·医学信息研究·

面向医学领域知识组织系统整合的缩略语构成方式及歧义性鉴别研究*

李晓瑛 李丹亚 钱 庆 李军莲 孙海霞 胡铁军

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 对医学缩略语的构成方式进行梳理与归纳，对各种类型缩略语的歧义性进行详细地鉴别与比对，并就缩略语的歧义性对医学领域知识组织系统整合中术语归并所带来的影响进行探讨。

〔关键词〕 医学领域；知识组织系统；缩略语；构成方式；歧义性

Research on Abbreviation Composition Form and Ambiguity Identification for Medical Knowledge Organization System Integration *LI Xiao-ying, LI Dan-ya, QIAN Qing, LI Jun-lian, SUN Hai-xia, HU Tie-jun, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper summarizes the composition forms of medical abbreviations, the ambiguity of each kind of abbreviations is identified and compared in detail, followed by the discussion of their effects to terminology merging for medical knowledge organization system integration.

〔**Keywords**〕 Medical domain; Knowledge organization system; Abbreviation; Structure; Ambiguity

1 引言

1.1 术语歧义性鉴别

网络环境下，构建整合知识组织系统是一种实现领域术语集成的重要方法，也是本体、叙词表、分类表等不同知识组织系统之间互操作的基础。知识组织系统整合中的一项关键内容即是归并字面形式不同、但表达同一概念的多个术语，包括缩略语，但词形完全相同的术语在不同上下文语境下往往具有不同的含义。因此，就提高知识组织系统整合的准确性与效率而言，术语歧义性鉴别显得尤为重要。

1.2 缩略语

缩略语通常是对全称进行成分简化后，所获得的可以代表原全称及意义的特殊语言单位[1]。由于缩略语避免了使用者对其冗长全称的拼写，至今已普遍应用于日常交流与正式文书中。医学领域知识组织系统中亦存在大量的缩略语，如DNA（Deoxyribonucleic Acid，脱氧核糖核酸），AIDS（Acquired

〔收稿日期〕 2013-09-05

〔作者简介〕 李晓瑛，博士，助理研究员，发表论文10余篇；通讯作者：李丹亚。

〔基金项目〕 国家科技支撑计划“科技知识组织体系的协同工作系统和辅助工具开发”（项目编号：2011BAH10B02）；中央级公益性科研院所基本科研业务费“面向知识组织系统整合的英文同义关系自动发现技术研究”（项目编号：12R0116）。

Immune Deficiency Syndrome，获得性免疫缺陷综合症），SARS（Severe Acute Respiratory Syndromes，严重急性呼吸综合征）等，这些缩略语具有拼写简洁、数量庞大、构成方式复杂、新词产生迅速等特点。医学领域知识组织系统的缩略语极大地简化了医师及研究人员的拼写与交流过程，是医学术语的一种重要表达方式，至今已被广泛应用于医学信息检索与期刊文献中。但缩略语的简化构成方式在表达简洁的同时，其歧义性容易引起使用混淆及理解偏差，同时对医学领域知识组织系统整合中基于字面形式的术语同义归并造成一定的干扰。本文主要对当前医学领域知识组织系统中缩略语的构成方式与歧义性，及其对医学领域知识组织系统整合中术语同义归并所带来的影响进行系统研究、梳理及归纳总结。

2 医学缩略语构成方式研究

2.1 概述

英文缩略语通常由若干个字母构成[2]，但医学缩略语的构成方式却具有特殊性。在本文所调研的 100 多个医学领域知识组织系统中，有 33 个系统的概念名称存在缩略语，总数达 38 万。结合其字面表现形式，这些缩略语的构成方式主要为以下 3 类：仅由字母构成的字母型缩略语、含数字及符号的混合型缩略语、由多个单词组成的短语型缩略语。

2.2 仅由字母构成的字母型缩略语

根据构成缩略语的字母来源，这类缩略语又可分为 3 类：首字母缩略语、截取缩略语以及首音缩略语[3]。首字母缩略语由短语中每个单词的第 1 个字母构成，因其对应的全称字串过长，为方便拼写及增加期刊文献的可读性而选取其中若干字母作为代表，是目前最为常见的一种缩略语构成方式。这类缩略语主要用于表示疾病或药物、化学物质名称，例如美国国立癌症研究所（National Cancer Institute，NCI）叙词表（NCI Thesaurus，NCIt）中的 COLD（Chronic Obstructive Lung Disease，慢性阻塞性肺疾病）[4]、医师数据查询数据库（Physician Data Query，PDQ）中的 FA（Folic Acid，叶酸）[5]。截取缩略语通过截取全称中的某一部分构成。根据截取部位的差异，这类缩略语又可分为 4 类：一是截取首词缩略语，即只截取全称的首部而构成的缩略语，一般用于表示药物或化学物质名称，如 PDQ 中的 AMR（Amrubicin Hydrochloride，盐酸氨柔比星）；二是截取中词缩略语，即只截取全称的中部而构成的缩略语，如 MedlinePlus 健康主题（MedlinePlus Health Topics，MedlinePlus）[6]中的 Flu（Influenza，流行性感冒）；三是截取尾词缩略语，即只截取全称的尾部而构成的缩略语，如 NCIt 中的 Phone（Telephone，电话）；四是截取首尾词缩略语，即同时截取全称的首部与尾部而构成的缩略语，如 NCIt 中的 AA（Aruba，阿鲁巴岛）。首音缩略语通过截取全称中第 1 个音节以及后面音节的关键字母而构成，通常用于对字串较长且不易拼写的化学物质名称的简称，如医学主题词表（Medical Subject Headings，MeSH）[7]中的 BAAM（bromoacetylalprenololmenthane）。

