Finding the Synonymous Terms from the English-Chinese Term Lists of Computer Science

Hu Junfeng
Institute of computational linguistics, Peking University

Abstract: 'Synonymous Term' is one of the most important features in term bank construction. Usually, this feature is described manually by field experts or terminologists. Work has been done to extract synonymous terms from the bilingual computer term lists that came from different sources. These rules will be helpful to standardize the English-Chinese term translation in the future.

1 Introduction

In establishing the term bank of computer science, different kind of computer term resources include 'computer encyclopedia', 'computer dictionary', CJK computer terms etc. were collected. Altogether 184,615 different English-Chinese bilingual term entries from ten different sources were included in the finial term list[1].

One of the important works in establishing term bank is to find out the synonymous terms. Because most of the computer terms originally came from English and then translated into Chinese, the English translation equivalence of the Chinese terms plays an important role in finding the Chinese synonymous terms.

2 The Mirror Term Set and Shadow Term Set

Definition 2.1 In a given term list, a form entry is defined as the unique form of the term entries.

For example, like in table 1.1, the Chinese term entry 时间片(time slice) occurred 6 times in the term list with different English equivalent terms. But the string '时间片' is treated as one 'form entry' in the term list. The table 1.2 shows the different Chinese equivalent terms of the form entry 'slot'.

As any term entry can only have one unique corresponding form entry, in this paper, when describe the corresponding form entry of the term entry x, 'form entry x' is used directly without further declaration.

In table1.1, all the English terms in the first column constitute the mirror set of the form entry '时间片'; In table1.2, all the Chinese term entries in the second column build up the mirror set of the form entry 'slot'.

Table 1.1 The different translation equivalent terms of the Chinese form entry '时间片'

English entry	Chinese entry
quantum	时间片
slice	时间片
slice of time	时间片
slot	时间片
slot duration	时间片
time slice	时间片

[1] The 'term list' mentioned in this paper means the induplicate Chinese English bilingual term entries list. When combining different lists into one, the duplicate bilingual term entries were purged.

Table 1.2 The different Chinese translation equivalent terms of the English form entry ‘slot’

English entry	Chinese entry
slot	槽
slot	槽口
slot	插槽
slot	插件槽
slot	存储槽
slot	存储区
Slot	存取窗口
Slot	裂口
Slot	切口
Slot	时间段
Slot	时间片

Table 1.3 The shadow term set of the form entry ‘书目’

S-set entries	Example of the M-set entries	Form entry
书目	Bibliographical	书目
文 献 目 录	bibliography	书目
书目学	bibliography	书目
书 目 提 要	bibliography	书目
目录学	bibliography	书目

Definition 1.2 In a given bilingual term list, all the equivalent terms of a form entry x build up a set of equivalent terms of x - Mirror Term Set, denoted as M-set(x).[2]

Definition 1.3 In a given bilingual term list, while a term entry B∈M-set(x) and x ∈ M-set(A), B is defined as a shadow term entry of the form entry A. All the shadow term entries of A constitute the Shadow term entry set of A, denoted as S-set(A).
Deduction1.2.1 A∈S-set(A)

If there is only one none empty element in the S-set(A) (| S-set(A)| = 1), A is defined as a uni-shadow term.

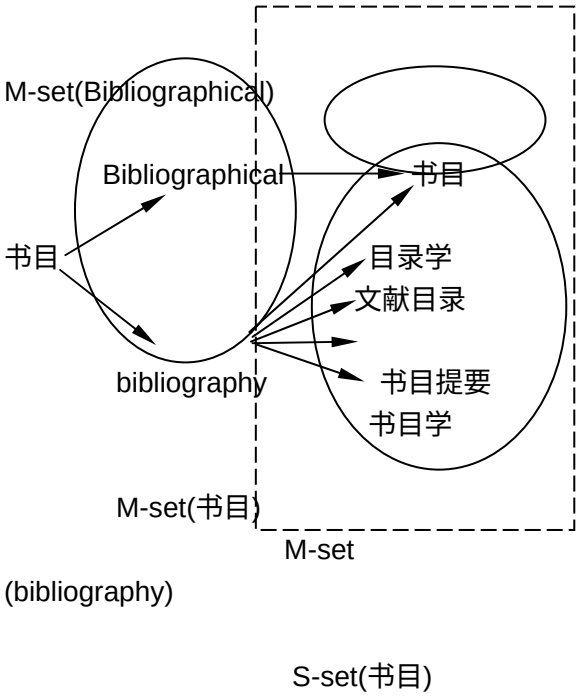

Figure 1.1 The illustration of the generate procedure of S-set(书目)

[2] Apparently, there are no duplicate term entries inside one Mirror term set. That means, inside one M-set, there have only one to one correspondences between the term entries and form entries. In this paper, M-set(x) is also used to refer to the ‘Mirror form entry set(x)’ that derived from the Mirror term set(x).

3 From S-set to synonymous term set

Hypothesis 3.1 While A is an uni-shadow term and |M-set(A)| >1, M-set(A) is likely to be a synonymous term set. In this paper, the M-set of an uni-shadow term is denoted as S-set.

Figure 3.1 illustrated the relation between uni-shadow term and its M-set terms. The related M-set terms are likely to be synonymous terms.

