Steffen Staab, Rudi Studer. Handbook on Ontologies. Springer, 2004
Part II Ontology Engineering
7 Building a Very Large Ontology from Medical Thesauri

7 从医学叙词表建立一个超大型本体

Udo Hahn and Stefan Schulz

曾新红译
2005.8

文摘 我们报告一个大规模知识转换和加工案例研究。从一个全面的、语义较浅的术语仓库 UMLS 中得来的医学知识，被转换为形式上严格的表达描述逻辑格式。这样，UMLS 的广泛覆盖度就和用于一致性和循环检测的推理机制结合在一起。它们不仅是将直接从 UMLS 引入的知识进行正确净化的关键，也是随后对大容量丰富而复杂的术语知识结构进行升级和精化的关键。形成的生物医学本体目前包含超过 24 万个概念实体，并且构建了迄今为止最大型的形式化知识库之一。

7.1 介绍

像疾病和医学过程编码这样的任务，或在书目数据库中搜寻医学文献，通常都需要参考共享的领域知识。典型地，这种共有的知识通过术语表（受控词表）、叙词表或分类号来提供。它们的作用在于，将单个概念的词汇或短语变体的使用统一起来（通过指向一个标引词），将术语经由语义关系链接起来（例如，X 是 Y 的上位词或下位词，X 是 Y 的同义词，X 是 Y 的一部分或 X 含有部分 Y），或产生一个数字或符号分类等级系统（按其所代表的术语的专指度递增排序）。

和许多其它学科不同，医学有一个长久以来遵守的传统，在各种各样的医学术语表、叙词表和分类系统中汇集和结构化其知识，疾病分类法（taxonomies）、医疗过程、解剖术语等等。这些工作典型地受限于规定上位词、下位词、相关词或同义（准同义）词。这在 UMLS（Unified Medical Language System, 一个伞状系统，覆盖了 60 多部医学叙词表和分类法，例如 MESH, ICD, SNOMED, DIGITAL ANATOMIST）[18]中最为明显。

从概念的角度来看，UMLS 可以分成两个主要部分。一方面，UMLS *Semantic Network (SN)*形成了上部本体，它包含由 54 种语义关系所链接的 134 种语义类型，从而产生了 7473 个制高点。另一方面，UMLS *Metathesaurus* 含有 875255 个概念（2003 年版），每个概念都被赋予了一个或多个 UMLS SN 类型。这些概念通过语义关系链接，也通过 UMLS SN 来提供。在这些 Metathesaurus 概念之间总共存在 10552299 个语义链接，它们大部分直接取自原始资料（sources），有一些则由 UMLS 开发者添加。这些链接的绝大部分反映了叙词表风格的上位/下位术语关系。

UMLS SN 和 Metathesaurus 构成了一个巨大的语义网络，其语义浅而且完全直观，这是因为它们最初是打算供人使用的，以支持与健康相关的知识管理。已知的大规模以及 UMLS 的演化多样性和固有的异构性，因缺乏一个形式化的语义基础而导致不一致和循环定义等就一点不奇怪了[6，5]。当人处于循环（loop）中而且其使用仅限于疾病或过程编码、会计（accountancy）或文档检索任务时，这也许不会引起非常严重的问题。如果期望将其应用于更知识密集型的应用，例如医疗决策[23]或医疗叙述的理解[11]，这些缺点就可能导致僵局。

因而，处理医学知识的形式化模型已被提出，如概念图、语义网络或描述逻辑[7，17，

22，34，10]。毫不意外的是，要获得更多的表达性和形式上的严格，必须付出代价，即不断增加的建模工作和随之增加的维护成本[20]。完全使用这种严格方法的可用系统，尤其是那些采用高端知识表示语言的系统，通常仅限于相当小的子领域内。这些资源中我们已知最复杂的是 GRAIL 编码的 GALEN 知识库，它覆盖了 9800 个概念[22]。有限的覆盖面妨碍了它们的常规使用（这在医学信息界始终是一个高回报的方面）。

