Domain Knowledge Engineering Based on Encyclopedias and the Web Text*

SUI Zhifang
Institute of Computational Linguistics, Peking University
szf@pku.edu.cn

CUI Gaoying
Institute of Computational Linguistics, Peking University
cuigy@pku.edu.cn

DING Wansong
Institute of Computational Linguistics, Peking University
dws@pku.edu.cn

ZHANG Qinlong
Institute of Computational Linguistics, Peking University
zql@pku.edu.cn

Abstract

Domain knowledge is the fundamental resources required by all intelligent information processing systems. With the upsurge of new technology and new products in various domains, the manual construction and updating of domain knowledge base can hardly meet the real needs of application systems, in terms of coverage or effectiveness.

Based on natural language text analysis, this paper intends to draw a basic framework for the construction of domain knowledge base. Using encyclopedia resources and text information resources on the Web, we focus on the method of constructing domain knowledge base through technologies in natural language text analysis and machine learning. Moreover, an open network platform will be developed, through which common users can work with domain experts to contribute domain knowledge.

The technology can be applied to the construction and updating of domain knowledge base for intelligent information processing, and it can also provide help for the knowledge updating of encyclopedias.

Keywords: Domain knowledge base, Natural language text analysis, machine learning, encyclopedia, open platform

1 Introduction

Domain knowledge is the indispensable resource for an intelligent information processing system. With the upsurge of technology, more and more new technology, new product and new techniques come into being. The manual construction and updating domain knowledge base can hardly meet the real needs of application systems, in terms of coverage or effectiveness. In order to improve the robust of an information system, we need to study a computer-aided method to solve the bottleneck of domain knowledge acquisition.

The Institute of Computational Linguistics, Peking University, now cooperates with Encyclopedia of China Publishing Hall on the project of human-machine interactive encyclopedia knowledge engineering. We want to study how to exploit, use and update encyclopedia resource properly. We will use the technique of natural language processing, machine learning and text mining to acquire domain knowledge semi-automatically from

* This research was funded by 973 Natural Basic Research Program of China 2004CB318102 and Natural Sciences Foundation of Beijing 4052019

both encyclopedia and the Web text. Furthermore, we set up an open platform for domain knowledge acquisition, so that common network users and domain experts can work together to contribute new domain knowledge. Based on this technology, we can build domain knowledge base on each domain. This technology can be used as an important method of constructing and updating the domain knowledge base for the intelligent information retrieval and extraction systems. In the same time, it can also provide help for the knowledge updating of encyclopedias.

2 Related works

The researches on knowledge acquisition can be divided into three parts: artificial construction, semi-automatic construction and automatic construction.

Artificial methods are usually used in constructing the common sense knowledge base, such as CYC[1], WordNet[2], EuroWordNet[3], HowNet[4], and CCD[5] etc. That's because common sense is steady comparatively and it can not be affected by the task, also it can be reused by various kinds of system when constructed. For instance, since the WordNet was established in 1985, it had been widely used in IR, Text categorization, QA system etc. Similarly, the HowNet is being used in many Chinese information procession systems. It's worthy of large-scale devotion for long-time using.

On the contrary, domain knowledge is tied with some concrete domain. Once the application domain changed, we need to re-construct the domain knowledge base. Furthermore, domain knowledge updates continually, so that the domain knowledge base should be updates frequently. So it's unrealistic to construct the domain knowledge base manually.

In the way of constructing domain knowledge base, the semi-automatic method is mainly used. [6][7][8][9][10] established the platform of human-computer interactively working for the construction of domain knowledge base. They use various kinds of text processing and language analysis tools, which have the functions of morphological analysis, partial syntactic analysis, partial semantic analysis, with the mode of online cooperation, helping knowledge engineers or domain experts to find the domain concepts and the relations among them. The acquired knowledge can be added into the domain knowledge base. All these methods try to use pattern matching or various layers of NLP technology to acquire domain knowledge from large-scale free text. Free text is easy to get, however, it comes from different kinds of domain, including complicated language phenomenon hence is hard to understand. It's difficult to extract knowledge reliably using current technology of NLP and machine learning from such free text. If there not exists a pre-defined basic domain knowledge architecture, it is difficult to acquire the concepts and the relations relative to the domain. Also, among the above-mentioned methods, the construction of domain knowledge base depends on the expert's point of view and opinions. However, it's very difficult to let experts to construct the "real-time" domain architecture objectively and roundly and hence express it clearly in the given time.

