UMLS多词表整合机制研究*

□ 李丹亚 胡铁军 李亚子 李晓瑛 孙海霞 李军莲 钱庆 / 中国医学科学院医学信息研究所 北京 100020

摘要：文章基于对UMLS超级叙词表、语义网、专家辞典的分析，及对UMLS系统结构和数据的解析，重点探讨UMLS多来源词表同义概念的归并机制、异构异质词表术语、关系、属性信息的集成方法，以及范畴体系的构建理念，最后就UMLS多词表整合策略进行讨论。

关键词：UMLS，词表整合，概念结构，语义网，词汇工具
DOI：10.3772/j.issn.1673—2286.2012.04.006

1 引言

UMLS（Unified Medical Language System[1]，一体化医学语言系统）是生物医学领域、跨语言多词表集成的知识组织系统。知识组织系统被用来组织信息资源，提供基于知识内容的各种服务。

UMLS由美国国立医学图书馆（NLM）创建，目的是使信息系统能够理解生物医学领域同一概念的不同表达形式，克服相同概念分散于不同信息系统导致的诸多使用障碍。UMLS试图在用户和信息资源之间建立具有相同知识内容的关联。UMLS在国外已被广泛应用于医疗信息系统、病案系统、自然语言处理、文本自动标注、智能检索及搜索引擎等领域。

UMLS由超级叙词表（Metathesaurus）、语义网（Semantic Network）、专家词典（SPECIALIST Lexicon）三个密切关联的知识源（UMLS Knowledge Sources）和工具组成[2]。超级叙词表是生物医学词表概念、术语、涵义、关系的广泛集成；语义网为超级叙词表中的概念提供统一的组织和分类，并揭示概念之间的语义关系；专家词典和工具是超级叙词表建立和维护所需的生物医学词汇库和一组工具集。

UMLS整合了上百个异构异质词表中的上千万个概念名称及相关信息，这些极度丰富的知识源的集合是复杂的。本文基于对UMLS超级叙词表、语义网、专家词典和工具的分析，及对UMLS系统结构及数据的解析，试图探讨UMLS的多词表整合机制，以期为我国网络环境下知识组织系统的建设与应用提供参考。

2 超级叙词表的构建

超级叙词表是一个非常庞大的术语库，集成了生物医学和健康方面的本体、叙词表、分类表、疾病编码集、专家系统、词汇表中的术语及相关信息。2011AA版的超级叙词表整合了160个词表，包括2404937个概念，10655002个来源词表的概念名称，21种语言[3]。超级叙词表构建过程中需解决两个

* 本文系国家“十二五”科技支撑计划课题“科技知识组织体系协同工作系统和辅助工具开发”（编号：2011BAH10B02）及“面向外文科技文献的超级科技词表和本体建设”（编号：2011BAH10B01）”的研究成果之一。

主要问题：实现不同来源词表同义概念的归并，解决异构异质词表术语、关系、属性信息的统一管理和集成。

2.1 多来源词表同义概念的组织

超级叙词表以概念（concept）为核心进行组织，所有来源词表具有同样涵义的词和短语组成概念或同义词类（synonym class），每个概念与其他概念之间以语义邻居（semantic neighborhood）[4]方式形成不同词表概念间的语义关联。

2.1.1 概念模型的构建

超级叙词表的概念结构（Metathesaurus concept structure）[4,5]起着连接不同来源词表相同概念交替名称及名称变体的作用，是实现多词表同义概念整合的基础。概念结构由概念名称、概念名称标识符、概念名称的关键特性构成，超级叙词表的概念结构如图1所示。

（1）概念和概念标识符：一个概念是一个涵义（meaning），

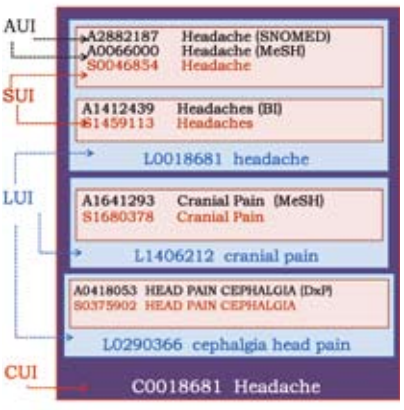

图1 超级叙词表的概念结构[6]

可以用许多不同的名称表示。超级叙词表的重要任务是明确来源词表每个名称的确切涵义，连接不同来源词表具有相同涵义的所有名称。超级叙词表中每个概念都分配一个唯一的永久的概念标识符（Concept Identifier，CUI），相同CUI代表概念的同义词集合。

