基于邻近概念信息的 FMA 本体概念名消歧法*

王浩茂 梁 铮 周小茜 罗凌云

（南华大学 衡阳 421001）

〔摘要〕 提出一种基于本体中概念间的父子关系以及词法信息的自动消歧法，以消除解剖学基础模型本体中含有介词时概念名中的歧义，具体介绍研究对象、技术基础和研究方法。结果显示该方法能成功解析 90.95% 含有单个介词 of 的概念名，消除其中原有的歧义。

〔关键词〕 本体；消歧；结构关系；词法信息

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2018.03.013

Disambiguation Method of FMA Ontology Concept Names Based on Close Concept Information *WANG Hao-mao, LIANG Zheng, ZHOU Xiao-xi, LUO Ling-yun, School of Computer Science and Technology, University of South China, Hengyang 421001, China*

〔**Abstract**〕 The paper puts forward a method that disambiguates automatically based on paternity and lexical information of ontology intermediate concept to disambiguate concept names containing prepositions in ontology of foundational model of anatomy, introduces study object, technical basis and study method in details. The result shows that the method is able to analyze 90.95% of concept names containing a single proposition and clears up the original ambiguity contained.

〔**Keywords**〕 Ontology; Disambiguation; Structure relation; Phrase information

1 引言

解剖学基础模型（Foundational Model of Anatomy, FMA）是生物医学信息领域一个常用的大型领域本体，由华盛顿大学结构信息学小组开发并维护[1-2]。由于 FMA 非常庞大且内部语义结构复杂，难免存在纰漏，因此仍处在不断完善中。当前，在医学本体的质量评估领域有不少针对 FMA 开展的研究工作[3-4]，其中不乏基于词法分析的方法，如根据反义词和语义关系的传递性寻找不合理的语义关系[5]，以及利用对称形容词和互模拟结构诊断 FMA 等[6]。

然而在 FMA 中存在一些单词数量众多（最多有 18 个单词）且结构复杂的概念名，由于英文本身的结构歧义[7]，难以解析[8]，给基于词法分析的本体研究带来困难。本研究试图消除由介词引起的定语修饰范围不确定而造成的结构歧义，如对于 FMA 概念名 Left surface of heart，由于 Left 的修饰

〔收稿日期〕 2017-09-30

〔作者简介〕 王浩茂，本科生；通讯作者：罗凌云，副教授，博士后。

〔作者简介〕 国家自然科学基金（项目编号：61502221）；湖南省教育厅优秀青年项目（项目编号：14B153）。

范围不确定，所以可以产生两种不同的解析树，见图 1。

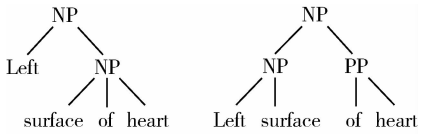

图 1 Left surface of heart 的两种解析树

在图 1 左侧的解析树中，Left 用来修饰 surface of heart；而在右侧的解析树中，Left 仅修饰 surface。通常对于这种既有修饰词（第 1 个词就是修饰词）又有介词 of 的概念名，修饰词的修饰范围不仅可以作用于 of 之前，也可延伸至整个字符串的末尾，因而造成结构歧义。当使用 Stanford Parser[9] 分析 FMA 概念名时，发现其结果更偏向于图 1 中右侧的情形。在实验中分析另一个名词 Mesothelial cell of parietal peritoneum 时，得出的结果也类似于图 1 中右侧的解析树。对于 Mesothelial cell of parietal peritoneum 来说，其解析结果是正确的，而对于第 1 个样例 Left surface of heart 来说，这种分法得出的解析树却是错误的，事实上 Left 应该用来修饰 surface of heart，而不是 surface。对于此类概念名，为寻找其正确的分词方式，本研究借鉴词义消岐中常见的基于上下文信息的方法[10]，提出一种基于邻近概念信息以及词法信息的自动消歧法，用来划分定语（修饰词）作用域，以消除歧义。本研究采用的邻近概念主要指父节点概念。

