电子病历自由文本实体关系抽取*

马敬东 梁力凡 夏晨曦

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 针对医学领域采用自举进行关系抽取的研究较少且国内面向医学领域的基础工具缺失问题，在一般自然语言处理技术的基础上，采用自举的算法框架，以最短依存路径构建关系模式，在过滤机制中引入候选实体的正向性评价，介绍新的算法优化策略，通过试验评价系统的性能，总结本研究的贡献与局限。

〔关键词〕 实体关系抽取；电子病历；自举；依存句法分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.12.001

Entity Relation Extraction for Free Text in Electronic Medical Records *MA Jing－dong, LIANG Li－fan, XIA Chen－xi, School of Medical and Health Management, Tongji Medical College, HUST, Wuhan 430030, China*

〔**Abstract**〕 Since there are few researches on relation extraction through bootstrap and few basic tools for medical field in China, on the basis of general natural language processing technology, the paper constructs a relation mode with the shortest dependency path and bootstrap algorithmic frame, introduces positive evaluation on candidate entities in the filtering mechanism and a new algorithm optimization strategy. It summarizes the contribution and limitation of the research by evaluating the performance of the system through tests.

〔**Keywords**〕 Entity relation extraction; Electronic Medical Record (EMR); Bootstrapping; Dependency parsing

1 引言

电子病历是医疗机构生成的针对医疗活动过程中文字、图表等数据的数字化信息，也是便于存储、管理和传输的医疗记录，其中包含着大量丰富的医疗知识，通过分析即可得到诸如疾病的患病特征、用药情况以及治疗方式等各项之间的潜在关系，这样的知识数据可以促进医疗知识创新、发展循证医疗、辅助临床和管理决策，并且还可以为用户建立个性化的健康模型。电子病历是结构化文本和非结构化文本相结合的一种知识数据，非结构化文本中包含着大量的专业术语。所以在其之上的信息提取便成为了挖掘知识的第一步，其中电子病历中的实体提取以及实体关系的抽取是核心内容。国内外均有大量研究探讨如何从电子健康档案（Eletronic Health Records，EHR）的自由文本中抽取实体及实体关系[1]，其研究对象主要是二元关系，即两个概念间的关系，如从“脑 CT 检查提示腔隙性脑梗死”中抽取＜脑 CT，腔隙性脑梗死＞这种检查与疾病的关系。所抽取的二元关系可以构建知识库[2]，为临床决策和循证医学提供基础支撑。

在抽取方法上，有研究通过手工制定的规则得到了较好的疾病本体学习性能[3]，也有研究采用信

〔收稿日期〕 2016－12－15

〔作者简介〕 马敬东，博士，副教授，发表论文多篇；通讯作者：夏晨曦，讲师。

〔基金项目〕 教育部网络时代的科技论文快速共享专项研究课题“基于互联数据的论文共享方法”（项目编号：0214516155）。

息抽取方法，利用分词、命名实体识别等分析工具[4-5]，或结合特定情境处理规则的自举（Bootstrap）方法[3]、监督学习（Supervised Learning）方法[6]直接进行实体关系抽取。归纳起来，医学领域二元关系的抽取方法主要可分为基于规则（Rule - based）、监督学习和半监督学习（Semi - supervised Learning）3 种。基于规则的抽取方法是当前应用的主流[7]，因为规则的准确率高[8]、易于解读并易于被非技术人员定制改造，其性能在特定领域也优于最高水平的监督学习方法[9]；但是，手工制定规则需要大量的领域知识和人力劳动，并且难以全面覆盖潜在概念关系，召回率偏低。至于监督学习，则需要足够的手工标注语料作为训练数据，同样耗费大量的时间和精力，也导致其在不同场景间难以移植。属于半监督学习的自举方法避免了基于规则和监督学习方法所需的大量人工劳动。而且与基于规则方法相比，自举方法具有较高的召回率，能自动覆盖到出现频次较少的实体关系模式；与监督学习方法相比，自举方法具有较好的领域可移植性。因此自举方法在电子病历关系抽取这一领域具有良好的应用前景。

