一体化医学语言系统及其在知识发现中的应用研究*

李晓瑛 李军莲 李丹亚
（中国医学科学院医学信息研究所，北京 100020）

摘要：人工智能时代下，知识发现已成为构建知识图谱、开发智能系统、提供知识服务的重要基础和保障。本文在深入分析一体化医学语言系统内容与结构的基础上，研究其应用于知识发现的基本原理与主要作用，以期为我国一体化语言系统在知识发现中的应用实践提供一些参考与借鉴。

关键词：知识发现；知识组织系统；一体化医学语言系统；自然语言处理
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2019.09.004

知识组织系统（Knowledge Organization System，KOS）是对人类知识及结构的规范组织和语义阐述[1]，包括一般意义上的同义词环或术语表，传统文献数据库系统所使用的叙词表，检索引擎中采用的分类表与自动扩展词表，网站导航辅助资源浏览的分类结构，用于信息系统互操作的一体化语言系统，网络环境下以概念为基本单元的本体，以及语义网络和知识图谱等[2]。其中，一体化语言系统又称一体化情报检索语言，能够通过统一整合若干部知识组织系统，实现相同知识内容的一致组织与有机关联，从而更好地发挥优势互补效应，标志着知识组织系统从单一类型走向多种类型的融合协同发展[3]。

早在1984年，美国国立医学图书馆（U.S. National Library of Medicine，NLM）便通过整合上百部生物医学领域的知识组织系统而形成一种大型知识组织工具，命名为一体化医学语言系统（Unified Medical Language System，UMLS）[4]。UMLS广泛用于不同健康医疗信息系统、生命科学数据、生物医学文献资源，以及相关领域计算机系统之间的语义互操作和知识内容关联。此外，NLM基于UMLS开发了语义知识发现工具（Semantic Knowledge Representation，SemRep）[5]，并在输入文本数据格式上与PubMed文献服务系统建立了无缝衔接，易于从MEDLINE大规模生物医学文献资源中发现隐含的知识，有助于构建知识图谱、开发智能系统和提供知识服务。在我国，代表性的一体化语言系统包括医学分类主题一体化系统[6]、中文一体化医学语言系统（Chinese Unified Medical Language System，CUMLS）[7]、中医药一体化语言系统[8]，以及国家“十二五”科技支撑计划课题“面向外文科技文献的超级科技词表和本体建设”资助构建的具有我国自主知识产权、覆盖理工农医4个领域的科技知识组织系统（Scientific & Technical Knowledge Organization System，STKOS）。然而，上述一体化语言系统目前大多应用于生物医学文献及外文科技文献的组织、主题标引与检索（如中国生物医学文献服务系统SinoMed、国家科技图书文献中心网络服务系统等），尚未大规模开展语义知识发现等相关研究与应用实践。本文对UMLS内容结构及其在SemRep知识发现中发挥的作用进行深入分析，以期为CUMLS在生物医学领域、STKOS在理工农医等科技领域的知识发现相关研究与应用实践提供一些有意义的借鉴与参考。

*本研究得到NSTL先期研发任务“STKOS超级科技词表内容建设机制和发展研究（医学部分）”（编号：XQYF0101-4）资助。

1 UMLS内容结构分析研究

就宏观而言，UMLS由超级叙词表（Meta thesaurus）、语义网络（Semantic Network）、专家词典和词汇工具（SPECIALIST Lexicon and Lexical Tools）3个知识源组成（见图1）。其中，超级叙词表是对生物医学领域内主要概念、术语及其语义关系的统一组织与整合，2019AA版共收录了214部知识组织系统、约380万个概念、超过1 440万条来源术语记录；语义网络涵盖了127种语义类型、54种语义关系，用于对超级叙词表的所有概念进行统一的范畴分类与层级组织；专家词典和词汇工具是一套支持超级叙词表创建和更新的、面向自然语言处理的大型词典和Java软件工具集。

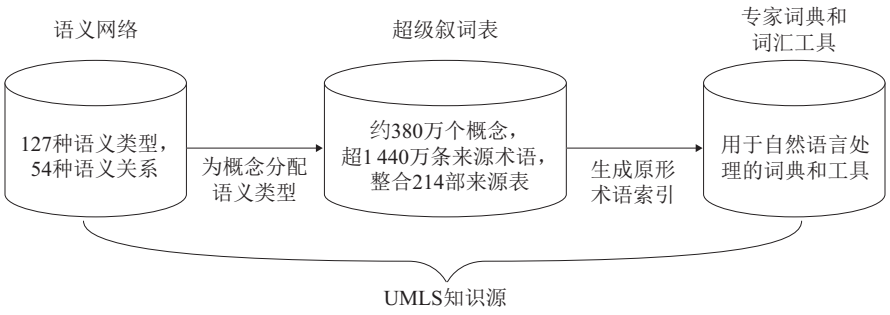

图1 UMLS组成结构

1.1 超级叙词表

超级叙词表是一部大规模的知识组织系统，收录了生物医学领域具有影响力的术语表、叙词表、分类法、本体、疾病编码集等来源词表中的重要术语及相关知识。类似于其他知识组织系统，概念及术语、语义关系、描述属性是超级叙词表的3个核心组成元素。