（2）概念名称及字符串标识符：概念名称是指概念的术语表示形式，超级叙词表中的每个概念的名称或字符串都被赋予一个唯一永久的字符串标识符（String Identifier，SUI）。字符串的任何改变、大小写或标点符号等方面的不同都被看做一个独立的字符串或名称，分配一个独立的SUI。

（3）来源术语及标识符：来源词表的概念名称被称为“来源术语（atoms）”，来源术语是构建超级叙词表的最基本单位，每个来源词表中的每个字符串都要分配一个唯一的永久的来源术语标识符（Atom Identifier，AUI），相同字符串的AUI被连接到一个SUI。

（4）原形化术语及标识符：超级叙词表中的英文条目，每个字符串都要进行词汇的原形化处理，如名词复数转换成单数、动词变化形式转换成动词原形、形容词比较级最高级转换为形容词原级等，其目的是将同一术语的各种词汇变体形式连接到一起。超级叙词表的每个原形化术语都分配一个唯一永久的词汇标识符（Lexical Identifiers，LUI）。

（5）CUI、LUI、SUI、AUI的关系：超级叙词表的概念结构通过CUI-LUI-SUI-AUI将不同来源词表一个概念的多个异形同义术语连同多个词串变体形式连接起来。图1中的“Headache（头痛）”是一个来源术语，出现在两个来源表（SNOMED和（MeSH），每次出现都分配一个不同的AUI，由于两个词串相同，连接到一个SUI；“Headaches”是“Headache”的复数形式，两者有不同的SUI，由于单数、复数是彼此间的词汇变形，两者连接到同一个LUI；同样“Cranial Pain”和“HEAD PAIN CEPHALGIA”分别可以有一个 LUI、多个不同的SUI和多个不同的AUI，由于上述三个LUI有同样的涵义，均连接到同一个CUI。

2.1.2 概念优选名称的推荐

概念优选名称是指一组同义词中概念的正式表达名称。为方便超级叙词表的构建，避免人工选词这类繁重的工作，超级叙词表建立了概念默认优选名称（Default Preferred Names）推荐机制，系统自动为同义词群指定优选名称，使用者也可根据特定需求，通过定制改变概念的优选名称。

概念默认优选名称的指定是基于超级叙词表所有来源词表的指定等级及术语类型的优先顺序确立的，优先顺序的确定主要考虑来源词表的覆盖范围、正规临床或生物医学领域使用这些概念名称的程度等因素。超级叙词表的一个专门文档（MRRANK）用于保存默认优先顺序队列表，内容包括顺序号、来源词表缩写名称、字符串类型，如“0027|MSH|MH|N|”。

2.1.3 概念名称文档元数据

超级叙词表概念结构及来源词表的一些重要信息保存在概念

名称文档（MRCONSO），元数据项包括：概念字符串，标识概念词形变化及词义变化的CUI、LUI、SUI、AUI，标识概念优选名称的LUI状态（TS）、SUI状态（STT）和AUI状态（ISPREF），以及来源表缩写名称、来源表唯一标识符、来源表术语类型（TTY）等，其中TTY取值涵盖了所有来源词表260多个术语类型的缩写名称[7]，图2是MRCONSO文档的一个实例[8]。

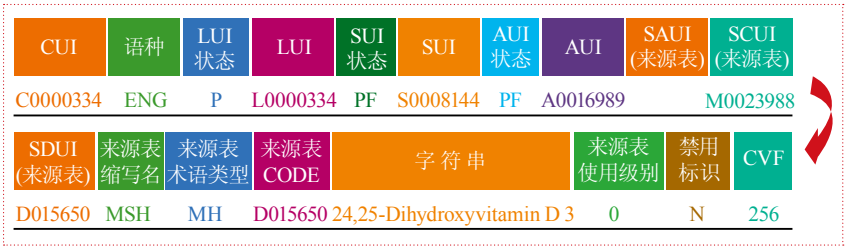

图2 概念名称文档（MRCONSO）实例

2.1.4 同形异义概念的处理

有时一个词串能表达两个或多个概念，如“COLD”在一个来源词表中是表示温度的名称，而在另一个词表中，却是“感冒”的交替名称，在第三个词表中，“COLD”是“Chronic Obstructive Lung Disease”的缩写名称。超级叙词表中的AMBIGSUI（SUI歧义文档）、AMBIGLUI（LUI歧义文档）专门用于鉴别这种同形异义词，文档包括一个SUI与多个CUI关联、或一个LUI与多个CUI关联的同形异义字符词串，如“S0828891|C0007018|”和“S0828891| C1171258|”表示同一个词串“CO”有两个概念，前者表示“Carbon Monoxide”，后者表示“Complication Aspects”。