然而，国内外针对电子健康档案和电子病历的信息抽取研究鲜有采用自举算法框架。而且与国外相比，国内的相关基础较为薄弱，一方面体现在我国电子病历中的书写表述尚不完全规范，成熟开放的医学术语词表缺位；另一方面体现在我国缺乏专门面向医学领域的自然语言处理工具，而国外已经有较为成熟的 MedLEE[10]、cTAKES[11] 等系统。这给中文电子病历的信息抽取带来了重要阻碍。本研究直接面向关系抽取的应用，在一般中文自然语言处理技术的基础上，以自举方法为框架，引入新的过滤机制和优化策略，所实现系统在电子病历的测试场景下得到了较好的应用效果。本文首先介绍自举算法框架，其次描述所采用的关系模式构建方法以及引入的过滤机制和优化策略，通过试验评价系统的性能，最后总结研究的贡献与局限，为未来的研究工作提供建议和参考。需要说明的是使用了“实体”、“关系”和“模式”3 个关键概念。由词语所组成的一个概念意义完整的名词短语被称为实体（Entity），实体可以是单个词语，如“冠心病”，也可以是多个词语的组合，如“风湿性关节炎”；而两个存在特定语义关系的实体则构成实体关系，如“《福尔摩斯》出自于柯南·道尔之手”这句话，“《福尔摩斯》”和“柯南·道尔”便组成了 < 书籍，作者 > 的实体关系，本文将这种实体关系简称为关系；关系模式（Relation Pattern）是在自动关系抽取技术中指示关系是否存在以及对应实体位置的上下文特征，本文简称为模式。

2 自举算法框架

2.1 基本框架

自举算法的思路由 Hearst 于 1992 年提出，在关系抽取方面的应用最早在 1996 年由 Riloff 实现[12]，1999 年用 DIPRE 系统[13]用于抽取互联网中的 < 书籍，作者 > 关系而得到广泛关注。在大规模语料情境下，自举被认为是一种高效的信息抽取方法。自举算法存在多种不同的实现方式和技术细节，本文提出的自举算法框架，见图 1，通过在文本中查找关系来抽取候选模式，符合过滤条件的候选模式将加入模式库，再将模式库中的模式与自由文本进行匹配来抽取候选关系，候选关系在过滤后加入关系库，如此循环，直到没有生成新的模式或关系。第 1 轮循环中所用的关系由人工选择，称为种子关系（Seed Relations）。在此框架下，本文研究的重点在于候选模式和关系的抽取方法及过滤机制。

