [11] proposed a method that can generate extraction patterns automatically by clustering, and applied the method to Chinese agricultural text.

2.Chinese IE system architecture

2.1 IE system architecture

Hobbs once proposed an generic IE system architecture, and hold that a typical IE system is composed of the following ten modules: text segmentation, pretreatment, filter, pre-analysis, analysis, the combination of the fragments, semantic interpretation, lexical disambiguation, coreference resolution or discourse analysis, template generation.

Of course, not all of the IE systems must definitely comprise these modules, and the order of them can change also, for example, the executive order of module 6 and module 7 can exchange, but a complete IE system must achieve all the functions that described in above modules.

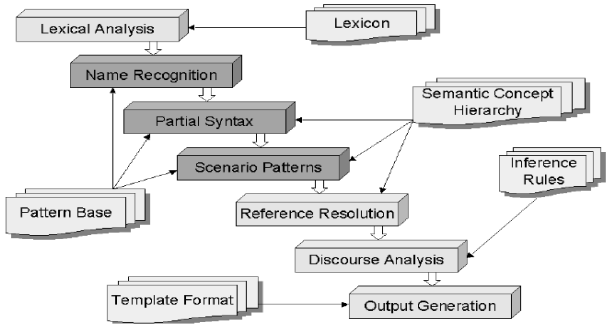

Fig. 2 IE system architecture

Figure 2 shows the architecture of New York University's Proteus IE system, and it is generic.

2.2 Chinese IE

The research on Chinese IE starts relatively late, and the characteristics of Chinese language also increase its difficulty, the following are the main characteristics:

(l)There are no break signs in Chinese. The Chinese sentence is composed of a string of characters without any space or other kinds of break sign, and the vocabulary has few morphological changes, so the basic task and special problem of Chinese IE is Chinese segmentation, since segmentation can't be absolutely correct, which will impact the effect of the following processing.

(2) There is no one-to-one relationship between the result of parsing and semantic analysis in Chinese, that is to say, the result of parsing can't directly be use to make semantic analysis.

(3) The structure is loose while its grammar is flexible. In Chinese, not only subject and object can be omitted, even predicate verb can be omitted too, so the comprehension of Chinese must take more consideration of background knowledge.

(4) The semantic meaning is very flexible. On the one hand, the flexibility of grammar is caused by that of semantic meaning, on the other hand, the same structure can express different semantic meaning and the same semantic meaning can be expressed with different structure.

Because IE system is a kind of application-oriented system, any change of its application target and language environment will result in the change of architecture. Generally speaking, Chinese IE system should include the following four key parts, as figure 3 shows.

.

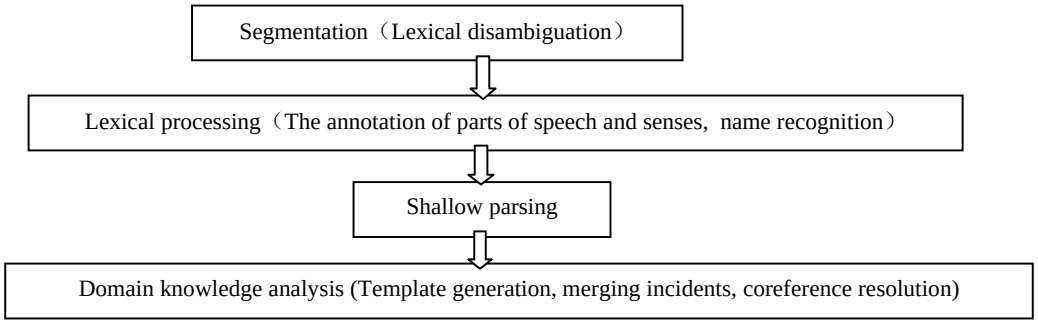

Fig. 3 The process of Chinese IE

3. The Opportunity and Challenge of IE: semantic annotation

Now, most of the research on Chinese IE is begin with the conceptions of parts of speech and syntactic function, while the semantic meaning is rarely considered, thus bring about following negative influence:

(1) The miss of semantic meaning. For example, the meaning of “plants” and “animals” is different, but their part of speech is same.

(2) In the process of parsing, it is difficult to extract a word from the adjacent ones when they are in a sequence of words with same part of speech.

(3) In the fields of few information expression ways, the sentences that describe different objects always have same grammatical structure, so, the difference of expression disappears in grammatical analysis.

