电子病历命名实体识别和实体关系抽取研究综述

杨锦锋 [1] 于秋滨 [2] 关 毅 [1] 蒋志鹏 [1]

摘 要 电子病历 (Electronic medical records, EMR) 产生于临床治疗过程, 其中命名实体和实体关系反映了患者健康状况, 包含了大量与患者健康状况密切相关的医疗知识, 因而对它们的识别和抽取是信息抽取研究在医疗领域的重要扩展. 本文首先讨论了电子病历文本的语言特点和结构特点, 然后在梳理了命名实体识别和实体关系抽取研究一般思路的基础上, 分析了电子病历命名实体识别、实体修饰识别和实体关系抽取研究的具体任务和对应任务的主要研究方法. 本文还介绍了相关的共享评测任务和标注语料库以及医疗领域几个重要的词典和知识库等资源. 最后对这一研究领域仍需解决的问题和未来的发展方向作了展望.

关键词 电子病历, 命名实体识别, 实体关系抽取, 共享评测任务

引用格式 杨锦锋, 于秋滨, 关毅, 蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, **40**(8): 1537−1562

DOI 10.3724/SP.J.1004.2014.01537

An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction

YANG Jin-Feng[1] YU Qiu-Bin[2] GUAN Yi[1] JIANG Zhi-Peng[1]

Abstract Electronic medical records (EMRs) are generated in the process of clinical treatments. Named entities and entity relations in EMRs reflect patients′ health conditions and represent patients′ personalized medical knowledge. Consequently, named entity recognition and entity relation extraction on EMR are important expansion of information extraction in the medical domain. In this paper, the language characteristic and structure features of EMR narratives are firstly discussed, and then general methods for named entity recognition and relation extraction are sketched out. Furthermore, this paper introduces and analyzes the tasks and corresponding methods for named entity recognition, entity assertion recognition and relation extraction of EMR in detail. Related shared evaluation tasks and annotated corpora as well as several important dictionaries and knowledge bases are also introduced. Finally, problems to be handled and future research directions are proposed.

Key words Electronic medical record (EMR), named entity recognition, entity relation extraction, shared task

Citation Yang Jin-Feng, Yu Qiu-Bin, Guan Yi, Jiang Zhi-Peng. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, 2014, **40**(8): 1537−1562

电子病历 (Electronic medical record, EMR) 是指医务人员在医疗活动过程中, 使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化信息, 并能实现存储、管理、传输和重现的医疗记录[1], 是由医务人员撰写面向患者个体描述医疗活动的记录. 通过分析电子病历能挖掘出大量与患者密切相关的医疗知识, 这种认识早已获得共识[2]. 比如, 某患者电子病历中, “头 CT 检查显示腔隙性脑梗死”. 在这句话中, “头 CT” 是检查手段, “腔隙性脑梗死” 是疾病, 这二者在电子病历信息抽取研究中被称为命名实体, 这两个实体间的关系是 “头 CT” 证实了 “腔隙性脑梗死” 的发生, 或者说 “腔隙性脑梗死” 可以通过 “头 CT” 这种检查手段得到确认. 从电子病历里自动挖掘这些知识就是要自动识别电子病历文本中与患者健康密切相关的各类命名实体以及实体间的关系[3], 电子病历命名实体识别和实体关系抽取是电子病历信息抽取研究的主要内容, 该研究在医学信息学 (Medical informatics) 中用于临床决策支持 (Clinical decision support, CDS) 研究服务于医疗专业人员[4], 同时在用户健康信息学 (Consumer health informatics) 中支持用户健康状况建模和个性化医疗健康信息服务研究服务于普通

收稿日期 2013-08-30 录用日期 2013-12-18
Manuscript received August 30, 2013; accepted December 18, 2013
国家自然科学基金 (60975077) 资助
Supported by National Natural Science Foundation of China (60975077)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 哈尔滨工业大学语言技术中心网络智能研究室 哈尔滨 150001 2. 哈尔滨医科大学附属第二医院病案室 哈尔滨 150086
1. Web Intelligence Laboratory, Language Technology Center, Harbin Institute of Technology, Harbin 150001 2. Medical Record Room, The 2nd Affiliated Hospital of Harbin Medical University, Harbin 150086

患者和用户[5]. 中文领域的临床决策支持系统早在上世纪 90 年代初就已经展开了研究[6].

电子病历主要有两类, 即门诊病历和住院病历. 门诊病历通常较短, 包含信息较少, 也缺乏对患者治疗情况的跟踪, 因而, 电子病历信息抽取和文本挖掘研究大多关注于住院病历. 如不明确说明, 本文所指的电子病历均指住院病历. 电子病历并不是完全结构化的数据, 还包括一些自由文本等复杂的无结构数据. 这种文本信息方便表达概念以及事件等, 但是同时也为搜索、统计分析等研究制造了障碍, 因此, 自然语言处理、信息抽取等相关技术在电子病历上的分析、挖掘中将发挥重要的作用. 自然语言处理应用于电子病历文本, 也叫医疗语言处理 (Medical language processing, MLP)[7], 其基础研究包括词性标注、分词、句子边界识别、句法分析、命名实体识别和实体关系抽取、共指消解等. 命名实体及其关系是电子病历医疗知识的主要载体, 同时命名实体识别和实体关系识别也是电子病历文本挖掘和信息抽取研究的主要内容.

由于电子病历是患者治疗经过的记录, 电子病历文本包含了大量的实体, 且实体类型较多, 主要有四大类实体. 1) 首先, 电子病历记录了患者、医生以及医疗机构的名称、编号等隐私信息 (Private health information, PHI). 在开放领域, 这些信息是普通的命名实体, 但是在医疗领域, 这些信息是患者和医生的隐私信息. 所以, 电子病历对外发布的先决条件就是去隐私化信息 (De-identification)[8]. 实现去隐私化信息首先要识别 PHI, 用替代信息替换病历中的 PHI, 以保持病历文本的完整性[9]. 2) 电子病历记录的是治疗经过, 因而疾病、症状、检查和治疗这四类与疾病治疗密切相关的实体是电子病历中数量最多的实体 (疾病和症状也被合并为医疗问题). 比如 “高血压”、“脑梗死” 是疾病, “眼震”、“言语笨拙” 是症状, “头 CT”、“彩超” 是检查, “抗血小板凝聚”、“胰岛素” 是治疗. 为了表达的一致性和准确性, 这四类实体通常有固定的表达, 作为专业术语使用. 为了使这些专业术语得到共识, 很多机构维护了大量的专业术语以及术语的变体. 在使用和维护过程中, 这些术语又被称作概念. 我们从信息抽取的角度, 把这些概念视为命名实体 (后续章节提到的实体等同于概念). 电子病历中的医疗问题, 也就是疾病和症状, 还存在着一些重要修饰成分 (或者叫上下文特征), 比如, “不排除缺血性疾病” 和 “双侧眼球运动自如, 无眼震”. 在这两个例子中, “缺血性疾病” 不是确定发生的, “眼震” 是肯定排除的. 如果不识别这些修饰成分, 抽取出来的医疗知识将谬以千里. 所以电子病历命名实体识别研究还需要识别疾病和症状的修饰. 3) 在治疗类实体中, 药物是重要的治疗信息, 不仅涉及药物名称 (包括通用名、商品名), 还包括剂量、施治方式、施治频次、施治持续时间等信息, 这些实体通常视为药物的属性. 4) 时间信息也是一类重要的实体数据. 因为患者的治疗和病情的发展有时序性, 所以在病历中很多表示事件的实体都与时间相关. 时间在病历中的表达方式没有固定的形式, 比如 “于 2012-06-08 11:24 步入病室”、“于入院前 5 小时无明显诱因出现右侧肢体麻木”. 电子病历命名实体识别就是要识别出病历文本中这些表达患者医疗信息或者健康信息的实体.

电子病历中实体之间不是孤立存在的, 相互之间存在着一定的关系, 实体关系正是医疗知识的主要体现. 电子病历中的实体关系主要有三大类. 1) 概念之间的关系: 电子病历中概念间关系包括疾病和症状的关系、疾病和疾病的关系、疾病和检查的关系以及疾病和治疗的关系. 这些关系是最重要的实体关系, 根据这些关系可以构造患者健康状况的简明摘要, 不仅体现了患者的健康状况, 而且体现了医疗知识. 2) 概念间的等价关系: 有些概念可能是其他概念的等价表达, 只是形式不同而已, 这种关系的识别是共指消解的主要研究内容, 通常作为命名实体识别的后续任务. 3) 概念和时间的关系: 表示事件的概念通常具有时间属性, 即什么时间发生的. 比如观察到的症状、所做的检查、给予的治疗都有明确的时间点. 抽取概念和时间的关系就是还原概念的时间属性信息, 这些时间信息用于建立事件发生的时间线, 据此跟踪患者的健康状况、查找病因、分析治疗的有效性和副作用等.

图 1 总结了电子病历中常见的实体类型和实体关系类型. 这些实体和实体关系从不同角度反映了电子病历中的医疗知识和患者健康状况. 在进行电子病历信息抽取研究时, 不同类型的实体和实体关系抽取作为不同的研究任务展开. 隐私信息的识别通常包含于去隐私化研究任务中[9]; 药品信息识别需要抽取药品的诸多属性数据, 因而也作为一个独立的信息抽取任务[10], 类似的多元组属性信息抽取研究也见于疾病信息的抽取[11]. 概念的识别和概念关系的抽取类似于开放领域实体识别和实体关系抽取, 是电子病历信息抽取研究的主要任务[12]. 医疗问题的修饰识别是电子病历信息抽取研究中独有的任务, 该任务由早期识别症状的否认信息发展而来[13], 现在已成为电子病历信息抽取研究的重要任务[12]. 共指消解关注实体间等价关系的识别, 虽然是处理一类特殊的实体关系, 但一直是信息抽取研究的重要任务, 在电子病历信息抽取中也作为独立的研究任务受到越来越多的关注[14−16]. 时间信息是病历中特殊的实体, 体现了健康状况的时间维度, 时间信息识别研究在开放领域受到了极大的关注[17−18], 在医疗领域也作为一个独立的研究任务吸引起了研究者的兴趣[19−20].

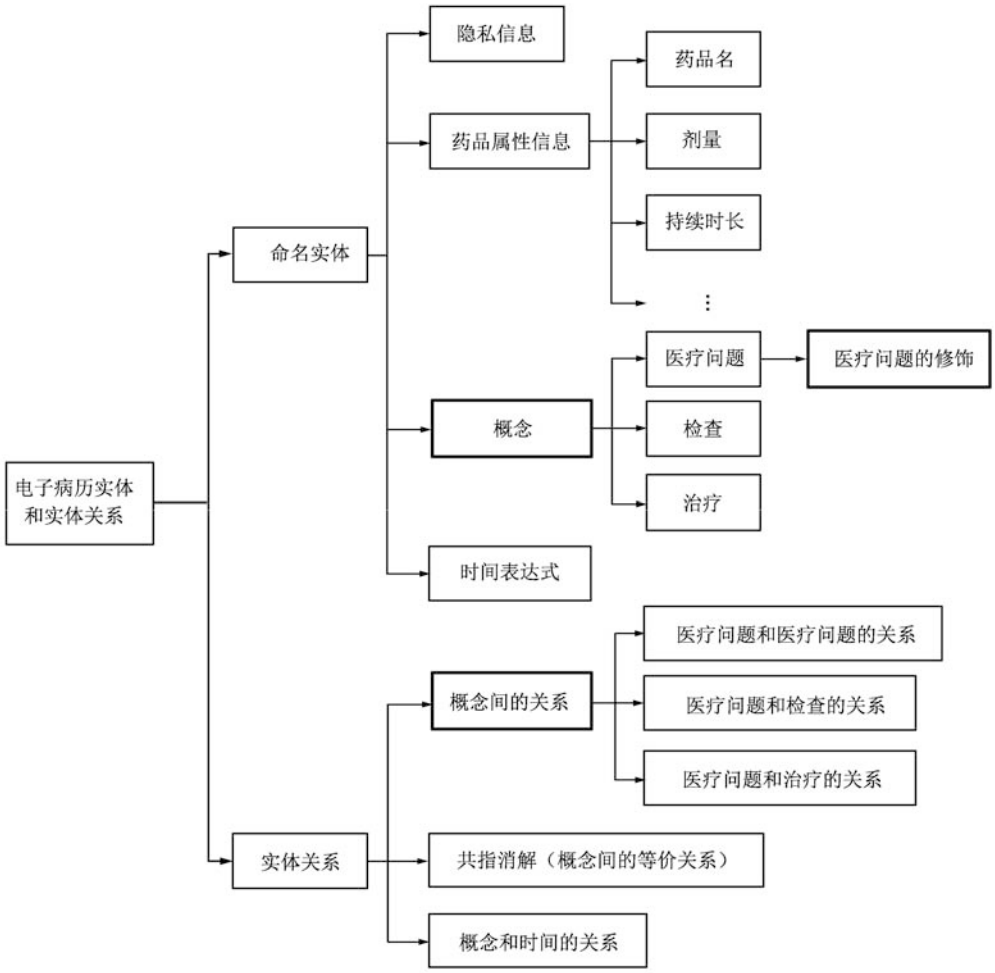

图 1　电子病历实体类型和实体关系类型

Fig. 1　Entity types and entity relation types in EMR

在上述各类任务中, 概念 (医疗问题、检查、治疗) 的识别、医疗问题的修饰和概念间关系的抽取是电子病历信息抽取研究的三个核心任务. 概念、修饰和概念间关系这三类知识体现了以医疗问题为中心的思想, 修饰体现了医疗问题和患者的关系, 检查是为了证实医疗问题, 治疗是为了改善医疗问题. 根据这三类知识, 我们可以把电子病历中抽取出来的实体按照实体关系组织起来, 系统地表示以医疗问题、治疗和检查为主体的医疗知识, 而且这些医疗知识与患者密切相关, 具有个性化特点. 鉴于此, 本文主要关注命名实体识别、疾病或症状的修饰识别、实体关系抽取这三类研究. 这三个任务围绕命名实体展开 (如图 2 所示), 涵盖了电子病历信息抽取的基本任务.

电子病历是重要的医疗临床数据, 不仅包含了医生的专业知识, 而且与患者的健康状况密切相关. 由于电子病历这种双重特性, 电子病历命名实体和实体关系识别研究成为命名实体识别研究在医疗领域的重要拓展, 同时也是电子病历信息抽取的重要研究内容, 该课题的研究吸引了国内外越来越多研究者的关注. 本文在接下来的部分首先分析电子病历文本的重要特点, 概述命名实体识别和实体关系抽取的任务和研究思路, 然后分别详细阐述电子病历命名实体研究的三个任务以及各自的研究方法, 接着介绍国内外主流的评测会议以及现有的资源建设情况, 最后对本文工作进行总结并展望电子病历命名实体和实体关系识别研究的发展趋势.