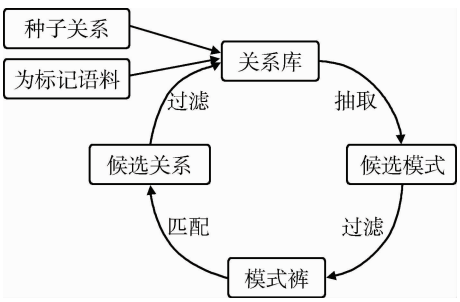


图 1 自举算法框架

2.2 候选模式抽取概述

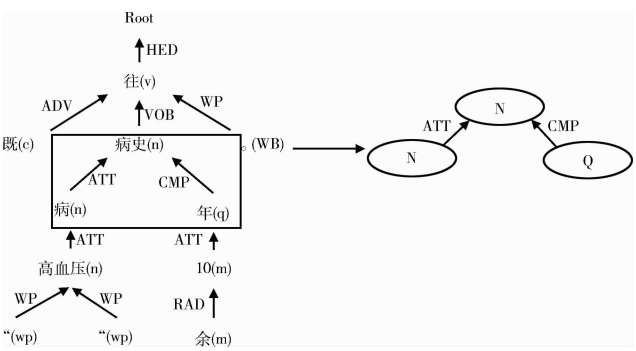
候选模式的抽取是指抽取二元关系所处上下文

的特征，常见的有直接以上下文字符串作为模式、在字符串基础上通过聚类而进一步提炼的模式以及基于依存句法的模式。不同的模式构建方式对上下文的抽象程度不一样，一般而言构建方式越抽象则覆盖到的文档越多（提高召回率），但是也越容易出现错误的关系抽取（降低准确率）。自举的一个重要问题是错误传播。假如错误的关系引入到循环过程中，则会生成错误的模板，进一步匹配到错误的关系，导致错误被不断放大，这种现象称为语义漂移（Semantic Drift）[14]。候选关系和模式的过滤机制便是防止错误传播的重要措施。自 Snowball 系统[15]之后，过滤机制主要是模式与关系之间相互依赖的评价方法。

3 研究方法

3.1 候选模式和关系抽取方法

3.1.1 *实例* 利用哈尔滨工业大学社会计算与信息检索研究中心研发的语言技术平台[16]对文本数据进行分词、词性标注和依存句法分析等预处理，选取实体间的依存句法路径以及节点上的词性作为模式抽取时所使用的特征。有研究指出[17]，实体间的最短句法路径是比全局句法路径更为有效的上下文特征，因此本研究采用最短依存路径作为模式。以“既往‘高血压病’病史 10 余年”为例，其模式，见图 2，将“高血压病”和“10 余年”之间最短依存路径上的依存关系信息和词性信息抽取出来作为模式。在图 2 左侧的依存句法树中，每个词都会以一定的关系依存于句子中的另外一词或节点；括号内表示分词节点的词性，如 n 表示名词，q 表示量词；箭头表示句法依存方向。



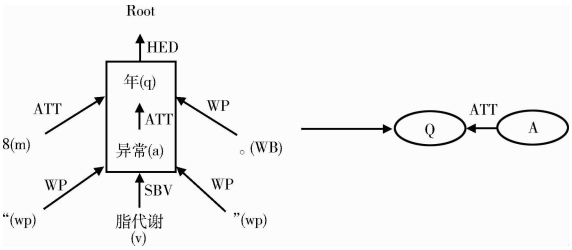


图 2 山形（上半部分）和一字形（下半部分）的模式抽取举例

3.1.2 *步骤* 本研究在此模式构建方式的基础上实现了根据关系抽取候选模式和根据模式匹配候选关系的过程。抽取候选模式的步骤为：（1）查找关系所在句子，将 <“高血压病”，“10 余年”> 这个关系直接与句子“既往‘高血压病’病史 10 余年”进行比对，句子符合查找条件则进行自然语言处理，构建如图 2 左侧所示的依存句法树，给出实体在句子中的始末位置。（2）系统根据实体的始末位置在句法路径中截取实体对应的节点集合，如图 2 中会将“高血压（n）”和“病（n）”这两个节点作为集合截取出来。（3）确定节点集合的根节点，如图 2 所示，实体“高血压病”分词集合以“病”为根节点，实体“10 余年”则以“年”为根节点。（4）截取实体根节点之间的语法路径以生成模式，在图 2 中，“病”与“年”之间的最短路径便是通过“病史”相连，于是将这 3 个节点的词性信息和句法路径信息抽取出来便是关系 <“高血压病”，“10 余年”> 在这个句子中所对应的模式。

3.1.3 *两种路径形式* 两个实体的共同父节点便是模式的根节点，如图 2 例子中上半部分的“病史”节点和下半部分的“年”节点，其位置不同导致模式存在如图 3 所示的两种形态：（1）一字形，模式根节点与其中一个实体的根节点重合。（2）山形，模式根节点不与实体根节点重合，而是两个实体根节点的依存路径上第 1 个重合的上位节点。不同形式的模式将影响候选关系抽取的方法。匹配候选关系的总体步骤包括：（1）根据模式根节点的词性在语料库中查找符合的句子及位置。（2）从模式根节点开始向下句子匹配句法依存树中的节点路径和词性，返回与模式完全匹配的实体根节点。（3）组合实体根节点下的词节点，形成实体。