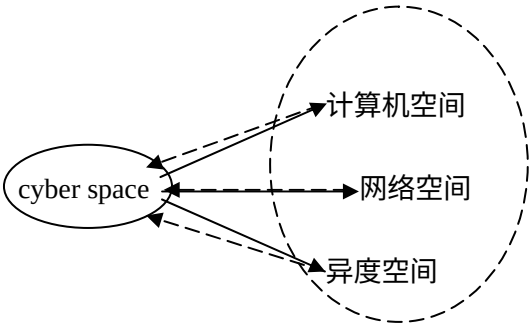

Figure 3.1 The uni-shadow term 'cyber space' and its M-set terms

Under this hypothesis, 9570 English synonymous term sets candidates and 24,098 Chinese synonymous term sets candidates were found in the collected 184,615 bilingual computer term entries.

Table 3.1a and 3.1b shows some examples of uni-shadow terms and their equivalent S-sets. In the tables, each S-set is separated with a blank line. Proof reading of the S-sets shows that apart from some typing errors, most of the term entries in a S-set are synonymous[3].

[3] The standard of synonymy is rather vague. Since the terms in the S-set share the same translation equivalence, they are treated as synonymous if not refer to completely different concept.

Table 3.1a English uni-shadow entries and their equivalent Chinese S-sets

English uni-shadow entries	Equivalent Chinese S-sets
0-type grammar	0 型文法
0-type grammar	0 型文法
critical error	重大错误
critical error	严重错误
critical error	关键性误差
critical error	临界错误
Customize	针对客户需求设计
Customize	专用化
Customize	自定义
Customize	针对客户需求修改
Customize	用户化
Customize	使符合客户需求
Customize	定做
Customize	订制
Customize	定制
cyberspace	信息空间
cyberspace	网络空间
cyberspace	异度空间
cyberspace	电脑空间
cyberspace	虚拟空间
cyberspace	赛伯空间

Table 3.1b Chinese uni-shadow entries and their equivalent English S-sets

Chinese uni-shadow entries	Equivalent English S-sets
网络节点	node of network
网络节点	network nodes
网络节点	network node

Chinese uni-shadow entries	Equivalent English S-sets
网络接口	network interface
网络接口	internetwork interface
网络技术	network technology
网络技术	network technique
网络管理	network administration
网络管理	network management
系统定义记录	system definition record
系统定义记录	system-defined record
系统程序	system program
系统程序	system routine
系统参数	system parameters
系统参数	parameter of a system
系统参数	System parameter

Compared with 9,570 English S-sets, there are 24,098 Chinese S-sets, more than twice as many. Some how it shows that there are more ambiguous expressions for the same term concept in Chinese than in English. Further more, the average entry number in Chinese S-sets is 2.638 and the average entry number 2.129. The data shows that there is a long way to go for Chinese terminology standardization.

4 Analysis of the S-set Terms

For English S-set terms, hyphen using responsible for one third (around 3000 cases) of all the S-set term sets. Like ‘end of job routine’ and ‘end-of-job routine’, the purge program do not identify them as one term entry, so they appear as synonymous in the next stage.

The ‘noun1+noun2’ term in English can often be written as the style ‘noun2 of noun1’ while noun1 behaves as a modifier. For example like ‘performance standard’ can be written as ‘standard of performance’. There are around 7 hundred synonymous pairs in this style.

Morphological inflection is another important reasons for the English synonymous terms. More than 500 hundred English S-sets are came from this reason, like ‘network node’ and ‘network nodes’.

Synonymous words used in the compound terms produce quit amount (around 1/3) of S-sets, like ‘system program’ and ‘system routine’.

Apparently, around half of the S-sets of the English terms are not really synonymous terms. Lemmatization and the orthographic standardization can eliminate most of the terminology ambiguity in this style.

But for Chinese S-set terms, the situation is much worse. Because most of the computer terms are translated from English, and the translation style are very different. That introduce quite amount of ambiguity of Chinese terms.

To solve this problem, the translation of the productive element of the terms in a specific field must be carefully standardized. For example, like the word ‘mode’, it can be translated mainly into ‘方

式’ or ‘模式’ or ‘型’. The difference among these words are subtle. Since this word is frequently used as an element of other terms, the ambiguous translation of this word produced more than 400 Chinese S-sets.

5 Conclusions

From the English-Chinese bilingual computer term list, more than 30 thousand S-sets were found, which in most of the cases are synonymous sets or orthographic variations.

Due to the very strict demand of the S-set defined in this paper, not all the synonymous terms are included in the S-sets. Future work needs to be done to find out the other synonymous term sets in the term list and also to gives out the recommended term for each synonymous term set.

Conclusions

1. Bai Xiaojing, Hu Junfeng, Zhan Hongying, Chen Yuzhong, Yu Shiwen, A corpus based approach to term bank construction, Workshop on "International Standards of Terminology and Language Resources Management" LREC 2002

2. Yu Xinli, Li Mingfei, Building a platform of automated terminology extraction and analysis, EAFterm Newsletter 2001 Vol.5

3. Feng Zhiwei, introduction of modern Terminology, Language Publishing House, 1997, BeiJing