3 Domain Knowledge Engineering Based on Encyclopedias and the Web text

In this paper, we propose a technology of domain knowledge engineering based on encyclopedias and the web text. Encyclopedia is the embodiment of the systematization and centralization of existed domain knowledge. The knowledge has been compiled and modified by many experts. Compared with free text, there are more canonical and NLP technologies can be used comparatively easily to extract knowledge from it. Since the knowledge in encyclopedia is more systematic, we can easily construct the basic frame of domain knowledge. So we will use NLP technology and machine learning method to construct the kernel of domain knowledge based on the analysis of the encyclopedia. Then based on the kernel of domain knowledge base, we can extract domain knowledge from other text resources.

There exist some researches on extracting knowledge from the encyclopedia [11] [12] [13] [14]. These researches use the encyclopedias as the only source to acquire knowledge. However, with the high-speed improvement in each

domain, there is severe knowledge lag in encyclopedias. So it is inadequate to use encyclopedias as the only source for knowledge acquisition. We need to learn more domain knowledge from other text resource besides encyclopedias.

With the surge of Internet, information in it is increasing exponentially. Abundant knowledge lies in this huge Web resource. If we can extract knowledge from the Web, we could update and expand domain knowledge base most efficiently. Standing in the computational linguistics' point, we focus on retrieving information from the content rather than from the structure of the Web.

This paper studies the technology of domain knowledge engineering. Using encyclopedia resource and text information resource on the Web, we focus on the method of constructing domain knowledge base through technologies in natural language text analysis and machine learning. Moreover, an open network platform will be developed, through which common users can work together with domain experts to contribute domain knowledge.

4 Strategy and Research Plan

4.1 Learning the style of knowledge-dense text from encyclopedias

The compilation of encyclopedias always follows some specified compilatory model. Encyclopedias have the relatively formal diction and different compilatory model for different kinds of entries. Because of that, the paraphrasable text in encyclopedias has clearer model to express the relation among concepts in most cases. For instance, "X is a kind of …Y", "X is composed of A, B, C and D", "A, B and C make up D", "X can be divided into A, B and C". Through recognizing the terms and partial parsing for the paraphrasable text in the encyclopedia, we could learn the patterns, which express the relations among concepts. Furthermore, we could learn the styles of knowledge dense text based on those patterns. Next step, we will follow the styles and combine the HTML target set to acquire more knowledge dense text fragments from the web. Based on such knowledge-dense text, some deeper natural language processing technologies could be used to extract domain knowledge reliably.

4.2 Automatic extraction of terms

Through analyzing the characters and expression forms of Chinese terms, we learn term knowledge from large-scale domain corpus and term bank. Using natural language processing method combing rule and statistics, we can automatically extract Chinese terms from corpus.

A term is a kind of phrases, whose components are close related. Further more, it has strong domain feature. The close relation of the components in a term can be captured through calculating the static association rate between the words that compose a term candidate. The linguistic feature can be captured through analysis the grammatical structural information of the terms. While the domain feature of a term can be captured through the domain component that has the possibility of composing a term. For example, "movable terminal" and "social economy" are both composed by the close related components. While, the former is a term in the domain of information science and technology, and the latter is just a common phrase instead of a term. The reason lies in that the former has the domain feature comes from one of its components" terminal", while the latter has not the domain feature.

We will use the above characteristic and representation forms of a term to perform automatic term extraction. The system of automatic term extraction includes two phases: learning stage and application stage.