2 研究对象与技术基础

本研究使用的 FMA 本体为 3.1 版，来源于开放生物医学本体（Open Biomedical Ontology，OBO）实验室[11]提供的网络本体语言（Web Ontology Language，OWL）[12]文件。在 FMA 中，存在两种主要的语义关系，即上下位关系和部分－整体关系，前者记为 IS－A 关系（如 Right Hand IS－A Hand），后者记为 Part－of 关系（如 Hand Part－of Free upper limb）。根据语义关系，概念名被划分成具备有向图形式的概念层次模型。通常将 IS－A 关系构成的有向图叫做 Taxonomy，Part－of 关系构成的有向图叫做 Partonomy。本研究的分析方法主要用到语义网技术（或称语义 Web）。在语义网技术中，用 RDF[13] 作为通用数据格式模型，它是一种有向图模型。而 Web 本体语言是一种对指定特定域有约束功能的形式语言。本研究采用 SPARQL[14] 作为 RDF 的查询语言。

3 研究方法

3.1 概述

本研究选用 FMA 中那些只存在 1 个介词 of，且 of 之前至少出现两个单词的概念名作为研究对象。在整个 FMA 本体中，一共出现 78 977 个概念名。其中含有介词 of 的概念名有 61 877 个。在这些含有 of 的概念名中，只出现一个介词 of 的有 35 174个，其中有 14 153 个概念名的 of 之前只存在 1 个单词，并且这个单词是名词。剩下的 21 021 个概念名（相当于整个 FMA 的 27%）即为本研究的语料集。语料集中的概念名仅含有 1 个 of，且 of 之前存在多个单词，能够确定的是第 1 个单词是作为定语存在的。对于该语料集中的每 1 个概念名，为确定其第 1 个单词的修饰范围在 of 之前还是 of 之后，即为确定其应该采用图 1 中的何种语义解析树，从它的父概念中寻找“证据”：对于概念名 S，观察它的某个子短语是否在 S 的某个父节点中出现。如果出现，该子短语就应该作为 1 个不可划分的整体，在 S 的解析树中作为 1 棵独立的子树存在。本研究中主要使用前文提到的两类父子关系：IS－A 父子关系和 Part－of 父子关系。具体地，在概念名 S = （$a_0 a_1 a_2 \cdots a_n$ of s_1）中，由前文提到的结构信息可知，a_0 作为定语共有两种修饰域，一种作用于 of 之前，另一种延伸至 of 之后。为消除因 a_0 修饰范围不确定而引起的歧义，将 S 分解成不同的子短语，构建 3 类不同的“证据集”，在此基础之上，分别设计 3 种测试方法，根据一定的规则，检测每 1 类证据集中的元素是否在 S 的父概念中出现。其中，方法 1 用来检测 a_0 的修饰范围是否延伸至 of 之后；方法 2 用来检测 a_0 是否仅修饰介词 of 之前的名词；若前两种方法均未成功，方法 3 利用祖先

关系继续进行检测。

3.2 提取语料集

通过 SPARQL 语句从 FMA 本体库中查询得到含有单个 of 的概念名集合，去除 of 之前只含有 1 个单词的概念名，筛选得到语料集，以下称语料集（I）。对于语料集（I）中的每 1 个概念名，在 FMA 本体库中查询其所有的 IS－A 和 Part－of 父概念名，分别构造 IS－A 父集合与 Part－of 父集合。