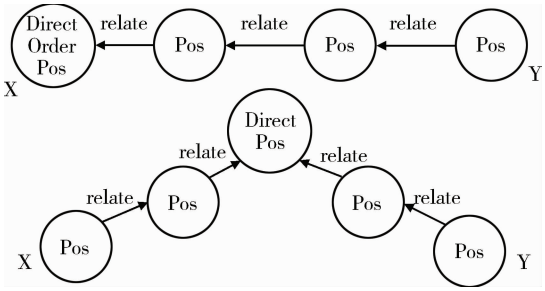


图 3 模式的两种路径形式：一字形和山形

3.2 过滤机制

3.2.1 *概述* 关系和模式的过滤是对二者分别评分，得分大于预设阈值的关系和模式被纳入下一轮循环。本文的过滤机制参考 Gupta 等的实体抽取方法[18]，基于关系和模式的相互依赖性加以改进，设计了如下对候选关系的正向性评价方法。

3.2.2 *实体的正向性评价* 如前文所述，关系由两个实体组成。本文将在已知关系中出现过的实体称为正确实体，否则称为候选实体。正向性评价用于评价抽取的候选实体是正确实体的可能性，包括编辑距离、语义优势比和分布相似度 3 个维度。需要注意的是尽管正确关系中的实体均是正确实体，但正确实体所构成的关系不一定是正确的关系。正向编辑距离测量两个词语之间达到完全相同所需要的对字或字母基本操作（即增删改）次数。本研究计算候选实体与每一个正确实体的编辑距离除以正确实体的字数，取其最大值作为编辑距离维度的得分。该维度反映候选实体与特定正确实体在字形方面的最大相似度。语义优势比计算候选实体的每一个字（不包括数字）在正确实体中出现的频次和正确实体总数之比。该维度反映候选实体与正确实体在字形方面的总体相似程度。分布相似度体现实体在上下文方面的相似程度。在本研究中的上下文采用实体的前后 3 个词，以余弦夹角计算上下文的相似度，根据分布相似度对实体进行聚类，继而以实体的正确与否作为因变量、以实体所属的类群作为自变量建立 Logistic 回归模型，所属类群的系数作为得分。最后，本文对上述 3 个维度分别进行 Softmax 标准化和最大最小标准化，以 3 个维度的平均分作为候选实体的正向性评价总得分。已知正确实体的得分设为 1，由于一个关系具有两个实体，关系的正向性评价取两个实体中得分较低的一个。

3.2.3 *候选模式的信度评价* 模式的评分如下式：

$$\mathrm{conf}(p) = \frac{|P_r|}{|P_r| + \sum_{e \in U_r}(1 - \mathrm{score}(e))} \lg(|P_r|)$$

式中：$|P_r|$为与候选模式匹配的已知关系数量；U_r 为模式匹配到的候选关系；score（e）为候选关系 e 的正向性评价得分。ExDISco 系统[19]和 Gupta 等[18]的模式评分方法与本研究类似，但 ExDISco 系统在分母部分没有计算正向性得分总和而是统计候选关系的数量，忽略了候选关系之间的信度差异；Gupta 等的方法还考虑了已知错误实体，但在数据量较大的情况下严重影响算法效率。本研究所使用的研究方式相当于二者的折中，兼顾评分的全面性与算法的开销。

3.2.4 *关系的信度评价* 结合关系的正向性评价和抽取关系的模式的信度。参照 Snowball 系统[20]中的评价方法，关系 r 的来源模板信度计算如下：

$$\mathrm{conf}(t) = 1 - \prod_{i=1}^{|P|} (1 - \{\mathrm{conf}(p_i)\} \cdot \mathrm{Match}(r, p_i))$$

式中：conf（p_i）为模式库中第 i 个模式的信度；Match（r，p_i）考察模式 p_i 是否能与关系 r 匹配，是则返回 1，否则返回 0。模式信度和正向性评价的平均得分即为关系的信度。

