

语义映射质量及影响因素分析

——以国家叙词库语义映射为例*

鲍秀林 吴雯娜

摘 要 国家叙词库采用以《汉语主题词表》为中心表、其他中文叙词表向中心表映射的模式进行基于概念的语义集成。为了研究学科交叉对国家叙词库语义映射造成的影响,本文选取八部叙词表中环境科学大类 5 120 个概念,与《汉语主题词表》环境科学大类 8 329 个概念进行映射实验。通过计算机自动匹配发现,多表映射存在一个概念等同映射到多个概念,多个概念等同映射到一个概念等逻辑问题。本文以实例形式分析这些逻辑问题的存在形式,并从概念专指度、概念语义模糊、语义关系结构、学科背景等方面分析产生这些逻辑问题的主要原因;在此基础上,提出解决这些逻辑问题的原则以及映射完整性的判定规则。图 16。表 5。参考文献 14。

关键词 语义映射 多表映射 等同映射 映射质量 国家叙词库

分类号 G254

Analysis of the Quality of Semantic Mapping and Its Influencing Factors: A Case of the National Thesauri Warehouse's Semantic Mapping

BAO Xiulin & WU Wenna

ABSTRACT

The National Thesauri Warehouse is an integrative service system of the thesaurus resources of China. Its construction is a large and complex project, which includes many aspects while its core part lies in the semantic integration of multi-thesaurus. Based on the concept of semantic integration, the National Thesauri Warehouse takes *Chinese Thesaurus* as the center vocabulary and uses other Chinese thesaurus mapping to the center vocabulary.

To explore the effect of interdisciplinary on the National Thesauri Warehouse semantic mapping, we select 5 120 concepts from the environmental science classes of 8 thesaurus and 8 329 concepts from the environmental science class of *Chinese Thesaurus* to take mapping experiments.

By analyzing the results of automatic matching, we find that the existence of quasi-synonyms, related words, hyponyms/hypernyms in words of equal relationships, especially inconsistent mapping methods of two thesaurus for synonyms, causing many logical problems in multi-thesauri mapping, such as one concept

* 本文系国家自然科学基金项目“国家叙词库构建方式与发展机制研究”(编号:13BTQ013)的研究成果之一。This article is an outcome of the project “Research on the Construction Mode and Development Mechanism of the National Thesauri Warehouse” (No. 13BTQ013) supported by National Social Science Foundation of China.

通信作者:鲍秀林,Email:baoxl926@163.com,ORCID:0000-0002-3427-8268(Correspondence should be addressed to BAO Xiulin,Email:baoxl926@163.com,ORCID:0000-0002-3427-8268)

equivalent mapping to multiple concepts or multiple concepts equivalent mapping to one concept. Since the notion of equivalent mapping is the foundation for recognizing and processing other mapping relationships, it can be further developed to hierarchical mapping and relevant mapping, and thus to enrich the network relationships among the conceptual nodes. Therefore, the main purpose of this paper is to study the logical problems generated by equivalent mapping.

In this paper, we analyze the forms of the logical problems and the causes at different dimensions, such as concept specificity, conceptual semantic fuzziness, semantic relationship structure, subject background. In view of these, the principles of solving these problems are put forward, and the decision rules of mapping integrity are also presented.

Semantic mapping of different thesaurus has many advantages including not only of great help to the compatibility and semantic specification, and improve the quality of the National Thesauri Warehouse, but also enriching and supplementing the semantic relation of *Chinese Thesaurus*, and promoting the integrated application of Chinese thesaurus. The analysis of the logical problems of semantic mapping and its influencing factors may have a reference value to the actual mapping work can be provided. The weakness of this paper lies in the limited experimental data, where there may be some logical problems that have not been completely found. 16 figs. 5 tabs. 14 refs.

KEY WORDS

Semantic mapping. Multi thesaurus mapping. Equivalent mapping. Mapping quality. The National Thesauri Warehouse.

0 研究背景

我国从 20 世纪 60 年代开始研究情报检索语言,迄今为止已经编制综合性或专业性汉语叙词表 150 多部^[1],部分叙词表在大的数据库知识组织中发挥着重要作用。伴随着计算机网络时代的到来,叙词表也从传统的应用模式扩展到网络应用。但从资源共享的角度来讲,叙词表资源种类越多,信息资源共享的障碍就越大。众多没有规律、不受控制增长的词表给文献检索新增了很多困难和麻烦。此外还有大部分叙词表以印刷本的形式存在,不为人知,且疏于管理和维护^[2]。因此,改变叙词表的存在状态,使之易于获取和利用,并在使用中发展和完善,成为很紧迫的任务。