几乎所有在形式表示语言基础上开发的知识库都是从零开始设计的——没有系统地使用包含在广泛传播的医学术语学中的大型知识体。因此，将由非形式化的术语学提供的大规模覆盖面与达到最新技术发展水平的知识表示系统支持的高等级表达力和演绎推理能力结合起来，开发更大规模的形式上可靠的医学本体，将会是一种很有吸引力的方法。这种观点已由 Pisanelli 等[19]所倡导，他们从 UMLS SN 和部分 Metathesaurus 中提取知识，并将其与基于逻辑的不同来源的顶层本体结合。另一个例子是 SNOMED[8]从一个多轴编码系统到一个形式上建立了良好基础的本体的重新设计[33，32]。但是这些工作完全集中在按照分类学（taxonomy）的基于归纳的推理上，缺乏合理的分元系统的覆盖面（医学知识库的另一个关键部分）。

7.2 根据部分-整体等级进行推理

医学知识主要是围绕归纳等级（按 is-a 关系进行分类推理（taxonomic reasoning）就是基于其上的）和部分-整体等级（允许按 part-of 和 has-part 关系进行分元推理）来组织的。和概念分类法（taxonomy）中基于归纳的推理不同，按部分-整体等级进行的推理到目前为止还没有完全的结论性的机制存在（不同方法的调查参见 Artale 等[1]）。

在描述逻辑范式之内（调查参见[2]），已经提出了若干种对表示语言的扩展，它们为部分-整体推理提供了特殊的构造器[22，14]。从医学角度来看，一个主要的挑战伴随着跨越部分-整体等级的属性的传播而出现，通常称为“跨越可传递角色的遗传”（例如 *inflammation-of* o *part-of* → *inflammation-of*）[4，21]。但是这种推理模式不能被归纳，因为它面临着大量例外。在一种类似的风格中，部分-整体关系的传递属性通常都假定有效（is assumed to hold）。但是，对于医学，和在具有常识的领域中一样[9，35]，这种观点已经是无效的了。那么，正规的传递使用的表示，以及对非传递性的 *part-of* 关系的例外处理，在一个同构的形式化框架内都必须得到同等的考虑。

由以前的方法所推动[26，28]，我们形式化了一个分元推理[12]，它考虑了上述因素。我们的解决方法没有超出易懂、简洁的概念语言 *ALL*[3]的表达能力，因为我们希望这个模型尽量简单。我们的建议集中在一个特定的概念编码数据结构上，即所谓的 *SEP triplets*。它们定义了 is-a 等级结构的字符模式，支持典型的传递性 *part-of* 关系的推理的模拟。用这种格式，*anatomical-part-of* 关系描述一个生物体物理部分之间的部分关系。

一个 SEP 三元组（见图 7.1）包括，首先是一个复合*结构（structure）*概念，称为 *S-node*（例如，*Hand-structure,* H_S）。每个*结构*概念一方面直接包含一个解剖学的*实体（Entity）*概念，另一方面还包含是那个实体概念的 *Part* 的所有事物的一个共有包含器（subsumer）。这两个概念分别称为 *E-node* 和 *P-node*，例如 *Hand-Entity*（H_E）和 *Hand-Part*（H_P）。当 E-node 代表在我们的领域中被准确地模型化的解剖学概念时，S-node 和 P-node 就构成了形式化重构系统模式和分元推理下的例外所需的人造表示物。

图 7.1 SEP 三元组：分类法中的部分关系（Partitive Relation）

更确切地说，作为一种存在的情况，*P-node* 是那些让它们的 *anatomical-part-of* 角色由相应的 *E-node* 概念填充的概念的共有包含器。例如，*Hand-Part* 包含那些其所有实例都有一

个 *Hand-Entity* 作为必要整体的概念。作为一个额外的限制，*E-node* 和 *P-node* 可以建模为多重不相交。对于大多数表示单个对象的概念来说这是一个合理的假设，其部分和整体不可能是同一类型的（一个红血球细胞不可能还是另一个红血球细胞的一部分）。相反的是，堆（masses）和集合（collections）则可能有同一类型的部分和整体，例如一个组织可能是另一个组织的一部分。[30]