3.3 优化策略

3.3.1 *概述* 在抽取既往史的应用场景下，上述系统的性能仍不能达到令人满意的水平。因此，针对自举算法在本场景下出现的问题，本文提出 4 种优化策略，即改变分句方法、实体匹配方法、抽取实体中间词和检查实体包含关系。

3.3.2 *分句方法* 既往史在电子病历中的表述方式与日常表述存在一定差异，其中最显著的差异在于断句。既往史往往将意义完整的几句话写成一个长句，如“‘高血压病’病史 12 年，最高血压达 180/100mmHg，现应用‘依那普利 1 片 bid、尼莫地平 1 片 tid’治疗……”，这导致了 LTP 工具的依

存句法分析存在较大的误差。经过考察，疾病或手术及其对应时间不会被逗号分隔开，因此在使用 LTP 工具处理之前，本研究将逗号全部替换为句号，使句法分析局限于一个更小的句子范围内，提高了句法分析的准确性。

3.3.3 *实体匹配方法* 由于 LTP 工具的命名实体识别方法不支持医学领域的疾病、手术、检查等专有名词的识别，即使分词工具内置词表具有部分医学词汇，也可能与电子病历中的表述不一致。电子病历中的医学术语，如疾病实体“高血压病”往往被切分成“高血压”和“病”两个词，因此在模式匹配的过程中有必要将词语重新组合为实体，不同的实体匹配方式对匹配结果有着重要影响。当模式呈山形时，本研究主要通过将实体根节点的下位节点串接起来的方法进行实体匹配，在串接中截去以标点符号（括号除外）隔开的字符串，得到结果为匹配实体。当模式呈一字形时，“年”既是实体根节点又是模式根节点，在句法路径中层级较高、下位节点较多，因而不适用于该方法。针对后一种情况，考察根节点前后的节点是否为前一个节点的子节点，是则组配到一起并继续往前后节点延伸，否则停止组配过程，返回已组配的实体。

3.3.4 *抽取实体中间词* 模式在匹配关系过程中经常出现抽取根节点过浅的问题。能正确匹配某一个句子的模式在另一句中就可能因为根节点过浅导致所匹配的实体过大。以图 2 为例，如果模式匹配过浅，以“病史”作为时间的实体根节点，则得到“病史 10 余年”这样的错误实体。这种情况下，实体在编辑距离、语义优势比、分布相似度 3 个维度均有着较高的评价，难以将其与正确实体区分开。为解决该问题提出实体中间词的概念。在特定关系的实体对之间，某些词会频繁出现在目标实体对的句法依存路径上，因此包含这些词的实体极有可能根节点匹配过浅，形成实体过大。本系统将实体中间词定义为“正确实体向上 3 个父节点中出现频率大于 0.3 的词”。系统在抽取模板的过程中，抽取实体中间词并存储，以后包含中间词的实体均被认为不正确，实体的评分为 0。

3.3.5 *实体包含关系的检查* 实体匹配错误的另一个主要原因是模式匹配经常出现实体根节点过深的问题。与前一个问题相反，这种匹配错误产生不完整的实体。如图 2 中，如果模式匹配过深，以“10”作为时间的实体根结点，则得到 <“高血压病”，“10 余” > 这样的错误概念关系，这种情况同样难以被正向性评价甄别。为解决该问题，引入实体包含关系的检查。所谓实体包含关系，即在词形上一个实体是否包含另一个实体，是则说明前者包含后者。被包含的实体，而且不是包含于后半部分，则评价得分为 0。例如，在已发现实体“类风湿性关节炎”的前提下，“类风湿性关节”将被判定为错误，而“关节炎”则通过包含关系的检查。每一轮循环后，正确实体内部也需要进行实体包含关系的检查。

4 应用实践

4.1 数据准备与试验方案

为了评价系统的性能，从青岛市某医院的电子病历数据库中随机选择了 1 000 条既往史信息作为语料，利用本研究提出的方法抽取 < 疾病/手术，时间 > 的特定关系。之后对该语料进行人工标注，对比标注文本与抽取结果，评价系统性能。研究中输入的种子关系共 11 对实体，见表 1。