超级叙词表保留来源词表的词义，如果两个来源词表使用同一名称表达不同的概念，超级叙词表将两种涵义都表达出来，并说明哪个涵义来自于哪个来源词表。如果不同来源词表中的两个概念相互冲突，两种观点都包含在超级叙词表中。

2.2 多词表关系属性的组织

超级叙词表除对多来源词表同义关系进行整合外，对来源词表中的其他关系主要采用保留和继承策略，如何将诸多来源词表的关系统一保存并能够还原每个来源词表的关系，是超级叙词表需要解决的重要问题。

图3 通用关系文档（MRREL）实例

NLM经过对所有来源词表的仔细分析，提取出词表中明确标识或暗指的各种关系，并将关系归纳为通用关系、完整层级关系、共现关系和映射关系四类，通用关系又分为11种。这里所涉及的关系不包括超级叙词表概念结构中的同义关系。

2.2.1 建立通用关系标签及标识符

超级叙词表中的每一种通用关系均分配一个基本关系标签（REL），描述关系的性质，如RB（广义关系）、RN（狭义关系）、PAR（直接上位关系）、CHD（直接下位关系）、同位关系（SIB）、RO（相关关系）、SY（同义关系）、RL（相似关系）、RQ（可能同义关系）、QA（限定关系）和QB（被限定关系），共11种。

超级叙词表中约1/4的关系带有来源词表的附加关系标签（RELA）[5]，更加确切地解释关系的性质，如“is a”、“branch of”、“component of”等。附加关系标签可以描述来源词表中的各种关系表示形式，目前超级叙词表的附加关系标签表述形式有600多种[7]。

超级叙词表采用统一的基本关系标签和来源词表特有的附加关系标签的表示方式，对来源词表的各种复杂关系进行组织并将其明确标识出来。图3是通用关系文档（MRREL）两者关系标签的应用实例[9]。

超级叙词表中的每个关系对都有一个关系标识符（Relationship Identifiers，RUI），关系标识符的出现或消失表示超级叙词表关系的变化。

2.2.2 建立来源词表层级结构路径

超级叙词表基于来源词表明确标识或暗指的等级关系，还原了每个层级结构的完整路径。图4是完整层级结构路径文档（MRHIER）的一个实例[10]。如果来源术语有多个层级，每个层级结构都被保留，采用AUI列表方式表示来源术语直接上位节点到层级结构顶层的完整路径，用数字表示属于第几个层级。完整层级结构路径的建立有助于超级叙词表方便地对每个来源词表进行浏览导航、查询、显示和调用。

图4 层级结构路径文档（MRHIER）实例

2.2.3 建立或保留共现和映射关系

超级叙词表中的共现关系可以通过某些信息源同一记录中共同出现的概念名称频次计算出来，如MEDLINE数据库，概念的共现关系是计算同一文献中主要主题词共同出现的频次。共现关系可能存在于类似概念之间，或存在于虽然很不同但具有某种重要关联的概念之间，也可能出现在没有其他明显关系的两个概念之间。映射关系文档包括MRMAP及精简版MRSMAP。共现和映射关系有些继承于来源词表，有些是超级叙词表构建过程中创建的。

2.3 多词表描述属性的组织

超级叙词表中的描述属性是指关于概念或来源术语的特征信息。来源术语属性来自某一个词表，有些属性是通用的，另一些属性仅与特定来源词表相关，如“Headache”的定义来源于MeSH词表，但它在超级叙词表中也适用于其他术语，相反，一个来源术语的使用日期则仅适用于该术语。概念属性是在超级叙词表建设过程中生成的，可应用到一个概念的所用名称，如语义类型“Sign or Symptom（体征或症状）”既是“Headache”概念优选名称的属性，也适用于连接到此概念的所有术语。

