A Corpus-based Approach to Term Bank Construction

Bai Xiaojing, Hu Junfeng, Zan Hongying, Chen Yuzhong, Yu Shiwen

Institute of Computational Linguistics, Peking University, China

E-mail: {baixj, hujf, zanhy}@pku.edu.cn

Abstract

In this paper, a corpus-base approach is presented in the construction of the information science and technology term bank in which domain classification, reference and part of the definition are extracted from corpus. Farther experiments show that the structure analysis of the terms can be helpful in the corpus-based domain classification of the terms.

1. Introduction

Currently, a joint project is under way between China National Institute of Standardization(CNIS) and the Institute of Computational Linguistics(ICL), Peking University to construct a term bank in the field of information science and technology. The project aims at :

1. an ontology system
2. a corpus for term bank construction
3. a corpus-based terminology extraction program
4. a constructed term bank and the related specifications and standards, and others for terminologies in the field of information science and technology

The implementation of the whole project features various approaches, among which the corpus-based one constitutes our present focus.

The corpus in this project consists of two parts, an essential corpus of 15 million Chinese characters and an extension corpus of 60 million and more, responsible for different tasks respectively. The corpus-based approach enables us to address the goals of our project by the following schemes:

1. Categorization of the terminologies in our term bank
2. Assistance for defining the terminologies in our term bank
3. Training and testing of the automatic extraction program

Now, initial plans have been made for the implementation of these schemes, with experiments conducted in support of our further efforts.

2. Knowledge Classification

An ontology system is very important for the standardization of the term bank establishment. Up to now, there still do not have a ready-made classification scheme of information science and technology, not to say to put each specific terminology into one specific domain category. So 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.

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 intend to set up an appropriate and practical classification while make it integrated with the some existed international or national standards. We have referred to the ACM Computing Classification System, ICS(the International Classification for Standards), CLC(the Chinese Library Classification), computer encyclopedias, and some technical dictionaries. After we have consulted many materials, we classify the

knowledge of information science and technology field into five subjects:

1. pandects of information science and technology
2. computer
3. automatization
4. telecommunication
5. electronics

under each subject we provide four subclass: theory, technology, application and product & material. We also have set up a mapping between ICS and our classification. For example, ICS:35:220 are integrate into our classification in data storage device(its classification number is 020403).

Generally, our classification is on the second level of subjects, and some detail on the third or fourth level. 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.

3. Corpus Compilation

For the essential corpus, we turn to experts in the field of information science and technology. All the texts are chosen and provided by experts of specified branches.

In the meaning time, with the help of a program, field experts will tag all the terms and the related information in the corpus, i.e., categorize them into the very branches of the field they belong to. The essential corpus is built for data training in the automatic extraction program.

For the extension corpus, the size is more than 60 million Chinese characters. In this corpus, we can get concordance and collocation information about the terms, as automatic processing will be possible for this part, and further, considerable amount of useful information, which can facilitate the definition of the terms, can be extracted from the corpus. Moreover, this corpus will serve as a test set for the terminology extraction program.

4. Corpus-based Categorization of Terminologies

Up till now, a basic framework has been drafted out for the purpose of categorization, while the terminologies available now are more than 70,000. Given the possibility that the initial framework can be developed to a sound system for categorization, locating the Terms into this system will still be a hard job.

It is in this consideration that we come up with the corpus-based approach. The essential corpus provided by various field experts carries the field information and the terminology tagging. Terminologies tagged by field experts are to be compared with the Terms. This is designed to be a process of matching, after which the Terms can be put into their respective categories. In other words, we try to classify the terms according to their distribution in the corpus. For the first step, as a test, we obtained 100 texts (258,045 characters in total) about Computer Network, with 2,486 different terms tagged out (i.e., 2,486 terminologies are regarded as valid). Considerable terms, which are unlikely network ones, proved otherwise in the corpus.

For example:缓冲/cache, which does not seem to be an OS term in Chinese, is a true network concept in the following sentence: "与我们熟悉的磁盘缓冲技术类似，Internet 缓冲是在一台本地服务器上开辟一块缓冲区，保存访问 Internet 时获得的数据，这样在以后的浏览过程中如果还是访问那些网页，就不需要再次访问Internet ，而直接从缓冲中获得数据就可以了".

That means corpus based categorization can give a more accurate description of the field information about the terms. This will benefit not only the term categorization, but also the definition of the terms. In some cases, it can

even give us clues to find out terms with different shades of meaning.