2013 年,中国科学技术信息研究所承担了国家社会科学基金项目“国家叙词库构建方式与发展机制研究”。国家叙词库除了具备集成词表提

供不同层次的服务与应用的功能外,“整理、保存我国各学科领域叙词表”也是其重要功能之一,对全国范围内不同学科、不同版本的中文叙词表提供保存和管理功能,具备收录范围“全”的特点。此外,国家叙词库也接纳不同时代、不同版本的叙词表,包括不断出现的新编叙词表。

国家叙词库的构建是一个庞大而复杂的工程,涉及众多方面,但其核心在于多表的语义集成。国家叙词库采用以《汉语主题词表》(以下简称《汉表》)为中心词表,其他词表向《汉表》映射的模式进行基于概念的语义集成^[3]。为了研究国家叙词库语义映射因为学科交叉带来的影响,笔者选取横跨社会科学、自然科学和工程技术三大领域的环境科学大类的八部中文叙词表^[4-11]、5 120 个概念,与《汉表》环境科学大类^[12]8 329 个概念进行映射实验(见表 1),借助计算机自动匹配的结果见表 2。通过分析这些映射结果,发现多表映射存在的问题,进一步提出解决办法和处理原则。

表 1 实验数据范围

来源表	简称	范畴号	范畴名	叙词(个)	非叙词(个)
大气科学主题词表	大气表	07	大气化学、大气污染	123	24
煤炭科技文献检索词典	煤炭表	14	矿区环境保护	125	16
钢铁工业主题词表	钢铁表	X5	环境污染及其防治	258	38
		X7	三废处理与综合利用		
中国有色金属工业主题词表	有色金属表	23	环境科学	258	64
化工汉语叙词表	化工表	27	化工污染与环境保护	1 519	125
交通汉语主题词表	交通表	CK	环境科学	222	28
环境科学叙词表	环境表	02	环境污染与净化	1 054	198
		15C	大气污染防治		
		15D	水处理与水污染防治		
		15F	固体废物处理与利用		
		15G	放射性污染防治		
		15H	废物处理工艺、设备与材料		
环境保护档案主题词表	环保档案表	SA3	环境污染及其防治	1 561	263
汉语主题词表	汉表	X	环境科学	8 329	17 674

表 2 八部叙词表概念与《汉表》自动匹配的结果

指标	叙词表	大气表	煤炭表	钢铁表	有色金属表	化工表	交通表	环境表	环保档案表	合计
抽取数据	叙词	123	125	258	258	1 519	222	1 054	1 561	5 120
	非叙词	24	16	38	64	125	28	198	263	756
环境科学大类匹配	SPT-OPT	32	69	123	108	294	100	303	339	1 368
	SPT-ONPT	29	17	42	45	158	35	145	151	622
	SNPT-OPT	6	4	7	11	19	6	48	34	135
	SNPT-ONPT	10	6	9	24	28	11	50	43	181
	词汇匹配	77	96	181	188	499	152	546	567	2 306
	叙词匹配	62	86	165	154	457	138	454	493	2 009

(注:①SPT和OPT表示来源表叙词与目标表叙词的匹配;SPT和ONPT表示来源表叙词与目标表非叙词的匹配;SNPT和OPT表示来源表非叙词与目标表叙词的匹配;SNPT和ONPT表示来源表非叙词与目标表非叙词的匹配。②由于有部分叙词是通过非叙词匹配上,所以存在“叙词匹配”的结果大于“SPT-OPT”与“SPT-ONPT”之和的情况。)

1 映射逻辑问题表现形式

借助计算机,自动映射虽可减少叙词表映射的工作量,降低映射的复杂度,但仍需要人工对叙词表术语的概念语义进行正确判断,这也是保证映射质量的关键因素。由于依赖计算机辅助的自动映射使得概念脱离原有的语义环境,同时不同的叙词表在词形、词义和结构等方面存在差异,使得多表映射不可避免存在一些逻辑问题。

在 ISO25964-2 提出的三种映射类型 (Equivalence Mapping、Hierarchical Mapping、Relative Mapping) 中,等同映射 (符号为 EM) 优先级最高,是识别和处理其他映射关系的基础,等级映射 (分为向上/宽泛映射和向下/限定映射) 和相关映射 (符号为 RM) 则可在等同映射的基础上建立,从而丰富概念节点之间的网络关系,因此本文主要研究等同映射产生的逻辑问题。