（1）概念及术语。超级叙词表以概念为中心整合与组织同义术语，所有来源词表中具有相同含义的术语组成了概念。概念具有4种标识符，即来源术语标识符（AUI）、概念名称字符串标识符（SUI）、原形化术语标识符（LUI）和概念标识符（CUI）。基于AUI-SUI-LUI-CUI的组织模式使超级叙词表能够将不同来源表中表达相同概念的多条异形同义术语连同多种词形变体整合到一个单元，有助于不同医疗信息系统之间的无障碍交互操作与集成融合。

（2）语义关系。超级叙词表不仅整合了来源词表中的同义关系，而且秉承统一保存且能还原追溯的原则，保留和继承了来源表中的其他语义关系。经过仔细分析，NLM将这些语义关系组织为4种类型，即常用关系、等级结构关系、共现关系和映射关系。其中，常用关系细分为广义（RB）、狭义（RN）、直接上位（PAR）、直接下位（CHD）、同位（SIB）、相关（RO）、同义（SY）、相似或相同（RL）、相关及可能同义（RQ）、限定（AQ）与被限定（QB）共11种；等级结构关系继承自来源表中建立概念等级结构的上下位关系，支持全面展示来源表的整个等级结构；共现关系来自MEDLINE、AI/RHEUM、CCPSS 3个外部信息源，经求取两条术语在相同数据源中同时出现的频次而获得；映射关系大多存在于不同来源表的编码与标识符之间，或继承于来源表，或在超级叙词表创建中产生。上述多种类型的语义关系，成为UMLS超级叙词表提供丰富生物医学知识的重要数据基础。

（3）描述属性。超级叙词表的描述属性是有关概念及术语的注释说明类信息，有些具有通用性（如定义），有些则仅适用于特定来源表（如创建日期）。其中，继承于来源表的概念定义为特性属性，记录在定义属性数据文档；语义网络为概念所分配的语义类型为共性属性，存储于概念语义类型数据文档；其他多种名称的特征描述信息，均被统一管理于属性数据文档。

1.2 语义网络

UMLS语义网络包括语义类型、语义关系两部分。语义类型是一套统领超级叙词表概念名称的范畴类目，共有127种；语义关系为一个设定于语义类型间的关系类型集合，共54种。在UMLS语义网络中，语义类型为节点，语义关系构成节点之间的连边，通过链接形成网状的语义关联（可视为小型的生物医学知识图谱）。

（1）语义类型。UMLS语义类型的最核心作用是

对超级叙词表的概念提供统一的顶层范畴类目。127种语义类型通过直接上下位的等级关系，按树形结构进行组织，顶层始于“实体（entity）”“事件（event）”两大支，依次逐级展开，层级最深达7级。

（2）语义关系。UMLS语义网络中的知识关联涉及54种语义关系，包括“等级关系（is_a）”和“相关关系（associated_with）”。等级关系即为语义类型树形结构和语义关系等级体系中的直接上下位关联。相关关系主要分为5种类型，即“物理相关（physically_related_to）”“空间相关（spatially_related_to）”“功能相关（functionally_related_to）”“时间相关（temporally_related_to）”及“概念相关（conceptually_related_to）”，并按树形结构逐级展开。

1.3 专家词典和词汇工具

UMLS专家词典是一部支持自然语言处理系统的通用英文词典，收录了道兰氏图解医学辞典、美国传统词频书、朗文当代高级词典等多种来源的常见英语单词及生物医学常用词语；词量规模大（约47万条），涉及86万条词形变体，含句法（syntax）、词法（morphology）、字法（orthography）及自然语言处理所需信息。词汇工具为一组开发于专家词典和英文词汇语法规则之上的Java程序集，主要针对自然语言常用词的高度变异性，核心工具包括原形化工具（Norm）[9]、词索引生成器（WordInd）及词变体生成器（Lexical variant generator，Lvg）。在超级叙词表构建与维护过程中，专家词典和词汇工具起到词形还原、词形归并、索引生成等作用。