1　电子病历文本特点

电子病历是患者在医疗机构就诊时产生的医疗记录. 电子病历数据的产生可以说是全民参与的结果, 每天都会产生大量的病历. 电子病历数据的形式主要有表格、自由文本、图像这三种, 自由文本形式的非结构化数据是电子病历中非常重要的数据, 主要有出院小结、病程记录、主诉、现病史、病历小结、医患沟通记录、医患协议、超声报告. 出院小结是对患者治疗过程和治疗效果的总结, 病程记录主要是阶段性记录患者临床表现、 经历的检查和治疗等医

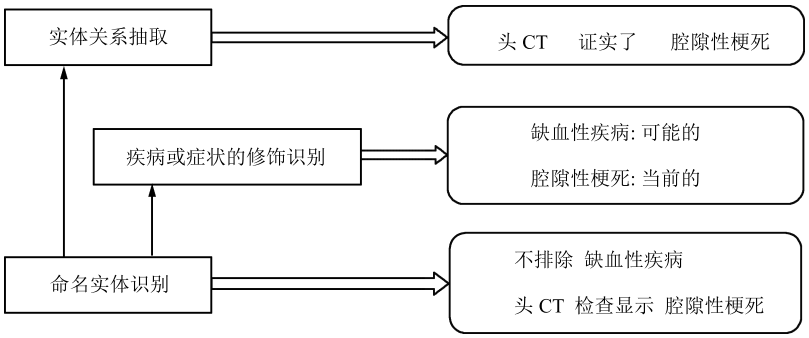

图 2 电子病历命名实体识别和实体关系抽取研究的三个任务

Fig. 2 The three tasks of named entity recognition and relation extraction on EMR

疗活动过程; 主诉、现病史和病历小结内容都包含在出院小结和病程记录里; 超声报告只涉及单项检查, 检查结果也包含在病程记录里; 医患沟通是医务人员就治疗的风险告知患者及家属, 医患协议主要是患者应遵守的纪律等. 出院小结和病程记录是电子病历中最重要的两类自由文本, 是电子病历信息抽取和文本挖掘关注的重点. 这些自由文本由医务人员撰写, 包含了患者的症状描述、检查结果的分析、做出的诊断、以及给予的诊疗方案, 是医务人员专业医疗知识的集中体现, 也是患者个性化健康信息的集中体现. 这些文本数据的输入便捷性、可理解性和呈现方式是电子病历研究的热点问题[21]. 输入便捷性是指输入方式应该方便医生输入电子病历内容, 尽量减少医生的负担, 包括医疗语言尽可能简洁、使用受控术语和模板等; 可理解性是指病历文本表达的意思准确无误并且在不同医疗机构之间可交换阅读; 良好的呈现方式指电子病历应该便于医务人员阅读, 快速找到重要的信息. 正因为这三个问题, 与传统医学文献中的文本相比, 电子病历中的文本不管是结构还是语言, 都具有一些新特点.

1.1 大数据特性

电子病历系统是信息化技术在医疗领域的重要应用, 是临床使用最早也是最主要的一个工具[2]. 由于国外发达国家信息化起步较早, 国外很多医疗机构早在上个世纪 70 年代就已经开始实施电子病历系统来管理和存储临床医疗数据, 积累了大量的电子病历. 以印第安纳大学医学中心的电子病历系统 (Regenstrief medical record system, RMRS) 为例, 该系统是最早的电子病历系统之一[22], 始建于 1972 年, 目前该电子病历系统为 1 300 000 个患者提供服务, 已产生 15 000 000 份电子病历.

2009 年国务院颁布了 “关于深化医药卫生体制改革的意见”, 2010 年卫生部出台了《电子病历基本规范 (试行)》和《电子病历系统功能规范 (试行)》等规范. 在国家一系列政策的推动下, 电子病历系统在各级医院广泛实施. 我国医疗机构数量庞大, 患者的就医需求也与日俱增, 门诊病历和住院病历急剧增长. 仅以哈尔滨医科大学附属第二医院病案室给出的近 10 年住院病历统计数据为例 (如图 3 所示), 就可了解电子病历数据量的庞大.

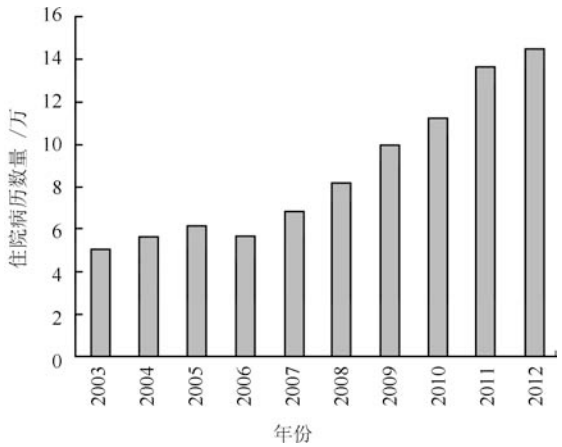

图 3 哈尔滨医科大学附属第二医院住院病历统计数据

Fig. 3 Statistics of in-patient records of The 2nd Affiliated Hospital of Harbin Medical University

海量的电子病历数据堪称医疗领域的大数据, 是座知识的宝库, 蕴含了大量的医疗知识和患者的健康信息[2]. 在当前大数据研究浪潮下, 电子病历信息抽取和文本挖掘越来越吸引人们的目光. 这些研究将为临床智能支持、循证医学研究和疾病监控等提供支持, 从而提高医疗服务质量.

1.2 结构特点

病程记录和出院小结是电子病历中最重要的两类自由文本, 电子病历信息抽取也主要关注这两类文本. 下面主要介绍病程记录和出院小结的结构特点.

1.2.1 病程记录结构特点

病程记录的撰写从传统面向来源的组织方

式, 演化出面向时间的组织方式 (Time-oriented medical record, TOMR)[23] 和面向问题的组织方式 (Problem-oriented medical record, POMR)[24], POMR 已成为病程记录方式的事实标准[25]. 这两种方式各有利弊[26], 在当下的电子病历撰写中混合使用, 即以医疗问题为中心组织内容, 同时按照时间轴展开记录每个时间点的医疗问题情况. 这种记录方式有助于对医疗问题的治疗情况和进展进行跟踪和分析. 面向问题的病程记录普遍采用 SOAP (Subjective, objective, assessment, plan) 格式撰写, 首先描述各种症状、体征以及重要检查结果, 然后对这些证据进行综合评估并做出诊断, 最后给出相应的诊疗计划. 以匹兹堡大学医学中心的病程记录样本为例[27], 我们抽取的病程记录结构如图 4 所示.

图 4　匹兹堡大学医学中心的病程记录结构

Fig. 4　The structure of a progress note from University of Pittsburgh Medical Center

国内电子病历的病程记录主要有三类: 首次病程记录、日常病程记录 (也叫查房记录)、上级医师查房记录[28]. 首次病程记录详细记录了患者的病例特点、诊断和诊疗计划, 下面分析首次病程记录文本特点. 图 5 是哈尔滨医科大学附属第二医院首次病程记录. 图 5 所示的首次病程记录按照内容可以划分为主诉、既往史、主观症状、客观检查、评估和诊断以及诊疗计划. 总体看来, 首次病程记录基本按照 SOAP 格式组织病历. 从文本结构形式看, 首次病程记录明显地分为几个章节 (Section), 每一个章节以名称 (Section name) 和冒号表示出来. 每个部分的内容以条目的形式罗列, 总体表现出明显的半结构化形式.

在内容方面, 每一个章节表达的内容都是独立的, 每章节的名称指示了该部分要表达的内容. 主诉部分描述患者此次就诊的主要症状; 病例特点详细描述患者的既往史、症状和体征, 以及与疾病密切相关的辅助检查; 临床初步诊断则直接表示患者可能所患疾病 (如果这部分是待查, 可结合后续病程记录或出院小结获取); 鉴别诊断部分描述的是与患者所患疾病相关但被排除的疾病及其主要区别性症状; 诊疗计划描述施加于患者的治疗措施. 可以看出, 病程记录是按照症状、疾病、检查和治疗这四大要素安排各部分内容. 虽然各部分都表达独立的内容, 但各部分之间存在着密切的联系. 病历特点中描述的症状和体征可以认为是临床初步诊断疾病的可能症状和体征, 辅助检查是为了确证疾病的诊断, 诊疗计划是针对疾病展开的施治措施. 因此, 在对首次病程记录进行信息抽取时, 可以针对每部分的特点, 设计适合该部分的抽取任务和抽取算法, 同时, 结合各部分之间的联系, 便于展开关系抽取的研究.

病程记录是治疗过程的记录, 症状、诊断、检查和治疗可能随着医疗活动的进展而发生改变, 更多的检查和疾病的关系、治疗和疾病的关系、治疗和症状的关系需要从后续的病程记录挖掘. 因此可以考虑把每一段病程记录作为挖掘的文档, 根据时间先后顺序, 整合产生完整的医疗知识.

1.2.2　出院小结结构特点

出院小结是指经治医师对患者此次住院期间诊疗情况的总结. 出院小结内容涵盖诊疗活动的各个方面, 各方面的内容分开描述, 形成明显的结构特征. 以 I2B2 (Informatics for integrating biology and the bedside) 2010 评测中使用的出院小结为例, 与病程记录结构类似, 出院小结各部分按照章节 (Section) 描述, 每一个章节以章节名称 (Section name) 开始, 包含的章节主要有入院日期、出院日期、主要诊断、主要治疗、现病史、既往史、入院药物、过敏史、个人史、体格检查、住院治疗过程、出院医嘱、出院药物等.

国内电子病历出院小结与国外病历中的出院小结在结构上基本相同, 主要包括入院日期、出院日期、入院情况、入院诊断、诊疗经过、出院诊断、出院情况、出院医嘱、医师签名等[28]. 哈尔滨医科大学附属第二医院出院小结基本遵循国家规范, 完整包含了各章节的内容. 以该医院出院小结为例, 诊断部分说明患者确诊的疾病; 入院情况描述患者入院时的各种重要的临床表现以及重要检查结果; 出院情况描述患者出院时各种重要的临床表现以及重要检查结果, 可以和入院情况对比, 分析出不良症状得到治愈或缓解; 诊疗经过部分简单列出了治疗措施; 治疗效果明确表示本次治疗是有效还是无效; 出院医嘱简单列出后续的治疗措施和注意事项. 和病程记录的结构形式类似, 出院小结各章节描述内容单一, 可根据不同章节制定不同的抽取策略; 同时, 每个章节之间的联系可用于关系抽取.

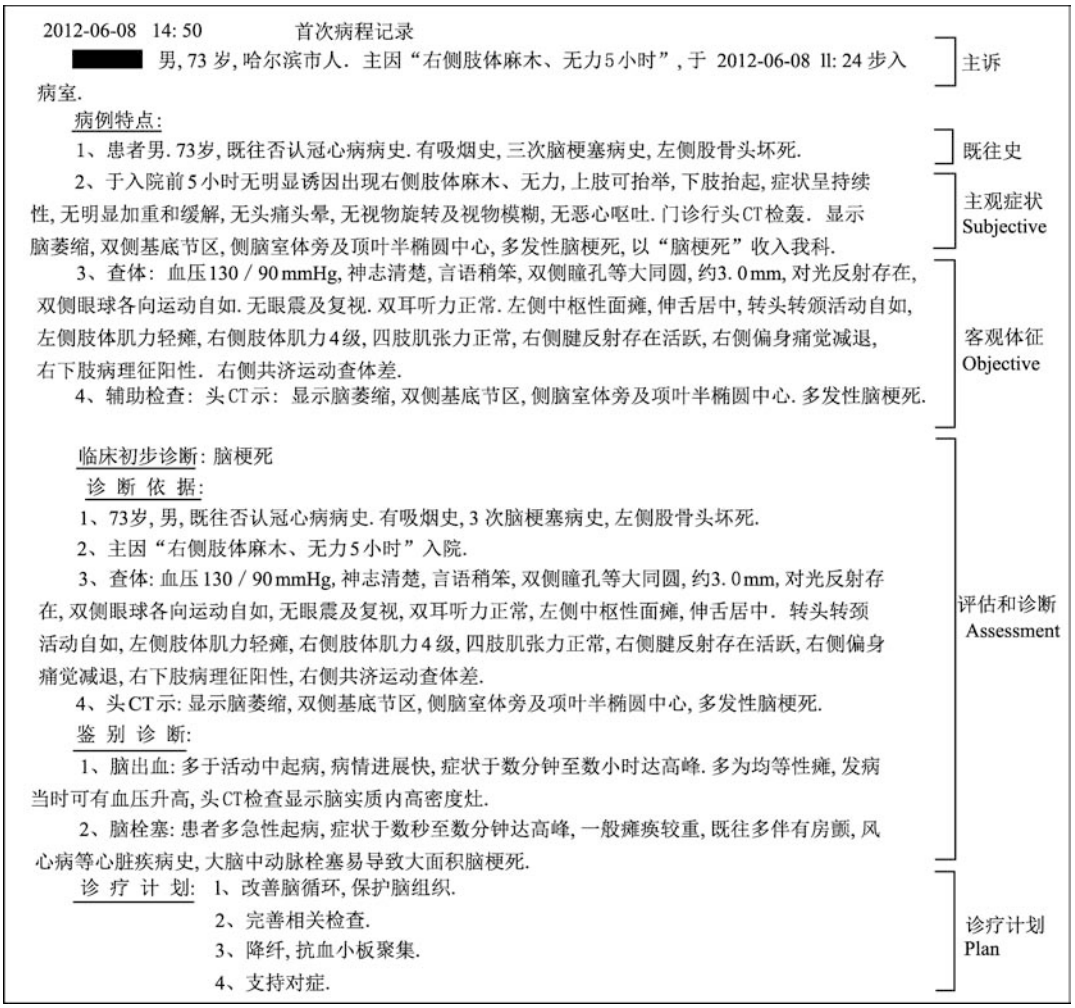

图 5 哈尔滨医科大学附属第二医院首次病程记录

Fig. 5 A progress note from The 2nd Affiliated Hospital of Harbin Medical University

1.3 语言特点

电子病历文本的自然语言处理研究属于生物医学领域的研究, 但电子病历文本与生物医学文献文本存在较大差异. 生物医学文献通常指的是书、论文、摘要等文本内容, 其语言是编辑良好且严格符合语法的, 而病历使用的语言则表现出独特的子语言 (Sublanguage) 特性[29−30], 这些特性包括: 1) 忽略隐含信息导致句子语法成分不完整, 比如缺少动词 (神志清楚, 言语稍笨); 2) 符号在医疗领域的特殊意义 (双下肢肌力 5+ 级); 3) 表达模式化并且不同的模式可能等价; 4) 包含大量术语和受控词汇; 5) 子语言和通用语言混合使用 (“于入院前 5 小时无明显诱因出现右侧肢体麻木”, 这句话就是通用语言的表达方式); 6) 存在固定的语义类型 (比如词的语义类型可能是疾病、症状或者检查等). 电子病历是医务工作者对临床治疗经过的记录, 因此, 电子病历文本必须要求语句精炼并且准确无歧义. 文献 [31] 对英文电子病历的文本特点进行了总结, 中文电子病历文本也具有类似的特点: 1) 包含大量专业术语 (如“共济运动”、“脑梗死”); 2) 医疗行业习惯用语大量出现 (如 “伴”、“否认”、“示”、“尚可”、“未见”); 3) 经常包含一些以数字和单位表示的检查结果 (如“100/70 mmHg”、“3.0 mm”); 4) 包含英文缩写词 (如 “CT”、“MRI”); 5) 句子语法结构不完整, 但模式化较强 (如描述症状是身体部位 + 描述 (“上肢可抬举”, “言语笨拙”), 排除症状是 “否认 – 无” + 描述 (“无发热”)); 6) 为了表达清晰, 用半结构化的方式组织各部分内容.