为了通过分类推理形式化重构 *anatomical-part-of* 关系，我们假设 C_E 和 D_E 代表 *E-node*，C_S 和 D_S 分别代表包含 C_E 和 D_E 的 *S-node*，C_P 和 D_P 分别代表经由 *anatomical-part-of* 角色与 C_E 和 D_E 相关的 *P-node*（见图 7.1）。这些约定可以通过下列术语学表达式来捕获：

$$C_E \sqsubseteq C_S \sqsubseteq D_P \sqsubseteq D_S \qquad (7.1)$$

$$D_E \sqsubseteq D_S \qquad (7.2)$$

P-node 定义如下（这里我们介绍 D_E 和 D_P 之间的不相交限制，即 D 没有一个实例可以是 *anatomical-part-of* 另一个 D 的实例）：

$$D_P = D_S \cap \neg D_E \cap \exists\, anatomical\text{-}part\text{-}of.D_E \qquad (7.3)$$

因为 C_E 由 D_P 包含（根据（1）），我们推断 *anatomical-part-of* 关系也保留在 C_E 和 D_E 之间：

$$C_E \sqsubseteq \exists\, anatomical\text{-}part\text{-}of.D_E \qquad (7.4)$$

图 7.2 在一个 SEP 编码的分元系统中允许/中止角色传播

经由 SEP 三元组进行的概念等级编码允许知识工程师开启和关闭部分-整体关系的角色传播属性，分别取决于是 E-node 还是 S-node 被处理为一个概念关系的目标。在第一种情况下，跨越整体-部分等级的角色传播被禁止，在第二种情况下则被允许。作为一个例子（见图 7.2），*Enteritis* 被定义为 *has-location* $Intestine_E$，即关系 *has-location* 的范围（range）被限制在 *Intestine* 的 E-node。这排除了 *Appendicitis* 作为 *Enteritis* 的分类，虽然 *Appendix* 也经由一个 *anatomical-part-of* 关系与 *Intestine* 相关。但是，在“开启”状态，*Glomerulonephritis*（*has-location* $Glomerulum_S$）被归类为 *Nephritis*（*has-location* $Kidney_S$），同时 *Glomerulum* 是一个 *anatomical-part-of* Kidney。同样，*Perforation-of-Appendix* 被归类为 *Intestinal-Perforation*（见[12]有一个深入的分析）。

7.3 本体工程工作流程：重新设计 UMLS

我们的目标是从 UMLS 的两个主要子领域（解剖学（anatomy）和病理学（pathology））中提取概念知识，以利用描述逻辑构建一个形式上可靠的本体。该知识转换工作流程包含四个不同的步骤（也参见图 7.3 中的图表）：

图 7.3 从 UMLS 构建一个 LOOM 知识库的工作流程

1. 描述逻辑格式的术语学公理从关系表结构（从 UMLS 源导入）自动生成。当解剖学和病理学部分的所有领域概念都考虑进来时，只有一个从 UMLS 仔细挑选的关系类型子集被并入。其中有诸如 *part-of/has-part*、*is-a* 或 *has-location* 这样的关系，因为我们认为它们分别是分元和分类等级以及空间知识的可靠的指示符。出于这一层次处理的更进一步考虑，我们排除了过度通用的关系，例如 sibling-of 或 associated-with，因为它们可能将噪声引入新的知识库的关系结构。
2. 然后这个“未经加工”的本体立刻由描述逻辑分类器（classifier）（详见[16]）自动检测，

看它是否包含定义循环或有不一致的地方。

3. 如果遇到不一致或循环的知识结构，则由一个生物医学领域专家手工解决不一致或循环。然后，再运行一次分类器以检测修改后的本体是否是一致和非循环的。在这一层次的知识库直接反映了 UMLS 的（依然浅的）表达力，但已经嵌入了一个有效的形式化框架。
4. 对于许多应用来说，UMLS 规范的表达力和粒度（专指性等级）是不充分的。因此，该本体需要额外的手工润饰。这里我们把那些在前面几轮没有考虑的关系（如 *sibling-of* 或 *associated-with*）合并成一个用于本体重建模的启发式支持，同时还有一些完全新的非常专指的关系被创建（例如 *inflammation-of*，*perforation-of* 或 *lineardivision-of*）。后者为更深的从医学叙述中自动提取知识所需，而这正是我们的主要应用[11]。

