Construction of The Term Bank

Zan Hongying Yu Jiangsheng Hu Junfeng Yu Shiwen
Institute of Computational Linguistics, Peking University

ABSTRACT: This paper prompt a method to construct the Chinese-English term bank in the information science and technology field. It is important to create a well knowledge classification system before we construct the term bank in this field. The Chinese lemma, English lemma, definition, category, relatives and its example are necessary for each item of the term bank. Besides the basic columns, we also contain the information of the lemma's synonym, area of use, time of use, source of use, hyper category, web site link and etc. Thus the term bank is a factual semantic net structure, it will improve the terms' standardization in machine translation and the results' veracity in information extraction in the information science and technology field.

Key Words: term, term bank, knowledge classification, corpus, knowledge base

In recent more than 15 years the Institute of Computational Linguistics (ICL), Peking University (PKU), focuses primarily on the fundamental researches and applications of Chinese language information processing. The research of ICL covers a wide range of areas, including language model, language parsing technology, computational lexicography, computational semantics and application system. The IR/IE-oriented Chinese Concept Dictionary (CCD), one of ICL's ongoing project, based on WordNet and more than 20,000 concepts has been recorded. CCD has already taken shape and is making notable contribution to Natural Language Processing. With more and more foreign language information swarming into China, some of the Chinese terms in the information science and technology field are divergent in their intension and extension. This situation has affected the information's understanding and spreading, and baffled the science exchange. Therefore it is imperative to build a standard usable term bank to ensure the understanding and the conveying of the knowledge in the information science and technology field.

1. Introduction

Language is the most important vehicle of cultural information. With the rapid development of computer, communication, network technology and their application, language in the field of information science and technology has also swiftly changed. This change is mainly embodied by the alteration of words. That is why there are new terms mushrooming while some former terms gradually dying out or getting new sense. In the meantime, some of the Chinese terms in the information science and technology field are divergent in their intension and extension due to the translation confusion and different language traditions in the mainland China, Hong Kong, Macao, Taiwan and other Chinese speaking areas. This situation has not only affected the information's understanding and spreading, baffled the science exchange, and also held up China's joint with the rest of world in this field. Furthermore, as the information science and technology has close connection with the local language and culture, new concepts and new terms coming from the Chinese information processing are even more in need of standardization. This will be propitious to the evolution and the prevalence of Chinese culture.

Therefore something must be done to standardize the words' intension and extension in the information science and technology field. The construction of the term bank in the information science and technology field will benefit to the term standardization in this field.

2. Knowledge Classification

The first thing for constructing the term bank in the field of information science and technology is to build an appropriate knowledge category system or concept system. But so far there has been no professional concept system that is publicly accepted to be normative, scientific, and integrated. Things in the field of science and technology are even worse. On the one hand, this discipline is rather short and its development is relatively slow in China, and the original source is, in most cases, foreign to the Chinese, which has brought about a large amount of new concepts coming from outside. On the other hand, because of the difference of various language traditions and knowledge background, the terms in the field are not unitive. There exists one concept that represents several terms. For example, "program" can be translated into "程序" or "程式". This situation has hindered the course of communication and the spread of technology. The establishment of the knowledge category system will ensure the validity and the application of the term bank, and will benefit the understanding of different words with the same sense and the disambiguation of one word with several senses.

The information science and technology field contains not only the computer and communication subjects. In general, this field includes all subjects relative to information. Now there is no acknowledged opinion that bounds this field. We think the information science and technology field covers the theory and the technology of information's collection, identification, extraction, transformation, storage, transfer, process, retrieval, test, analysis, use and so on. It includes the subjects of computer, communication, multi-media, automation, video, remote sensing, etc. After we have consulted many materials, we find that no existing category standard can be easily used without change. So to ensure the validity and the integrity, we refer to the information science and technology category boundary of the National Information Industry Board, and adjust them according to the Chinese Library Classification, relative encyclopedia, and some technical dictionaries. Our knowledge classification system commonly process to the second level sections, some of them details to the third level sections.

Frankly, Our knowledge classification system has fewer hierarchical levels. The reason is that we plan to get a more general and shallow classification and to avoid the frequent modification of the structure of the term bank due to the slight change of term category. The change of terms' intension and extension will be reflected through some attributes in our term bank. The attributes in the term bank are very easily modified or expanded. We have got some successful experience through the construction of the Grammatical Knowledge Base of Contemporary Chinese. Furthermore, the other ongoing project of ICL PKU, Chinese Concept Dictionary (CCD) contains more than ten levels hyponymy inherited from WordNet, and practice shows it is not necessary for the machine translation and the information extraction.

3. The Construction of the Term Bank

● The Structure of the Term Bank

A term is the abstract of a concept in the certain knowledge. Terms' information in term bank must cover enough fields. According to our design and the GB/T 13725 (the generic principles and methods to construct term bank), we decide that our term bank structure conforms to the principle of category plus attributes. The information of each term item include follows: category coding, Chinese lemma, Chinese phonetic notation, English lemma, English abbreviation, concept definition, synonymy terms, hypernym terms, relative terms, information of principal / recommendable / allowed / unused, using area, article source sort, article author, article time, usage example, web site link and so on.