一般来说,两个概念间的等同映射应该是一对一的,由于等同关系的词语间有完全同义词也有准同义词、相关词、上/下位词等,当来源表和目标表对同义词的处理方式不一致时,基于计算机自动映射产生的各种逻辑问题在所难免,比如单个概念等同映射到多个概念或多个概念等同映射到一个概念等情况。模型表示见图 1 (注:以下符号“D”表示代项词,“F”表示下位词,“C”表示相关词,“S”表示上位词,“BM”表示上位映射,“NM”表示下位映射,“RM”表示相关映射)。

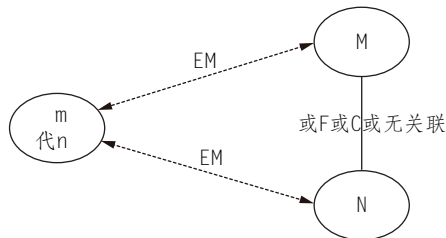


图 1 映射的逻辑问题表示模型

图 1 中,概念 m 与概念 n 在内涵上或为完

全同义词或为准同义词 (n 可能是 m 的完全同义词、近义词、下位词、上位词、相关词), 而其等同映射的两个概念 M、N 在另一部表中, 表现为相关关系、属分关系, 或没有建立起任何关联, 即 $m=n, m=M, n=N$ 时, 出现 $M \neq N$ 的逻辑问题。具体有以下三种表现形式。

(1) 当概念 m 与 n 来自来源表, M 与 N 为目标表概念时, 表现为一对多的映射问题。

(2) 当概念 m 与 n 来自目标表, M 与 N 为来源表概念时, 表现为多对一的映射问题。

(3) 当一对多和多对一交叉并存时, 衍生出多对多的映射问题; 当概念 M 不存在时, 而 n 为 m 的下位词或相关词时, 衍生出一对一的映射问题。

通过对表 2 等同映射的结果统计, 各来源表存在一对多和多对一的等同映射逻辑问题, 数量见表 3。其中, 两部环境专业叙词表的数量最大, 例如《环境表》中有 60 个概念与《汉表》中 140 个概念有一对多的映射关联, 有 116 概念与《汉表》中 53 个概念间存在多对一的映射关联。下面以实例来分析这些逻辑问题产生的原因及其影响因素, 从而提高映射的准确性并制定统一的处理原则。

表 3 来源表一对多和多对一映射数量统计结果

来源表	一对多		多对一	
	概念量	映射量	概念量	映射量
大气表	5	11	14	7
煤炭表	7	14	0	0
钢铁表	11	23	14	7
有色金属表	16	36	20	9
化工表	22	51	66	30
交通表	6	15	4	2
环境表	60	140	116	53
环保档案表	67	148	148	66

由于一对多的等同映射问题和多对一的问题是互逆的, 二者产生的原因以及二者交叉或

者衍生出来的逻辑问题,其原因也基本一致,所以只分析其中一个方向的问题即可。

2 映射的逻辑问题分析

2.1 来源表代项为同义词

当来源表代项本身是同义词或近义词,表达一个概念,但在目标表中却将它们建立了其他关系或者没有建立任何关联,使得它们在目标表中表达两个概念,这样来源表中的概念与目标表建立映射时,就会产生一对多的映射问题。

(1) 同义词在目标表为属分关系

一般来说“电除尘器”都是“静电除尘器”,《有色金属表》中二者为等同概念,《汉表》中编制者依据字面理解将它们处理为属分关系,实际上从《汉表》的“电除尘”下位词来看也只不过是“各种形式的除尘器”而已(见图2与图3)。

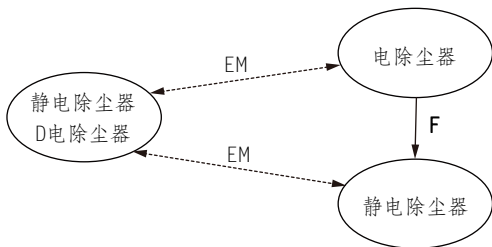


图2 来源表的同义词在目标表为属分关系示例

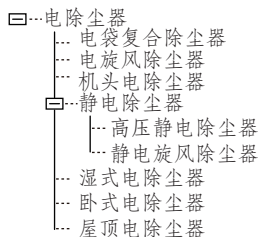


图3 《汉表》“电除尘器”等级关系

(2) 同义词在目标表为相关关系

如图4所示,《环境表》中“水污染”和“水体污染”本身为等同关系,《汉表》中将“水体污染”和“污染水体”两个不同语义的概念合并为同义词,与“水污染”建立了相关关系。