步骤 1：描述逻辑表达式的自动生成。概念和关系的来源是 1999 年版的 UMLS SN 和 1999 年版的 Metathesaurus 中的 *mrrel*、*mrcon* 和 *mrsty* 表。*mrrel* 表提供两个 UMLS CUIs（概念唯一标识符）之间的语义链接，参见表 7.4。这些表是 ASCII 文件，被导入 Microsoft Access 关系数据库，并使用嵌入于 VBA 程序语言的 SQL 进行操纵。对于 mrrel 子集中的每个 CUI，其字母数字代码被英语标引词（preferred term）代替。

对 UMLS SN 的顶层概念进行手工重建模之后（以不同的深度，按照目标领域），我们从总共 85899 个概念中提取了 38059 个解剖学概念，从 Metathesaurus 中提取了 50087 个病理学概念。包含进这些集合之一的标准是对预先定义的语义类型的指派。并且有 2247 个概念被发现同时包含在两个集合中（解剖学和病理学）。因为我们想让这两个子领域严格不相交，我们复制了这些交迭的概念，并按照它们分别的子领域给所有的概念加上 *ana-*或 *pat-*前缀。通过观察这确实是合理的，这些混合概念确实表现出多重含义。例如，*tumor* 一方面有一种恶性疾病的含义，另一方面也有一种解剖结构的含义。

作为解剖领域的目标结构我们选择 SEP 三元组，每个解剖学概念一个。这些都用术语学语言 LOOM[15]来表示，之前我们已用一个专用的 *deftriplet* 宏将此语言进行了扩展（实例见表 7.1）。建立分类和分元等级时只有 UMLS 提供的 *part-of*、*has-part* 和 *is-a* 关系属性被考虑（见图 7.3）。结果是一个混合的 *is-a* 和 *part-whole* 等级结构。

对于病理学领域，我们将 UMLS 中的 *CHD*（child）和 *RN*（narrower relation）视为标志分类（*is-a*）链接。没有考虑部分-整体关系，因为这个范畴没有应用于病理学领域。对于所有包含在病理学概念的定义声明中的解剖学概念，指派给解剖学概念的 S-node 是它们通过 *has-location* 关系链接到的那个默认概念，从而允许跨越部分-整体等级进行角色传播（见 7.2 节）。

表 7.1 生成的 LOOM 格式三元组

作为一个基础假设，在此过程中产生的所有角色都被认为是存在量化的（existentially quantified）。这意味着在两个概念 *A* 和 *B* 之间保持的任何关系 *r*（*part-of, has-location* 等）都被映射到一个角色 $\exists r.B$，它在概念 *A* 的定义中是一个必要的条件。一个概念定义的所有概念限制都被映射到一个限制合取。

在两个子领域中，像非常常见的 *sibling* 关系（*SIB*）这样的浅关系都被保留（作为代码行中的注释），用来为随后的手工精化阶段提供启发式引导（参见第 4 步骤）。

步骤 2：由描述分类器进行自动一致性检查。UMLS 概念的导入产生了解剖概念的 38059 个 *deftriplet* 表达式和病理学概念的 50087 个 *deftriplet* 表达式。每个 *deftriplet* 都扩展为三个

defconcept（S-, E-和 P-node）和两个 *defrelation* 表达式（*anatomical-part-of-x, inv-anatomical-part-of-x*），总共达到 114177 个概念。这产生了总共 240764 个定义（definitory）LOOM 表达式（和来自 UMLS SN 的概念一起）。

