Connotation Description of Terms and Corresponding Automatic Method

Zhang Yunliang[1], Zhu Lijun[1], Qiao Xiaodong[1], Zhang Quan[2]
1 Institute of Scientific & Technical Information of China, Beijing 100038, P. R. China
2 The Institute of Acoustics, CAS, Beijing 100080, P. R. China
E-mail:zhangyl@istic.ac.cn

ABSTRACT The interoperability problem between knowledge organization systems has become more critical because of the complication and reuse of KOS. Enlightened by dictionary complication and the characteristics of Chinese language, we use connotation description instead of relationships between items in KOS. We use the conceptual primitives and related instance base of Hierarchical Network of Concepts (HNC) theory and develop both semi-automatic and full-automatic methods for different applications. It is an exploration and should be realized and revised in the future.

Key Words: connotation description; conceptual primitives; Hierarchical Network of Concepts (HNC) theory; Chinese information processing; interoperation;

1. Introduction

In a typical knowledge organization system, such as thesaurus, there are a lot of items and connections between items. When the KOS is the only one used in an application system, it usually works perfectly. But more and more applications need 2 or more knowledge organization systems. In these cases, the relations sometimes are different in types and granularity, we must integrate or merge or mapping the systems with interoperation techniques. Recall what experts do in dictionary compilation, they select a limited amount of words or characters (in Chinese), may be 1000~2000, which are commonly understood by the public. All the words are interpreted by these selected words or characters. But there is a big problem of the dictionary compilation model, namely the interpretation is very difficult for computer to understand. To cope with this problem, there are at least two approaches. The first one is to improve the natural language understanding techniques and make the computer has the common sense of a human being with basic education experience, which is not very practical now. The other approach is to design a limited primitive set, and the primitives are unambiguous and linked by some joint marks that provide the exact meaning of a word.

There are at least three theories for connotation description of Chinese words for computer applications. They are conceptual Graph theory developed by professor Lu (Wu, Hu and Lu,2008), HowNet theory developed by professor Dong(Dong and Dong, 1999), HNC (Hierarchical Network of Concepts) theory of Professor Huang(Huang,1998). The first theory is still in variation, the other two theories are relatively mature and the primitive sets are basically fixed. Now the HowNet and HNC both have several tens of thousands instances for general information processing constructed by manual work. In this paper, we adapt HNC theory to do some experiment on automatic method of connotation description of some items in our research knowledge organization system.

2. The characteristics of Chinese language

Chinese is very different from western languages, Most Chinese characters are not only form units and phonetic units of information, but also semantic units. In traditional Chinese linguistics, Chinese language has the composition character of: 1) Assemble characters into word, assemble words into clauses, assemble clauses into chapter; 2) From the characters understand the word, from the words understand the clauses, from the clause understand the chapter. So the traditional Chinese linguistics pays more attention to characters, and there is a discipline named Chinese Exegetics. Especially on characters and words, the meanings of characters are usually primitives of the meanings of words composed with these characters. For example the Chinese word 政府 (government)has two Chinese characters 政 and 府 , 政 means politics related and 府 means office or residence. Now the amount of the common used Chinese characters is less than 4 thousand. Considering the polysemy and synonym phenomena, there are several thousand primitives of meanings. So some scientists, such as Lu, Dong and Huang, propose some expression methods of the semantic primitives. The semantic primitives of cause can be expressed with Chinese characters, and can also be expressed with a new symbol system, which is composed with characters and figures.