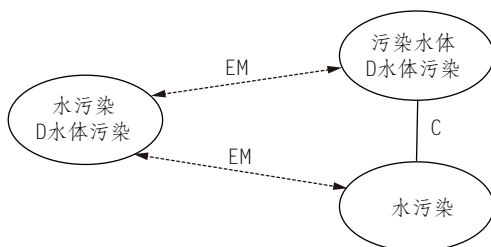


图4 来源表的同义词在目标表为相关关系示例

(3) 同义词在目标表为兄弟关系

如图5所示,《环保档案表》中“飞机噪声”的代项“机场噪声”含义为“飞机噪声的汇合”,其本质就是“飞机噪声”,二者为同义词。但是《汉表》中依据字面理解为两个不同内涵的概念,在“环境噪声”族下互为兄弟关系。

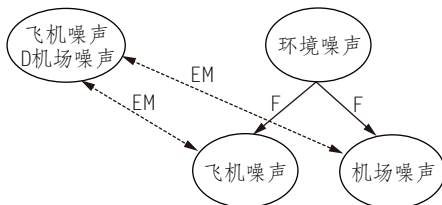


图5 来源表的同义词在目标表为兄弟关系示例

(4) 同义词在目标表中无关联

如图6所示,《钢铁表》中等同关系的“防潮剂”和“干燥剂”两个近义概念,等同映射到《汉表》的“防潮剂”和“干燥剂”,但二者在《汉表》中分别属“防护剂”和“除水剂”,分属两个不同词族,没有建立任何关联。

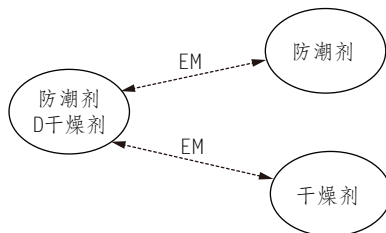


图6 来源表的近义词在目标表中无关联示例

2.2 来源表代项为上位词

如图7所示,《大气表》将“本底浓度”限定在“大气科学”范围内,与“大气本底值”建立了用代关系,但“本底浓度”实际含义为“人类活动

干扰前的环境状态下,地球生物圈中的大气、水体、土壤、生物等环境要素在自然形成和发展过程中,其本身原有的基本化学组成和能量分布”。从综合词表的角度看,“本底浓度”应该是“大气本底值”的上位词。在《汉表》中“本底浓度”恢复本义与“大气本底(D大气本底值)”建立了属分关系。

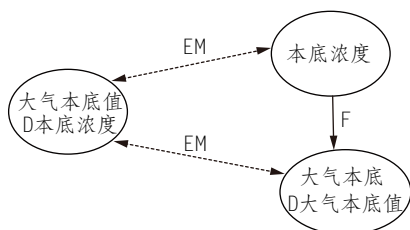


图7 来源表的代项在目标表为上位词示例

2.3 来源表代项为下位词

如图8所示,在《煤炭表》中从其专业的角度考虑,“水源保护”就是对“地下水保护”,所以二者建立了用代关系,但从综合词表的角度看,“水源保护”除了“地下水保护”还有“地上水源的保护”,所以在《汉表》中二者是上下位关系。

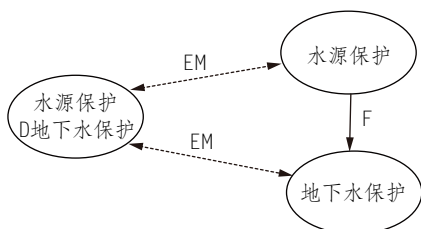


图8 来源表的代项在目标表为下位词示例

2.4 来源表代项为相关词

如图9所示,《环保档案表》中“γ射线辐

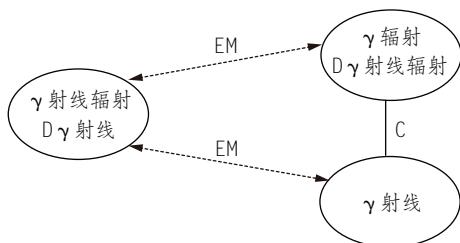


图9 来源表的代项在目标表为相关词示例

射”与其核心概念“γ射线”建立了用代关系,自动匹配到《汉表》的“γ辐射”和“γ射线”,二者为相关关系更合适,《汉表》的处理方式更合理。

2.5 衍生的逻辑问题

图1中,当M不存在时,m代项n为下位词或相关词时也会衍生同样的逻辑问题。

(1) 目标表代项为下位词的多对一问题衍生

如图10所示,《化工表》中有概念“苯酐生产尾气”,基于词形精确匹配到《目标表》“尾气”,原因是《汉表》中“尾气”代“苯酐生产尾气”,推导的“苯酐生产尾气=尾气”是错误的,二者实为属分关系。