3.2 方法 1

对于语料集（I）中的概念名 S = （$a_0a_1a_2\cdots a_n$ of s_1），为验证 S 能否使用图 1 中左侧的方法解析，从图 1 左侧解析树的根出发，将 S 分为 a_0 和［$a_1a_2\cdots a_n$ of s_1］两个部分。由于 a_0 作为定语存在，所以 S 的证据集集中在后半部分，即其范围为［$a_1a_2\cdots a_n$ of s_1］到［a_n of s_1］。如若概念名 S 为 Posterior papillary muscle of left ventricle，它能分解出两个子短语：papillary muscle of left ventricle 和 muscle of left ventricle，形成 S 的第 1 类证据集。如果该证据集中的某个子短语在 S 的父节点中出现，那么该子短语应该成为 S 的语义解析树的 1 棵独立子树，表明 a_0 的修饰范围延伸至 S 的末尾。以此方法对语料集（I）中的每个概念名进行测试，相关实现见伪代码描述：

```
for 概念名 S in 语料集（I）：
    设置 mark_ isa 和 mark_ partof 为未标记
    分解得到概念名 S 的证据集
    过程 ①：证据集中元素范围为［a1a2…an of s1］~［an of s1］
    for 证据 E in 证据集：
        分别判断证据 E 是否在概念名 S 的 IS－A 和 Part－of 父集合中出现
        根据是否在父节点中出现设置 mark_ isa 和 mark_ partof 标记
        if 标记 mark_ isa 与 mark_ partof 均被标记：
            结束循环
    if 标记 mark_ isa 或 mark_ partof 为 True ：
        依据两个标记分类（分类标准见 3.5）
    else：
        过程 ②：不能分类的概念名形成语料集（II），等待方法 2 测试
```

3.4 方法 2

对于语料集（II）中的概念名 S = （$a_0a_1a_2\cdots a_n$ of s_1），采用方法 2 继续处理，依据图 1 右侧解析树的分解方法提取证据集。此时其证据集包括两个元素：［$a_0a_1\cdots a_n$］和［of s_1］。如果 S 的父节点包含［$a_0a_1\cdots a_n$］，但不包含［$a_0a_1\cdots a_n$ of］（表明［$a_0a_1\cdots a_n$］独立于 of 存在），或者 S 的父节点包含［of s_1］但不包含［a_n of s_1］（表明［of s_1］是独立于［$a_0a_1\cdots a_n$］存在），那么子短语［$a_0a_1\cdots a_n$］应该成为 S 的解析树的子树。如概念名 S：Basivertebral foramen of thoracic vertebra 是 Basivertebral foramen 的 IS－A 子节点，同时也是 Body of thoracic vertebra 的 Part－of 子节点，那么 Basivertebral foramen 将在 S 的解析树中作为 1 棵独立子树存在。方法 2 的具体算法与方法 1 相似，将遍历语料集（II）中的概念名及过程①所使用的证据集修改为方法 2 的证据集，且判断其中的元素是否严格出现在概念名 S 的父节点中。最后将过程②修改为构造语料集（III），等待方法 3 测试。需要注意的是，上文所提到的严格出现指：倘若［$a_0a_1\cdots a_n$］出现在概念名中，那么［$a_0a_1\cdots a_n$ of］则不应该出现；相应地，若［of s_1］出现在概念名中，［a_n of s_1］则不应该出现，以此确保 a_0 的修饰范围在介词 of 之前而没有延伸到 of 之后。

3.5 方法 3

对于语料集（III）中的概念名 S = （$a_0a_1a_2\cdots a_n$ of s_1），由于前两种方法均失败，设计 1 个新的证据集，它只包含 1 个元素，即子短语［$a_0a_1\cdots a_n$ of］。方法 3 要求 S 的父概念名包含这个证据，其原因是：虽然［$a_0a_1\cdots a_n$ of］不支持图 1 中任何一种解析树形式，暂时无法确定 S 的正确分词方式，但若根据方法 3 的规则不断向上寻找祖先节点，［$a_0a_1\cdots a_n$ of］的存在能确保它们仍然在本研究的语料集（I）之中，一旦某祖先节点不再在语料集（III）之中，意味着该祖先节点能够被前两种方法解析。例如

Spinous process of tenth thoracic vertebra 拥有一连串的 IS－A 关系：Spinous process of tenth thoracic vertebra→Spinous process of thoracic vertebra ——→Spinous process of vertebra ——→Process of vertebra。在方法 1 的帮助下，可以在节点 Spinous process of vertebra 处确定 spinous 修饰的是 process of vertebra 而不是 process，由此推导，在 Spinous process of tenth thoracic vertebra 中，第 1 个单词 Spinous 修饰的是 process of tenth thoracic vertebra。方法 3 的算法与前两个算法类似，只需要修改相应的证据集，此处不再赘述。