电子病历的输入方式也对电子病历文本产生较大的影响. 目前电子病历的输入方式主要有两种: 1) 先口授后录入; 2) 医生直接输入到电子病历系统中. 考虑到输入病历文本的效率, 目前国内外大部分电子病历系统的输入方式是第 2 种, 采用基于模板的输入方式或者在一份旧的病历之上修改产生一份新的病历, 可以充分利用拷贝粘贴等快捷手段[32].

这种输入方式虽然高效, 但导致的结果是电子病历文本中重复内容较多, 并且出现信息不一致和不及时的问题[32−33]. 基于计算机辅助的病历智能生成系统是电子病历输入的新趋势[34−35], 实现该目标的基础是对已有电子病历进行分析处理以及信息抽取.

医生的医疗知识融合于描述性的自由文本中, 为计算机自动处理制造了障碍, 因而自然语言处理、信息抽取等相关技术在电子病历的分析和挖掘中将发挥重要的作用. 同时病历文本的半结构化特点和语言特点给自然语言处理技术的应用带来新的挑战和机遇. 挑战主要在于电子病历文本行文风格与开放领域文本或其他领域文本迥然不同, 因此, 已有的基础处理工具如分词、词性标注、句法分析在电子病历文本上都有可能失效. 如果充分利用电子病历文本这些特点, 也会带来一些机遇, 比如将医疗领域大量的知识库用于信息抽取, 较强的模式化表达习惯使得模式挖掘相对容易, 语言精练准确无歧义使处理难度大为降低, 半结构化的组织形式使得信息抽取方法可以灵活地实施, 并且可以利用结构之间的联系进行信息的归纳和推理. 这些特点如果充分利用将非常有利于电子病历文本信息抽取的研究.

2 命名实体识别和实体关系抽取研究概述

2.1 命名实体识别研究概述

命名实体 (Named entity, NE) 最初是在 MUC-6[36] (Message Understanding Conference) 上提出的. 在语言使用中, 命名实体具有独立的意义, 常常作为一个整体出现在语句中. 命名实体识别 (Named entity recognition, NER) 是指识别文本中具有特定意义的实体, 主要包括人、地名、机构名、专有名词等[36], 其扩展任务还包括名实体的单复数识别任务[37]. 从语言分析的全过程来看, 命名实体识别属于词法分析中未登录词识别的范畴. 命名实体识别本质上是一个模式识别任务, 即给定一个句子, 识别句子中实体的边界和实体的类型. 该任务通常把边界信息和类型信息组合成一系列的标记, 那么识别任务就是对句子中的每一个词赋予一个标记. 标记一般用 B_C 和 I_C 的形式给出, 其中 C 是类别标记, B 和 I 是位置标记. B 表示一个实体的开始, I 表示实体的继续. 对于不属于任何实体的词一般用 O 来表示. 该标记不但给出了对应的词语所属的实体类别, 而且还指明了它在实体中的位置. 命名实体识别方法主要是两类, 一类是基于分类的方法对每一个词在多个标记上进行分类, 选择分类概率最大的标记, 另一类方法是基于序列化标注的方法对多个词同时标记选择联合概率最大的标注序列.

[1] http://www.nist.gov/speech/tests/ace/index.htm

序列标注方法在命名实体识别研究中受到较多青睐, 并且在线序列标注方法因其可扩展性和适应性强等优点在自然语言处理研究中逐渐得到发展[38].

自 MUC 提出命名实体识别任务之后, 一系列国际会议均把该任务作为其中一项指定任务, 尤其是自动内容抽取 (Automatic content extraction, ACE) 评测[1] 把该任务作为主要任务, 并对命名实体的定义进行了完善和细化[39], 成为开放领域命名实体识别研究的参考. 命名实体识别研究方法主要有基于规则和词典的方法、基于机器学习的方法. 基于规则和词典的方法是命名实体识别中最早使用的方法, 该方法多采用语言学专家手工构造规则模板, 选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词 (如尾字)、中心词等方法, 以模式和字符串相匹配为主要手段, 这类系统大多依赖于知识库和词典的建立. 王宁等[40] 利用规则的方法进行金融领域的公司名识别, 该系统对知识库的依赖性强, 同时开放和封闭测试的结果也显示了规则方法的局限性. 机器学习方法是从样本数据集合中统计出相关特征和参数, 以此建立识别模型. 目前研究的重点主要是基于机器学习的方法. 命名实体识别任务包括识别实体边界和实体类型, 可以看作是词的分类问题, 因此可以采用基于分类的方法如贝叶斯模型、支持向量机 (Support vector machine, SVM)、最大熵 (Maximum entropy, ME) 等. Lin 等[41] 应用支持向量机在 863 NER 2004 语料上识别中文命名实体取得 95 % 的准确率. 赵健[42] 应用最大熵来完成该任务, 任务中需要识别的目标类别包括中文姓名、地名、机构名、其他专有名词四种, 把每一种类别的命名实体又细分为开始部分、中间部分、结尾部分和整体四种情况, 把不属于以上四种类别中任何一种的词语都归为一类, 这样类别标记集合总共包含 17 ($= 4 \times 4 + 1$) 个标记, 应用最大熵对每个词进行多分类从而实现命名实体识别, 识别结果的 F 值达到 77.87 %. 命名实体识别还可以看作为一个序列化数据的标记问题, 因此可采用隐马尔科夫模型 (Hidden Markov model, HMM)、最大熵马尔科夫模型 (Maximum entropy Markov model, MEMM) 和条件随机域 (Conditional random field, CRF) 等模型. Finkel 等[43] 把命名实体识别视作序列标注问题, 采用 CRF 建立自动标注模型, 考虑的特征主要包括词特征、前后缀、词性序列和词的形态等. 赵健[42] 采用 CRF 模型实现了该任务, 并与隐马尔科夫模型、最大熵马尔科夫模型的实验结果进行了比较, 条件随机域模型在中文命名实体上的性能最好, 平均准确率达到 84.53 %. 值得注意的是, Finkel 等[44] 把命名实体识别和短语结构句法

分析联合起来研究训练一个联合模型, 是多任务学习研究, 结果表明命名实体识别结果和句法分析结果的 F 值均有显著提升. 机器学习方法依赖于大量的标注语料, 只需要少量标注语料的半监督学习方法成为命名实体识别研究的新方向[45]. Ke 等[46] 应用半监督学习协同训练 (Co-training) 算法实现中文组织名的识别, 在少量标注语料的情况下, 结合大量的未标注语料, 协同训练 CRF 模型和 SVM 模型, 取得的 F 值比单个模型的 F 值高出约 10 %. Nadeau[47] 应用半监督学习识别 100 种不同类型的命名实体, 其研究表明少量的标注语料也可以构建高精度的命名实体识别系统. Ando 等[48] 提出一种新的半监督学习方法, 该方法采用多任务学习, 把在标注语料上的命名实体识别任务和在无标注语料上的两个辅助性无监督任务联合起来, 在 CoNLL 2003[2] 数据上取得 F 值为 0.8931, 超过了之前在该数据集上最好的研究结果. Collobert 等[49] 提出一个统一的神经网络框架和学习算法联合解决自然语言处理中的序列标注问题, 包括词性标注、组块分析、命名实体识别和语义角色标注, 是深度学习在自然语言处理中的典型应用, 也是多任务学习的典型范例, 该研究首先基于神经网络语言模型无监督地从无标注文本中学习词的向量表示, 实现命名实体识别时还引入了一个地名词典, 实验在 CoNLL 2003 数据上, F 值达到 0.899, 超过了对比系统[48] 的 F 值.

开放领域命名实体识别研究主要以基于标注语料的有监督机器学习方法为主, 基于标注语料和未标注语料的半监督学习和联合相关任务的多任务学习越来越受关注, 备受瞩目的深度学习框架应用于以命名实体为代表的自然语言处理研究中的优秀表现更加强化了这一趋势.

2.2 实体关系抽取研究概述

实体和实体之间存在着语义关系, 当两个实体出现在同一个句子里时, 上下文环境就决定了两个实体间的语义关系. 如雇员和公司间的雇佣关系、商品和类目之间的类属关系、药品和疾病之间的治疗关系等. 完整的实体关系包括两方面: 关系类型和关系的参数. 关系类型说明了该关系是什么关系, 如雇佣关系、类属关系等; 关系的参数也就是发生关系的实体, 如雇佣关系中的雇员和公司. 关系的参数至少是两个, 两个参数的关系叫二元关系, 两个以上参数的关系是多元关系. 关系有对称关系和非对称关系, 对称关系的参数不考虑参数的顺序, 非对称关系的参数要考虑顺序, 不同的顺序表达不同的关系. 有时候实体关系还会有时间属性, 即实体关系存在的有效期[50]. 大多数关系识别研究只考虑非对称二元关系, 形式化表示为 $R = r(e_1, e_2)$, R 表示实体关系, r 表示关系类型, e_1 表示关系的第 1 个参数, e_2 表示关系的第 2 个参数. 实体关系识别只能发现实体间显式的并且预定义的实体关系, 实体间有些关系并不明显但可以通过其他的显式关系推导出来, 假设生成研究就是为了发现实体间这种隐含关系, 是对关系识别的一个重要补充[51−52]. 实体关系抽取是命名实体识别的后续任务, 同一个句子范围内两个实体间的关系比较明确, 通常大部分研究都只考虑一个句子中的两个实体之间的关系, 而不考虑跨越句子的实体之间的关系[12, 45, 53−54]. 不同领域实体类型的定义不同, 实体间的语义关系也不同. 语义关系的类型取决于实体类型和对应领域的特点, 甚至还取决于抽取目的. 在进行语义关系抽取前, 需要定义好待抽取的语义关系类型, 然后根据两个实体的上下文特征预测实体间概率最大的语义关系, 通常采用分类方法来实现实体间关系的识别[45, 53−54].

开放领域实体关系抽取是 1998 年最后一次 MUC[55] 会议引入的, 随后同实体识别一起转入 ACE[39] 评测中, 2008 年 ACE 评测中的关系抽取任务中包括 7 个大类关系和若干子关系. 关系抽取最初采取的是基于知识库的方法[56], 基于统计的机器学习方法逐渐受到研究者们的青睐[45], 大量有监督机器学习方法能成功地抽取实体关系, 但是需要大量标注语料, 人们的研究目光又转向了只需少量标注语料的半监督机器学习方法[57−59], 甚至是无监督学习方法[60−62]. Zhang 等[63] 将中文实体名和关系识别看作一系列分类问题. 整个过程可被分成两个阶段: 第一阶段是学习过程, 用标注数据训练若干分类器; 第二阶段是抽取过程, 通过使用学习得到的分类器抽取中文实体名和它们的关系. Suchanek 等[64] 应用 kNN 和 SVM 分类方法从 Web 文档中识别实体关系, 该研究基于链接语法[65] 分析待抽取关系的模式, 把模式作为分类的特征, 实验结果证明了深层语法结构在关系识别中的重要性. Brin[66] 和 Agichtein 等[57] 是最早应用半监督学习方法在小样本上迭代地抽取模式进而根据模式识别实体关系. Zhang[59] 提出基于 SVM 的弱监督关系分类系统应用 SVM 算法进行关系抽取, 弱监督学习过程包括两个组件: 一个底层有监督学习器和一个在其上的 Bootstrapping 算法. 宁海燕[67] 对比研究了基于特征的有监督、半监督和无监督的实体关系抽取方法, 有监督学习方法采用的是最大熵和 SVM, 特征词位置信息能有效提高实体关系分类准确率; 半监督学习采用的是 Bootstrapping 方法, 对实体特征和种子集规模进行了对比研究, 研究表明半监督

[2] http://www.cnts.ua.ac.be/conll2003/ner

实体关系抽取能够提高实体关系抽取的准确率; 无监督实体关系抽取方法主要采用的是聚类方法, 主要研究了聚类算法以及合并策略对实体关系抽取的影响, 对比三种聚类算法, 即 K-means、自组织映射和 Affinity propagation 算法, 以及两种合并策略 (DCM 和 Cosine), Affinity propagation 算法能够获得较高的精度, 自组织映射算法在运行时间上更有优势.

在开放领域中, 面向海量网页文档的关系抽取不同于标准数据集上的研究, 由于数据量极其庞大, 蕴含的实体及实体关系也极为丰富, 所以有些研究并不关注实体的类型以及关系的类型; 另一方面, 由于容易获得海量的数据, 基于少量种子的半监督抽取方法可以实现持续不断的学习; 一些百科网站的页面都有固定的格式, 因而基于规则的方法很容易从这些页面抽取出实体及实体关系. 这方面工作的典型代表分别是华盛顿大学开发的 Reverb[3]、卡内基梅隆大学开发的 NELL (Never-Ending Language Learner)[4] 和德国马克斯普朗克研究所构建的 YAGO (Yet Another Great Ontology)[5]. 华盛顿大学的 Anthony Fader[68] 等开发的 Reverb 系统旨在从海量开放域文本中抽取对象间的关系, 不考虑关系的类型, 该系统以词性标注和组块分析处理后的语句作为输入, 基于句法约束和词汇约束抽取包含动词的关系短语, 然后识别关系短语最左和最右的名词短语作为抽取的候选关系对, 最后用一个逻辑斯蒂回归模型预测候选关系对的置信度, 实验结果优于同类方法. 目前 Reverb 已经从网页中抽取到约 500 万个关系. 卡内基梅隆大学 Carlson 等[69] 构造了一个 "永不停止的" 语言学习框架 NELL, 持续不断地从网页中抽取指定类别的实例间关系, 并对学习算法不断优化; 该框架以一个小规模本体 (包含少量的类别及实例、类别关系及关系实例) 为种子, 采用四种不同的方法分别从语句中、表格和列表中、名词短语中以及已经抽取到的关系实例中抽取类别的实例及实例间关系, 原型系统运行 67 天, 抽取到的实例及实例间关系共 242 000 个, 准确率达到 0.74. 目前 NELL 已经抽取到高置信度的实例及实例间关系数量超过 200 万. 德国马克斯普朗克研究所的 Suchanek 等[70] 基于 Wikipedia[6] 和 WordNet[7] 构建了一个轻量级可扩展的开放领域本体 YAGO, 该系统联合规则方法和启发式方法从维基百科数据中抽取实体和实体关系, 并增加到 WordNet 中, 准确率达到 0.95; YAGO2 s[71] 是其重构版本, 抽取组件分为 30 个独立的模块, 能并行地完成抽取、检查、演绎和合并等操作, 目前 YAGO2 s 包含约 1 千万个实体和 1 亿个实体关系.