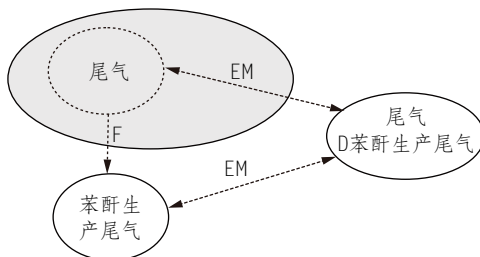


图10 代项为下位词导致的一对一映射问题

(2) 目标表代项为相关词的多对一问题衍生

如在《有色金属表》中,“毒物脱除”自动匹配到《汉表》“毒物控制”,二者实质为相关关系,在来源表中没有“毒物控制”一词(即阴影部分不存在),“毒物脱除”与目标表的“毒物控制”建立相关映射更合适(见图11)。

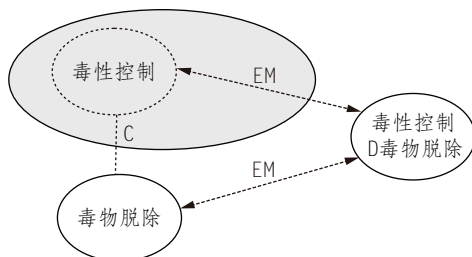


图11 代项为相关词导致的一对一映射问题

3 映射影响因素分析

《希望工程指南》(*Desire Project Handbook*)中提出,因不同的叙词表编制规则及应用等方面存在差异,映射是一个复杂、耗时的过程。Doerr^[13]认为术语选择、先组和后组程度、学科覆盖范围、语义概念和语义关联的不同是影响叙词表映射的主要问题。叙词表的映射需要考虑词表的学科背景(应用方向)、编制规则、专指度等异质性问题,综合实际需要,才能达到预期的映射效果。

3.1 概念专指度

一般来说,某一学科范畴的核心概念基本是确定的,每部词表根据自己的学科特点和编制规则来选词收词。通常情况下,综合叙词表的概念专指度相对较低,专业叙词表概念专指度相对较高;小型叙词表概念专指度相对较低,大型叙词表概念专指度相对较高。而叙词表概念专指度高低直接影响映射结果。例如,《化工表》的“有害废弃物海上焚烧”“空心毛细管填充柱”等,《环境表》“电子射线脱硫”“玻璃棉过滤器”等,在《汉表》仅收录了其上位或相关概念。

这是因为《化工表》仅收录化工相关的专业词汇,所以可收录更为专业的词汇,而《汉表》是一部综合词表,考虑其收词量有限,一般收录的词较为宽泛。从以上几例可以看出,《化工表》中过于专指的概念在《汉表》中没有能直接与其等同映射的概念。

3.2 概念语义模糊

在研究背景中提到同义词处理,由于代项中加入下位词、相关词等,使得概念的语义模糊,这会直接导致映射的准确性问题。

如表4中,前面4例为代项中代入下位词而不宜采用用代推导建立等同映射,建立上位映射更妥。除了《汉表》对低频词的处理规则,也存在编制人员对概念的不同理解而做的处理,第5个例子来源表将“大气监测”等同于“大气污染监测”,但在《汉表》中这两个词处理为上下位关系,处理时可依据《汉表》的思维理解为此处是以大代小;第6个例子也是目标表以大代小,但是这个大的概念却被上位词“辐射剂量”做了限定,而来源表没有对该词做学科限定,这样既不能建立上位映射更不能建立等同映射,可以建立相关映射。

表4 概念大代小示例

编号	来源表	概念	《汉表》概念
1	大气表	工业污染云	烟云(D 工业污染云)
2	大气表	光化学烟雾	化学烟雾(D 光化学烟雾)
3	化工表	苯酐生产尾气(S 空气污染物)	尾气(D 苯酐生产尾气)
4	大气表	体积排放速率(C 质量排放率)	排放速率(D 体积排放率)
5	大气表	大气监测(D 大气污染监测)	大气监测(F 大气污染监测)
6	有色金属表	最大允许剂量(S 最大允许水平 Z 标准)	允许剂量(D 最大允许剂量 S 辐射剂量)

我国词表编制高峰期在20世纪90年代以前,当初收录的词在随后的时间中可能被用得少,或有新生词出现,所以来源表中有一部分叙词,因其低频使用,《汉表》在编制时被作为上位词的入口词处理。也即源词表中某一叙词,在

《汉表》中为非优选词,且对应的优选词概念范围比该词大(上位词)。

3.3 语义关系结构

叙词表中的概念并非孤立存在,与其所处

的语义关系息息相关。概念映射需要综合考虑词语以及词间关系,除了等同关系,还要考虑上、下位和相关等语义关联。每一部词表都是一套独立的知识组织体系,因编制规则、编制人员知识背景以及收词专指度等的不同,编制的语义关系也会不一样,从而影响概念的映射。