表 1 种子关系

实体 1	实体 2
高血压病	15 年
高血压病	多年
冠心病	10 余年
骨量减少	1 年
急性肝炎	8 年
咳嗽	3 年
十二指肠憩室	2 年
脂代谢异常	3 年
脂代谢异常	多年
慢性肾功能不全	1 年
椎基底动脉供血不足	1 年

4.2 试验结果

实际运行中发现过滤机制可能过于严格，为提高召回率，评分阈值设置比较宽松。研究中将模板

评价的评分阈值设置为 0.1，关系的评分阈值设置为 0。经过 7 轮循环后没有新的模板和关系而终止，最终抽取得到 1 804 个关系，其中有 33 个错误关系，文本经人工标注共有 3 532 个可抽取的实体关系。系统的准确率（Precision）为 98.2%，召回率（Recall）为 50.1%，F_1 指标为 0.664。表 2 列举了结果中的 17 个关系。本系统的性能瓶颈主要在于一般自然语言处理工具在医学领域的适用性，其问题集中体现在分词、词性标注和句法分析 3 方面。工具的分词性能主要影响了本研究的准确率。在 33 个错误关系当中，有 26 个是 LTP 分词错误导致的，如将“史 9”作为单独一个词，其余则主要是动词性质的不完整实体，如将“心律失常”中的“失常”划分为一个单独的动词。工具的词性标注和句法分析性能影响了本研究的召回率。文本中未被识别的正确关系以实体间的并列关系为主。以“‘脂肪肝’、‘脂代谢异常’病史 1 年”这句话为例，由于“异常”具有动词性质，在依存句法依存于“病史”而非“脂代谢”，导致“脂代谢异常”无法作为一个实体被抽取。

表 2 结果举例

实体 1	实体 2
单纯甲状腺肿	2 年
不宁腿综合征	2 年
梅尼埃病	6 年
支气管哮喘	3 年
过敏性哮喘	21 年
胆囊结石	8 年
慢性胃窦炎	33 年
高血压病	8 年
胃溃疡	20 余年
风湿性心脏病	51 年
高血压病	11 年
左肺陈旧性肺结核	7 年
慢性支气管炎	40 余年
帕金森病	8 年
精神分裂症	4 年
高血压病	13 余年
肥厚性心肌病	23 年

4.3 分析

4.3.1 概述 试验结果表明本研究设计的方法在中文电子病历自由文本的关系抽取问题上取得了较优秀的性能，其准确性显著优于 DIPRE[13]、ExDISco[20]、Snowball[21] 等经典的网络文本关系抽取系统，与近年来的一般化中文关系抽取系统比较也处于领先水平。对此 Zheng 等做的中文网络文本关系抽取研究[21]，除了 <国家，首都> 关系的抽取准确性略高于本研究之外，其他关系的抽取准确性均低于本研究；准确率、召回率和 F 指标均高于刘丹丹等[22]的系统性能；准确率高于郭喜跃等[23]和王逯姚[24]的基于监督学习的抽取方法，召回率相对较低。总体来说本研究的抽取系统存在以下几点创新之处。

4.3.2 直接进行关系抽取 大部分采用自举算法的研究均在命名实体识别的基础上进行[1]；然而，加入命名实体识别等技术，将使系统性能进一步受限于预处理的准确性，而且中文医学领域也缺乏相关的方法或工具。直接进行关系抽取实现了较高的准确率，对国内进行的类似研究有重要借鉴意义。

4.3.3 创新的关系模式过滤机制 在常用的基于模式和关系相互依赖性的评价方法以外，加入了候选关系的正向性评价，这种做法在实体抽取中有所尝试，现有基于自举的关系抽取算法中未见应用。本研究对此方法加以改进，应用到关系抽取领域，取得不错的效果。有研究采用多分类互斥的方法解决自举算法的语义漂移问题，但该方法过于严格，会进一步降低召回率，更适用于海量语料[14]。