3.6 概念名分类

3.6.1 结果分类 对于每 1 种测试方法，设被测的概念名为 S，其证据集为 T，根据 T 中的元素在 S 的 IS－A 父节点和 Part－of 父节点中出现的方式，可将 S 分为 4 类。（1）类别 A。存在 1 个证据 $s \in T$，且 s 出现在 S 的 IS－A 父节点中，同时 1 在一个 $s' \in T$，且 s′出现在 S 的 Part－of 父节点中。（2）类别 B。存在 1 个 $s \in T$，且 s 出现在 S 的 IS－A 父节点中，但是 T 中的所有元素均不出现在 S 的任何 Part－of 父节点中。（3）类别 C。T 中的所有元素均不出现在 S 的 IS－A 父节点中，但是至少存在一个 $s \in T$，且 s 出现在 S 的 Part－of 父节点中。（4）类别 D。S 的所有 IS－A 父节点和 Part－of 父节点均不包含 T 中的任意元素。4 种结果分类，见图 2。

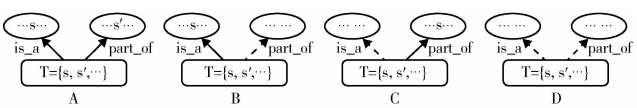

图 2 4 种结果分类

3.6.2 测试方法与分类 对于每个概念名先后使用 3 种方法对其进行测试，见图 3。首先用方法 1 对语料集（I）中的概念名进行分类。如果方法 1 测试得到的分类是 D，这些概念名形成语料集（II），将会通过方法 2 继续检测，方法 2 检测得到的结果将被分为 DA、DB、DC、DD 4 类。被归类为 DD 的概念名，即对语料集（III），将会通过方法 3 继续检测，得到 DDA、DDB、DDC、DDD 4 个分类结果。经过上述 3 种方法分类后，所有语料集（I）中的 FMA 概念名将会被分为如下几个类别：A、B、C、D、DA、DB、DC、DD、DDA、DDB、DDC 和 DDD。只有被分类为 DDD（占总测试数据的 9%）的概念名没有找到消除分词歧义的证据。如概念名 S：Spinous process of tenth thoracic vertebra 有 1 个 IS－A 父节点 Spinous process of thoracic vertebra 以及两个 Part－of 父节点 Tenth thoracic vertebra 和 Tenth thoracic vertebra arch。在方法 1 中，其证据集中只有 1 个元素，即 process of tenth thoracic vertebra，因为它没有在 S 的任意 1 个父节点中出现，所以 S 被划分为 D 类，通过方法 2 进行检测。方法 2 的证据集包含 Spinous process 和 of tenth thoracic vertebra，虽然 S 的 IS－A 父节点包含 Spinous process，但它同时也包含 Spinous process of，所以仍然不符合方法 2 的要求，会被划分为 DD 级。在方法 3 测试时，证据集中只有 Spinous process of 1 个元素，该子短语在 S 的 IS－A 父节点中出现，但并没有在 S 的 Part－of 父节点中出现，所以 S 最终将会被划分至 DDB 类。

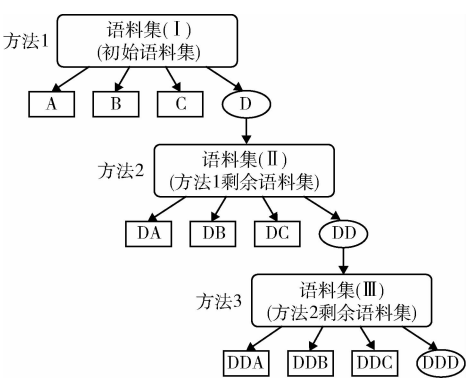

图 3 3 种测试方法与分类

4 测试结果及分析

根据图 3 所示的 10 个类别对语料集（I）中的所有概念名进行分类，其结果，见表 1（其中的百分比均基于总量 21 021 计算）。语料集（I）中有一半以上的概念名属于 DDB 类，说明那些被划分为 DD 的概念名（以下简称 S），虽无法在直接父节点