从 38059 个解剖学三元组中，1219 个 *deftriplet* 声明包含一个*:has-part* 子句，其后跟随着三元组变量号码的列表（a list of a variable number of triplets），在 823 种情况下带有多于一个的参数（平均基数 3.3）。4043 个 *deftriplet* 声明包含一个*:part-of* 子句，其后跟随着多于一个的参数（平均基础 1.1）（只在 332 种情况下）。然后结果知识库被提交给描述分类器检查术语循环和一致性。在解剖学子领域，发现了一个术语循环和 2328 个不一致概念。在病理学子领域虽然没有发现一个不一致概念，但却判定了 355 个术语循环（见表 7.2）。

表 7.2 从知识转换过程得出的经验数据

步骤 3：一致性的手工恢复。被分类器识别出的本体解剖学部分中的不一致，全部都可以通过 *is-a* 和 *part-of* 链接追溯到两个三元组的同时链接，一个编码引起了冲突（由于我们默认要求相应的 P-和 E-node 不能相交）（参见表达式（3））。在大多数这种情况下，受影响的父节点属于一类明显不能正确建模为 SEP 三元组的概念，例如 *Subdivision-Of-Ascending-Aorta* 或 *Organ-Part*。这些概念中每一个概念的含义几乎都解释了一个 P-node 的含义，这样，SEP 内在的不相交条件冲突可以通过用简单的 LOOM 概念替换所涉及的三元组而得到解决（通过将它们与已存在的 P-node 相比较，或中断 *is-a* 或 *part-of* 链接）。

在该本体的病理学部分，我们料想会出现大量的术语循环，这完全是按照分类归类（taxonomic subsumption）(*is-a*)翻译语义较弱的 *narrower term* 和 *child* 关系的结果。考虑到该本体的规模，我们认为 355 个循环是一个可以容忍的数字。那些循环主要是由非常类似的概念（例如 *Arteriosclerosis* 与 *Atherosclerosis, Amaurosis* 和 *Blindness*）以及未被说明的范畴（“other”，“NOS”=*not otherwise specified*(不另加规定)）引起。这些都直接继承自源术语表，而且众所周知从它们的定义上下文进行翻译非常困难，例如 *Other-Malignant-Neoplasm-of-Skin* 与 *Malignant-Neoplasm-of-Skin-NOS*。在许多情况下，决定哪些关系可以保留和哪些关系必须删除是不定的，因为在生物医学术语表中常常不能获得有关术语确切含义的共识。从进一步的分析中，我们得到了一个否定列表，含有 630 个概念对。在随后的提取循环中，我们将这个表并入 LOOM 概念定义的自动结构（construction）中，并且，带着这些新的限制，生成一个完全一致的知识库。

步骤 4：人工校正和精化。基于前面提到的工作步骤建立这个大容量的本体需要一个人三个月的工作。第 4 步——如果对整个本体操作的话——可以预计是非常费时的而且要求广而深的医学专门知识。来自两个子领域的随机样本，100 个解剖学概念和 100 个病理学概念，由第 2 作者（一个领域专家）进行分析。这花了一个人大约 1 个月的时间。从我们所获得的经验来看，可以得出以下工作流程：

• *检查分类等级和分元等级的正确性*。手工添加或删除分类和分元链接。在任何可能的时候用非初始的归类（subsumption）替代初始的归类。这是一个关键点，因为自动生成的等级只包括父概念和必要条件。作为一个例子，自动生成的 *Dermatitis* 的定义包括这样的信息：它是一种 *Inflammation*，并且角色 *has-location* 必须由概念 *Skin* 填充。然而，一个 *has-location skin* 的 *Inflammation*，不能自动分类为 *Dermatitis*。

结果：在*解剖学*样本中，100 个概念中只有 76 个可以毫不含糊地分类为属于“规范的”解剖学。（其余的，例如 *ana-Phalanx-of-Supernumerary-Digit-of-Hand*，涉及病理解剖学，被立