例如“废物综合利用”一词,在《化工表》《环境表》《有色金属表》《汉表》中的关系结构见表5。其中《化工表》中“综合利用”与“废物综合利用”没建立任何关系,《环境表》中二者为用代关系,《有色金属表》中“废物利用”与“综合利用”是属分关系,而在《汉表》中“废物利用”与“废物综合利用”为用代关系,“废物利用”的上位词“资源利用”和“综合利用”都指引参到“利用”。《环境表》的下位词基本可直接与《汉表》中相应下位词映射,但是《化工表》的下位词都不可以。《有色金属表》的“废物利用”与“余热利用”“尾矿综合利用”为同级关系,但在《汉表》中三者分别在不同词族中,意味着映射到三个不同词族。

3.4 学科背景

叙词表中任何一个叙词的内涵,除了受其上下位词、相关词的限定外,还需要考虑所在词表的学科背景和叙词表应用的范围,词形相同的概念有可能表达不同的含义,因而不能直接建立等同映射。

如上例《环境表》中“废物利用”代“综合利用”,意味着“综合利用”的内涵限定在“废物利用”的范畴内,而《汉表》中“综合利用”是一个通用性概念,外延比《环境表》的“综合利用”要宽,所以尽管《环境表》“废物利用”和《汉表》“综合利用”通过非叙词词形匹配上了,但不能建立等同映射。而《有色金属表》的“综合利用”没有任何上位限定,其代项中近义词也不具备太多限定意义,所以可以和《汉表》“综合利用”建立等同映射。

综上所述,虽然基于词形或同义词匹配的

表5 “废物综合利用”语义结构示例

《化工表》	《汉表》
废物综合利用	废物利用(摘录)
S 化工污染与环境保护	D 废物资源利用率
F1 废物	D 废物综合利用
F1 废物处理	D 三废综合利用
F1 废物处理装置	D 变废为宝
F1 废物管理	D 废弃物利用
F1 废液回收	D 废品利用
	S 资源利用
综合利用 F27	F1 废气利用
S 生产管理	F1 污泥利用
	F1 废液利用
《环境表》	F2 废水利用
废物利用	F2 废酸利用
D 综合利用	F2 废油利用
D 三废综合利用	F1 固体废物利用
F1 废气利用	F2 废渣利用
F1 废水利用	F2 废纸利用
F1 废物资源化	F2 废玻璃利用
F1 粪便利用	F2 废塑料利用
F1 固体废物利用	F2 垃圾发电
F1 污泥利用	F1 污泥利用
F1 余热利用	C 利用
《有色金属表》	综合利用(摘录)
废物利用	F1 固体废物综合利用
S 废物处理	F1 矿产资源综合利用
S 综合利用	F2 矿石综合利用
	F3 尾矿综合利用
综合利用 X75	F3 煤矸石综合利用
D 综合回收	F1 煤气综合利用
F1 废物利用	F1 木材综合利用
F1 尾矿综合利用	F1 水资源综合利用
F1 余热利用	C 利用
C 炉渣处理	C 深加工

匹配度很高,但是并非词形相同或相似就等于概念内涵相同。总之,影响映射的因素有很多,可能不止笔者上述列出的几类情况,如要建立正确的映射关系,需要根据该术语的语义环境进一步判断其正确的含义,从而确保映射的准确性。人工对概念内涵判定时,需要从以下四个方面考虑^[14]:①词表学科范畴;②术语的释义;③术语的语义结构;④术语的用代关系。从范畴、释义、语义结构、用代关系的角度来确定来源表概念的真正内涵与外延,继而与目标表建立正确的映射关系。

4 映射中逻辑问题的处理规则

4.1 一对多的逻辑问题处理规则

不论是叙词表的编制者对词间关系处理不一致,还是来源表或目标表本身对概念处理错误所导致的一对多映射问题,因为来源表与目标表构建好的词间关系是不可改变的,在处理这类逻辑问题时,能做的就是修改映射类型。针对一对多的映射情况,统一的处理规则为以优选词表达的概念为准,断开基于非优选词词形自动匹配的映射关系。