4.3.4 针对中文临床文本场景提出的优化策略 所提出的 4 种优化策略在一定程度上弥补了 LTP 工具对于临床文本的不适应，同时增强系统对候选实体的识别能力，显著提升系统性能。

5 结语

在自举算法框架下，结合实体的语义特征制定了新的过滤机制，有效应对语义迁移问题。针对新型过滤机制难以处理的错误类型，在断句、实体匹

配、实体包含关系等方面针对性地提出了一系列优化策略，将本系统的准确率提升到 98% 以上，显著优于以往研究。所使用的方法对于中文医学领域的信息抽取具有重要参考意义。

然而，本文所介绍的信息抽取方法仍存在一定局限，有待未来工作解决：（1）召回率偏低，主要原因是一般自然语言处理工具缺乏医学领域专业词表或语料支持。未来需要专门针对自然语言处理技术在医学领域的应用进行优化，如引入医学专业词表，或利用临床文本语料，提升分词、词形标注和句法分析的准确性。（2）中文医学领域缺乏自然语言处理的成熟工具，未来研究应完善相关术语词表或改进医学命名实体识别的方法以弥补该缺陷，同时建议参考国外面向医学领域且较为成熟的自然语言处理工具，针对中文医学领域特点开发类似工具。（3）算法开始所输入的种子关系对算法性能的影响在文献中没有太多的讨论。选取种子关系时考虑了种子关系所处上下文的特殊性以及分词后的词性，从而加强关系模式的覆盖面，提高算法的召回率。未来研究需要进一步探讨种子实体关系对抽取效果的影响。（4）仅以患者既往史为例探讨了算法的可行性。有研究[18]指出不同的应用场景对自举算法的性能有着一定影响。尤其是针对 < 疾病/手术，时间 > 关系抽取所提出的 4 种性能优化策略，虽然可供中文医疗领域的信息抽取研究借鉴，但仍不能确定其可推广性，未来需要在更多应用场景下考察其过滤机制和优化策略的适用性。

参考文献

1 Uzuner Ö, South BR, Shen S, et al. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text [J]. Journal of the American Medical Informatics Association, 2011, 18 (5): 552 – 556.

2 杨芬. 本体学习中概念和关系抽取方法研究 [D]. 重庆：重庆大学, 2010.

3 陈莺莺. 病历信息抽取方法的研究与实现 [D]. 杭州：浙江工业大学, 2010.

4 贺海涛，郑山红，侯丽鑫，等. 基于中文文本的疾病领域本体学习的研究 [J]. 吉林大学学报：信息科学版, 2014, 32 (1): 76 – 81.

5 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述 [J]. 自动化学报, 2014, (8): 1537 – 1562.

6 吴嘉伟. 电子病历实体关系抽取研究 [D]. 哈尔滨：哈尔滨工业大学, 2014.

7 Chiticariu L, Li Y, Reiss FR. Rule – based Information Extraction is Dead! Long Live Rule – based Information Extraction Systems! [C]. Proceedings of the Conference on Emperical Methods in Natural Language Processing, 2013: 827 – 832.

8 Gupta S, Maclean DL, Heer J, et al. Induced Lexico – syntactic Patterns Improve Information Extraction from Online Medical Forums [J]. Journal of the American Medical Informatics Association Jamia, 2014, 21 (5): 902 – 909.

9 Nallapati R, Manning CD. Legal Docket – entry Classification: where machine learning stumbles [C]. Hawaii, USA: Proceedings of the Conference on Emperical, Methods in Natual Language Processing, 2008: 438 – 446.

10 Sevenster M, van Ommering R, Qian Y. Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE [J]. Journal of Digital Imaging, 2012, 25 (2): 240 – 249.

11 Savova GK, Masanz JJ, Ogren PV, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications [J]. Journal of the American Medical Informatics Association, 2010, 17 (5): 507 – 513.

12 Riloff E. Automatically Generating Extraction Pattern from Untagged Text [A]. AAAI – 96 Proceedings, 1996: 1044 – 1049.

13 Brin S. Extracting Patterns and Relations from the World Wide Web [J]. Lecture Notes in Computer Science, 1998, (590): 172 – 183.