3. Connotation description method

In this paper, we use HNC theory to do the connotation description work. HNC is a comprehensive theory for Chinese language processing with consideration of traditional linguistics and modern computer techniques (Jin, 2006), which includes theories of words expression with HNC symbols, sentence category analysis, and sentence group analysis and so on. HNC symbol is the conceptual connotation expression of the natural language words (including words with only one character or a lot of characters). A HNC symbol of a specific meaning of a word is a character string of letters, figures and special signs such as #, $ and so on. The conceptual primitives of HNC are organized in a tree structure. Though in fact it is a forest with 18 categories, 101groups and 456 trees, of course, it can be considered as a tree with some imaginary nodes. In a primitive tree, the depth is 2.The amount of the genuine primitives in all levels is 6580. All the 6580 nodes in the tree, whether they are leaf nodes or not, can be used to construct the HNC symbol. The connotation description process of HNC theory is analyze the connotation of a specific meaning of a word, then select proper HNC conceptual primitives and at last linked the conceptual primitives with proper operator(expressed as special signs)(Huang, 1998). In table 1 follows, we give the usage of the special sign in HNC theory, especially on the meanings and examples.

The merit that with rich operators is very important for connotation description. For example, in table 1, the word 拥戴 has two primitives: “v943e61” and “v71101”, means “support” and “respect” respectively, but “support” is not equal to “v943e61”, it is only an approximate natural language expression. In this HNC symbol, the two conceptual primitives are both important for description of word 拥戴 and without any of them, the connotation cannot be integral. So we use a operator “and” two express the equivalent importance. Then about another example of 巡警. At first 巡警 is “pa41”, that is a human being who belong to a militarized organization. In some scenarios, this connotation is enough for application. But in another scenario, maybe we also to know “va11”, namely that what patrols do is a professional activity serve for government and to avoid sins. Without the “va11”, the basic connotation is integral roughly, so a refinement operator”+” is used to link the two primitives. It is a little similar to series expansion in mathematics, the former is important than the latter, and with more expansion, the connotation is more accurate.

Table 1 The operator in HNC symbol expressions

Operator	Meaning	Example	HNC symbol with operator
#	作用(act)	暗算(plot to do something)	v9332#v322
$	效应(effect)	罢免(dismiss somebody from some position)	va01e22$v3629
&	对象(bbject)	报捷(declare the victory)	v9239ea2&gw30ae71
|	内容(content)	备耕(prepare for ploughing)	v11e21|v661
,	逻辑并(and)	拥戴(support with respect)	(v943e61,v71101)
;	逻辑选(or)	波折(reverse or stagate)	r1392;r139e522
!	非(not)	不得(forbid)	!jlvu12c32
^	反(anti)	属于(belong to)	^v461
(,lmn,)	一般逻辑组合(general logic composition)	直播(live broadcast)	(vc23aa,l11,su1021)
/	偏正(modification)	执照(license)	va018/gw23
||	主谓(subject-predicate)	自费(at one's own expense)	r4005|| (v010|gz93ae02)
+	展开(refinement)	巡警(patrol)	pa41+va11
( )	作用范围(scope)	诱发(to induce)	((va71,v9441)#(v900;v80))

4. Semi-automatic and full-automatic method

There are some constraints when the words expressed by connotation description. The words should be traditional Chinese words or loan words from free translation. If from transliteration, it doesn’t work.

The basic of the automatic connotation description is material or instances of the word-HNC symbol pairs. Now the pairs are constructed by manual work, now the instances divided into two parts. The first part has about 3,000 pairs of monosyllabic words and HNC symbols. The second part has about 40, 000 pairs of polysyllabic words and HNC symbols. All the words are collected as a dictionary.

The processing procedure is as follows:

1) Resolute the input word with the dictionary derived from pair instances. In this step a forward maximum matching method is used and the word is separated into words in dictionaries, or words and separated Chinese characters. We name this kind of words as word-components, and characters as character-components.
2) List the HNC symbol of word-components with the nature sequence and link them with “and” operator.
3) If there are separated Chinese characters. Select the words which contain these characters and count all the semantic primitives, give the most 3 frequent conceptual primitives as candidate and insert them into the sequence of step 2 in proper position. If there are no such characters in the instance base, use “null” as the result.
4) Now Check the candidate sequences for knowledge engineers and they decide which sequence is better, and revise the operator. Perhaps some revision on primitives is also necessary.