实体关系抽取研究目前主要关注二元关系的抽取, 研究方法主要以基于分类的方法为主, 半监督学习方法能有效利用未标注语料因而受到较多关注. 句法结构以及实体上下文特征对实体关系有重要指示作用, 因而高精度句法分析对实体关系抽取有着重要影响. 在大数据浪潮下, 面向海量文档的关系抽取研究已受到更多的关注, 研究方法可以因抽取目的不同而各异, 但是由于大数据特点, 实现方法的一个共同点就是要能持续不断的学习, 这些研究为开放领域知识图谱的构建奠定了坚实的基础.

2.3 命名实体识别和实体关系抽取的跨领域研究

命名实体识别和实体关系抽取是信息抽取中的重要任务, 对信息抽取技术的研究与应用都有重要的意义, 同时它也是篇章理解的一个核心技术, 对信息检索、问答系统、信息过滤、机器翻译等有非常积极的意义. 正因为这两个任务的重要性, 很多研究者在不同的领域都展开了研究. 生物医学文献和临床电子病历的主要特征是未登录词数量庞大、文本充斥着大量命名实体、新的命名实体不断出现、很多命名实体拥有多个不同的书写形式、命名实体以及实体间关系承载着丰富的专业知识. 因此生物医学领域和临床医疗领域是命名实体识别和实体关系抽取研究的两个重要领域. 生物医学领域的研究主要从大量生物医学文献中识别生物医学实体 (如基因名字、蛋白质名字、疾病名字等) 及实体关系, 该领域应用最广泛的评测数据集有 GENIA 语料库[72] 和 GENETAG 语料库[73], 主要有两个评测会议: JNLPABA (Joint Workshop on Natural Language Processing in Biomedicine and Its Applications)[74] 和 BioCreAtIve (Critical Assessment of Information Extraction Systems in Biology)[75]. 针对电子病历实体识别和关系识别的系统研究是 I2B2 在 2010 年的公开评测任务[8] 中发起的, 主要识别医疗问题、治疗和检查三类实体及其关系, 这次评测对电子病历实体类型和实体关系进行了系统的定义, 并发布了基于真实电子病历的评测数据[12]. 本文的主要工作就是对该领域命名实体识别和实体关系抽取研究进行调查和综述.

实体识别模型从一个领域移植到另一个领域通

[3] http://reverb.cs.washington.edu/
[4] http://rtw.ml.cmu.edu/rtw/
[5] http://www.mpi-inf.mpg.de/yago-naga/yago/
[6] http://www.wikipedia.org
[7] http://wordnet.princeton.edu/
[8] https://www.i2b2.org/NLP/Relations/

常会遇到严重的性能下降问题, 主要原因在于命名实体的类型定义不同、不同领域有着不同的语言特点[76]. 尽管不同领域间的差异较大, 但研究思路基本相同, 一般采用基于规则、基于词典或者机器学习方法. 在应用机器学习方法时, 一般把命名实体识别和实体关系抽取转化为分类问题或序列标注问题, 采用有监督学习[45]、半监督学习[57−59] 或针对特定实体的无监督学习方法[60−62]. 实践表明, 基于机器学习的方法的效果要好于基于字典和基于规则的方法, 因为它可以利用更多的特征, 甚至可以把基于字典和基于规则的方法的结果也作为特征. 因而越来越多的研究者开始尝试使用机器学习的方法研究实体识别和实体关系抽取. 基于机器学习的方法依赖于大量的标注语料, 而且标注语料也是评价识别效果的黄金标准. 因此当在一个新的领域进行实体识别和实体关系识别研究时, 首先面临的问题就是如何以小的成本获得足够支持机器学习模型训练和评价的标注语料. 目前解决这个问题的思路有两个: 1) 探索新的标注模式[77] 和半自动语料标注方法, 如应用主动学习 (Active learning) 减小标注成本[78−80]; 2) 探索基于少量标注语料的机器学习模型, 如应用半监督学习 (Semi-supervised learning) 模型减小标注训练语料的数量[81−83], 探索基于部分标注语料 (Learning from partial annotations) 的机器学习模型[84−87] 或基于迁移学习 (Transfer learning) 的机器学习模型[88−89].

3 电子病历命名实体识别

3.1 电子病历命名实体识别任务

电子病历文本中命名实体主要涉及患者接受医疗诊治的记录中表示特定意义的实体, 如疾病名、症状、药品名、检查名、医疗手段等. 下面几个示例列举了可能的实体.

1) 门诊以脑梗死、皮质下动脉硬化性脑病收入我科 (“脑梗死” 和 “皮质下动脉硬化性脑病” 是疾病);

2) 患者 1 年前开始出现记忆力减退 (“记忆力减退” 是症状);

3) 患者彩超结果汇报轻度脂肪肝 (“彩超” 是检查, “脂肪肝” 是疾病);

4) 糖尿病皮下注射胰岛素控制 (“胰岛素” 是药物, 属于医疗手段).

不同的研究对电子病历命名实体的定义各有不同, 主要体现在实体类型的粒度上. 具有代表性电子病历命名实体定义是 I2B2 2010 评测任务中给出的. 该评测首次对电子病历命名实体进行了系统的分类[12], 该分类依据参照 UMLS[90] 定义的语义类型, 把命名实体分为三类: 医疗问题 (包括疾病和症状)、治疗、检查. 这种分类充分体现了面向问题的思想, 医疗手段是为了治疗医疗问题, 检查是为了确认医疗问题. 这三类实体的定义以及对应的 UMLS 语义类型见表 1 所示.

电子病历命名实体识别任务就是自动识别病历文本中的在医疗上表达独立意义的各类命名实体, 该任务包含两方面的工作: 1) 识别命名实体的边界; 2) 确定命名实体的类型, 实体间不重叠不嵌套. 如前所述, 该任务通常把边界信息 (B, I, O) 和类型信息 (Problem, treatment, test) 组合成一系列的标记, 识别任务转化成对每个词赋予一个标记. 该任务涉及到的可能的标记一共有 $7 (= 2 \times 3 + 1)$ 个. 该任务的评价不仅要考虑识别命名实体的边界, 还要考虑命名实体的类型, 评价指标采用精确度、召回率和 F 值.

考虑到电子病历命名实体识别研究起步较晚, I2B2 2010 定义的实体类型的粒度较大, 这些识别的实体如果用于后续研究的话, 可能需要对实体类型继续细分, 比如医疗问题可能需要进一步分为疾病和症状, 治疗也可能需要进一步分为药物和处置. 事实上, Uzuner 在 I2B2 2010 评测之前的研究中就是把医疗问题拆分为疾病和症状[3]. 在这个分类中, 疾病的定义主要对应于 UMLS 语义类型中的疾病名、综合症等, 简单地说, 疾病就是医生作出的诊断; 症状的定义主要对应于 UMLS 语义类型中的症状和体征, 也就是疾病引起的各种不适或异常的表现. 由此可看出, 疾病和症状并不是并列的, 症状是由疾病导致的, 同时, 不同的疾病也可能有相同的症状. 中文电子病历有其自身的特点, 病程记录 (如图 5 所示) 中病例特点部分和出院小结中入院时情况部分主要描述患者的症状和体征, 这些症状和诊断疾病存在着对应关系. 同时, 中文电子病历对患者的症状描述非常详实, 并且按照患者自诉症状和医生检查体征分开描述. 由于电子病历半结构化特点, 我们可以制定有针对性的策略抽取患者的症状. 因此, 在我们的中文电子病历命名实体识别研究中, 我们把医疗问题拆分为疾病和症状, 把实体类型定义为疾病、症状、检查和治疗四个主要类型, 并且把症状再细分为自诉症状、检查结果两个子类型.

3.2 电子病历命名实体识别方法

早在 1968 年, Weed[24] 提出面向问题组织电子病历就是为了医务人员便于诊断推理, 基于电子病历文本的临床决策支持研究倍受关注[4], 该研究首先需要应用自然语言处理、信息抽取等技术对电子病历文本进行处理, 识别文本中实体和实体关系. 电子病历实体抽取一般采用基于词典和规则的方法或机器学习方法.

表 1 电子病历实体类型

Table 1 Entity types of EMR

实体类型	类型的定义	对应 UMLS 的语义类型
医疗问题 (Problem)	医生给出的诊断或名称是 ICD-10 定义的术语或由疾病导致的不适表现或异常表现	Pathologic functions (病理功能), disease or syndrome (疾病或综合症状), mental or behavioral dysfunction (精神或行为障碍), cell or molecular dysfunction (细胞或分子功能障碍), congenital abnormality (先天性畸形), acquired abnormality (获得性异常), injury or poisoning (受伤或中毒), anatomic abnormality (解剖异常), neoplastic process (肿瘤进程), virus/bacterium (病毒/细菌), sign or symptom (体征或症状)
检查 (Test)	为了发现、证实医疗问题而施加给患者的检查过程、仪器等	Laboratory procedure (化验程序), diagnostic procedure (诊断程序)
治疗 (Treatment)	为了解决医疗问题而施加给病人的治疗程序、干预措施、给予物品	Herapeutic or preventive procedure (治疗或预防过程), medical device (医疗器械), steroid (类固醇), pharmacologic substance (药物物质), biomedical or dental material (生物医学或牙科材料), antibiotic (抗生素), clinical drug (临床用药), drug delivery device (药物输送设备)

3.2.1 基于词典和规则的方法

在医疗领域, 大量的词典以受控术语的形式得到维护, 最著名的受控术语词典包括 ICD-10[9]、UMLS 和 SNOMED CT[10], 因而早期电子病历命名实体识别多采用基于词典的方法, 并形成了医疗领域三个代表性的通用工具, 即 MedLEE、MedKAT 和 cTAKES. 1994 年 Friedman 等[91] 开发了最早的 MedLEE (Medical language extraction and encoding) 系统, 基于词汇和语法的规则识别 X 射线报告里的疾病名和疾病的修饰成分, 并对疾病名规范化后, 与医疗实体词典 MED (Medical entities dictionary)[11] 建立映射, 对疾病进行编码. 基于词典识别命名实体的两个著名的系统是 IBM 的 MedKAT (Medical knowledge analysis tool)[92] 和梅奥诊所的 cTAKES (Clinical text analysis and knowledge extraction system)[93], 这两个系统均基于 UIMA[94] 框架, 对病历预处理基本相同, 包括句子边界识别、词性标注、浅层句法分析. MedKAT 利用癌症 ICD (ICD-O[12]) 识别癌症病历中的疾病概念, 并提出癌症疾病知识表示模型 (Cancer disease knowledge representation model, CDKRM) 表示抽取的疾病以及疾病关系的知识; cTAKES 利用 UMLS、SNOMED CT 和 RxNORM[13] 识别病历中的概念并编码. 由于电子病历里文本包含大量专业术语, 因而词典是非常重要的资源, 单纯基于词典的方法难以适应复杂的语言现象, 所以基于机器学习的方法受到更多的关注, 并融合各种术语词典, 或是把上述三个系统的处理结果作为机器学习方法的特征.

3.2.2 基于机器学习的方法

命名实体识别任务可以转化为对文本中的词赋

[9] http://www.who.int/classifications/icd/en/
[10] http://www.ihtsdo.org/snomed-ct/
[11] http://med.dmi.columbia.edu/
[12] http://www.who.int/classifications/icd/adaptations/oncology
[13] http://www.nlm.nih.gov/research/umls/rxnorm/

予给定类型的标记, 因而, 基于机器学习的方法主要是基于分类的方法和基于序列标注的方法, 基于序列标注的方法能联合考虑相邻词的标注结果, 因而得到较多关注. 叶枫等[95] 采用条件随机场模型对疾病、临床症状、手术操作 3 类中文病历中常见的命名实体进行智能识别, 融合人工整理的词典辅助抽取特征, 该研究是首次在中文电子病历中的命名实体识别研究, 但是对实体类型的定义不完整, 语料规模较小, 并且语料只包含了现病史和既往史. Li 等[96] 对比了 CRF 和 SVM 在电子病历命名实体识别中的性能, 并结合 SNOMED, 对比结果显示 CRF 识别结果的 F 值为 0.86, 而 SVM 只有 0.64. Uzuner[12] 概述了 I2B2 2010 评测中各小组实现电子病历命名实体抽取的方法, 大部分采用 CRF 模型, 并结合 UMLS 作为知识库, 辅助抽取特征, 也有少数采用 SVM 的对比研究. Jiang 等[97] 分别采用 CRF (CRF ++) 和 SVM (Tiny SVM) 实现了命名实体识别, 引入的特征主要有三类: 上下文特征、UMLS 和 NLP 系统 (如 MedLEE) 的处理结果以及章节信息 (如 Section name), 并结合手工整理的规则进行后处理, CRF 取得的最好结果是 F 值为 0.8475, SVM 取得的最好结果是 F 值为 0.8326, 在该研究中 CRF 模型优于 SVM 模型. Jonnalagadda 等[98] 采用了一个半监督 CRF 模型识别命名实体, 半监督体现在该模型应用分布式语义 (Distributional semantics) 方法从未标注语料中利用向量空间模型表示词的上下文特征并计算语义相似度, 把与待识别的对象有相似的上下文的词作为特征词, 该方法的 F 值达到 0.823. de Bruijn 等[99] 采用半马尔科夫模型 (一种隐马尔科夫模型) 对词进行序列化标注实现命名实体的识别, 用到的标记只有四类 (Outside, problem, treatment, test), 引入的特征主要有上下文特征和 UMLS 以及 cTAKES 的结果, 并采用 Self-training 方法扩大训练语料, 取得最好结果是 F 值为 0.8523, 词典资源和高维度特征对系统性能提升有重要贡献. de Bruijn 的研究在评测中取得了最好成绩.

表 2 对上述电子病历命名实体识别研究方法进行了总结. 由于电子病历具有包含大量术语这一显著特点, 以 UMLS 为代表的词典资源以及 MedLEE 和 cTAKES 等基于词典的 NLP 工具发挥重要作用, 因此词典资源的构建工作是一项重要的基础工作; 同时, 以 CRF 为代表的序列标注方法成为主流, 并且半监督方法能充分利用大量未标注的语料也开始受到关注. 从 I2B2 2010 最好结果看, 命名实体识别的精度达到比较高的程度, 但还有较大的提升空间, 病历文本中出现的大量缩写术语是目前最大的挑战. 另一方面, 针对病历文本的词法和句法分析技术还没有得到深入研究, 大部分工作仍然是简单利用现有工具, 这方面研究的开展也将极大地促进电子病历命名实体识别研究. 目前没有调研到应用多任务学习方法研究电子病历命名实体识别.