14 Curran JR, Murphy T, Scholz B. Minimising Semantic Drift with Mutual Exclusion Bootstrapping [C]. Proceedings of the 10th Conferece of the Pacific Association for Computational Linguistics, 2007.

15 Agichtein E, Gravano L. Extracting Relations from Large Plain – Text Collections [C]. Proceeding of the Fifth ACM Conference on Disital Libraries, 2000: 85 – 94.

16 Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform [C]. Beijing: Proceedings of the 23rd Intemational Conference Computational Linguistics, 2010.

17 Bunescu R C, Mooney R J. A Shortest Path Dependency Kernel for Relation Extraction [C]. Vancouver B C, Canada: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005: 724 – 731.

18 Gupta S, Manning C. Improved Pattern Learning for Bootstrapped Entity Extraction [C]. Baltimore, Maryland, USA: Eighteenth Conference on Computational Natural Language Learning, 2014: 98 – 108.

19 Yangarber R, Grishman R, Tapanainen P, et al. Automatic Acquisition of Domain Knowledge for Information Extraction [C]. Uppsala, Sweden: Conference on Computational Linguistics Association for Computational Linguistics, 2010: 940 – 946.

20 Agichtein E, Gravano L. Snowball: extracting relations from large plain – text collections [C]. Queensland, Australia: ACM Conference on Digital Libraries, 2010: 85 – 94.

21 Agichtein E, Gravano L. Snowball : extracting relations from large plain – text collections [C]. Queensland, Australia: ACM Conference on Digital Libraries, 2010: 85 – 94.

22 刘丹丹, 彭成, 钱龙华, 等. 词汇语义信息对中文实体关系抽取影响的比较 [J]. 计算机应用, 2012, 32 (8): 2238 – 2244.

23 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取 [J]. 中文信息学报, 2014, 28 (6): 183 – 189.

24 王逯姚. 基于多核学习的肿瘤—药物—基因语义关系抽取 [D]. 北京: 北京协和医学院, 2015.

2017 年《医学信息学杂志》编辑出版重点选题计划

2017 年本刊将继续以“学术性、前瞻性、实践性”为特色，及时追踪并深入报道国内外医学信息学领域前沿热点，反映学科研究动态，展示学科应用成果，引领学科发展方向。现对 2017 年度编辑出版重点选题策划如下：

一、医药卫生体制改革与医药卫生信息化

1“十三五”卫生信息化建设的创新与发展；2“互联网 +”环境下医药卫生发展的新方向、新举措；3 医药卫生信息化发展规划与战略；4 区域卫生、公共卫生、基层卫生信息化建设；5 人口健康信息平台建设；6 医疗卫生信息相关标准研究与应用；7 医疗卫生信息化相关法律法规。

二、医学信息技术

1 健康医疗大数据的采集、存储与管理、深度挖掘与应用创新；2“互联网 +”医疗模式的技术实现；3 精准医学与个性化医疗技术；4 可穿戴设备、远程医疗服务与健康管理；5 物联网、智慧医疗技术与实现；6 各类医学信息系统互联互通，更新与改造升级；7 医疗信息共享及安全监管。

三、医学信息研究

1 医学信息学基础理论及方法研究；2 医学科技创新体系和发展战略；3 公民健康素养培养及健康促进；4 医药卫生信息分析评价、舆情监测；5 医药卫生知识发现技术与实现。

四、医学信息组织与利用

1“互联网 +”环境下医学图书馆的新形态与新功能；2 大数据驱动的健康知识服务与决策咨询服务；3 医学知识组织的关键技术与发展方向；4 医学信息交互及存取；5 医学图书馆区域合作及资源共享模式研究。

五、医学信息教育

1“互联网 +”环境下医学信息专科、本科、研究生教育及继续教育面临的挑战创新；2 医学信息素养教育；3 医学信息课程改革与实践；4 国外医学信息学教育的先进经验借鉴。

（《医学信息学杂志》编辑部）