即从分析中排除。）除了对 UMLS 语义类型的赋值，只有 27 个（直接）分类链接被发现。83 个 UMLS 关系（大部分是 *child* 或 *narrower* 关系）被手工升级为分类链接。12 个（直接）*part-of* 关系和 19 个 *has-part* 关系被发现。4 个 *part-of* 关系和一个 *has-part* 关系不得不删除，因为我们认为它们难以让人相信。51 个 UMLS 关系（大部分是 *child* 或 *narrower* 关系）被手工升级为 *part-of* 关系，并有 94 个 UMLS 关系（大部分是 *parent* 或 *broader* 关系）被升级为 *has-part* 关系。在将这些较浅的 UMLS 关系整理和升级为语义上更专指的关系之后，又再次检查了样本的完整性。结果有 14 个 *is-a* 和 37 个 *part-of* 关系仍被认为是缺失了。

在病理学样本中，对病理学子领域的指派 100 个概念中有 99 个被认为是可信的。在 12 个概念定义中有总共 15 个伪 *is-a* 关系被识别出来。有 24 个 *is-a* 关系被发现缺失。

• *检查:has-part 参数（假定 "real anatomy"）*。在 UMLS 源中，*part-of* 和 *has-part* 关系被认为是对称的。按照我们的转换规则，角色 *has-anatomical-part* 对一个 E-node B_E 的附属，其范围（range）限制于 A_E，这就意味着有一个概念 *A* 为概念 *B* 的定义而存在。另一方面，A_E 的归类通过 P-node B_P 来分类，B_P 经由角色 *anatomical-part-of*（被限制到 B_E）来定义，这意味着假如 A_E 存在，则 B_E 也存在。

这些限制并不总是符合“真正的”解剖学，因为解剖学的结构可能表现出病理学的变体。图 7.5（左）勾画出概念 A 必须 anatomical-part-of 概念 B，但其存在并不是 B 的定义所要求的。这代表了外科手术干预的结果，例如，没有阑尾的大肠，没有牙齿的口腔。

图 7.5 使用 SEP 三元组进行部分-整体推理的模式

结果：样本中通过自动导入和手工整理而得到的所有 112 个 has-part 关系都检查过了。分析表明，为了不回避病理上变异的解剖学对象的耦合归类，其中有超过半数的关系（62 个）应该删除。例如，将 *has-anatomical-part.Thumb* 作为一个存在的限制保留在 *Hand* 的定义中，将会导致不允许将所有那些因先天或后天畸形而没有拇指的实体归类为 *Hand*。作为一个例子，*Ileum* 的大部分实例不包含一个 *Meckel's Diverticulum*，但是 *Meckel's Deverticulum* 的所有实例都必须 *anatomical-part-of Ileum*。许多手术干预切除了解剖结构（阑尾，胆囊等），产生了类似的模式。在我们的形式化体系中，这对应于一个 S-node 和一个 P-node 之间的单个分类链接（参见图 7.5 左边部分）。相反的情况也是可能的（参见图 7.5 右边部分）：A_E 的定义并不意味着角色 anatomical-part-of 由 B_E 填充，但 B_E 也不意味着相反的角色由 A_E 填充。例如，一个 *Lymph-node* 必须包含 *Lymph-follicles*，但是存在不是 *Lymph-node* 一部分的 *Lymph-follicles*。这种模式代表了宏观（可数）对象（例如器官）与多重统一的微观现象之间的部分元素包含（mereological）关系。（参见[30]）

• *分析 sibling（同属）关系和将概念定义为不相交*。在 UMLS 中，SIB 关系链接在分类或分元等级中分享同一个父母的概念。同属概念对可能有也可能没有共同的后代。如果没有，它们就构成了两个不相交子树的根。在一个分类等级中，这意味着一个概念暗示着另一个概念的非（例如，良性肿瘤不可能是恶性肿瘤，反之亦然）。在一个部分等级中，这可能解释为空间不相交，即一个概念不可能在空间上与另一个概念交迭。例如，*Esophagus* 和 *Duodenum* 是空间上不相交的，但 *Stomach* 和 *Duodenum* 不是（它们共享一个共同的过渡结构，称为 *Pylorus*），就像所有的相邻结构都有一个共同的表面或区域。空间不相交可以建模，这样概念 *A* 的 S-node 的定义就隐含着概念 *B* 的 S-node 的非（参见[31]有一个更详细的讨论）。
*结果：*我们发现，在解剖学领域每个概念平均有 6.8 个同属，在病理学领域则平均有 8.8 个。到目前为止，还只对解剖学领域实行了对同属关系的分析。从总共 521 个同属关系中，有 9