3.3 疾病和症状的修饰识别任务

电子病历命名实体识别还有一个额外的任务, 就是识别疾病和症状的修饰成分, 包括当前的、否认、既往史、非患者本人、待证实等. 疾病和症状的修饰是用来说明疾病或者症状与患者的关系, 表明疾病或者症状是发生于患者本人还是患者亲属, 是肯定发生还是可能发生等信息, 这些修饰信息对患者的健康状况度量有重要影响. 修饰信息识别是电子病历文本实体信息抽取中的一个独特研究. 如下面的例子:

1) 双侧眼球运动自如, 无眼震 (“无” 修饰症状

表 2 电子病历命名实体识别方法总结

Table 2 Summarization of methods for named entity recognition of EMR

作者	方法	用到的资源	数据	评价 (F 值)
Friedman 等[91] (MedLEE)	词典和规则	语法规则、MED	射线报告	0.7
Coden 等[92] (MedKAT)	词典	ICD-O	结肠癌病理报告	0.82
Savova 等[93] (cTAKES)	词典	UMLS、SNOMED	梅奥诊所电子病历[100]	0.715
Li 等[96]	SVM	SNOMED	梅奥诊所电子病历	0.64
Jiang 等[97]	SVM	UMLS、MedLEE	I2B2 2010 评测数据	0.8326
叶枫等[95]	CRF	自建词典	现病史和既往史、自建语料	0.9506
Li 等[96]	CRF	SNOMED	梅奥诊所电子病历	0.86
Jiang 等[97]	CRF	UMLS、MedLEE	I2B2 2010 评测数据	0.8475
Jonnalagadda 等[98]	半监督 + CRF	/	I2B2 2010 评测数据	0.823
de Bruijn 等[99]	半监督 + HMM	UMLS、cTAKES	I2B2 2010 评测数据	0.8523

“眼震”, 表示否认症状);

2) 脑梗死病史 10 年 (“脑梗死” 是以前发病的, 表示是既往病史);

3) 临床初步诊断: 脑出血 (“临床初步诊断” 表示疾病还需进一步确认, 是可能的).

该任务在命名实体识别基础上展开, 对识别出来的医疗问题 (包括疾病和症状) 在表 3 给出的 6 个修饰类型上进行分类, 每个医疗问题只属于一个分类. 因此, 该任务的实现通常采用基于分类的方法实现, 评价指标采用精确度、召回率和 F 值.

表 3 医疗问题的修饰及其意义

Table 3 Assertion types of medical problem

疾病的修饰	意义
当前的	患者正在遭受某种疾病或正表现出某种症状
不存在的	患者否认某种疾病、症状, 或未观察到某种症状, 或者疾病是既往史
非患者本人的	患者亲属或配偶的
有条件的	该疾病是否是在特定条件下才会发生
可能的	根据当前症状做出可能的疾病诊断
待证实的	该疾病是以后可能会发生的

自 Chapman 首次识别疾病是现在存在的 (Present) 还是过去存在的 (Absent)[13], 该项研究受到了持续关注. I2B2 2006 年评测[101] 和 2008 年评测[102] 均开展了此项研究任务. I2B2 2010 评测对需要识别的修饰进行了归纳和重新定义[12], 如表 3 所示. 从表 3 可看出, 任务新增了有条件的和可能的两种修饰, 并且把既往史、未观察到症状和患者否认某种症状这三类修饰合并为一个. 中文电子病历中, 疾病和症状的修饰类型基本可以参照表 3 来设计.

3.4 疾病和症状的修饰识别方法

疾病和症状的修饰识别研究早期多采用基于规则的方法, 后期多采用基于机器学习的方法或者机器学习和规则相结合的方法, SVM 受到较多关注.

3.4.1 基于规则的方法

比较典型的针对疾病和症状的修饰识别始于 Chapman 等[13] 的工作, 该研究使用正则表达式识别某种疾病是当前发生的 (Present) 还是当前未发生的 (Absent), 他们提出的 NegEx 算法在 1 000 多句病历的 1 235 种疾病相关的概念中取得了 0.945 的准确率和 0.778 的召回率. Aronow 等[103] 在对病历文本的分类研究中实现了基于规则 “否定” 识别 NegExpander, 该工具识别带有 “否定” 修饰的名词短语和连词, 并把 “否定” 修饰扩展到连词连接的其他名词短语. 在 I2B2 2010 评测数据上, Goryachev 等[104] 专门针对 “否定” 或者 “当前未发生的” 修饰识别对比了 NegEx、NegExpander 以及 SVM 和朴素贝叶斯方法, 识别结果的 F 值分别为 0.894、0.912、0.8595 和 0.7817, 从该结果可看出, 在识别 “否定” 修饰任务上, 基于规则的方法要优于基于分类的方法. Negfinder 除了使用规则以外引入了句法分析的特征来识别 “否定” 修饰的概念, 获取了 0.977 的准确率和 0.953 的召回率[105]. 针对 I2B2 2010 “否定” 的识别, Sohn 等[106] 构建了 DepNeg, 在小规模样本上构造动词短语的依存模式, 然后根据依存模式抽取依存路径, 基于依存路径识别医疗问题的 “否定” 修饰, 获得的最好精确度达到 0.9884. Harkema 等[107] 基于 NegEx 提出了 ConText 算法, 从电子病历文本中识别疾病的修饰, 这些修饰包括否认、既往史、非患者本人、待证实. 研究表明, 疾病修饰的分布变化很大, 算法在不同风格的文本上表现不好, 这一问题在 I2B2 2010 评测中也有所表现[12].

3.4.2 基于机器学习的方法

基于机器学习的方法通常把该任务视为分类问题, 并把词典和规则系统的结果作为分类特征. 在 I2B2 2010 评测之前, Uzuner 等[108] 对医疗问题的修饰识别分别采用两类方法进行了比较研究: 1) 基于规则的方法实现了一个扩展的 NegEx 系统 ENegEx, 增加了一些规则识别 “非患者本人” 的修饰; 2) 基于机器学习的方法采用了 SVM, 用到的特征主要有上下文特征 (上下文窗口为医疗问题前后 4 个词)、句法特征、医疗问题所在的章节的标题. 通过对比试验, 两类方法均能取得较好的结果, 并且都能用于不同机构的电子病历文本, 基于分类的方法要优于基于规则的方法, 主要原因是由于基于分类的方法能充分利用医疗问题所处的上下文信息. I2B2 2010 评测中几个最有效的方法都使用 SVM 构建分类器[97, 99, 109−110], 同时, 结合上下文信息和一些表示否定、不确定以及家族史的词典融入分类器中[110], 或者把规则系统的结果作为分类特征[97, 109]. de Bruijn 采用两阶段分类方法取得最优结果: 在第一阶段, 使用 SVM 对概念的每个词进行预测; 在第二阶段, 使用第一阶段的分类结果对概念整体进行分类[99]. Clark 等联合使用 CRF 和最大熵模型以及状态规则取得了很好的结果, 平均 F 值达到 0.9343, 在该方法中, 使用 CRF 识别概念的提

示范围, 结合概念状态规则, 构建最大熵分类器, 实现对修饰的分类[111].

表 4 对上述疾病和症状的修饰识别方法进行了总结. 由于病历文本独特的语言现象, "否定" 修饰大量出现, 基于规则的方法能有效识别 "否定" 修饰, 但对其他的修饰识别能力有限, 这也是基于规则的方法的不足; 基于机器学习的分类方法, 除了整合规则外, 还可以充分考虑上下文特征、句法特征以及病历文本的章节名称等结构化信息, 能有效的识别各种类型的修饰, 被证明是有效的方法. 所以, 机器学习和规则相结合的方法是修饰识别研究的重点, 同时为了充分发挥机器学习方法的优势, 应加强病历文本的词法和句法分析的研究.

4 电子病历实体关系抽取

4.1 电子病历实体关系抽取任务

电子病历命名实体关系抽取主要研究从电子病历中抽取疾病、症状、检查和治疗这几类实体间的关系. 这些实体关系体现了患者健康状况信息和针对患者的医疗处置措施, 也体现了医生的专业知识. 如下面的例子:

1) 头 CT 检查显示腔隙性脑梗死 (检查 "头 CT" 证实了疾病 "腔隙性脑梗死");

2) 患者彩超结果汇报轻度脂肪肝、慢性胆囊炎, 给予饮食指导, 继续治疗方案 ("彩超" 证实了 "轻度脂肪肝" 和 "慢性胆囊炎", "饮食指导" 施加于 "轻度脂肪肝" 和 "慢性胆囊炎").

电子病历实体关系抽取任务在命名实体识别基础上展开, 对病历文本中同一个语句中的两个命名实体赋予预定义的关系类型, 因而该任务转化为分类问题, 通常采用基于机器学习的方法实现, 评价指标采用精确度、召回率和 F 值. 目前电子病历实体关系只限于一个句子范围内两个实体之间的关系.

Uzuner 首先对医疗实体关系抽取进行了开创性的研究, 详细定义了六大类医疗实体关系: 当前疾病和治疗的关系、可能的疾病和治疗的关系、疾病 (包括当前的和可能的) 和检查的关系、疾病和症状的关系、当前症状和治疗的关系、可能的症状和治疗的关系[3]. 如果已经定义了修饰识别任务, 实现了疾病和症状的修饰识别, 那么在关系抽取时, 可以不考虑修饰的影响, 直接抽取实体间的关系, 然后借助实体的修饰, 可以得到文献 [3] 定义的上述各类关系. 所以, 在 I2B2 2010 评测中, 实体关系的定义没有考虑修饰的因素. I2B2 2010 首次对电子病历命名实体关系进行了系统的分类[12], 这些关系包括医疗问题和医疗问题之间的关系、医疗问题和检查之间的关系、医疗问题和治疗之间的关系. 这三类关系以医疗问题为中心, 反映了电子病历面向医疗问题的信息组织方式. 这三类关系只限于一个句子范围内两个实体之间的关系. 表 5 详细列出了医疗问题、检查和治疗这三类实体间的关系.

针对中文电子病历特点, 医疗问题被拆分为疾病和症状, 那么在定义实体关系时, 也应作相应调整. 主要体现在两方面: 1) 医疗问题和治疗 (或检查) 的关系转变为疾病和治疗的关系以及症状与治疗的关系, 医疗问题和检查的关系也转变为疾病和检查的关系以及症状和检查的关系; 2) 医疗问题之间的关系替换为疾病和症状的关系 (疾病导致了症状)、疾病和疾病的关系 (疾病导致了另一个疾病)、症状和症状的关系 (症状伴随另一个症状).

自动抽取这几类实体间的关系可以构造患者健康状况的简明摘要, 医生可以预先快速浏览病人的信息, 后续再关注特定的细节. 除了可以用作医疗研

表 4 疾病和症状的修饰识别方法总结

Table 4 Summarization of methods for assertion classification

作者	方法	用到的资源	数据	评价 (F 值)
Chapman 等[13] (NegEx)	规则	正则表达式规则	出院小结	0.853
Mutalik 等[105] (Negfinder)	规则	正则表达式规则、句法规则	自建语料	0.965
Sohn 等[106] (DepNeg)	规则	依存规则	I2B2 2010 评测数据	0.838
Harkema 等[107] (ConText)	规则	正则表达式规则、触发词	6 种类型的病历文本	0.76 ~ 0.93
Uzuner 等[108]	SVM	/	三个机构的病历	0.35 ~ 0.98
Grouin 等[110]	SVM	NegEx	I2B2 2010 评测数据	0.931
Jiang 等[97]	SVM	MedLEE	I2B2 2010 评测数据	0.931
de Bruijn 等[99]	SVM	cTAKES	I2B2 2010 评测数据	0.936
Clark 等[111]	CRF、最大熵	语义分类词典、状态规则	I2B2 2010 评测数据	0.934

表 5　医疗问题、检查和治疗这三类实体间的关系

Table 5　Relations among entities of medical problems, tests and treatments

关系大类	具体关系	关系的意义
医疗问题和治疗之间的关系	治疗改善了医疗问题 (TrIP)	治疗改善或治愈了医疗问题
	治疗恶化了医疗问题 (TrWP)	治疗没有改善也没有治愈医疗问题, 或者恶化了医疗问题
	治疗导致了医疗问题 (TrCP)	治疗不是针对该医疗问题的, 而是导致了该医疗问题
	治疗施加于医疗问题 (TrAP)	治疗是施加于该医疗问题的, 是结果没有提及
	因为医疗问题而没有采取治疗 (TrNAP)	因为该种医疗问题而不能采取治疗
	治疗和医疗问题间不存在上述关系	/
医疗问题和检查之间的关系	检查证实了医疗问题 (TeRP)	检查结果证实 (否认) 了该医疗问题
	为了证实医疗问题而采取检查 (TeCP)	为了证实医疗问题而采取检查, 但结果未知
	检查和医疗问题间不存在上述关系	/
医疗问题和医疗问题之间的关系	医疗问题导致另一个医疗问题 (PIP)	医疗问题是另一个医疗问题的不同表现或者医疗问题导致了另一个医疗问题
	医疗问题和医疗问题间不存在上述关系	/

究, 关系抽取还可用作研究受不同疾病折磨病人之间的共性、发现药物的禁忌症或提供药物性能的客观的度量. 同时, 以医疗问题为中心, 我们可以把电子病历中抽取出来的实体按照实体关系组织起来, 系统地表示以医疗问题、治疗和检查为主体的医疗知识. 图 6 是医疗知识的表示结构. 在图 6 中, 肺炎是患者的现病史, 发病原因在上下文中未提及, 与肺炎相关的医疗问题 (症状) 是呼吸窘迫, 通过胸部 X 光检查结果证实了肺炎, 经过抗生素治疗取得有效的治疗结果.

图 6　以医疗问题、治疗和检查为主体的医疗知识

Fig. 6　Medical knowledge represented by medical problems, tests and treatments

4.2　电子病历实体关系抽取方法

电子病历实体关系抽取的研究方法主要采用基于机器学习的分类方法, 分类模型一般采用 SVM、最大熵等.