In the stage of learning, we use a series of machine learning methods to get various kinds of integrated knowledge for automatic term extraction from a large-scale corpus and a term bank. These knowledge includes the inner structural knowledge of terms, the statistical domain features of term component, the statistical mutual information between the components of terms, the outer environment features of terms and the distinct text-level features of term recognition etc... In the stage of application, through an efficient model, we use all these various types of knowledge into automatic term extraction.

4.3 Design and implementation of partially analysis technology oriented for knowledge-dense text

The knowledge-dense text fragments (including encyclopedia and the Web text segments whose style is similar to encyclopedias) is relatively simple. Therefore, it's possible to implement deeper analysis on it using the natural language processing technology.

We use comprehensive language knowledge and statistical technique together to design language analysis method oriented for knowledge-dense text. We will use the comprehensive language knowledge resource, such as The Grammatical Dictionary-base of Contemporary Chinese, Chinese Semantic Dictionary, Chinese Concept Dictionary (CCD), and Termbank of Information Technology, which was developed by Institute of Computational Linguistics (ICL) of Peking University. Moreover, we will design a natural language partial parsing and understanding technology combining statistic technology base on the 80,000,000 words IT corpus. Concretely, on the base of the developed software such as word segmentation, POS tagging, term extraction and identification, skeletal dependency analysis of sentence, we will combine semantic restrict information with syntax rules, so that during syntax analysis we can get the semantic restrict information between syntax components at the same time. We will label semantic roles for predicate-head and its central valency components using Chinese Semantic Dictionary. So we can get shallow case frames of sentences after natural language partial parsing and understanding for text sentences. And domain knowledge will be extracted in the later stage from this analysis result.

4.4 Establishing the basic knowledge description frame based on the encyclopedia

The knowledge of encyclopedia is relatively systematic, mature and intensive. On this foundation, it is easier to set up a basic domain knowledge base which includes the kernel of domain knowledge. In the encyclopedia, every subject is described by attributes, and different subjects are organized hierarchically.

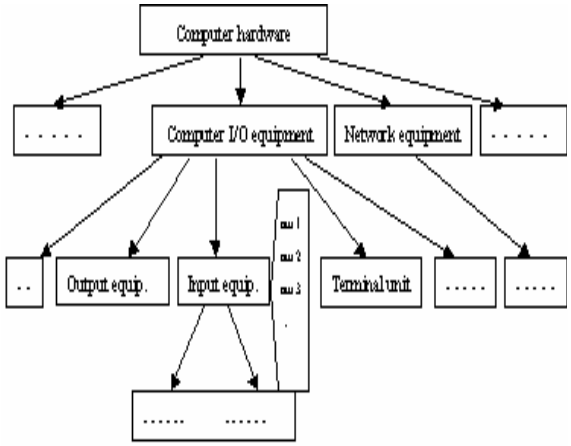

Figure1: An example of the fragments of domain knowledge framework

For example as showed in Figure1, aiming at the subject "Input equipment" in the domain of computer hardware, the encyclopedia describes the basic knowledge around the subject from many of point of views such as components, function, classification etc, which we call attributes here. On the other hand, "Input equipment", "Output equipment", "Terminal unit" constitute the subject "computer I/O equipment"; furthermore, "Input equipment", "Computer storage equipment", "Network equipment" are also components of the "Computer hardware". This paper will make the "classification + Attribute" as the basic knowledge description method for constructing the basic domain knowledge base. When the basic knowledge description method is set up, we take every entry in the encyclopedia as a subject, through analyzing the correlative sentence and recognizing the key terms in the paraphrase text and the relations among the terms, we can describe the basic knowledge on this subject. In the next step, we may couple several subjects in the same domain gradually in order to construct the basic domain knowledge base in this domain.

4.5 Using bootstrapping method to expand domain knowledge base

The structure of the web text is incompact, and the diction is not canonical enough. However, the web text is easy to get and contains a great amount of new knowledge. So based on the

basic domain knowledge base, we can select the knowledge dense text fragments from the web resource as the source to acquire more new knowledge.