超级叙词表依据多来源词表属性的共性和特性对来源词表的定义、语言类型等进行单独管理，对其他多样化描述属性信息则进行统一组织和管理。每个概念或来源术语可以有多条属性信息，每条属性信息都用一条单独的记录描述，通过设置属性名称（ATN）和属性内容（ATV）元数据项，将不同来源词表的属性缩写名称标识符和属性内容全部继承下来，目前超级叙词表中的属性缩写名称近600种[11]。图5是一般属性文档（MRSAT）有关MeSH词表“ABO Blood-Group System”的一条标引注释实例[12]。超级叙词表的每个属性都被分配一个属性标识符（Attribute Identifiers，ATUIs），属性标识符的出现或消失表示超级叙词表内容的变化。

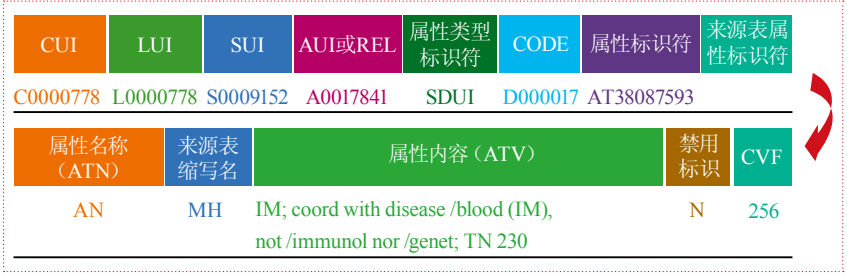

图5 一般属性文档（MRSAT）实例

2.4 异构异质词表的整合架构

2.4.1 超级叙词表的框架

超级叙词表收录了本体、叙词表、分类表等多种不同类型的知识组织工具，这些来源词表的元数据项和存储格式各不相同，NLM基于对诸多来源词表信息的分析和梳理，设计了能够充分兼容各来源

词表的超级叙词表的架构及元数据框架。图6为超级叙词表的框架结构图。

超级叙词表由词表内容管理、词表版本管理、数据管理和概念名称索引四部分组成。词表内容管理包括术语描述、关系描述、其他属性描述三个主要部分，术语描述包括对概念名称及歧义术语的管理，关系包括通用关系、等级关系、映射关系、共现关系等；属性描述信息包括定义、历史信息、语义类型、一般属性信息等。

2.4.2 超级叙词表的数据文档

超级叙词表提供两种格式的数据文档：RRF格式（Rich Release Format）和ORF格式（Original Release Format）[13]。RRF格式功能更强，采用简单的数据模型，利用文本格式行和列的方式组织数据，如“C0018681|T184|A2.2.2|Sign or Symptom |AT17639733|3840|”，经过长时间的运行表明这种方式能很好支持超级叙词表复杂数据结构的表述[14]。

超级叙词表由20多个文档构成。词表内容方面的文档以术语、关系属性、描述属性为对象对多来源词表内容进行组织和描述，版本管理文档对术语变更（MRCUI、MRAUI）、删除（DELETEDCUI、DELETEDLUI和 DELETEDSUI）及合并（MERGEDCUI、MERGEDLUI）信息进行管理，数据管理文档包括对超级叙词表本身的各种文档、元数据、缩写名、来源表、列表等进行管理，此外还包括20多个不同语种的概念名称词索引，及英语的原形化词索引和原形化字符串索引。