Frunza 等[112] 在 Medline[14] 摘要数据上研究了疾病和治疗之间的三种关系 (治愈、抑制、导致副作用) 的识别, 对比了三种分类方式, 其中构造一个分类模型实现三个关系类型上的分类的方式是最好的, 分类模型以朴素贝叶斯和 SVM 为主, 特征的选取上除了词特征和短语特征外, 还引入了 UMLS 语义类型, 取得最好的分类结果在治愈、抑制和导致副作用这三种类型上分别是 0.9855、1.0 和 0.8889. 在 I2B2 2010 评测之前的关系抽取预先研究中, Uzuner 等[3] 以句子为单位识别电子病历实体关系, 实体关系分成 6 个大类, 训练 6 个 SVM 分类器实现疾病、症状、检查和治疗之间的关系识别, 分类结果在各关系子类上单独评价, F 值介于 0.62 和 0.89 之间, 用到的特征主要包括实体在句子中的顺序和距离、词汇特征 (包括组成实体的词、出现在实体上下文中的动词等) 以及链接语法 (Link grammar)[65] 分析结果, 对比结果表明词汇特征能发挥重要作用. 该研究为电子病历实体关系抽取研究提供了重要借鉴. Rink 等[113] 采用 SVM 在 I2B2 2010 评测数据上识别电子病历中实体间预定义的语义关系, 病历文本使用 GENIA[15] 预处理, 抽取的特征主要包括词汇特征、上下文特征以及上下文文本之间的相似度 (采用

[14] http://medline.cos.com/
[15] http://www.nactem.ac.uk/GENIA/tagger/

编辑距离算法计算), 并使用 Wikipedia、WordNet 和 General inquirer[114] 辅助特征抽取, 取得评测中该任务最好结果 F 值为 0.737. 该研究还对比分析了特征对分类模型的影响, 结果表明词典和上下文特征在关系识别中发挥了重要作用: 在没有词汇特征和上下文特征情况下的抽取结果的 F 值下降了 3.7 %, 同时在没有上下文文本相似度特征情况下 F 值下降了 1.1 %. 但是有些实体所处的上下文信息不够丰富, 上下文特征缺乏. 针对该问题, Demner-Fushman 尝试使用 UMLS 中概念间关系作为分类特征解决上下文信息不足的问题[109], 取得结果的 F 值为 0.666. de Bruijn[99] 采用最大熵作为分类模型研究关系抽取, 利用 cTAKES 对病历文本进行依存分析, 抽取上下文特征, 特征还包括 UMLS 中概念语义类型和语义关系, 该研究对比研究了有监督分类和基于 Self-training 的半监督分类在关系识别中的表现, 结果表明 UMLS 和依存句法分析结果以及未标记数据对关系识别有显著影响, 取得结果的 F 值为 0.731, 略低于最好成绩. 电子病历中有些类型的关系远多于其他类型的关系, 同时大部分实体间不存在语义关系, 针对这种不平衡的现状, Ryan[115] 分析了电子病历中实体关系的偏置特性, 提出了一种节省标注语料的半监督实体关系识别方法, 该方法采用 SVM 作为分类器, 用到的辅助资源主要有 UMLS、链接语法分析器, 对标注的样本先进行预测, 把置信度低的样本加入到训练集中, 这种方式类似于主动学习, 比一般的半监督学习节省 10 % 的标注语料. 除了采用分类方法识别关系, 计算实体的共现也是发现实体关系的有效手段. Wang 等[116] 从电子病历中计算疾病和症状的共现来发现两者间的关联关系, 因为关系类型简单, 抽取结果 F 值达到 0.91. Chen 等[117] 从医学文献和电子病历中计算疾病和药品实体的共现来发现两者间的关联关系, 获取疾病和药品的潜在医疗知识, 抽取结果的准确率较高. 另外, Roberts 等[118] 在临床信息抽取系统 CLEF 中采用 SVM 在肿瘤患者病历上实现了医疗命名实体关系的识别, 对比研究了语句内和跨语句范围的关系识别, 结果表明适用于语句内实体关系识别的方法在跨语句范围的实体关系识别上准确率较低.

表 6 对上述电子病历实体关系抽取方法进行了总结. 电子病历实体关系抽取的一个最大特色是医疗领域的词典发挥了重要作用, 这主要是因为电子病历文本简短精炼, 上下文信息不充分, 而体现医疗知识的实体关系可以部分地从词典得到; 同时, 上下文特征的提取依赖于病历文本的句法分析, 因而大部分研究都对病历文本进行了句法分析的预处理, 对比研究表明, 词典知识和句法分析能有效提升电子病历实体关系抽取的性能; 基于半监督的方法能充分利用大量未标注的语料, 并且能够一定程度克服实体关系分布不平衡的问题, 在电子病历实体关系抽取中已初步表现出其优势; 如果关系抽取任务只关注简单的关系, 那么基于共现的方法在大规模数据上就能取得很好的效果. 电子病历实体关系抽取除了显式关系抽取研究外, 基于大量已抽取的命名实体和实体关系的隐含关系发现研究能挖掘出大量隐式实体关系[119−120], 这是对当前电子病历实体关系抽取研究的重要补充.

表 6 电子病历实体关系抽取方法总结

Table 6 Summarization of methods for entity relation extraction of EMR

作者	方法	用到的资源	数据	评价 (F 值)
Frunza 等[112]	SVM	UMLS	Medline	0.8889 ~ 1.0
Uzuner 等[3]	SVM	Link grammar parser	两个机构的病历	0.62 ~ 0.89
Rink 等[113]	SVM	Wikipedia、WordNet、General Inquirer、GENIA	I2B2 2010 评测数据	0.737
Demner-Fushman 等[109]	SVM	UMLS	I2B2 2010 评测数据	0.666
de Bruijn 等[99]	半监督 + 最大熵	UMLS、cTAKES	I2B2 2010 评测数据	0.731
Ryan[115]	半监督 + SVM	UMLS、Link grammar parser	I2B2 2010 评测数据	0.8
Wang 等[116]	统计共现	MedLEE	自建病历语料	0.91
Chen 等[117]	统计共现	MeSH、UMLS、MedLEE	生物医学文献和电子病历	/

5　电子病历信息抽取研究语料资源

医疗领域的电子病历是信息抽取研究的一个新的领域, 由于电子病历具有数据量庞大、富含医疗知识和患者的个性化健康信息这两大特点, 电子病历信息抽取研究变得越来越重要, 引起了广泛的关注. 电子病历信息抽取研究仍然以统计机器学习方法为主要方法, 现行的统计机器学习方法非常依赖于标注语料, 一方面是为了训练模型, 另一方面是为了评价模型的学习效果, 电子病历标注语料库对医疗领域信息抽取研究有着同样的作用[31]. 但由于电子病历涉及隐私信息, 获取电子病历存在较大障碍, 而且电子病历是专业文档, 涉及较多专业知识, 构建标注语料成本太高. 发起和组织共享任务被证明是一种解决电子病历语料匮乏的有效途径[121]. 国外以 I2B2 为代表的医疗领域共享任务及语料库大大促进了电子病历信息抽取的研究, 其他共享任务的语料库包括病历分类语料、自杀笔记情感分析语料和病历文本检索语料.

5.1　I2B2 评测及语料库

I2B2 从 2006 年开始组织电子病历信息抽取研究的评测, 建设了面向特定任务的语料库, 这些任务围绕实体识别和关系抽取展开, 在电子病历信息抽取研究领域树立了标杆, 不仅推动了该领域的研究, 而且为其他语言电子病历信息抽取研究的语料库建设提供了有益的借鉴. 到 2012 年, I2B2 一共组织了 6 次电子病历信抽取评测, 发布了 7 个语料库, 如表 7 所示.

1) I2B2 2006

I2B2 于 2006 年首次组织电子病历信息抽取研究评测. 电子病历对外发布之前必须首先去掉患者和医生的隐私信息, 所以评测中第一个任务就是去隐私信息 (De-identification challenge)[9], 第二个任务是识别患者吸烟状态 (Smoking challenge)[101]. 评测使用的电子病历来源于美国联盟医疗 (Partners Health Care), 文本类型是出院小结.

去隐私信息研究首先要识别隐私信息, 然后用替代信息替换隐私信息, 这些隐私信息主要有 8 类, 包括患者姓名、医生姓名、医疗结构名称、各类标示、诊疗日期、位置、电话号码、患者年龄. 去隐私信息任务提供了 889 份标注的病历文本, 其中 669 用于训练, 220 用于测试.

患者是否吸烟以及吸烟状态是很重要的临床依据. 识别患者吸烟状态研究主要是根据患者的出院记录, 对患者的吸烟状态进行分类, 吸烟状态类型有 5 个, 这 5 个类型分别是过去吸烟、现在吸烟、有吸烟史但无法区分是现在吸烟还是过去吸烟、从不吸烟、未提及与吸烟有关的信息. 识别患者吸烟状态任务提供了 502 份肺病专家标注的病历文本, 其中 398 用于训练, 104 用于测试.

2) I2B2 2008

由于临床研究需要统计肥胖的泛滥程度, 因此 2008 年的评测任务是根据患者的出院小结识别患者的肥胖及综合症的状态 (Obesity challenge)[102]. 该任务类似 I2B2 2006 评测中的识别患者吸烟状态任务. 识别患者的状态类型有 4 个, 分别是当前正处于肥胖状态、可能肥胖、不存在肥胖、病历中未提及.

该任务的 1 237 份电子病历来源于美国联盟医疗超重患者或者糖尿病患者的住院病历, 由专家标注, 遵循两种标注方案. 一种方案是只考虑出院小结字面上有没有出现肥胖的 (Obese) 或者糖尿病的 (Diabetic) 等字样, 另一种方案不仅考虑字面意思, 还要考虑文本蕴含的意思, 比如 “患者身高 5 英尺 2 英寸, 体重 230 磅”, 那么该患者标注为肥胖.

3) I2B2 2009

药品是电子病历里重要的医疗知识,包括药品的

表 7　I2B2 历次电子病历信息抽取评测及语料库

Table 7　Shared tasks of information extraction and corpuses in I2B2

年份	任务	病历文本类型	语料规模 (份)
2006	去隐私信息	出院小结	889
2006	识别患者吸烟状态	出院小结	502
2008	实别患者肥胖及综合症的状态	出院小结	1 237
2009	药品属性识别	出院小结	251
2010	识别病历中的概念、概念的修饰、概念间关系	出院小结、病程记录	871
2011	概念共指消解	出院小结、病程记录	978
2012	抽取事件和事件发生的时间	出院小结	310

商品名、通用名、给药方式、给药剂量等重要的信息. 这些信息可以看作是药品的属性, 2009 年评测任务就是从病历文本中识别药品的属性信息 (Medication challenge)[10], 主要包括药名、剂量、模式、频率、持续时间等信息.

药品属性识别任务提供了 251 份标注的出院小结, 病历来源于美国联盟医疗. 值得注意的是, 此次评测采用了团体标注模式构建标注语料[77]. 实验表明这种标注方法比专家标注更为高效 (成本低、速度快), 且标注质量相当.

4) I2B2 2010

2010 年评测任务扩展了往届的任务, 并关注病历中最重要的医疗知识, 包括疾病、症状、检查和治疗的识别, 以及这些实体间的关系的抽取 (Relations challenge)[12]. 此次评测有三个任务, 即识别病历中的概念、概念的修饰、概念间关系, 要识别的概念主要是医疗问题、检查和治疗这三类 (见表 1), 概念的修饰主要是医疗问题的修饰 (见表 3), 概念间的关系主要是医疗问题和医疗问题的关系、医疗问题和检查的关系以及医疗问题和治疗的关系 (见表 5).

此次评测的电子病历来源于三个不同的医疗机构: 美国联盟医疗和贝斯以色列女执事医疗中心 (Beth Israel Deaconess Medical Center) 提供了出院小结, 匹兹堡大学医学中心 (University of Pittsburgh Medical Center) 提供了出院小结和病程记录. 871 份标注的病历文本中, 397 份用于训练, 477 份用于测试, 877 份未标注的文本用于评测. 评测首次引入不同机构的电子病历和不同类型的病历文本, 保证了数据的多样性, 使得研究避免了特殊性.

5) I2B2 2011

2011 年评测关注实体识别的后续任务共指消解 (Coreference challenge)[16], 该任务实际上是抽取实体间的等价关系. 评测使用的电子病历仍然保持了机构多样性和类型多样性的特点, 来源于美国联盟医疗、贝斯以色列女执事医疗中心、匹兹堡大学医学中心和梅奥诊所 (Mayo Clinic), 病历文本包括出院小结、病程记录、射线报告和外科病理报告等. 标注的病历一共 978 份, 590 份用于训练, 388 份用于测试.

6) I2B2 2012

时间是病历中的重要信息, 因为患者病情进展、治疗过程都与时间相关. 2012 年评测关注时间这一特殊实体的识别 (Temporal challenge)[20]. 该研究的目的是要抽取事件和事件发生的时间, 建立事件发生的时间线, 据此跟踪患者的健康状况、查找病因、分析治疗的有效性和副作用等. 此次评测任务有三个:

1) 抽取表示医疗事件的概念, 包括医疗问题、检查、治疗、医院科室, 还包括与时间有关的事件, 比如入院、科室转移、出院等;

2) 抽取时间表达式包括日期、时间、频率、持续时长等信息, 并格式化为 ISO 规范格式;

3) 抽取时间表达式和事件的关系.

电子病历来源于美国联盟医疗和贝斯以色列女执事医疗中心, 病历文本是出院小结. 在 310 份标注的出院小结中, 190 份用于训练, 120 份用于测试.

从以上对 I2B2 电子病历信息抽取研究评测的介绍可以看出, 这些评测围绕实体识别展开, 其中 2010 年的评测对这些任务进行了系统的归纳, 该研究任务将为医疗知识图谱的研究提供支持, 因而我们主要关注这一届的评测任务, 后两届的评测任务可以看成是 2010 年评测任务的延伸.

5.2 其他语料库和共享任务

除了 I2B2 针对信息抽取研究构建的一些标注语料外, 还有用于电子病历文本自然语言处理其他方面研究的语料库和共享任务, 如自动疾病编码和病历文本检索.

1) 病历分类

疾病编码主要用于医疗补偿, 给患者的疾病赋予编码的精确程度直接影响到医疗补偿的多少. 对病历文本进行分类, 自动赋予疾病编码是电子病历信息抽取的又一个重要任务. 辛辛那提大学医学中心 (University of Cincinnati Medical Center) 和辛辛那提儿童医院医疗中心 (Cincinnati Children′s Hospital Medical Center) 于 2007 年组织了该任务的评测[16], 评测电子病历来源于辛辛那提儿童医院医疗中心, 病历文本是射线报告. 语料构建工作由 Pestian 等[122] 完成, 标注的病历文本一共有 1 954 份, 978 份用于训练, 976 份用于测试.

2) 自杀笔记的情感分析

该任务于 2011 年由 I2B2/VA 和辛辛那提儿童医院医疗中心联合举办, 目的是分析每篇自杀笔记的情绪, 可能的情绪包括愤怒、抱怨、恐惧、负罪、绝望等 16 种情绪. 任务收集到自杀笔记一共 1 319 份, 时间跨度从 1950 年到 2011 年. 参考文献 [123] 可进一步了解该任务和语料库的详情.