With the rapid development of language processing technology, the importance and urgency of semantic analysis is more and more prominent. Since the mid-1980's, many countries of the world have invested heavily in developing semantic dictionary for computer, such as: EDR concept dictionary of Japan, SenseWeb of Singapore and the following three of USA: Wordnet (Fellbaum, 1998), Mindnet(Richardson, 1998) and Framenet (Fillmore, 1998). A succession of research and development in Chinese semantic dictionary is carried out also, for example: Contemporary Chinese Semantic Dictionary in Information Processing (Liwei Chen,Qi Yuan, 1995) and The machine Tractable Dictionary of Contemporary Chinese Predicate Verbs in the 9th five year plan program, Hownet(Zhendong Dong, 1999), Chinese Concept Dictionary(CCD)(Jiangsheng Yu, Shiwen Yu, 2002)and Hierarchical Network of Concepts(HNC) theory(Zengyang Huang, 2005)[12] etc. Most of the methods above are based on the classifying of lexical meaning, with some of them have a few attribute descriptions, and the lexical meaning is not considered in certain composite frame, hence, these methods are working little in NLP system.

The development of NLP technology, with the establishment of HNC theory especially, brings about a good opportunity for developing Chinese IE. To processing IE, the text must be understood to a certain degree, but it is differs from real text understanding. The advantage of IE is that it simplifies the process of NLP by only pay attention of important information while neglects irrelevant or minor information, so IE is a kind of shallow or simplified text understanding technology.

4. Conclusions

The purpose of information extraction is to extract appointed information. It breaks through the limitation that the work of reading, comprehending, information extracting must all been done by people in information retrieval, and can automatically retrieve, comprehend and extract information. IE can further refine the result of IR and the relationship between IE and related conceptions is shown in Figure 4.

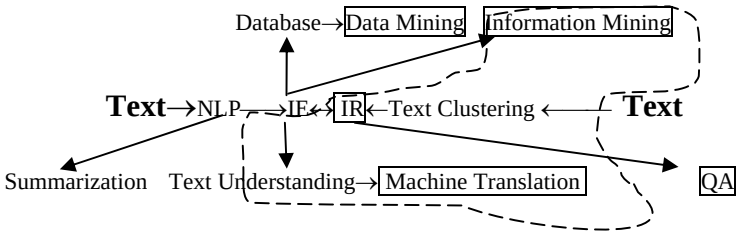

Fig. 4 The relationship between information extraction and related conceptions

There is Knowledge Discovery(or Text Processing) in the framework with dotted line，and the content in the rectangular frame is the output of text processing. As the input of text processing, there are two ways that text can take. As we can see from the Figure，IE is the kernel of Text Processing, while Natural Language Processing and Machine Learning is the basic of IE. Information Mining Technology based on IE is the research trend of Knowledge Discovery.

Acknowledgements

This dissertation is supported by Graduate Innovative Project of Jiangsu Province under Grant No.CX09B_204Z

References

[1] Yunbo Xiong.The Research on Several Key Techniques in Text Information Processing [D].Shanghai: Fudan University, 2006.
[2] Grishman R, Sundheim B. Message Understanding Conference-6:A Brief History[C]. In: Proceedings of 16h International Conference on Computational Linguistics COLING-96), 1996-08.
[3] TIPSTER SE/CM. How to get information about TIPSTER. In: Proc. of a workshop on held at Vienna. Virginia, 1996, 476-479.
[4] Rohini K. Srihari, Wei Li, Cheng Niu, Thomas Comell. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. In Proceedings of HLT/NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems(SEALTS), 2003: pp.52-59
[5] Hobbs, Jerrv R, Douglas E, etal. FASTLS. A cascaded finite — stat transducer for extracting information from natural — language text. Finite State Devices for Natural Language Processing. Cambridge, MIT Press, M A .1996.
[6] Kim J, Moldovan D. Acquisition of Semantic Patterns for information Extraction from corpora. In Proceedings of the ninth IEE Conference on Artificial Intelligence for Applications, Los Alarmitos, CA, IEEE Computer Society Press, 1993: pp.171 — 176.
[7] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group． Algorithms that learn to extract information BBN: Description of the SIFT system as used for MUC-7[R]． Proceedings of 7th Message Understanding Conference.1998.
[8] Zhang Y M, Zhou J F. A Trainable Method for Extracting Chinese Entity Names and Their Relations. In: Proceedings of the Second Chinese Language Processing Workshop, Hong Kong, 2000-10.
[9] Che Wanxiang , Liu Ting , Li Sheng. Automatic Entity Relation Extraction [J]. Journal of Chinese Information Processing.2005, 19 (2): 1-6.
[10] Li Lei, Zhou Yanquan, Wang Jinghua. Comprehensive Information Based Chinese Information Extraction System and Application. Journal of Beijing University of Posts and Telecommunications, 2005,28(6):48-51.
[11] Zhang Jiaheng, Wang Xingyi,LI Fei. Research on Automatic Generation of Extraction Patterns[J]. Journal of Chinese Information Processing. 2004,18 (1):48 — 54.
[12] Yaohong Jin. Language understanding technology and the applications of Hierarchical Network of Concepts [M].Beijing：Science Press,2006

Zhu Qian received the B.S. and M.S. degrees in computer science from Jiangsu University in 2000 and 2003 respectively. Her research interests include natural language processing, information extraction.
Now she is a doctoral candidate, lecturer.