概念名称和来源信息文档（MRCONSO）是超级叙词表的核心，通过唯一标识符，与关系类文档、属性类文档及其他文档相互关联，将所有来源词表的数据有序组织起来。

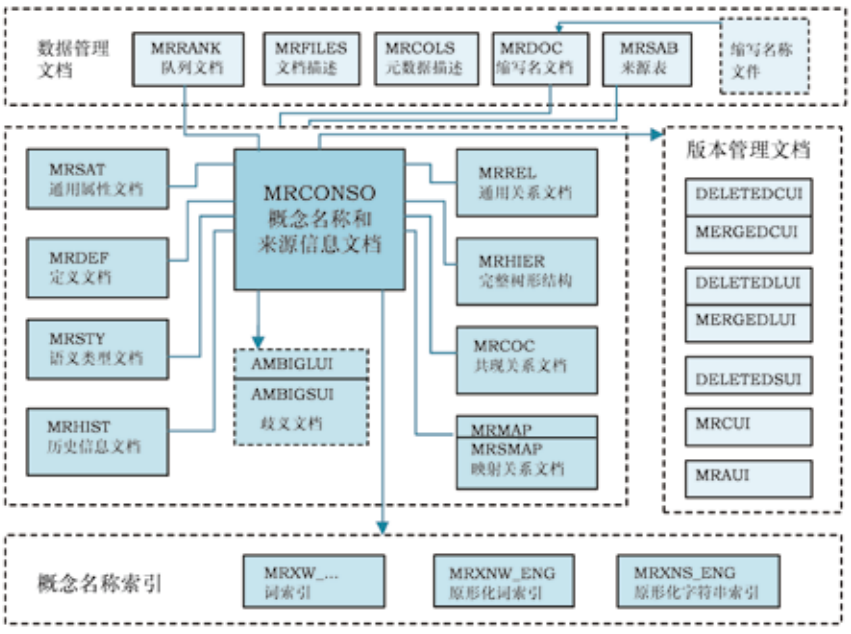

图6 超级叙词表的框架结构

3 UMLS语义网的体系结构

UMLS语义网将超级叙词表中的概念紧密关联起来，起到组织和统领超级叙词表的作用。语义网由两部分构成：语义类型是一组较为宽泛的范畴类目，提供超级叙词表中所有概念的统一分类；语义关系则是一组存在于语义类型之间的关系。语义类型是语义网中的节点，语义关系是节点之间的链接。

虽然超级叙词表的概念来源于不同词表，它们有着各自的范畴体系结构，但NLM将全部概念进行聚类汇集，遴选出概念范畴（即语义类型），建立了语义类型间特定的语义关系，从而将超级叙词表中的所有概念在统一范畴体系下进行组织，并揭示出较传统叙词表更为丰富的语义关系。语义网试图使计算机理解通用意义的知识，以便使其能根据已赋予实体的类别属性进行推理，并得出结论[4]。

3.1 基于语义类型的高层类目

UMLS语义网设计了135种语义类型[15]，涵盖了许多领域宽泛词汇的范畴，为超级叙词表不同来源词表搭建起统一的范畴。语义类型的高层分为“实体”和“事件”两大类，由此逐级展开形成树形结构，最深可达6级，见图7。超级叙词表中的每个概念至少被分配一个最专指的语义类型，语义类型的分配基于来源词表的信息确定。

A Entity(实体)<T071>
A1 Physical Object(物理对象)<T072>
A1.1 Organism(生物体)<T001>
A1.2 Anatomical Structure(解剖学结构)<T017>
A1.2.1 Embryonic Structure(胚胎结构)<T018>
A1.2.2 Anatomical Abnormality(解剖学异常)<T190>
A1.2.2.1 Congenital Abnormality(先天性异常)<T019>
A1.2.2.2 Acquired Abnormality(获得性异常)<T020>
A2 Conceptual Entity(概念性实体)<T077>
B Event(事件)<T051>
B1 Activity(活动)<T052>
B2 Phenomenon or Process(现象与过程)<T067>

图7 语义类型等级结构示例

3.2 基于语义关系的概念关系网络

UMLS语义网设计了54种语义关系[16]，包括等级关系链（isa）和一组非等级的相关关系链。“isa”关系是指语义类型和语义关系树形结构中的等级关系，关系具有继承性，一些情况下语义类型与继承的子类发生冲突，此时链接会被阻断。相关关系链包括五类：物理上相关、空间上相关、功能上相关、时间上相关、概念上相关，并依次细分，如图8所示。

语义网的54种语义关系虽然尚不能表达概念语义类型之间可能存在的一切关系，但比传统叙词表所能表达的等级、相关关系要细致和丰富得多。语义类型之间的关系表示一种可能、允许的关系，并非所有关系存在于两个特定概念对之间。如“治疗”虽然是语义类型“药用物质”和“疾病或综合征”之间存在的有效关系，但对于某种专指的药物如“青霉素”来说，就不能用于治疗某个专指的疾病，如“艾滋病”。

H ISA (等级关系链) <T186>
R Associated_with (相关关系链) <T166>
R1 Physically_related_to(物理上相关)<T132>
R1.4 connected_to(与…连接) <T174>
R2 Spatially_related_to(空间上相关)<T189>
R2.1 location_of (位于…) <T135>
R3 Functionally_related_to(功能上相关)<T139>
R3.1 affects (影响) <T151>
……
R3.1.3 disrupts(干扰) <T132>
R3.1.4 complicates(并发) <T149>
R4 Temporally_related_to(时间上相关)<T136>
R4.1 co-occurs_with(与…同时发生) <T137>
R5 Conceptually_related_to(概念上相关)<T158>
R5.1 evaluation_of(…的评价) <T161>