如图12所示,不论目标表中M、N之间是何种关系,都以来源表优选词m为准,断开非优选n映射关系链。

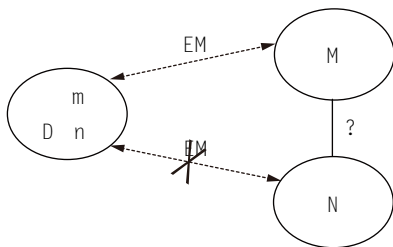


图12 一对多映射问题处理规则

4.2 多对一的逻辑问题处理规则

当目标表中同义或近义的两词在来源表中未构建等同关系或建立其他关系时,则保留自动推荐的映射关系。当目标表代项为下位词或相关或核心概念的方面词时,处理情况则以《汉

表》的优选词(或核心词)表达概念为准,由自动推荐的等同映射改为上位映射或相关映射。如图13所示,n为下位词时,与M建立上位映射;N为相关词时,与M建立相关映射。

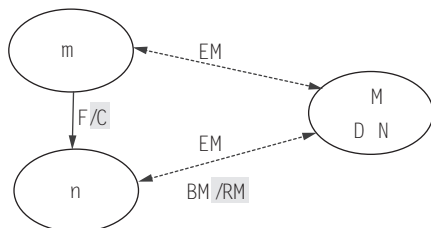


图13 目标表代项为下位词/相关词的处理规则

4.3 衍生的逻辑问题处理规则

衍生的多对多映射问题处理规则为断开因非优选词等同映射的关系链即可。在来源表概念m不存在的情况下,当n为M下位概念,与M的等同映射关系改为上位映射;当n为M相关概念,与M的等同映射关系改为相关映射。

5 结语

映射过程中为了确定正确的映射类型,需要对等同映射涉及的概念及词群进行语义分析,对于语义不明确的词群一般以优选词表达的概念为主。等同映射的判断依据为两个概念的位置互换后,原来概念体系的语义关系仍然成立。在等同映射的基础上,可对各自语义关系中无等同映射的概念建立等级映射或相关映射。

一般来说,来源表的一个概念只能等同映射目标表的一个概念,但等级映射和相关映射都有可能出现一对多映射情况,从逻辑上是允许存在的。在没有等同映射的前提下,在进行等级映射或相关映射时,判断该概念映射完整的依据如下。

(1)来源表的概念S在目标表中至少要有一个对应的上位概念M和一个对应的下位概念N,这样才能在目标表的知识组织体系中找到该概念的确定位置(见图14)。如《交通表》中“鸟兽害”在《汉表》中没有与之等同映射的概念,将

其上位映射到“动物灾害”,下位映射到“鸟害”,该概念在《汉表》的知识组织体系中准确定位在“动物灾害”和“鸟害”之间,如此才算完整,两者缺一都不能算完成映射。

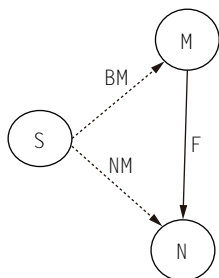


图 14 中间完整映射

(2) 来源表概念 S 的上位映射概念 N 是目标表中最小的(即等级关系链的最底端),目标表中不存在一个概念比 N 小,同时又比 S 大的情况(见图 15)。

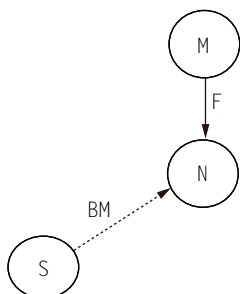


图 15 上位映射的完整性

(3) 来源表中概念 S 下位映射上的概念 M 是目标表中最大的(即等级关系链的顶端,族首词),即目标表中不存在一个概念比 M 大,同时又比 S 小的情况(见图 16)。

来源表中一个概念等级映射到目标表,在上述三种情况下才表示该概念的映射工作完成。

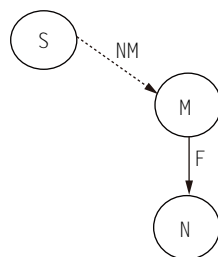


图 16 下位映射完整性

但是实际上存在如图 16 中有上位映射概念 M,没有下位映射概念 N,而概念 S 又不在等级链最低端的情况,这表示目标表的等级关系构建不完整,此种情况除了上位映射到概念 M,还应该建立 1—2 个相关映射才算完整。例如,《环境表》中“玻璃棉过滤器”上位映射到“纤维过滤器”,其与“纤维过滤器”下位词“纤维球过滤器”又不能建立上位映射,此种情况需要再建立合适的相关映射来表达正确的概念内涵,例如 RM“纤维棉”。

(4) 相关映射的完整性:相关关系的无方向性,使得相关映射也无方向性,相关映射的完整性也很难从逻辑角度进行判断,需要对概念进行语义分析。

总之,通过不同词表的语义映射,不仅对不同词表间的兼容性和语义规范以及国家叙词库的质量提升有很大帮助,还能够丰富和补充《汉表》自身的语义关系,促进我国词表的集成应用。而通过分析语义映射的逻辑问题及影响因素,可以为实际映射工作提供参考和借鉴。也就是说,概念的映射既要考虑其语义环境,判断映射的准确性,也需要判断概念映射的完整性。本文不足之处在于受实验数据的限制,可能还有一些逻辑问题并未完全发现。