2 UMLS在知识发现中的应用研究

NLM将UMLS应用于生物医学文本的知识发现，并结合自然语言处理技术开发出知识发现工具SemRep[10]。下文首先对SemRep发现知识的基本原理进行概述，而后对UMLS在SemRep处理过程中所发挥的作用进行归纳总结。

2.1 SemRep知识发现基本原理

SemRep本质上是一个自然语言处理系统，旨在应用通用句法分析和结构化领域知识库UMLS，发现生物医学文本中的语义关系和知识。如图2所示，输入文本后，SemRep将展开一系列处理及操作，最终返回所发现的知识。例如，从“Body composition analysis suggested that anamorelin is an active anabolic agent in patients with NSCLC”句子中，SemRep将挖掘出以语义关系三元组“anamorelin|TREATS|Non-Small Cell Lung Carcinoma”所表示的医学知识。

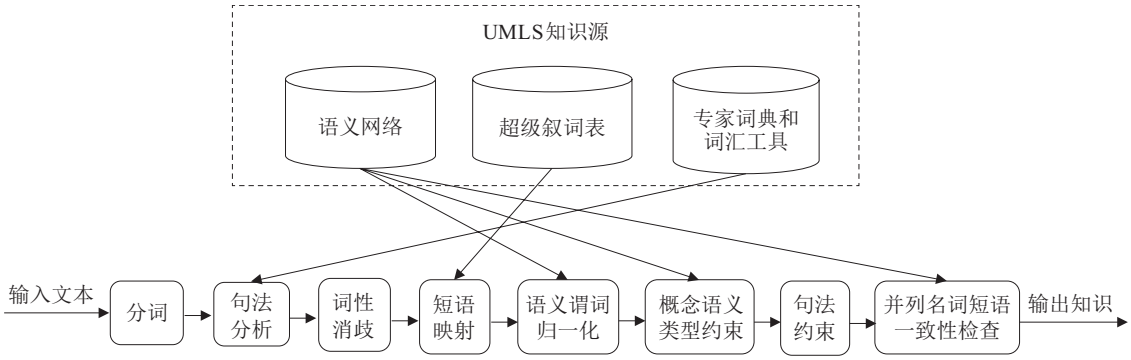

图2 基于UMLS知识源的知识发现工具SemRep处理流程

基于UMLS知识源的知识发现工具SemRep，基本原理与流程如图2所示。

（1）对输入文本进行分词。简言之，利用标点符号将段落切分成句子，再利用空格等分隔符将句子切分为短语。

（2）句法分析。基于专家词典中的语法信息进行句法分析，包括识别句子的语法结构，确定句子中短语之间的依存关系。

（3）词性消歧。基于前后短语的词性和句法分析结果，对具有多个词性的短语给出最确切的词性。

（4）短语映射。利用映射工具MetaMap，在句子所切分出的名词短语和UMLS超级叙词表中的概念名

称建立映射，映射规则涉及简单匹配、复合匹配及部分匹配3种类型，权重计算综合利用了向心度、变异度、覆盖度和内聚度4个参数[11]；此外，结合句法分析结果，初步提取出语义关系三元组“概念a|语义谓词|概念b”；其中，概念a与概念b均为UMLS超级叙词表中的规范概念（概念优选名称）。

（5）语义谓词归一化。基于UMLS语义网络所定义的语义关系类型，分析生物医学领域常见的语义动词，并采用动词原形进行归一化，最终确立了58种语义谓词标签（Label）。除了“is_a”“compared_with”，其他语义标签可根据其感情色彩划分为4类，即激发、抑制、肯定及否定，详见表1。

表1 SemRep语义谓词标签及其分类

类 别	SemRep语义谓词标签
激发	AUGMENTS，CAUSES，COMPLICATES，PREDISPOSES，PRODUCES，STIMULATES，NEG_DISRUPTS，NEG_INHIBITS，NEG_PREVENTS
抑制	NEG_AUGMENTS，NEG_CAUSES，NEG_COMPLICATES，NEG_PREDISPOSES，NEG_PRODUCES，NEG_STIMULATES，DISRUPTS，INHIBITS，PREVENTS
肯定	ADMINISTERED TO，AFFECTS，ASSOCIATED WITH，COEXISTS WITH，CONVERTS TO，DIAGNOSES，INTERACTS WITH，LOCATION OF，MANIFESTATION OF，METHOD OF，OCCURS IN，PART OF，PRECEDES，PROCESS OF，TREATS，USES，higher_than，lower_than，same_as
否定	NEG_ADMINISTERED TO，NEG_AFFECTS，NEG_ASSOCIATED WITH，NEG_COEXISTS WITH，NEG_CONVERTS TO，NEG_DIAGNOSES，NEG_INTERACTS WITH，NEG_LOCATION OF，NEG_MANIFESTATION OF，NEG_METHOD OF，NEG_OCCURS IN，NEG_PART OF，NEG_PRECEDES，NEG_PROCESS OF，NEG_TREATS，NEG_USES，NEG_higher_than，NEG_lower_than，NEG_same_as