图8 语义关系组织示例

3.3 语义网的框架结构

语义网由2个主文档、2个语义关系继承文档和2个管理文档构成。语义类型和语义关系文档（SRDEF）包括语义类型/语义关系名称、唯一标识符、树形结构号、定义、实例、使用注释等信息，如“STY|T100|Age Group|A2.9.4|An individual or individuals classified according to their age.|Adult; Infant, Premature; Adolescents; Aged,80 and over|”。语义网结构文档（SRSTR）通过“语义类型1|关系|语义类型2|链接状态”形式表示语义网的结构，如“Acquired Abnormality|result_of|Behavior|D|”。语义关系完整继承集文档（SRSTRE1/SRSTRE2），是对SRSTR的扩展描述，SRSTRE1用标识符表示语义网结构，如“T020|T186|T190|.....”，SRSTRE2用名称表示，如“Acquired Abnormality|isa|Anatomical Abnormal-ity|.....”。数据管理文档SRFIL和SRFLD分别用于文档和元数据描述。

语义网通过SRDEF文档的语义类型唯一标识符、语义类型名称、语义类型树形号与超级叙词表的MRSTY（语义类型文档）相互关联，见图9。

4 专家辞典和词汇处理工具

专家辞典和词汇处理工具用于超级叙词表同义概念的自动归并。专家辞典是在NLM自然语言专家处理系统（NLP）项目基础上研

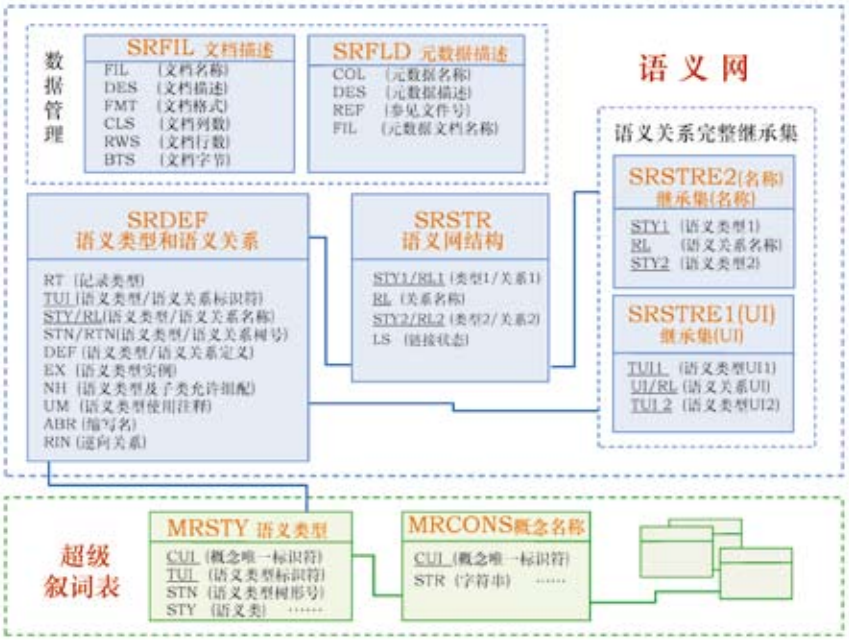

图9 语义网框架结构

发的，含有约45万条生物医学及常用英语词汇条目、80余万条词汇变化形式[17]，包括句法（syntax）、词法（morphology）、字法（orthography）方面的信息。句法指如何将若干词语组合起来，创建出某种含义；词法指词的变化规则、派生词、组合词；字法指词拼写方面的变化，如“oe/e”。

词汇常常通过一些变化以适应不同的语境，如动词时态的变化、形容词比较级的变化、名词单复数的变化，词汇处理工具有助于忽略这些变化，发现类似术语、发现词汇的变体形式、匹配术语到超级叙词表的概念，以及创建索引。词汇处理工具集是一组基于UMLS专家辞典和词汇处理规则的Java程序，常用工具包括以下几种。

LVG（Lexical Variant Generator，词形变化生成器）是词汇工具中最重要的一个，它像一个工具箱，里面有许多工具，包括62个流组件，36个命令选项[18]。每个流组建有不同的功能[19]，流组件可以任意选择或组合以满足自然语言词汇转换处理的特定需求。