上找到证据来选择正确的分词方式，但若顺着 S 的父节点向上查找，仍有机会在其祖先节点发现有效的证据。最终除 1 903 个被归类为 DDD（9.05%）的概念名之外，其他名词都能通过本研究提供的方法找到有效的证据以帮助消歧。在分类过程中，不同的类别对确定分词方式所提供的帮助程度有所不同：类别 A 要求证据同时出现在 IS－A 父节点和 Part－of 父节点中，虽然很有说服力，但是成立条件过于苛刻，那些没有 Part－of 父节点关系的概念名只能被归类到 B 或 D。此外，在类 B 和类 C 中，虽然证据只需出现在一类父节点中，然而由于 IS－A 关系往往走向更抽象的概念，所以利用 Part－of 关系来对概念名消歧也许更为可靠，也就是类 C 可能会比类 B 提供更可靠的信息。对于12 709个被划分为 DD 的概念名，方法 1 和方法 2 失败的原因可能是算法的局限性，在这两种方法中只使用了直接相邻的父节点而不是所有祖先节点。

表 1 各个分类的数量和百分比分布

类别	类别数量	百分比（%）
A	321	1.52
B	3 025	14.39
C	1 044	4.97
DA	396	1.88
DB	2 217	10.55
DC	1 309	6.22
DDA	245	1.17
DDB	10 547	50.17
DDC	14	0.07
DDD	1 903	9.05

5 局限性与不足

（1）只考虑包含 1 个 of 的概念名，在今后的工作中，将把研究工作扩展到含有多个介词 of 的概念名中。（2）可能存在一些概念名，经方法 1 和方法 2 测试均能成功，即存在两类不同的证据，分别支持图 1 中的两种解析方式。在这种情况下，需要获取进一步的信息才能帮助消歧。（3）一些概念名可能需要从祖先节点中获取证据，而不仅仅是父节点。实验结果中超过一半的概念名属于 DDB 类可以说明这一点，见表 1。如果把搜索范围从父节点扩展到祖先节点，甚至是整个 FMA 概念名集合，就能够对更多概念名进行消歧，但相应的计算复杂度也会提高。在相关文献[15]中，作者还使用了子串之间的关系来分析基因本体中的术语名称。与本研究不同的是其研究重点在于术语之间的关系，而不是每个术语的内部结构。

6 结语

本研究提出一种基于邻近概念（主要指父概念）信息以及词法信息的自动消歧法，用来消除 FMA 本体中含单个介词 of 的概念名的结构歧义。此研究是对未知领域的一种初步尝试，不仅可以帮助提高词法分析工具的准确性，也为学者研究大型领域本体中的复杂概念名做了必要的铺垫。

参考文献

1 University of Washington. Foundational Model of Anatomy [EB/OL]. [2017－09－15]. http://www.si.washington.edu/projects/fma.

2 Rosse C, Mejino Jr JLV. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy [J]. J Biomed Inform, 2003, 36 (6): 478－500.

3 Zhu X, Fan J W, Baorto D M, et al. A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies [J]. J Biomed Inform, 2009, 42 (3): 413－425.

4 Gu HH, Wei D, Mejino Jr JL, et al. Relationship Auditing of the FMA Ontology [J]. J Biomed Inform, 2009, 42 (3): 550－557.

5 Zhang GQ, Luo L, Ogbuji C, et al. An Analysis of Multi－type Relational Interactions in FMA Using Graph Motifs [J]. AMIA Annu Symp Proc, 2012, (2012): 1060－1069.

6 Luo L, Mejino JL Jr, Zhang GQ. An Analysis of FMA Using Structural Self－bisimilarity [J]. J Biomed Inform, 2013, 46 (3): 497－505.

7 许睿. 英语结构歧义成因的认知分析 [J]. 英语广场·学术研究, 2014, 40 (4): 46－47.

（下转第 80 页）

信息行为的影响不显著，经逐步 Logistic 回归，未被纳入回归模型，与现有一些研究结果有异。在以后的研究中可扩大样本人群范围，进一步验证该两种因素对微信健康信息行为是否存在影响。

参考文献

1 Maher C A, Lewis L K, Ferrar K, et al. Are Health Behavior Change Interventions That Use Online Social Networks Effective? A Systematic Review [J]. Journal of Medical Internet Research, 2014, 16 (2): e40.