We collect language patterns, which are known showing some kind of domain knowledge from encyclopedia. Using the language patterns as the seed set, we could learn more language patterns from the web text using boot strapping machine learning method. Using the expanded seed set, we could learn more language patterns from the larger text. This technique can expand domain knowledge base iteratively.

The system structure is as Figure2.

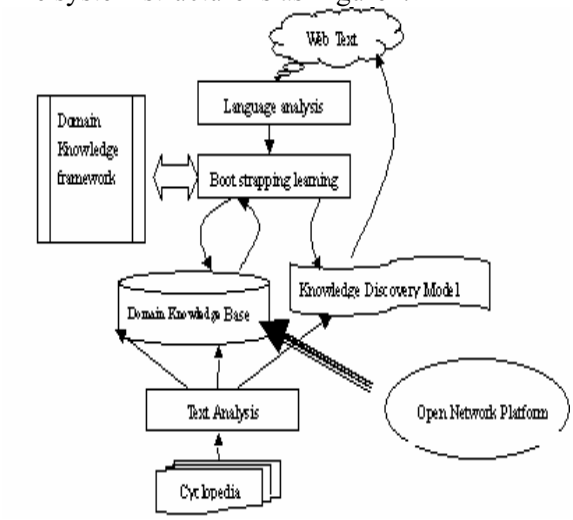

Figure2: The system structure for domain knowledge acquisition.

4.6 Developing open platform for domain knowledge collecting

With the rapid development of Internet, people could communicate and collaborate without face to face. They can share work and collaborate through the web. So we constructed an open human-computer interactive platform to call on domain knowledge experts and spacious common network users to collaborate together and contribute new domain knowledge. This platform could also assist the experts of encyclopedia in editing and managing new domain knowledge.

5 Current result

5.1 Automatic extraction of terms

We have exploited a term extraction system, including term extraction, human-computer interactive updating etc. The system is made up of basic source layer, learning layer, application layer and service layer.

We select the texts from 16 representative Chinese journals in the field of science and technology to construct the testing set.

The Principles for Testing

First of all, we manually tagged the terms in the testing texts. The principles we used in term tagging is very strict, that is, we only tag the longest terms in the texts, while ignore any of the term fragments in a longest term. For example, for the word sequence "接口 技术 规范 (Interface Technology Specification)", we only select "接口 技术 规范 (Interface Technology Specification)" as a term, while ignore "接口 技术 (Interface Technology", although it may be also a term in other context.

Similarly, for word sequence "数字 电视 信号 (Digital Television Signal)", we only select "数字 电视 信号 (Digital Television Signal)" as a term, while ignore "数字 电视 (Digital Television)".

The above testing principles may result in a great decrease of the precision and recall of term recognition. However, through these principles, we can find more problems existed in the term recognition algorithm.

The Testing Results

Based on the above testing principles, we get the precision and recall of term recognition as Table 1.

THE JOURNALS OF THE TESTING TEXTS	RECALL %	PRECISION %
Semi-Conductor Technology (1999-01)	65.6	55.1
Telecom Science (1998-01)	52.9	60.4
Computer and the Peripheral Equipments (1999-01)	52.9	71.2

The Research and Progress of Solid Electronics (2000-01)	57.6	62.4
Compute-Aided Design and (1999-01)	60.0	54.6
Computer Engineering (1999-04)	65.1	73.4
Computer Application (1999-01)	57.2	68.9
Automatic Measure and Control (1998-02)	51	59.5
Control Theory and Application (1999-01)	49.4	64.1
Software (1998-01)	52.6	54.1
Micro-Electronics (1999-01)	65.6	55.0
Wireless Communication Technology (2000-01)	57.7	69.6
Remote Sensing (1999-01)	67.1	62.1
System Emulation (1999-01)	64.9	75.4
Motional Communication (1999-01)	61	51.3
Chinese Cable Television (2000-01)	60.0	57.0
AVERAGE	**57.8**	**62.2**

Table 1: Testing Result

There is no unique standard for term's determination. What is a term? What is a common word? What is a term fragment? It is difficult to give an objective and unique standard that is operable for computers. Therefore, what the automatic term recognition system find can only be taken as the term candidates attached with the confidences. We still need the human terminology experts to give a final confirmation of the terms.