2.3 含数字及符号的混合型缩略语

蛋白质、基因及化学物质名称中往往含有数字或符号，导致其缩略语中亦含有数字或符号。例如，MeSH 中的 A1PI（alpha 1 Proteinase Inhibitor α 1，蛋白酶抑制剂），在线孟德尔人类遗传数据库（Online Mendelian Inheritance in Man，OMIM）[8]中的 FA1（FERTILIZATION ANTIGEN 1，受精抗原 1），NCIt 中的 5 - ALA（5 - Aminolevulinic Acid，5 - 磺基水杨酸）、4′ - HPP（4 - Hydroxypyrazolo［3，4 - d］pyrimidine，4 - 羟基吡唑并［3，4 - d］嘧啶）。

2.4 由多个单词构成的短语型缩略语

短语型缩略语是一种非常特殊的缩略语构成方式，短语中的部分单词为简写，在医学领域知识组织系统中大量存在。例如，国际疾病分类法，第 9 版，临床修订版（International Classification of Diseases，Ninth Revision，Clinical Modification，ICD9CM）[9]中的 Abdmnal pain unspcf site（Abdominal pain，unspecified site；腹痛，部位不明），MeSH 中的 2 3 CYCL NPD

(2′, 3′- Cyclic - Nucleotide Phosphodiesterases; 2′, 3′-环核苷酸磷酸二酯酶)。

3 医学缩略语歧义性鉴别研究

3.1 概述

英文缩略语的主要构成单元是字母，但字母通常只有音和形，不具有内在含义[3]，加之缩略语的构建过程中人为因素较大，最终导致同一个缩略语经常可表达多种含义。虽然医学领域知识组织系统中的缩略语具有学科局限性，但其歧义性仍不可忽视。例如，仅由一个字母组成的缩略语 C 在 NCIt 这一知识组织系统内部就有多种不同的理解：热量单位卡（Calorie）、电量单位库仑（Coulomb）及摄氏温度数（Degree Celsius），脱离了上下文语境，很难分辨 C 究竟表示什么意思，而这种歧义性也对医学知识组织系统整合等研究工作带来很大的困扰。本文对上述 3 种类型缩略语的歧义性逐一进行了鉴别与研究。

3.2 字母型缩略语歧义性鉴别

在本文所调研的医学领域知识组织系统中，仅由字母构成的字母型缩略语共有 11 617 个，约占缩略语总数的 3%，这类缩略语的字串长度最大为 33，最小为 1。通过对具有多种含义的歧义缩略语进行统计，发现其中 1 371 个缩略语具有歧义性，占此类缩略语总数的 11.8%。例如，在 NCIt 中 IPS 是特发性肺炎综合征（Idiopathic Pneumonia Syndrome）的缩写，而在 OMIM 中 IPS 代表鱼鳞病早产综合征（ICHTHYOSIS PREMATURITY SYNDROME）。

此外，本文对字母型缩略语在不同字串长度下的数量分布及歧义性进行了研究，结果见表 1。由表 1 可知，在医学领域知识组织系统中，81.8% 的字母型缩略语长度不超过 5，而 92.6% 的歧义性缩略语也分布在这个长度范围。因此，在面向医学领域知识组织系统整合的术语同义归并中应特别关注串长不超过 5 的字母型缩略语，以保证术语同义归并的准确率及可靠性。

表 1 字母型缩略语在不同串长下的数量分布与歧义性统计结果

缩略语串长	1	2	3	4	5	≥6
缩略语个数	42	588	3 119	3 976	2 376	2 111
具多义缩略语个数	14	243	723	238	52	101

3.3 混合型缩略语歧义性鉴别

就本文所分析的医学领域知识组织系统而言，含数字及符号的混合型缩略语共有 45 196 个，占缩略语总数的 11.9%，混合型缩略语的字串长度最大为 52，最小为 1，而其中的 824 个缩略语具有歧义，占总数的 1.8%。例如缩略语 HA2，在 NCIt 中为血吸附病毒 2 型（Hemadsorption Type 2 Virus）的简称，而在 OMIM 中的含义同次要组织相容性抗原（MINOR HISTOCOMPATIBILITY ANTIGEN）。表 2 列出混合型缩略语在不同字串长度下的数量分布及歧义性统计结果。从中可知，89.2% 的混合型缩略语长度不超过 10，而这个阈值也正是所有歧义性缩略语的最大长度。可见，具有歧义的混合型缩略语所占比例很小，而字串长度大于 10 的混合型缩略语完全没有歧义，在基于字面形式的医学领域术语同义归并中全然可忽略此类混合型缩略语的歧义性。