（6）概念语义类型约束。以语义关系三元组“概念a|语义谓词|概念b”表示的医学知识，对于既定的语义谓词而言，概念a与概念b存在一定的范围约束。例如，“治疗（TREATS）”存在于药物与疾病之间，并不会出现于药物与解剖结构中。因此，借助UMLS语义网络的语义类型，对每种语义谓词标签所关联的概念进行约束与筛选，从而排除不合理的语义关系三元组，提高所发现的医学知识的准确性。经选取语义类型为“化学物质与药物（Chemicals and Drugs）”的数据进行评测，SemRep知识发现的准确率可达到83%[10]，且其优化算法持续更进中。

（7）句法约束。从依存语法等自然语言处理的角度，对所发现的语义关系三元组进一步约束和检查。

（8）并列名词短语一致性检查。UMLS超级叙词表中的所有概念都被赋予一个或多个专指的语义类型（来自UMLS语义网络）。通过检查文本句中连续出现的名词短语（已利用MetaMap映射到UMLS概念）是否具有相同的语义类型，SemRep将精准地提取出并列名词短语，进而发现一组具有相同语义谓词的医学知识。以“a new class of anti-inflammatory drugs that have clinical efficacy in the management of asthma，allergic rhinitis，and inflammatory bowel disease”为例，鉴于名词“asthma”“allergic rhinitis”及“inflammatory bowel disease”为并列短语，且其映射到UMLS超级叙词表的3个概念具有相同的语义类型“Disease or Syndrome”，SemRep便可快速发现一组语义知识“Anti-Inflammatory Agents|TREATS|Asthma”“Anti-Inflammatory Agents|TREATS|Allergic rhinitis，NOS”及“Anti-Inflammatory Agents|TREATS|Inflammatory Bowel Diseases”。

（9）输出以三元组“概念a|语义谓词|概念b”表示的语义知识。

2.2 UMLS在SemRep知识发现中的作用

UMLS三大知识源（超级叙词表、语义网络、专家词典和词汇工具）在SemRep知识发现中至少发挥了以下4点重要作用。

（1）提供语法信息，支持句法分析。UMLS专家词典和词汇工具为SemRep分析文本的句法结构，提供了主、谓、宾、定、状、补等语法信息。

（2）提供大规模的同义术语，支持概念识别与提取。基于映射工具MetaMap，文本中的名词短语将与UMLS超级叙词表中的术语进行语义相似度匹配，并

借助术语之间的同义关系，以概念为中心进行汇聚，最终达到概念的识别与提取。

（3）支持语义谓词的归一化与标准化。UMLS语义网络所定义的语义关系，为SemRep知识发现中语义谓词的归一化与标准化（如将“treatment”“treated by”“treat”统一命名为“TREATS”），提供了依据与参考。

（4）提供概念范畴分类，支持对所发现知识的语义合理性进行检查，提升知识发现的准确性。UMLS语义网络中的语义类型为概念提供了范畴分类，便于对SemRep所发现的知识“概念a|语义谓词|概念b”，从语义合理性角度展开检查（如语义谓词“TREATS”不会出现在语义类型为“药物”与“解剖结构”的概念a与概念b之间），在一定程度上提高了SemRep知识发现的准确率和效率。

2.3 局限与不足

基于UMLS知识源的SemRep知识发现工具，为自动深度揭示大规模文献数据中隐含的生物医学知识提供了基础与平台。然而，通过对其知识发现的语义三元组结果进行多角度分析后，不难察觉出UMLS及SemRep的一些局限与不足。

（1）UMLS概念覆盖范围相对有限。尽管UMLS超级叙词表的收录范围和规模体量已经非常庞大，但在处理自然语言文本时仍显不足。例如，SemRep在挖掘“This result indicates that PGG-PTX was substantially less toxic in vivo than PGA-PTX”的语义知识时，将输出三元组“Paclitaxel|lower_than|Paclitaxel”；原因在于，超级叙词表并未收录“PGA-PTX”，经MetaMap映射后错误识别为“Paclitaxel”。