Our software includes human-computer interactive updating interface besides automatic term extraction. The interactive updating interface is as Figure3:

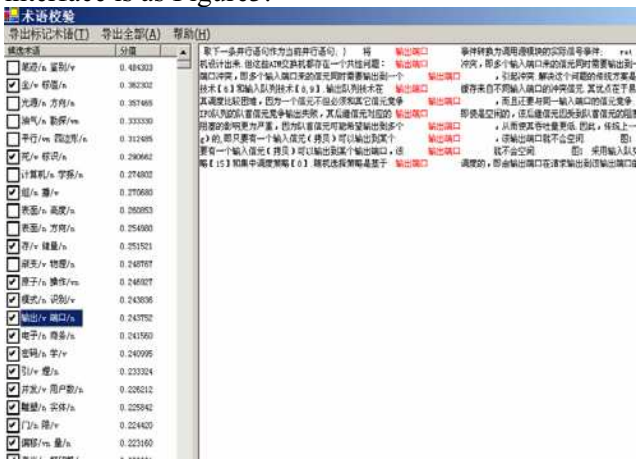

Figure3: The interactive updating interface after term extraction

5.2 Set up the basic database of encyclopedia

A lot of key concepts in the encyclopedia are well-marked with hyperlinks, titles, bookmarks and other Html tags according to different kinds of information respectively in the paraphrasable text. Using the information supplied by "China encyclopedia" e-press, we put encyclopedia subject information, relationship between subjects and term hierarchy into database to form an encyclopedia database for a primary domain knowledge base.

The core structure of encyclopedia database is presented as (main entry, relation term, relationship). Main entry is the entry that is listed in the encyclopedia. Relation terms are the hyperlink, bookmark, subtitle and so on. The relationship between main term and these elements now is null that need to be added with human assistance.

For example, the paraphrasable text of term "frequency divider" is showed in the database as Table2.

Main entry	Relation term	Relationship
Frequency divider	Crystal oscillator	Unknown
Frequency divider	Impulse frequency divider	Unknown
Frequency divider	Trigger	Unknown
Frequency	Regenerate	Unknown

divider	frequency divider	
Frequency divider	Trigger	Unknown
Frequency divider	Regenerate frequency divider	Unknown

Table2: the database fragment for the term "frequency divider"

5.3 Attribute relation template extraction

attribute relation type	template example
definition	xxx 家/学者/专家/奠基人/发明人/创始人/先驱
substitutable name mark	xxx，/。字/号 xxx
country	xxx（国名）xxx 家
nationality	xxx 族/族人
native place or home place	诞生地/生于 xxx/xxx 人
experience	毕业于 xxx；或 xxx 学位
literature	所著/著有/著作/发表 xxx/代表作 xxx/出版了 xxx/主要著作 xxx
working experience	xxx 年任 xxx；xxx 起任 xxx；任 xxx；兼任 xxx；历任 xxx；曾作过 xxx；曾任 xxx；被聘为 xxx
achievement and influence	被誉为 xxx/获 xxx 称号/xxx 创始人

Table3: The examples of the attribute relation template of human entry

We have semi-automatically extracted several attribute relation templates for human entries from encyclopedia text. The attribute relation template of human entry examples are as Table3.

5.4 Open platform for domain knowledge collection

We design the open platform for domain knowledge collection using ASP.NET network programming technology. We establish interactive working relation among domain knowledge engineers, domain experts and common users through the platform. The functions including:

- Domain knowledge requirement collection: on-line collection of new term entries of current domain, which are needed by the users.
- Domain knowledge supply collection: on-line collection of more detailed attribute information of the new terms.
- On-line management: system administrators manage new term information, which were submitted on line by the users.