表 2 混合型缩略语在不同串长下的数量分布与歧义性统计结果

缩略语串长	1	2	3	4	5	6	7	8	9	10	≥11
缩略语个数	7	57	571	5 319	10 151	8 937	7 817	4 955	1 724	794	4 846
多义缩略语个数	1	9	47	380	266	97	14	6	1	3	0

3.4 短语型缩略语歧义性鉴别

相比前两种类型，短语型缩略语共有 324 597 个，在整个医学领域缩略语中的比重最大，占缩略语总数的 85.1%，其中 3 349 个缩略语具有歧义性，比例仅为 1%。由于短语型缩略语中含有多

个单词，本文对不同组配单词个数下缩略语的数量分布与歧义性进行了统计分析，见表 3。从中可发现 87.9% 的短语型缩略语所含单词个数不超过 6，且这个范围覆盖了 99% 的多义短语型缩略语。此外随着组配单词个数的增加，不仅因短语型缩略语的简洁性逐渐降低，缩略语个数呈下降趋势，而且因单词间相互限定作用的加强，多义缩略语个数亦极大减少。

表 3 短语型缩略语在不同组成单词下的数量分布与歧义性统计结果

组配单词个数	2	3	4	5	6	7	8	9	10	≥11
缩略语个数	66 862	61 709	73 888	56 707	26 236	12 436	9 054	7 334	5 290	5 081
多义缩略语个数	413	1 152	1 241	439	70	20	8	3	3	0

为了定量地比较医学领域中 3 种类型缩略语的歧义性，本文提出一种缩略语歧义性的量化公式：

$$\text{歧义性} = \frac{\text{具有歧义的缩略语个数}}{\text{缩略语总数}} \quad (1)$$

利用公式 1，可分别算出这 3 种类型缩略语的歧义性。结合其数量比例，见表 4，发现尽管混合型与短语型缩略语占缩略语总数的比重很大，但其歧义性却非常小，因此在实际研究中可忽略这两种类型的缩略语，而将其看作普通医学术语。相对而言字母型缩略语的总数最少，但歧义性最大，且字母型缩略语的字串长度越小，简洁性越高，但歧义性越大（特别是长度不超过 5 的字母型缩略语）。在进行医学领域知识组织系统整合等研究时，对于出现在两个不同医学知识组织系统中词形完全相同的字母型缩略语，需进一步借助全称、同义词、定义、注释或其他辅助信息来判断其是否含义相同，从而提高术语同义归并的准确率。

表 4 3 种类型的缩略语
数量比例与歧义性比较结果（%）

类型	字母型	混合型	短语型
数量比例	3	11.9	85.1
歧义性	11.8	1.8	1

4 结论

缩略语的简洁性与歧义性是一对相互矛盾的特性。就本文所调研的医学领域知识组织系统中的缩略语而言，受字串长度与组配单词个数的影响，混合型与短语型缩略语的简洁性相对较差，但因其所含数字（符号）或多个组配单词之间的相互限定作用，这两类缩略语的歧义性最小。字母型缩略语因其最小的字串长度而具有最高的简洁性，但其歧义性却不容忽视，特别是字串长度不超过 5 的字母型缩略语。因此，为了减少字母型缩略语的歧义性对医学领域知识组织系统整合中术语同义归并工作带来的干扰，提高归并结果的准确性，建议研究者借助相应的全称、同义词、定义、注释或其他辅助信息来准确断定缩略语的具体含义。

参考文献

1 韩昉. 缩略语的产生特点及其规范性 [J]. 语文学刊, 2007, (11): 101 - 103.

2 Wikepdeia [EB/OL]. [2012 - 10 - 02]. http://en.wikipedia.org/wiki/Abbreviation.

3 张曲. 英语缩略语刍议 [J]. 当代教育理论与实践, 2010, (2): 118 - 120.

4 NCI Thesaurus [EB/OL]. [2012 - 05 - 18]. http://ncit.nci.nih.gov/.

5 PDQ [EB/OL]. [2011 - 07 - 11]. http://www.cancer.gov/cancertopics/pdq/cancerdatabase/.

6 MedlinePlus [EB/OL]. [2012 - 10 - 09]. http://www.nlm.nih.gov/medlineplus/healthtopics.html/.

7 MeSH Browser [EB/OL]. [2012 - 09 - 04]. http://www.nlm.nih.gov/mesh/MBrowser.html.

8 OMIM [EB/OL]. [2012 - 10 - 14]. http://www.ncbi.nlm.nih.gov/omim/.

9 ICD9CM [EB/OL]. [2012 - 10 - 04]. http://www.cdc.gov/nchs/icd/icd9cm.htm.