3) 病历文本检索

文本检索会议 (Text Retrieval Conference, TREC) 于 2011 年和 2012 年连续两届组织了电子病历的检索评测[124], 评测所用电子病历来源于匹兹堡自然语言处理数据库.

[16] http://computationalmedicine.org/challenge/previous

5.3 中文电子病历语料库构建

当前大多数电子病历信息抽取研究是针对英文电子病历的, 中文电子病历方面的研究屈指可数, 我们调研到的几个研究并没有公开所用的语料库, 因此系统构建中文电子病历语料库是我们开展研究的前提. 我们的中文电子病历来源于哈尔滨医科大学附属第二医院病案室, 从各科室随机选取 2 000 份病历用于构建语料库, 病历文本类型为首次病程记录和出院小结, 并经过去隐私信息处理. 由于病历文本语言的独特性, 通用领域的词法分析和句法分析等自然语言处理工具并不适合于病历文本, 而这些又是电子病历信息抽取的基础, 因此面向中文电子病历文本的分词、词性标注和句法分析也是我们研究的重点. 我们拟构建的语料库包括分词和词性标注、句法分析、实体识别、实体关系抽取四个语料. 目前我们已完成分词和词性标注的语料构建[125], 句法分析、实体识别和实体关系抽取语料的标注规范已制定完毕, 标注工作正在进行中.

我们将开放领域公认的宾州中文树库 (Penn Chinese treebank, PCTB) 标注规范[126−127] 作为基础规范, 在医生的指导下, 制定了适用于中文电子病历的分词标注规范[17]及词性标注规范[18]. 我们分别从病程记录和出院小结抽取语句, 构建了 1 094 个句子的分词和词性标注语料.

中文电子病历句法分析标注规范[19] 的制定以中文宾州树库标注规范[128] 为基础, 结合中文电子病历的实际标注情况及电子病历信息抽取的需要对标注规范进行迭代修订, 最后通过人机互助的方式进行病历标注工作. 构建工作重点解决由于标注人员领域知识不足导致的标注错误问题, 以及中文电子病历中省略句法成分引发的标注一致性问题. 我们预计初步标注 100 份病历文本 (病程记录和出院小结各 50 份), 形成大约 2 000 句标注的语法树.

中文电子病历实体识别任务的定义参见第 3 节, 该任务主要识别疾病、症状、检查和治疗这四类实体, 其中症状细分为自诉症状和检查结果两个子类, 并且识别疾病和症状的修饰. 中文电子病历实体关系抽取任务的定义参见第 4 节, 该任务主要识别疾病和检查的关系、症状和检查的关系、疾病和治疗的关系、症状和治疗的关系、疾病和疾病的关系、症状和症状的关系以及疾病和症状的关系. 在医生的指导下, 中文电子病历实体标注规范和实体关系标注规范[20] 参照 I2B2 2010 标注规范制定, 并结合中文电子病历特点进行调整. 为了辅助语料标注工作, 我们开发了图形界面的标注工具[21].

[17] http://wi.hit.edu.cn/dev/YuLiao/Seg.pdf
[18] http://wi.hit.edu.cn/dev/YuLiao/POS.pdf
[19] http://wi.hit.edu.cn/dev/YuLiao/Bracketing.pdf
[20] http://wi.hit.edu.cn/dev/YuLiao/NER.pdf
[21] https://github.com/yangjinfeng/emrproject

6 医疗领域词典及知识库

医疗领域一个显著的特点就是大量使用受控词汇, 这些受控词汇在医疗行业已经得到广泛认可, 因而以词典或知识库的形式得到维护和使用. 在医疗领域, 使用最广泛的知识库是一体化医学语言系统 (Unified Medical Language System, UMLS)、医学主题词表 (Medical Subject Headings, MeSH)、医学系统命名法 — 临床术语 (Systematized Nomenclature of Medicine — Clinical Terms, SNOMED CT) 和国际疾病分类 (International Classification of Diseases, ICD).

6.1 UMLS

UMLS 是由美国国家医学图书馆建立的一个集成的生物医学领域词典知识库, 其主要目的是集成不同机构维护的医学词典解决相同概念具有不同名称的问题并且建立术语的标准化格式, 支持计算机高效的检索[90]. 自 1986 年开始建设以来, UMLS 现在还在不断发展和完备中. UMLS 由四部分组成:

1) 超级叙词表 (Metathesaurus): UMLS 的核心数据表, 整合了来自不同源词表的概念和术语, 以及在源词表中已存在的概念间关系;

2) 语义网络 (Semantic network): 由语义类型之间的语义关系形成的语义网络, 通过为概念赋予一个或多个语义类型, 语义网络可以建立概念间语义关系;

3) 专家词典 (Specialist lexicon): 为解决术语变体问题而建立的词典, 在生物医学领域, 术语的各种变体大量存在, 专家词典就是一个为生物医学领域自然语言处理提供词汇信息的词典, 旨在解决词汇和术语的高度变异问题.

4) UMLS 支撑性软件包: 为了方便地使用 UMLS 而开发的支撑性工具.

前三部分是 UMLS 作为知识库的核心部分, 可以看出 UMLS 以整合概念和概念间关系为核心任务. UMLS 整合概念时, 建立了概念和表达相同概念的不同术语之间的对应关系, 专家词典则是考虑了术语的变体, 从词形上建立术语之间的对应关系, 同时有些概念在源词表中已经按照叙词的组织方式建立了语义关系, 因而这些概念和概念间关系就被整合到了 UMLS 中. 对于那些在源词表中没有建立关系而实际上存在语义关系的概念, 则通过语义网

络建立起语义关系. UMLS 对各种源词表的集成方式采取的是保守的策略, 甚至词表之间的冲突也都保留了下来, 所以, 对 UMLS 的使用首先要识别对研究问题有用的源词表, 并设计定制策略从 UMLS 中定制或者裁减知识[129]. UMLS 在生物医学和医疗领域得到广泛的应用, 如识别概念和关系[130]、语义消歧[131]、语义标注[132]、语义相似度计算[133]、本体构建[134−136] 等.

6.2 MeSH

MeSH[22] 是美国国家医学图书馆编制的受控专业词表, 主要用于生物医学文档的索引、分类和检索, 已经成为 UMLS 的源词表. 2012 年版本的 MeSH 包含 25 000 个主题词 (Descriptors, main headings). 大部分主题词都对应一个简短的定义和同义词列表, 所以 MeSH 也可以被看作一个同义词词典. 在 MeSH 中, 主题词是按树状层次结构组织的, 同一个主题词可能出现在层次树的不同地方. MeSH 词典中除了主题词以外, 还包含 83 个限定词 (Qualifiers) 和 139 000 个补充概念 (Supplements). 限定词可以用来和主题词组合表达更细的意义, 补充概念可以关联到其最相关的主题词. MeSH 已经被翻译成不同的语言, 中文版本的 CMeSH 由中国医学科学院医学信息研究所翻译, 但没有免费开放, 已被用于跨语言检索研究[137].

6.3 SNOMED CT

SNOMED CT 国际卫生术语标准开发组织发布的临床术语集, 是一部经过系统组织编排的、便于计算机处理的医学术语集, 涵盖大多数方面的临床信息, 如疾病、操作、微生物、药物等. 采用该术语集, 可以协调一致地在不同的学科、专业和照护地点之间实现对于临床数据的标引、存储、检索和聚合. 同时, 它还有助于组织病历内容, 减少临床照护和科学研究工作中数据采集、编码及使用方式的变异. SNOMED CT 的术语主要有三部分内容: 概念、描述、关系. 概念有不同的粒度, 也就是一般性概念和具体概念. 具体概念与一般性概念之间存在 is-a 关系. 每个概念赋予一个唯一标识. 描述是用来命名概念的短语, 可以理解为概念的同义表达, 每一个描述赋予一个唯一标识, 同一个概念的多个描述与概念的标识关联. 描述有三种: 概念的全称、概念最常用的名称、概念的同义词. 关系是相关概念间的语义关系, 主要关注 is-a 关系和属性关系, 这两种关系反映了概念间的定义关系, 通过这两种关系可以用一个概念逻辑的表示另一个概念. SNOMED CT 已经成为 UMLS 的源词表. 目前, SNOMED CT 还没有中文版发布.

6.4 ICD

ICD 是国际卫生组织制定的疾病分类体系, 用于流行病学、健康管理和临床诊断. 疾病分类与代码规定了疾病、损伤和中毒及其外部原因、与保健机构接触的非医疗理由和肿瘤形态学的分类与代码. 本标准适用于医疗卫生服务、医疗保障、民政、公安等部门对疾病、伤残、死亡原因分类的信息收集、整理、交换、分析等. ICD 当前版本是第 10 个版本. 为了便于迁移到 SNOMED CT 编码, 美国国家医学图书馆已经发布了 ICD-9 和 SNOMED CT 之间的相互映射工具. 电子病历系统一般采用 ICD 对疾病编码, 为了对电子病历描述的疾病进行精确的编码, 国际卫生组织发布了 ICD-9 的临床修订版 (ICD-9-CM)[23]. 我国卫生部于 2012 年发布了中文版的 ICD-10.

标准化的临床词典及知识库在医疗领域起着共享知识和标准化的作用, 在医疗领域信息抽取研究中是重要的词典资源. 上述四类主要的知识库中, UMLS 和 SNOMED CT 都缺乏对应的中文版本. UMLS 是医疗知识库的集大成者, 对 UMLS 的汉化工作不仅有益于临床医疗, 对中文医疗领域信息抽取研究也大有裨益. 为促进 UMLS 在中文电子病历信息抽取研究上的应用, 沈彤[138] 对 UMLS 的汉化工作展开了初步的探索.

7 结束语

本文对医疗领域电子病历命名实体识别和实体关系抽取研究进展进行了综述, 主要内容包括电子病历文本特点分析、电子病历实体识别、疾病和症状的修饰识别、实体关系抽取、电子病历信息抽取研究语料资源等研究现状. 考虑到电子病历涉及大量医疗领域词典及知识库并且这些知识库在电子病历信息抽取中的重要作用, 本文对主要的词典及知识库进行了介绍. 因为电子病历文本富含医疗知识和各种专业术语, 是典型的知识密集型文本, 从电子病历抽取医疗知识是很自然的研究课题, 所以围绕实体识别的电子病历信息抽取研究并不是一个新兴的研究方向, 研究方法也与通用领域研究方法大同小异, 主要是基于标注语料库的统计机器学习方法, 并结合词典和知识库. 但由于电子病历文本明显的子语言特点和内容极强的专业性, 以及电子病历的隐私性, 电子病历信息抽取研究非常不同于其他领域包括通用领域信息抽取研究, 远没有达到成熟的状态, 主要体现在缺乏面向电子病历语言的自然语

[22] http://www.nlm.nih.gov/mesh/
[23] http://www.cdc.gov/nchs/icd/icd9cm.htm

言处理工具、标注语料不够丰富, 现有研究方法没有充分考虑病历文本的特点并且只是针对英文病历等. 基于本文的调研, 我们认为下面几个方面是未来电子病历信息抽取研究中值得关注的方向:

1) 电子病历文本的语言有着鲜明的行业特色, 并且文本内容呈半结构化, 这些重要特点应充分应用于电子病历信息抽取研究中. 因此, 病历文本的子语言特点和内容的半结构化应整合到机器学习模型中.

2) 当前大多数电子病历信息抽取研究是针对英文电子病历的, 中文电子病历方面的研究刚刚起步, 还没有形成明确而系统化的研究任务, 缺乏公开的标注语料库. 这种一穷二白的局面大大制约了中文电子病历信息抽取的研究. 随着中文电子病历的广泛实施, 中文电子病历数量的急剧增长, 中文电子病历实体和实体关系识别研究势在必行, 而中文电子病历标注语料库的构建是首当其冲的.

3) 电子病历充斥着大量的医疗术语, 医疗领域的词典和知识库对电子病历信息抽取有着重要作用. 已发布的中文医疗领域词典和知识库较少, 这也制约了中文电子病历信息抽取的研究. 因此, 基于英文医疗知识库构建中文医疗知识库值得关注.

4) 为了充分利用电子病历的语言特点, 有必要开展针对病历文本的分词、词性标注以及句法分析等基础研究.

5) 现行大多数电子病历录入系统智能化程度较低, 为了追求效率, 拷贝–粘贴的输入模式被广大医务人员采用, 但这种输入模式导致电子病历质量低下. 因此智能电子病历输入系统的研究是必要的. 高质量的电子病历能促进信息抽取的研究, 同时信息抽取的研究结果又能促进病历系统智能化程度的提高.

6) 通用领域的知识图谱研究引起了人们的极大关注, 电子病历中挖掘出的实体和实体关系实则形成了医疗领域的知识图谱, 医疗领域知识图谱的研究, 包括知识图谱的构建和应用, 在大数据浪潮下, 吸引了越来越多研究者的关注.