5. Corpus-based Reference for Terminology Definition

Accuracy and standardization in defining terminologies also attract our attention and efforts. In the database of our term bank, there is a field named Reference, storing contexts of the Terms from the whole corpus, which are deemed as competent reference. Reference for terminology definition can be at various levels, namely, it can be sentence(s), paragraph(s) or even full text(s). Here the role of the corpus is significant, as it contains all the information that will be filled into the Reference field, and what is more, we are expecting templates for terminology reference or even for terminology definition, to be learned from the essential corpus and then applied to the extension part, thus achieving the corpus-based automatic referencing. In addition to category and terminology tagging, our field experts also have to tag the text contents that they regard as the competent references for terminologies. A program is designed to extract a language unit bearing a reference tag (starting with <Reference> and ending with </Reference>) containing or following a terminology tag (starting with <Term> and ending with </Term>), which is recognized as the reference information for the tagged terminology and will then be stored in the Reference field accordingly. The following are three examples.

Example 1: (a single sentence)

<Reference><Term>Vo I P **</Term>**可以定义为以IP 包交换的方式传输话音。**</Reference>**

Example 2: (a paragraph)

<Reference><Term>Vo I P 网关**</Term>**主要提供PSTN 电话通信网络与IP 网络的接口和转换。目前，一般采用H.323 作为IP 网络信令和SS7 作为PSTN 的信令。在这个市场的设备提供商中既有传统的数据网络公司如3Com 、Cisco 等，也有老牌的电信设备提供商如Alcatel 、Ericsson 、Nortel 、Lucent 等，以及Sonus 、Clarent 、convergent network 、Nuera 等公司。**</Reference>**

Example 3: (a full text)

<Reference>何谓**<Term>**DHC**P</Term>**？动态主机配置协议（Dynamic Host Configuration Protocol ，DHCP ）从原有的BootP 协议发展而来，原来的目的是为无盘工作站分配IP 地址的协议，当前更多地用于对多个客户计算机集中分配IP 地址以及IP 地址相关的信息的协议，这样就能将IP 地址和TCP/IP 的设置统一管理起来，而避免不必要的地址冲突的问题，因此常常用在网络中对众多DOS/Windows 计算机的管理方面，节省了网络管理员手工设置和分配地址的麻烦。中继代理服务器必须知道DHCP 服务器的地址，还要知道如何把接收到的报文转发给该服务器**</Reference>**

Sufficient data will avail us of the opportunity to learn reference templates, like “XX 可以定义为/can be defined as XX” in Example1; “XX 主要提供/is mainly for XX” in Example 2 etc. These are sample templates that can be used to extract the definition of the terms from corpus. Surely there can only have small number of the terms that can find definition directly from corpus, but the corpus-based contextual information, such as concordance and collocation are also helpful for experts to analysis the meaning and give the proper definition of the terms.

6. Automatic Extraction of Terminologies from Corpus

The third scheme is based on the understanding that the internal structure of terminologies is also a source of valuable knowledge for term bank construction. In this project, the internal structure of a terminology consists of three elements: 1) term constituents, including prefixes, suffixes, words and phrases that are frequently used in related technical documents, e.g., “性” and “接口”; 2) POS; and 3) semantic categories, each describing the common feature of a group of term constituents, like

the semantic category “equipped with/without a system of wires” derived from “无线” and “有线 ”. Patterning the internal structure of terminologies is a prerequisite to the automatic extraction of terminologies from the corpus. On the one hand, we analyze the Terms, together with those from the essential corpus and tagged by our field experts, and pattern their structures, using term constituents and POS information, e.g., “noun + 接口”. On the other hand, we generate new terms, replacing term constituents of the same categories in exiting terms with the other.

With “有线通讯”, “有线电视”, “有线电报”, for instance, we generate “无线通讯”,“无线电视”, “无线电报”. The automatic extraction program will then use the structure patterns and the new terms generated to extract terminologies from the extension corpus, either by character matching or by POS matching, or both. Large in amount as they are, the terminologies we have obtained r reach up till now. In this sense, the extension corpus is both a test set for the automatic extraction program and a source for additional terminologies by using the program. It therefore are still far from being enough. Considering the limited sources, we have to rely on the extension corpus for the automatic extraction of terminologies that remain out of oucalls our attention to the competence and performance of our corpus, and especially, the extension part.

7. Conclusion

We have devised the initial schemes for the application of the corpus-based approach to

1. the categorization of existing terminologies in our term bank
2. the learning of reference templates and the extraction of reference information from the corpus
3. the modeling of automatic terminology extraction

Experiments show that corpus can be very useful to illuminate the meaning of terms, which will help a lot to standardize the terms in the future.

References

1.Angelo, Robert. *A Synopsis of Wittgenstein’s Logic of Languag*e. http://www.roangelo.net/logwitt.

2.Feng, Zhiwei, , (1997). *An Introduction to Modern Terminolog*y. Yuwen Press

3.Sinclair John, (1991). *Corpus, Concordance ,Collocation.* Oxford University Press.

4.Kennedy Graeme, (2000). *An Introduction to Corpus Linguistic*s. Foreign Language Teaching and Research Press

5. http://www.acm.org/class/1998/

6.http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList

7.Chinese Library Classification, Version4.0, Beijing library press, China

8.Zan Hongying, Hu Junfeng, et al (2002) , Construction of the Term Bank, TAHK2002