For example，情报业(Information industry) is a word we want to describe the connotation, and get a word-component 情报 and a character-component 业. The corresponding HNC symbol of 情报 is jw03|(ga4;ga2;ga1;ga6) . In this expression, jw03 means 基本信息物（a basic object type of what human communicate）. The vertical bar means that what in the brackets is the concrete content of jw03 The symbol a means 专业活动(professional activities）and a4,a2,a1,a6 means 军事(military field),经济(economy field),政治(politics field),科技(science & technology field). The symbol g means a 静态表示(statistic description of a concept, in English we can say describe is a dynamic description and description a statistic description). And there is not a word 业, so we do a statistics, find 39 words that has this character, and find the most frequent semantic primitives are “ga219”, “g661”and “l630a8”. And “a219” means 建造（produce and build）, “661”means 生命体基本劳作（labor of livings）, “l630a8”means 表示过去的时态说明符(past tense indicator).So the candidate series are:

a) jw03|(ga4;ga2;ga1;ga6),ga219
b) jw03|(ga4;ga2;ga1;ga6),g661
c) jw03|(ga4;ga2;ga1;ga6),l630a8

And knowledge engineer will choose a) from the understanding to “情报业”, usually the HNC symbol as a) is enough, but it is not perfect. So knowledge engineer may use another operator “|”

to replace "," and replace "a219" with a more specific primitive "ga219\25". The final HNC symbol of 情报业 is "ga219\25|(jw03|(ga4;ga2;ga1))" . Though the connotation description needs human interaction with the computer, the result of the full-automatic method is acceptable in common applications.

5. Problem analysis

There are also some problems in the full-automatic or semi-automatic method of connotation description with HNC primitives. The words they can processed is limited to traditional Chinese words or loan words from free translation, but it is a difficult problem for computer to recognize the proper words from the improper ones. If the input is not suit for the method, the output will be meaningless.

The statistics usually give good output, but there are cases which should adopt the low-frequency conceptual primitives. We count the primitives in a HNC symbol, but only some primitives are mapping of the specific Chinese characters, there are primitives from other Chinese characters in these words. The appearance position is also need to be considered. In the example of 情报业, primitive"l630a8" only for Chinese character 业 in the beginning location of the words, and it should be excluded in the statistics.

This method may be also effective for English and other western languages. Some English etyma and words also have the compositionality, just like Chinese characters and words. Though HNC conceptual primitives derived from Chinese, the kernel is suit for all Languages. But to do this for English, the concept primitive forest should be customized for English, that is, some primitives should be expanded and some should be compressed. Of course an English connotation description instance base should be manually built or mapped from Chinese base at first.

6. Conclusion

In knowledge organization system, we use relationships to reveal the relevance between items. If we want to make the KOS more useful, more and more thinning relationships should be designed and added into items. And some knowledge organization systems will be used together. But the inconsistency will also increase. Enlightened by the dictionary compilation approach and the characteristics of Chinese language, we try to give the connotation description of a new item in a knowledge organization system from the component of Chinese characters and words. In some applications, the result of full-automatic connotation description method is acceptable. But there are still a lot of problems, which should be resolved or explored in future study.

References:

1. Wu Baosong, Hu Yi and Lu Ruzhan. (2008), "Research on Recursive Conceptual Graph Based Text Retrieval Model", Journal of the China Society for Scientific and Technical Information, Vol. 27 No. 6, pp. 825-831.
2. Zhendong Dong and Qiang Dong.(1999)."HowNet", available at: http://www.keenage.com/zhiwang/e_zhiwang.html (accessed 21 April 2009).
3. Huang Zengyang. (1998),HNC Theory, Tsinghua University Press, Beijing.
4. Jin Yaohong. (2006), Language processing technologies and applications based HNC theory, Sciences Publication Inc, Beijing.