2 2017 微信数据报告 [EB/OL]. [2017 - 12 - 11]. http://www.sohu.com/a/203522712_118792.

3 郭冬阳. 从健康类公众号看社交媒体中健康信息的传播 [J]. 东南传播, 2016, (5): 105 - 106.

4 李东晓. 微屏时代谁在传播健康？——对微信平台健康养生信息兴起的传播学分析 [J]. 现代传播, 2016, 38 (4): 21 - 26.

5 马天娇. 大数据时代微信虚假健康信息传播现状及治理 [J]. 新闻世界, 2017, (1): 56 - 58.

6 邓小昭. 网络用户信息行为研究 [M]. 北京: 科学出版社, 2010: 8 - 19, 23 - 26.

7 Rieh S Y. Judgment of Information Quality and Cognitive Authority in the Web [J]. Journal of the American Society for Information Science and Technology, 2002, 53 (2): 145 - 161.

8 Marton C, Choo C W. A Question of Quality: the effect of source quality on information seeking by women in IT professions [J]. Proceedings of the American Society for Information Science and Technology, 2002, 39 (1): 140 - 151.

9 Austvoll—Dahlgren A, Falk R S, Helseth S. Cognitive Factors Predicting Intentions to Search for Health Information: an application of the theory of planned behavior [J]. Health Information and Libraries Journal, 2012, 29 (4): 296 - 308.

10 吴丹, 李一喆. 老年人网络健康信息检索行为实验研究 [J]. 图书情报工作, 2014, 58 (12): 102 - 108.

11 邓胜利, 管弦. 基于问答平台的用户健康信息获取意愿影响因素研究 [J]. 情报科学, 2016, 34 (11): 53 - 59.

12 李欣颖, 徐恺英, 崔伟. 移动商务环境下 O2O 用户信息行为影响因素研究 [J]. 图书情报工作, 2015, 59 (7): 23 - 30.

13 莫秀婷, 邓朝华. 基于社交网站采纳健康信息行为特点及其影响因素的实证研究 [J]. 现代情报, 2014, 34 (12): 29 - 37.

14 成瑾, 何斯煦, 白海青. 微信用户信息分享行为研究: 基于计划行为理论的综合模型 [J]. 现代广告, 2016, (21): 39 - 47.

15 谢新洲, 安静, 王尧. 基于技术接受模型的微信用户信息发布行为研究 [J]. 情报学报, 2015, 34 (8): 801 - 808.

16 李晨, 黄灿. 微信用户信息分享行为动机研究 [J]. 现代情报, 2015, 35 (5): 57 - 62.

17 冯花朴. 潜在信息需求转化为信息行为的机理分析 [J]. 现代情报, 2009, 29 (10): 11 - 13.

(上接第 64 页)

8 Vadas D, Curran J R. Parsing Noun Phrases in the Penn Treebank [J]. Computational Linguistics, 2011, 37 (4): 753 - 809.

9 The Stanford Natural Language Processing Group. The Stanford Parser: a statistical parser [EB/OL]. [2017 - 09 - 15]. https://nlp.stanford.edu/software/lex - parser.shtml.

10 杨陟卓. 基于上下文语境的词义消歧方法 [J]. 计算机应用, 2015, 35 (4): 1006 - 1008.

11 OBO Technical WG. OBO FOUNDRY [EB/OL]. [2017 - 09 - 15]. http://www.obofoundry.org/.

12 Hitzler P, Krötzsch M, Parsia B, et al. OWL 2 Web Ontology Language Primer [J]. W3C recommendation, 2009, 27 (1): 123 - 244.

13 w3c. RDF [EB/OL]. [2017 - 09 - 15]. http://www.w3.org/RDF/.

14 w3c. SPARQL [EB/OL]. [2017 - 09 - 15]. http://www.w3.org/TR/rdf - sparql - query/.

15 Ogren P V, Cohen K B, Acquaah - Mensah G K, et al. The Compositional Structure of Gene Ontology terms [C]. Pacific Symposium on Biocomputing, 2004: 214.