图10 术语原形化处理

Norm（Normalized String Generator，原形字符串生成器）由 LVG的11个流组件组成，用于超级叙词表同义概念的归并。图10中所列的所有词形变化术语分别经过所有格、停用词、大小写、标点、词的非变化形式及排序等处理后生成统一的表示形式“disease hodgkin”，这种词汇的原形化处理方法有助于实现不同变体形式术语的词形规范，并可借助专家词典提供的异形同义词实现术语的词义规范。Norm还用于生成超级叙词表的原形化字符串索引（MRXNS）。

WordInd（Word Index Generator，词索引生成器）可将超级叙词表的字符串切分成词，转换为小写形式，生成词索引（MRXW），将原形化字符串索引（MRXNS）的字符串切分成词，生成原形化词索引（MRXNW）。

5 UMLS多词表整合策略

UMLS从整体规划、系统设计、构建过程，直至每个细节的处理，处处体现着模块化、可扩展、自动化、方便操作、简单适用的建设原则，尽可能减少人工构建的复杂程度。

5.1 基于概念的超级叙词表的建设理念

超级叙词表以概念为核心实现了不同来源词表同义词的汇聚，超级叙词表概念结构的设计为同

义概念的组织、归并搭建了基本架构，通过CUI-LUI-SUI-AUI将同一概念的不同名称和不同变体形式连在一起，起到词形规范和词义规范的作用，每个概念的优选名称由计算机自动推选确定。

基于概念构建超级叙词表与传统的以叙词为中心的构建理念有所不同，叙词本身是一个同义、近义、下位乃至反义概念的表达集合，而概念则表达一个明确的意思（meaning）。一个概念在涵义表达、关系表达方面较叙词更为精准和清晰，计算机也更容易理解和处理。以概念为核心组织术语，知识颗粒度更容易得到控制，人为干预程度较小，构建过程也相对简单。

5.2 概念的计算机辅助归并机制

NLM建立了专门用于自然语言处理的专家辞典和词汇处理工具，词汇工具基于专家辞典可以比较精确地识别词汇的各种变体和拼写变化形式，自动匹配不同词表各种词形变化术语到超级叙词表的同一个概念。超级叙词表还基于各来源词表术语间的同义关系和专家辞典提供的同义关系，较好地实现了词形不同但词义相同术语的归并。超级叙词表概念结构中LUI、CUI的设计及处理算法，也从不同层面反映了同一概念不同字符串从词形匹配到语义匹配的构建过程。基于计算机辅助的同义词自动推荐功能，减少了繁重的人工干预过程，提高了词表的建设效率及准确性。

5.3 充分继承丰富来源词表的语义关系

超级叙词表在构建过程中特别注重概念间语义关系的组织，在对来源词表同义概念进行归并的基础上，充分保留来源词表建立的各种关系，并进行很好的揭示、描述和组织。对于一些孤儿概念为其创建关系，保证超级叙词表中的每个概念都能与其他概念相互关联。超级叙词表还通过构建语义网使概念间语义关系的表达从线性模式向网状模式转变，以揭示概念间客观存在的更为丰富和复杂的语义关系。

5.4 充分保留整合多种知识源的信息

本体、叙词表、分类表本身就是人类知识的结晶，UMLS集成了众多知识源的信息，成为专业领域具有很高权威性的知识库，并具有网络百科的很多特征和功能。

超级叙词表通过精心设计，巧妙地实现了异构异质词表的同构化表示，这些知识源既可整合利用，也方便单独使用。NLM基于对各种来源词表信息的分析、梳理和组织，将这些知识组织工具的内容大致归为术语描述、关系描述、其他属性描述三类，通过对数据文档、元数据的合理布局和设计，以及对来源词表各种标识名称的管理，将纷繁庞杂的来源词表的信息充分保留并有序整合起来。据统计2011AA版超级叙词表中，来源词表的术语类型有260多种，关系表达形式有600多种，属性描述信息近600种，还含有大量的定义和注释信息。

5.5 搭建统一的超级叙词表范畴体系

UMLS语义网通过语义类型为超级叙词表搭建了统一的高层范畴类目，将各种来源叙词表、分类表、本体等知识组织工具在统一的范畴体系下进行组织，同时也保留了原有词表自身的范畴体系。语义类型之间的语义关系为类目间提供了更加丰富的关系，通过传导和继承极大丰富了来源词表概念之间的语义关系。UMLS语义网的建设理念、简便的构建方法和使用策略带给了我们新的启示。