The interface of the platform is as Figure4.

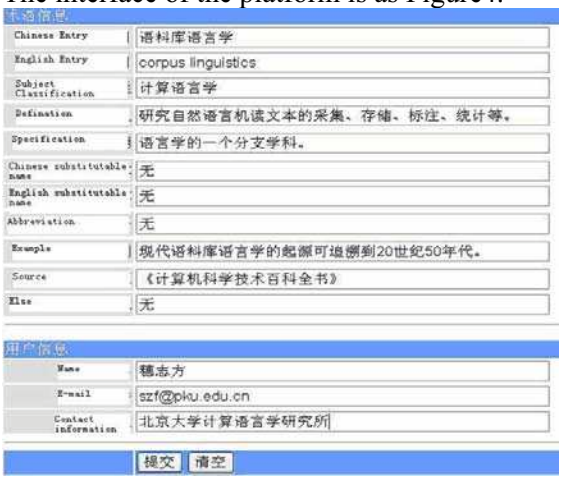

Figure4: The interface of the open platform for domain knowledge collection

6 Conclusion

The construction of domain knowledge base is a kind of high intelligent knowledge engineering. Since there is still have big gap between current level of technological development and real need, it is unrealistic to build domain knowledge base using automatic method or manual method only. However, in the human-computer interaction process, how to sufficiently absorb the knowledge resource which human being has already mastered and use it to supervise

automatic acquisition of new knowledge? How to call together knowledge engineers, domain experts and common network users and realize multi-member collaboration during the updating and extending process of domain knowledge base? These are the key problems to be settled in the knowledge engineering domain. This paper tries to do some exploration on these aspects.

References

[1] http://www.opencyc.org/

[2] http://www.cogsci.princeton.edu/~wn

[3] http://www.illc.uva.nl/EuroWordNet/

[4] http://www.keenage.com

[5] Yu Jiangshen, Liu Yang, Yu Shiwen, The specification of the Chinese Concept Dictionary, Journal of Chinese Language and Computing, Vol.13, 2003.

[6]A.Maedche, S.Staab, Semi-Automatic Engineering of Ontologies from text, Proceedings of International Conference on Software Engineering and Knowledge Engineering (SEKE' 2000), Chicago, IL, USA, 2000

[7] Szpakowicz,S., Semi-automatic acquisition of conceptual structure from technical texts, International journal of Man-machine Studies, 33(4),385-397,1990

[8] Biebow, B., Szulman, S., TERMINAE: a linguistic-based tool for the building of domain ontology. In Dieter fensel, Rudi Studer (eds.), Knowledge Acquisition, Modeling and Management, pp.49-66, 1999

[9] Lapalut, S., How to handle multiple expertise from several experts: a general text clustering approach. In F. Maurer (Ed.), Proc. 2nd Knowledge Engineering Forum (KEF'96), Karlsruhe, Jan., 1996.

[10] Mikheev, A., Finch, S., A workbench for acquisition of ontological knowledge from natural text. In proc. Of the 7th conference of the European Chapter for Computational Linguistics (EACL'95), Dublin, Ireland, pp. 194-201, 1995

[11] Richard Hull, Fernando Gomez, Automatic acquisition of biographic knowledge from encyclopedic texts, ExpertSystems with Applications, 16(1999), pp.261-270, 1999

[12] Fernando Gomez, Richard Hull, Carlos Segami, 1994, Acquiring Knowledge from Encyclopedic Texts, Proceedings of the 4th ACL Conference on Applied Natural Language Processing, Stuttgart, Germany, 1994

[13] Song Rou, Xu Yong, An Experiment on Knowledge Extraction from an Encyclopedia Based on Lexicon Semantics, pp.101-112, 2002

[14] Gu Fang, Cao Cungen, Biological Knowledge Acquisition from the Electronic Encyclopedia of China, Proceedings of ICYCS'2001, pp.1199-1203, 2001