参考文献

- [1] 鲍秀林,吴雯娜. 中文叙词表发展概况和性能测评(1980—2009)[J]. 图书馆论坛,2012(4):101-106.
(Bao Xiulin, Wu Wenna. Development situation and performance assessment of Chinese thesaurus (1980-2009) [J]. Library Tribune, 2012(4):101-106.)
- [2] 周杰,丁迺劲,吴雯娜,等. 网络环境下国家叙词库的构建研究[J]. 图书情报工作,2013,57(16):6-10.

- (Zhou Jie, Ding Qiuji, Wu Wenna, et al. Research on the construction of National Thesaurus System in the network environment[J]. Library and Information Service, 2013, 57(16): 6-10.)
- [3] 吴雯娜, 鲍秀林. 国家叙词库的体系结构与数据模型[J]. 中国图书馆学报, 2016(2): 81-96. (Wu Wenna, Bao Xiulin. The architecture and data model of National Thesauri Warehouse[J]. Journal of Library Science in China, 2016(2): 81-96.)
- [4] 中国气象科学研究院情报所. 大气科学主题词表[M]. 北京: 气象出版社, 1998. (Information Institute of China Meteorological Science Research Institute. Atmospheric Sciences thesaurus[M]. Beijing: China Meteorological Press, 1998.)
- [5] 《煤炭科技文献检索词典》编写组. 煤炭科技文献检索词典[M]. 北京: 煤炭工业出版社, 1986. (Establishment of Coal Science and Technology Literature Retrieval Dictionary. Coal Science and technology literature retrieval dictionary[M]. Beijing: China Coal Industry Publishing House, 1986.)
- [6] 冶金部情报标准研究总所. 钢铁工业主题词表[M]. 北京: 冶金部情报标准研究总所, 1991. (Research Institute of Information Standards of the Ministry of Metallurgy. The iron and steel industry thesaurus[M]. Beijing: Research Institute of Information Standards of the Ministry of Metallurgy, 1991.)
- [7] 中国有色金属工业信息研究所. 中国有色金属工业主题词表[M]. 北京: 机械工业出版社, 1993. (Information Research Institute of China Nonferrous Metals Industry. Chinese nonferrous metals industry thesaurus[M]. Beijing: Machinery Industry Press, 1993.)
- [8] 《中国化工文摘》编辑部. 化工汉语叙词表[M]. 北京: 中国化工信息中心, 1996. (China Chemical Abstracts Editorial Department. Chemical Chinese thesaurus [M]. Beijing: China National Chemical Information Center, 1996.)
- [9] 交通部科学研究院. 交通汉语主题词表[M]. 北京: 人民交通出版社, 2007. (Scientific Research Institute of the Ministry of Communications. Traffic Chinese thesaurus[M]. Beijing: China Communications Press, 2007.)
- [10] 《环境科学叙词表》编制组. 环境科学叙词表[M]. 北京: 中国环境科学出版社, 1989. (Establishment of Environmental Science Thesaurus. Environmental Science thesaurus [M]. Beijing: China Environmental Science Press, 1989.)
- [11] 国家环境保护局. 环境保护档案主题词表[M]. 北京: 中国环境科学出版社, 1995. (State Environmental Protection Administration. Environmental protection archives thesaurus [M]. Beijing: China Environmental Science Press, 1995.)
- [12] 中国科学技术信息研究所. 汉语主题词表(工程技术卷)[M]. 北京: 科学技术文献出版社, 2014. (Institute of Scientific and Technical Information of China. Chinese thesaurus (Engineering Volume) [M]. Beijing: Science and Technology Document Press, 2014.)
- [13] Doerr M. Semantic problems of thesaurus mapping[J/OL]. Journal of Digital Information, 2001, 1(8) [2015-06-24]. <https://journals.tdl.org/jodi/index.php/jodi/article/view/31/32>.
- [14] 陈辰. 叙词表映射语义判定和自动推理规则探析[J]. 图书情报工作, 2014, 58(2): 126-131. (Chen Chen. The research of semantic decision rules and inference rules in thesaurus mapping[J]. Library and Information Service, 2014, 58(2): 126-131.)

鲍秀林 中国科学技术信息研究所硕士研究生。北京 100038。

吴雯娜 中国科学技术信息研究所副研究馆员。北京 100038。

(收稿日期: 2016-03-17)