（2）SemRep对输出结果的逻辑性检查有待加强。例如，从“As patients with nonsmall cell lung cancer（NSCLC）and wildtype epidermal growth factor receptor（EGFR）are resistant to treatment with erlotinib or gefitinib，potential chemosensitizers are required to potentiate wildtype EGFR NSCLC cells to erlotinib/gefitinib treatment”句子中，鉴于自然语言句式结构复杂，SemRep发现2条互相矛盾的语义关系三元组“gefitinib|TREATS|Non-Small Cell Lung Carcinoma”和“gefitinib|NEG_TREATS|Non-Small Cell Lung Carcinoma”。显然，在优化过程中，通过对从同一条文本句中发现的多条语义知识三元组进行逻辑性检查，可有效地避免上述这类明显的错误。

3 结语

一体化语言系统是对特定领域多来源概念、术语及其语义关系的统一集成与有机融合，为领域知识发现提供了以概念为中心的同义术语汇聚、概念及实体识别、基于范畴类目及等级体系的概念分类组织等信息，有助于构建知识图谱、开发智能系统及提供知识服务。本文对一体化医学语言系统UMLS的内容结构进行了深入分析，研究了SemRep知识发现工具的基本原理与处理流程，并归纳总结了UMLS知识源在SemRep知识发现中所发挥的重要作用以及局限和不足，为我国开展基于CUMLS、STKOS等一体化语言系统的知识发现相关研究与应用实践提供了一些理论依据与实战经验。然而，就目前中文自然语言处理和知识发现工作而言，在及时更新补充一体化语言系统的主题覆盖范围、充分发挥知识组织系统的作用之余，还需深度分析中文自然语言的结构和句法特征，加快高效的句法分析和人工智能技术的应用研究，促进文本隐含知识的揭示、知识间潜在关联的发现更上一层楼。

参考文献

[1] 李丹亚，李军莲，李晓瑛，等. 医学知识组织体系发展现状及研究重点[J]. 数字图书馆论坛，2012（12）：13-21.

[2] 曾蕾. 用于标引浏览检索的语义工具[EB/OL]. [2019-09-01]. http://www.libnet.sh.cn/upload/htmleditor/File/071213121547.pdf.

[3] 李丹亚，李军莲. 医学知识组织系统：术语与编码[M]. 北京：科学出版社，2019.

[4] Unified Medical Language System[EB/OL]. [2019-09-01]. http://www.nlm.nih.gov/research/umls/.

[5] Semantic Knowledge Representation[EB/OL]. [2019-09-01]. https://semrep.nlm.nih.gov/.

[6] 李军莲，李丹亚，阮学平，等. 医学分类主题一体化系统建设[J]. 医学信息学杂志，2011，32（3）：69-73.

[7] 李丹亚，胡铁军，李军莲，等. 中文一体化医学语言系统的构建与应用[J]. 情报杂志，2011，30（2）：147-151.

[8] 尹爱宁，张汝恩. 建立《中医药一体化语言系统》[J]. 中国中医药信息杂志，2003，10（3）：90-91.

[9] 李晓瑛，李丹亚，胡铁军. 基于UMLS专家词典与工具的词形归并算法研究 [J]. 情报科学，2013，31（4）：134-138.

[10] RINDFLESCH T C，FISZMAN M. The interaction of domain knowledge and linguistic structurein natural language processing：interpreting hypernymicpropositions in biomedical text [J]. Journal of Biomedical Informatics，2003，36（6）：462-477.

[11] 李晓瑛，夏光辉，孙海霞. MTI自动文献标引系统研究 [J]. 医学信息学杂志，2015，36（3）：52-57.

作者简介

李晓瑛，女，1982年生，博士，副研究员，研究方向：知识组织，E-mail：lixiaoying@imicams.ac.cn。
李军莲，女，1972年生，硕士，研究馆员，研究方向：资源建设。
李丹亚，女，1954年生，学士，研究员，研究方向：知识组织。

Research on the Unified Medical Language System and its Application to Knowledge Discovery

LI XiaoYing LI JunLian LI DanYa
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: In the age of artificial intelligence, knowledge discovery has become a fundamental basis to construct knowledge graph, develop intelligent system and provide knowledge services. In this paper, the content and structure of Unified Medical Language System is analyzed and investigated in depth, as well as its application to knowledge discovery, which will be useful and meaningful for applying the well-developed unified language systems to discover semantic knowledge in China.

Keywords: Knowledge Discovery; Knowledge Organization System; Unified Medical Language System; Natural Language System

（收稿日期：2019-08-06）