个标识为 is-a，14 个标识为 part-of，17 个标识为 has-part，但有 404 个指向了拓扑分离概念。

• *解剖学-病理学关系的完善和修正*。令人惊奇的是，只有很少的病理学概念含有一个显式的对相应解剖学概念的引用。因此，这些关系必须由领域专家来添加。在每种情况下都必须做一个决定，即 E-node 或 S-node 是否必须处理为修正的目标概念，以决定跨越部分-整体等级的角色传播是禁止还是允许。

结果：在样本中我们发现了 522 个解剖学-病理学关系，领域专家从中判断出有 358 个（即 69%！）是不正确的。在 36 种情况下缺失了恰当的解剖学-病理学关系。所有 164 个 *has-location* 角色都按照它们是否被一个解剖学三元组的 S-node 或 E-node 所填充而进行了分析。在 153 种情况下，S-node（它允许跨越部分-等级关系的传播）被认为是恰当的，在 11 种情况下 E-node 被推荐。对 100 个病理学概念的分析表明，只有 17 个被链接到解剖学概念。有 15 种情况，对 S-node 的默认链接被认为是正确的，有 1 种情况对 E-node 的链接被推荐，在另一种情况这种链接被判定为错。

不可信限制的高数量指向了 UMLS 源中的 *has-location* 链接的轻量级语义。当我们按照一个用于此重要日常事务（routine）的连接词对它们进行翻译时，一种转折的含义似乎悄悄地流行于许多顶层概念（例如 *Tuberculosis*）的定义中。在这个例子中我们发现，所有的解剖学概念都可能通过 *has-location* 链接到这种疾病而受其影响。所有这些限制（例如 *has-location Urinary-Tract*）被遗传给子概念（例如 *Tuberculosis-of-Bronchus*）。对顶层病理学概念进行彻底分析是必要的，并且限制的连接词在必要的时候必须用转折词来替代。

7.4 讨论和结论

在医学界，领域知识必须更大规模地提供。为了不再从零开始开发复杂的医学本体，我们在这里提出了一种“保守的”方法——重用现有的大规模资源，但从这些资源中精化数据，以满足更具表达力的知识表示语言所施加的先进的建模要求。最后产生的本体可以用于那些要求形式上可靠推理的复杂应用（如文本理解）。

将概念知识从语义较弱的规范转换到严格的知识表示形式体系，其好处和问题都已由 Pisanelli 等[19]描述。他们从 UMLS 语义网络以及部分 Metathesaurus 提取知识，并将其转换为描述逻辑系统。Spackman 和 Campbell[33]介绍了 SNOMED 术语表如何从一个多轴代码系统演化为一个具有形式化基础的本体。他们的大致目标是要避免复合概念的不确切或无效的表示。然而，这两种方法都没有为分元关系提供特定的推理机制。

在 GALEN 的形式化框架中，Read Thesaurus 的一个片断被翻译成了 GRAIL，一种也是基于描述逻辑的知识表示系统[24]。它在一个交叉验证（cross-validation）研究中进行了检验，一方面检查 Read Thesaurus 中包含的定义逻辑上是否一致，另一方面，检查 GRAIL 的领域模型为它们编码是否足够丰富。虽然 GRAIL 有一个专门用于分元系统的专用推理机制，但是其改编局限于简单的类属等级，因为只有这些构成了 Read Thesaurus。