● The Implementation of the Term Bank

First we collected terms from existing technical dictionaries and encyclopedia, then filtered and merged them into our term bank through computer aided method. Of course, we need to make up for the absent information and part of them has been finished manually. By now our term bank contains about 30,000 term lemmas. Further goal is to reach 100,000 term lemmas. Through this method we can collect a majority of existing terms in the information science and technical field.

Next step we will extract the new terms in the field from Corpus term-tagged by field experts. The software for the automatic extraction of new terms is in developing.

Terms' information in our term bank ought to be accordant to their real usage. So we will update the term bank with their change. This will rely on the construction of a large-scale real contemporary corpus in the information science and technology field.

4. The Construction of The Balanced Corpus

With the rapid progress of the information science and technology, the language especially the terms in this field are changing continuously. The correct meaning of the words must be extracted from the current corpus considering the context. The terms' research should be on the basic of the usage and structure analysis instead of the researchers' intuition or experience. The large-scale real corpus in this field can provide a large amount of language material, and will benefit the terms' analysis such as the structure or the usage frequency statistics. We can extract new terms or new sense of some old terms to update our term bank in time. Furthermore, we can also use the corpus to train or test our new terms' automatic extraction software.

The international committee of terminology standardization has set up the fourth branch committee, language resource committee, which is responsible for this particular work. So the research about terms' corpus has been valued internationally, and it is just starting in China. We are beginning the construction of balance corpus in the information science and technology field. We have learned the successful experience through constructing the 26,000,000 characters corpus of People's Daily (1998 all year). Now we start to build the balanced corpus to support the term bank construction in the information science and technology field. According to our knowledge classification system, we will build an approximately balanced corpus including about 60,000,000 characters, while giving the attention to the productive and the academic contents and controlling the Chinese-English proportion maintained at about 12:1. The corpus will be tagged at the article level including the article's type, category, author and age, and all the terms will also be well tagged. The identification of terms needs corresponding academic knowledge, so we turn to the experts in relative subjects for articles and term-tagged corpus on various subjects and we are responsible for the supervision.

5. The Knowledge Base and The Automatic Extraction of Terms

New concepts and new terms emerge endlessly with the progress of science and technology. We can not collect all of them manually. Identifying the new terms is somewhat the same as identifying the undefined words in natural language understanding, but they are different in some aspects. New terms in certain field have more special rules. In general, people can not create a new word for each new concept. Most new terms are compounded with some existing terms or the certain formal abbreviation of some existing terms. Statistics show that the number of compound terms in larger field's term set will exceed 80 percent in terms' total. Therefore, we can dig out some useful rules through analyzing our collected terms in the field of information science and technology, sequentially build the terminology knowledge base within field, and support the automatic extraction of the new terms in this field.

In his monograph “An Introduction to Modern Terminology”, Professor Feng Zhiwei put forward the economical law for the formation of terms. He indicated that in certain knowledge terminology system there are a large amount phrasal terms which consist of fewer basic words. He also pointed out that in a term system the high-frequent words account for a small percentage of the total. So we can infer that through analyzing the collected terms' syntactic and semantic structure we can get the high-frequent basic words set in the information science and technology field. In the meantime we can also get the statistics of some common prefixes and suffixes in this field such as “性” or “机” etc. Further more, we can extract the constructive rules based on words category or just certain words, and build the knowledge base of terminology in this field. This is the essential pre-research for the development of new terms automatic extraction software. Our first version of automatic new term extraction software will be release in Sep, 2002.

6. Conclusion

What we have been doing contains the design and demonstration of system architecture; the knowledge classification; the establishment of the term structure criterion, term bank criterion and the corpus construction standards; terms collection and analysis; and the improvement of some existing software of Chinese segmentation and part-of-speech tagging to adapt to the terminology extraction.

The construction of the term bank in the information science and technology field will contribute to progress of standardization, promote the dissemination of information, and speed up the step of internationalization in this field. Further more, correct term translation (creation of term list) is the guarantee of translation quality. The normalization of terms and the construction of the plentiful information term bank will make it possible to achieve the terms standardization in machine translation. The creation of terms' classification in the information science and technology field will make the information retrieval and information extraction could improve some intelligence, reduce the losing of useful information or the occurring of rubbish information, and benefit the further concept retrieval.

Reference

① Yu Shiwen, Zhu Xuefeng, E. Kaske, Feng Zhiwei,1996, English-Chinese Lexicon of Computational Linguistics, Peking University Press
② the GB/T 13725 : The Generic Principles and Methods to Construct Term Bank
③ Feng Zhiwei,1997, An Introduction to Modern Terminology, Chinese press
④ R. Basili, L. Bordoni & M. T. Pazienza, 1997, Extracting Terminology from Corpora
⑤ Yu Shiwen, Zhu Xuefeng, Wang Hui, Zhang Yunyun,1998,The Grammatical Knowledge Base of Contemporary Chinese – A Complete Specification, Tsinghua University Press
⑥ Li Yun, Wang Qiangjun, Zhang Pu,2001, Study on Automatic and Dynamic Updating of IT Terminology, Proceedings of Conference of the 20th Anniversary of CIPSC