6 结束语

UMLS通过精心的规划和设计，构建了超级叙词表、语义网、专家辞典和工具，实现了生物医学领域跨语言多词表的整合。从UMLS的设计理念我们可以看出，它结合了许多领域学科专家的共同努力，包括生物医学、图书馆学、语言学、信息学及计算机领域；集成了多个机构和组织的研究成果，通过良好的组织，使不同检索语言实现兼容转换，使原来分散在不同知识源的概念、术语、关系得以统一，既保留了每个来源词表的体系结构和属性信息，又形成具有统一结构、表达规范、词汇丰富、关系丰富、注释详尽的超级叙词表。UMLS为网络环境下知识组织系统的建设和应用提供了全新的理念和构建模式，值得我们借鉴与思考。

参考文献

[1] Unified Medical Language System® [EB/OL]. [2011-09-01]. http://www.nlm.nih.gov/research/umls/.
[2] Fact Sheet: Unified Medical Language System® [EB/OL]. [2010-08-22]. http://www.nlm.nih.gov/pubs/factsheets/umls.html.
[3] National Library of Medicine. Statistics - 2011AA Release [EB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html.
[4] NELSON S J, MD, POWELL T, HUMPHREYS B L. The Unified Medical Language System® (UMLS®) Project [EB/OL]. [2010-08-22]. http://www.nlm.nih.gov/mesh/umlsforelis.html.
[5] 2 Metathesaurus [M/OL].[2009-09-01]. http://www.ncbi.nlm.nih.gov/books/NBK9684/.
[6] KLEINSORGE R, WILLIS J. Unified Medical Language System Basics [EB/OL]. [2008-12-05]. http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf.
[7] National Library of Medicine. Abbreviations Used in Data Elements - 2011AA Release [EB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html.
[8] National Library of Medicine. 2011AA UMLS Files: MRCONSO.RRF [DB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.
[9] 2011AA UMLS Files: Metathesaurus MRREL.RRF [DB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.
[10] 2011AA UMLS Files: Metathesaurus MRHIER.RRF [DB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.
[11] Attribute Names-2011AA Release [EB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/attribute_names.html.
[12] 2011AA UMLS Files: Metathesaurus MRSAT.RRF [DB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.
[13] 3 Metathesaurus-Rich Release Format(RRF) [M/OL]. [2009-09-01]. http://www.ncbi.nlm.nih.gov/books/NBK9685/.
[14] SPERZEL W D, TUTTLE M S, OLSON N E.The Meta-1.2 Engine: A Refined Strategy for Linking Biomedical Vocabularies [M/OL]. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care, 1992: 304-308. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html.
[15] Current Semantic Types in the Semantic Network [EB/OL]. [2011-06-23]. http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.
[16] Current Relations in the Semantic Network [EB/OL]. [2011-06-23]. http://www.nlm.nih.gov/research/umls/META3_current_relations.html.
[17] SPECIALIST LEXICON Growth – Statistics [EB/OL]. [2011-09-08]. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/statistics/index.html.
[18] LU CJ. The SPECIALIST NLP Tools [EB/OL]. [2010-12-01]. http://lexsrv3.nlm.nih.gov/LexSysGroup/Docs/Presentations/NlmAssociate2010/nlpTools-nlmAssociateLecture2010.pdf.
[19] An example of the normalization process for the term "Hodgkin's diseases, NOS." [DB/OL]. [2011-05-05]. http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_005.htm.
[20] UMLS® Reference Manual [M/OL]. [2009-09-09]. http://www.ncbi.nlm.nih.gov/books/NBK9676/.

作者简介

李丹亚，研究员。研究方向：医学知识组织系统，医学信息资源管理与利用。E-mail: ldy@imicams.ac.cn

Integration Mechanism of Multi-vocabulary in UMLS

Li Danya, Hu Tiejun, Li Yazi, Li xiaoying, Sun Haixia, Li Junlian, Qian Qing / Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, 100020

Abstract: This paper analyzes the metathesaurus, the semantic network, and the SPECIALIST Lexicon of UMLS, and explains the infrastructure and the data structure of UMLS. It also illustrates the algorithm of merge synonyms from multi-vocabularies. Further it explains the methods and mechanism to integrate the heterogeneous concepts, relationship, and attributes that are from different vocabularies, as well as approach on constructing the category system. Finally, this paper discusses the mechanism that enables UMLS to integrate vocabularies from multi sources.

Keywords: UMLS, Multi-vocabularies integration, Metathesaurus concept structure, Semantic network, Lexical tools

（收稿日期：2012-01-18）