VOXEL-MAN[27]（一个解剖学多媒体教学辅导系统）的开发者，以及 DIGITAL ANATOMIST FOUNDATIONAL MODEL(UWDA FM，一个解剖学语义网络)[25]的开发者，都强调了部分等级（partitive hierarchy），虽然是在一种非形式化的层次上。在 VOXEL-MAN 中，勾画出了一个细密的分元关系本体，它考虑了在解剖学领域发现的各种部分-整体关系。UWDA FM 的开发者将他们自己限制在一个较小的关系集合中，导致了分元和分类等级的精确分割。他们的胜过他人之处是高粒度的描述和广泛的覆盖面。

我们的方法试图将广泛的覆盖面、UWDAFM 的细密概念描述与形式严格的描述逻辑结合起来。我们用医学领域不可缺少的部分-整体专用推理能力升级了被导入的知识，尽管这已被描述为术语（即描述逻辑）语言的难题[13]。

一个稳定语言平台的保守的结构扩展是否能够延续至分元推理的许多不同的变体和不同的部分-整体关系，或是否需要新设计运算符或其他基础语言扩展，都还要视情况发展而定。至少在医学领域，对一个 *part-of* 子关系，即 *anatomical-part-of* 的限制是充分的，一个相对简单的“数据结构”扩展（像 SEP 三元组）已经产生了令人满意的结果，没有必要诉诸于深奥的语言扩展。我们有证据证明我们这里提出的三元组机制可以直截了当地扩展来覆盖 mereotopological 和（受限的）空间推理，如同[31，29]。

我们的研究表明，恢复产生自 UMLS 的本体的一致性相对简单明了，但因为所需的巨量手工工作，几乎不可能同时达到高度的充分性和完整性。但是，恢复充分性不应该首先被认为是消除 UMLS 源中含有的明显“错误”，而应该看作是在含义有微妙差异（由于知识源的异构）的医学术语的候选概念化（conceptualization）之间进行选择。另一个方面是需要校正因严格的公理假设驱动自动输出过程而变得不正确的概念定义（例如定义属性的连读），它不是在所有情况下都正确，因此，很有必要进行单独的手工规范。

一个实际的工作流程可能存在于明显不恰当声明的手工删除中，随后是那些来自焦点子领域的概念定义的填充。在这些重复的手工精化循环中我们发现了以下暗示：使用术语分类器，即计算包含关系的推理引擎，具有最为重要和最为杰出的启发式价值。因此，知识精化循环是真正半自动化的，加进站在一旁的人类知识工程师的医学专门知识，但也由推理系统（它理清正确（不正确）概念定义的因果关系）来驱动。

从我们方法的具体细节抽象出来，这种知识转换和知识加工的一些更概括的方法学问题出现了：

- *知识集成*：当若干个知识源必须合并时，一些知识部分可能会交迭，有些则可能相距太远需要定义合适的概念桥接。甚至那些彼此互补性很好的知识源也需要合适的接口，以便可以从一个转换到另一个。
- *粒度*：不同的知识源，有时甚至是同一个知识源，常常会有一些展示出非常细密的子领域描述，而有一些则是以低得多的专指度来处理的。在那些知识表示的不同粒度层次之间进行调停就成为恰当使用知识的一个重要要求。另外，也可能有必要故意为一个子领域的描述提供不同的抽象层次。
- *视角（views）*：在特定领域知识上没有单一的、规范化的视角，例如，肿瘤，同时是一个解剖学结构也是一个病理学现象。这种歧义在其概念表示上有直接的暗示，推论可从中诱导出来。甚至不同的思想流派在组织同一个子领域时其方法也不同。因此，必须提供在同一主题支持不同概念视角的形式化技术方法，而不是以一种独断的方式强求一致。
- *顶层本体*：通过一个统一的本体伞去集成一个大学科领域（如生物学，医学）的子学科的尝试，不可避免地导致了构建基础概念系统的抽象、“上层”部分的需求。在这一层次，像“organism”（如 animal, plant, virus）、“process”（光合作用，消化作用）、“substance”（叶绿素，血）和“structure”（动物或植物解剖学，细胞形态学等）这样的高层概念已被正确组织和概念化表示了，这样它们可以以一种有效的方式链接更专指、更具体的领域描述。