References

1 Ministry of Health of the People's Republic of China. The basic specifications of electronic medical records (trial) [Online], available: http://www.gov.cn/zwgk/2010-03/04/content_1547432.htm, December 27, 2013
(中华人民共和国卫生部. 电子病历基本规范 (试行) [Online], available: http://www.gov.cn/zwgk/2010-03/04/content_1547432.htm, December 27, 2013)

2 Wasserman R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. *Academic Pediatrics*, 2011, **11**(4): 280−287

3 Uzuner O, Mailoa J, Ryan R, Sibanda T. Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, 2010, **50**(2): 63−73

4 Demner-Fushman D, Chapman W W, McDonald C J. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 2009, **42**(5): 760−772

5 Eysenbach G. Recent advances: consumer health informatics. *British Medical Journal*, 2000, **320**(7251): 1713−1716

6 Lin Dong, Shao Jun-Li. A general and practical diagnosing and treating expert system of medicine. *Acta Automatica Sinica*, 1995, **21**(3): 380−382
(林东, 邵军力. 医学诊疗领域通用专家系统设计与实现. 自动化学报, 1995, **21**(3): 380−382)

7 Sager N, Friedman C, Lyman M S. Review of Medical language processing: computer management of narrative data. *Computational Linguistics*, 1989, **15**(3): 195−198

8 National Institutes of Health. Research Repositories, Databases, and the HIPAA Privacy Rule [Online], available: http://privacyruleandresearch.nih.gov/pdf/research_repositories_final.pdf, December 27, 2013

9 Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 2007, **14**(5): 550−563

10 Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 2010, **17**(5): 514−518

11 Xu Yong-Dong, Quan Guang-Ri, Wang Ya-Dong. Research of electronic medical record key information extraction based on HL7. *Journal of Harbin Institute of Technology*, 2011, **43**(11): 89−94
(徐永东, 权光日, 王亚东. 基于 HL7 的电子病历关键信息抽取技术研究. 哈尔滨工业大学学报, 2011, **43**(11): 89−94)

12 Uzuner O, South B R, Shen S, DuVall S L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 552−556

13 Chapman W W, Bridewell W, Hanbury P, Cooper G F, Buchanan B G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 2001, **34**(5): 301−310

14 Zheng J P, Chapman W W, Crowley R S, Savova G K. Coreference resolution: a review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 2011, **44**(6): 1113−1122

15 Tian Y H. Coreference Resolutionon Entities and Events for Hospital Discharge Summaries [Master dissertation], Massachusetts Institute of Technology, USA, 2007

16 Uzuner O, Bodnari A, Shen S Y, Forbush T, Pestian J, South B R. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 2012, **19**(5): 786−791

17 Filannino M. Temporal expression normalisation in natural language texts. *ArXiv Preprint*, ArXiv Preprint arXiv: 1206.2010, 2012

18 UzZaman N, Llorens H, Allen J, Derczynski L, Verhagen M, Pustejovsky J. TempEval-3: Evaluating events, time expressions, and tem-poral relations. *ArXiv Preprint*, ArXiv Preprint arXiv: 1206.5333, 2012

19 Zhou X J, Li H M, Lu X D, Duan H L. Temporal expression recognition and temporal relationship extraction from Chinese narrative medical records. In: Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering. Wuhan, China: IEEE, 2011. 1−4

20 Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 I2B2 challenge. *Journal of the American Medical Informatics Association*, 2013, **20**(5): 806−813

21 Tange H J, Hasman A, Robbe P F, Schouten H C. Medical narratives in electronic medical records. *International Journal of Medical Informatics*, 1997, **46**(1): 7−29

22 McDonald C J, Overhage J M, Tierney W M, Dexter P R, Martin D K, Suico J G, Zafar A, Schadow G, Blevins L, Glazener T, Meeks-Johnson J, Lemmon L, Warvel J, Porterfield B, Warvel J, Cassidy P, Lindbergh D, Belsito A, Tucker M, Williams B, Wodniak C. The regenstrief medical record system: a quarter century experience. *International Journal of Medical Informatics*, 1999, **54**(3): 225−53

23 Fries J F. Time-oriented patient records and a computer databank. *Journal of the American Medical Association*, 1972, **222**(12): 1536−1542

24 Weed L L. Medical records that guide and teach. *New England Journal of Medicine*, 1968, **278**(12): 593−600

25 Jacobs L. Interview with Lawrence Weed, MD — the father of the problem-oriented medical record looks ahead. *The Permanente Journal*, 2009, **13**(3): 84−89

26 Bossen C. Evaluation of a computerized problem-oriented medical record in a hospital department: does it support daily clinical practice? *International Journal of Medical Informatics*, 2007, **76**(8): 592−600

27 Tilstra S. In Search of the Holy Grail: How to Ensure the Perfect Progress Note [Online], available: http://www.im.org/Meetings/Past/2012/2012APDIMSpringConference/Presentations/Documents/Spring%20Meeting/Wksp%20202_Tilstra.pdf, December 27, 2013

28 Ministry of Health of the People′s Republic of China. The basic specifications of medical records writing (trial) [Online], available: http://www.gov.cn/gzdt/2010-02/04/content_1528415.htm, December 27, 2013
(中华人民共和国卫生部. 病历书写基本规范 [Online], available: http://www.gov.cn/gzdt/2010-02/04/content_1528415.htm, December 27, 2013)

29 Lynette H, Sager N. Automatic information formatting of a medical sublanguage. In: Proceedings of the 1982 Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin, German: Walter de Gruyter, 1982. 27−80

30 Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 2002, **35**(4): 222−235

31 Meystre S M, Savova G K, Kipper-Schuler K C, Hurdle J F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 2008, **47**(Suppl 1): 128−144

32 O′Donnell H C, Kaushal R, Barron Y, Callahan M A, Adelman R D, Siegler E L. Physicians′ attitudes towards copy and pasting in electronic note writing. *Journal of General Internal Medicine*, 2009, **24**(1): 63−68

33 Hammond K W, Helbig S T, Benson C C, Brathwaite-Sketoe B M. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. In: Proceedings of the 2003 American Medical Informatics Association 2003 Annual Symposium. Washington DC, USA: AMIA, 2003. 269−273

34 Wilcox L, Lu J, Lai J, Feiner S, Jordan D. ActiveNotes: computer-assisted creation of patient progress notes. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems. New York, USA: ACM Press, 2009. 3323−3328

35 Wilcox L, Lu J, Lai J, Feiner S, Jordan D. Physician-driven management of patient progress notes in an intensive care unit. In: Proceedings of the 28th International Conference Extended Abstracts on Human Factors in Computing Systems. New York, USA: ACM Press, 2010. 1879−1888

36 Grishman R, Sundheim B. Message Understanding Conference-6: a brief history. In: Proceedings of the 16th conference on Computational linguistics-Volume 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996. 466−471

37 Lang Jun, Qin Bing, Liu Ting, Li Zheng-Hua, Li Sheng. Number type recognition of Chinese personal noun phrase. *Acta Automatica Sinica*, 2008, **34**(8): 972−979
(郎君, 秦兵, 刘挺, 李正华, 李生. 中文人称名词短语单复数自动识别. 自动化学报, 2008, **34**(8): 972−979)

38 Tang Bu-Zhou, Wang Xiao-Long, Wang Xuan. Confidence-weighted online sequence labeling algorithm. *Acta Automatica Sinica*, 2011, **37**(2): 188−195
(汤步洲, 王晓龙, 王轩. 置信度加权在线序列标注算法. 自动化学报, 2011, **37**(2): 188−195)

39 Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The automatic content extraction (ACE) program tasks, data, and evaluation. In: Proceedings of the 2004 International Conference on Language Resources and Evaluation. Lisbon, Portugal: European Language Resources Association, 2004. 837−840

40 Wang Ning, Ge Rui-Fang, Yuan Chun-Fa, Wong K F, Li Wen-Jie. Company name identification in Chinese financial domain. *Journal of Chinese Information Processing*, 2002, **16**(2): 1−6
(王宁, 葛瑞芳, 苑春法, 黄锦辉, 李文捷. 中文金融新闻中公司名的识别. 中文信息学报, 2002, **16**(2): 1−6)

41 Lin X D, Peng H, Liu B. Chinese named entity recognition using support vector machines. In: Proceedings of the 2006 International Conference on Machine Learning and Cybernetics. Guangzhou, China: IEEE, 2006. 4216−4220

42 Zhao Jian. Research on Conditional Probabilistic Model and Its Application in Chinese Named Entity Recognition [Ph. D. dissertation], Harbin Institute of Technology, China, 2006
(赵健. 条件概率模型研究及其在中文名实体识别中的应用 [博士学位论文], 哈尔滨工业大学, 中国, 2006)

43 Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. 363−370

44 Finkel J R, Manning C. Joint parsing and named entity recognition. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 326−334

45 Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 2007, **30**(1): 3−26

46 Ke X, Li S Z. Chinese organization name recognition based on co-training algorithm. In: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering. Xiamen, China: IEEE, 2008. 771−777

47 Nadeau D. Semi-supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision [Ph. D. dissertation], University of Ottawa, Canada, 2007

48 Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 2005, **6**: 1817−1853

49 Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2011, **12**: 2493−2537

50 Zhang Qi. Research on Entity Relation Recognition in Information Extraction [Ph. D. dissertation], University of Science and Technology of China, China, 2010
(张奇. 信息抽取中实体关系识别研究 [博士学位论文], 中国科学技术大学, 中国, 2010)

51 Swanson D R. Complementary structures in disjoint science literatures. In: Proceedings of the 14th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1991. 280−289

52 Cohen A M, Hersh W R. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 2005, **6**(1): 57−71

53 Chen J X. Automatic Relation Extraction Among Named Entities from Text Contents [Ph. D. dissertation], National University of Singapore, Singapore, 2006

54 Che Wan-Xiang, Liu Ting, Li Sheng. Automatic entity relation extraction. *Journal of Chinese Information Processing*, 2004, **19**(2): 1−6
(车万翔, 刘挺, 李生. 实体关系自动抽取. 中文信息学报, 2004, **19**(2): 1−6)

55 Chinchor N. MUC-7 named entity task definition (Version 3.5). In: Proceedings of the 7th Message Understanding Conference. Fairfax, Virginia, USA, 1998. Appendix E [Online], available: http://newdesign.aclweb.org/anthology/M/M98/M98-1028.pdf

56 Aone C, Ramos-Santacruz M. REES: a large-scale relation and event extraction system. In: Proceedings of the 6th Conference on Applied Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 76−83

57 Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections. In: Proceedings of the 5th ACM conference on Digital libraries. New York, USA: ACM, 2000. 85−94

58 Bunescu R C, Mooney R J. Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL′ 07). Prague, Czech Republic, 2007. 576−583

59 Zhang Z. Weakly-supervised relation classification for information extraction. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2004. 581−588

60 Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. 415

61 Chen J X, Ji D D, Tan C L, Niu Z Y. Unsupervised feature selection for relation extraction. In: Proceedings of the 2005 International Joint Conference on Natural Language Processing. Jeju Island, Korea: Springer, 2005. 262−267

62 Zhang Zhi-Tian. The Research of Relation Extraction with Unsupervised Method [Master dissertation], Harbin Institute Technology, China, 2007
(张志田. 无监督关系抽取方法研究 [硕士学位论文], 哈尔滨工业大学, 中国, 2007)

63 Zhang Y, Zhou J. A trainable method for extracting Chinese entity names and their relations. In: Proceedings of the 2nd Workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 66−72

64 Suchanek F M, Ifrim G, Weikum G. Combining linguistic and statistical analysis to extract relations from web documents. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2006. 712−717

65 Sleator D, Temperley D. Parsing English with a Link Grammar, Technical Report CMU-CS-91-196, School of Computer Science, Carnegie Mellon University, USA, 1991

66 Brin S. Extracting patterns and relations from the world wide web. *The World Wide Web and Databases*, 1999, **1590**(2): 172−183

67 Ning Hai-Yan. Comparative Study of Automatic Entity Relation Extraction [Master dissertation], Harbin Insititute of Technology, China, 2010
(宁海燕. 实体关系自动抽取技术的比较研究 [硕士学位论文], 哈尔滨工业大学, 中国, 2010)

68 Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 1535−1545

69 Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka E R, Mitchell T M. Toward an architecture for never-ending language learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. Georgia, USA: AAAI, 2010. 1306−1313

70 Suchanek F M, Kasneci G, Weikum G. YAGO: A core of semantic knowledge unifying wordnet and Wikipedia. In: Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM, 2007. 697−706

71 Biega J, Kuzey E, Suchanek F M. Inside YAGO2s: a transparent information extraction architecture. In: Proceedings of the 22nd International Conference on World Wide Web Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. 325−328

72 Kim J D, Ohta T, Tateisi Y, Tsujii J. GENIA corpus — a semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003, **19**(Suppl 1): 180−182

73 Tanabe L, Xie N, Thom L H, Matten W, Wilbur W J. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 2005, **6**(Suppl 1): S3

74 Kim J D, Ohta T, Tsuruoka Y, Tateisi Y, Collier N. Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the 2004 International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. 70−75

75 Arighi C N, Roberts P M, Agarwal S, Bhattacharya S, Cesareni G, Chatr-Aryamontri A, Clematide S, Gaudet P, Giglio M G, Harrow I, Huala E, Krallinger M, Leser U, Li D, Liu F, Lu Z, Maltais L J, Okazaki N, Perfetto L, Rinaldi F, Sætre R, Salgado D, Srinivasan P, Thomas P E, Toldo L, Hirschman L, Wu C H. BioCreative III interactive task: an overview. *BMC Bioinformatics*, 2011, **12**(Suppl 8): S4

76 Xu Wei, Fu Bin, Liu Liu, Yuan Chun-Fa, Li Wen-Jie. Domain extension of Chinese named entity recognition. In: Proceedings of the 9th Chinese National Conference on Computatinal Linguistics. Dalian, China, 2007. 503−508
(徐薇, 付滨, 刘柳, 苑春法, 李文捷. 中文命名实体识别系统的领域扩展, 第九届全国计算语言学学术会议. 大连, 中国, 2007. 503−508)

77 Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the I2B2 medication challenge. *Journal of the American Medical Informatics Association*, 2010, **17**(5): 519−523

78 Baldridge J, Osborne M. Active learning and the total cost of annotation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: Association for Computational Linguistics, 2004. 9−16

79 Settles B, Craven M, Friedland L. Active learning with real annotation costs. In: Proceedings of the 2008 NIPS Workshop on Cost-Sensitive Learning. Vancouver, Canada, 2008. 1−10

80 Tomanek K, Wermter J, Hahn U. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic, 2007. 486−495

81 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. New York, USA: ACM, 1998. 92−100

82 Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 1995. 189−196

83 Zhu X J, Goldberg A B. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009, **3**(1): 1−130

84 Fernandes E R, Brefeld U. Learning from partially annotated sequences. In: Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases (Volume Part I). Berlin, Heidelberg: Springer-Verlag, 2011. 407−422

85 Lou X H, Hamprecht F. Structured learning from partial annotations. *ArXiv Preprint*, ArXiv Preprint, arXiv: 1206.6421, 2012

86 Hovy D, Hovy E. Exploiting partial annotations with EM training. In: Proceedings of the 2012 NAACL-HLT Workshop on the Induction of Linguistic Structure. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 31−38

87 Tsuboi Y, Kashima H, Oda H, Mori S, Matsumoto Y. Training conditional random fields using incomplete annotations. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchster, UK: ACM, 2008. 897−904

88 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

89 Torrey L, Shavlik J. Transfer learning. *Handbook of Research on Machine Learning Applications*. Hershey, PA: IGI Global, 2009

90 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 2004, **32**(suppl 1): D267−D270

91 Friedman C, Alderson P O, Austin J, Cimino J J, Johnson S B. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1994, **1**(2): 161−174

92 Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen P C. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *Journal of biomedical informatics*, 2009, **42**(5): 937−949

93 Savova G K, Masanz J, Ogren P V, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen Piet C. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Information Association*, 2010, **17**(5): 507−13

94 Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 2004, **10**(3−4): 327−348

95 Ye Feng, Chen Ying-Ying, Zhou Gen-Gui, Li Hao-Min, Li Ying. Intelligent recognition of named entity in electronic medical records. *Chinese Journal of Biomedical Engineering*, 2011, **30**(2): 256−262
(叶枫, 陈莺莺, 周根贵, 李昊旻, 李莹. 电子病历中命名实体的智能识别. 中国生物医学工程学报, 2011, **30**(2): 256−262)

96 Li D C, Kipper-Schuler K, Savova G. Conditional random fields and support vector machines for disorder named entityrecognition in clinical texts. In: Proceedings of the 2008 Workshop on Current Trends in Biomedical Natural Language Processing. Morristown, NJ, USA: Association for Computational Linguistics, 2008. 94−95

97 Jiang M, Chen Y, Liu M, Rosenbloom S T, Mani S, Denny J C, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 601−606

98 Jonnalagadda S, Cohen S T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 2012, **45**(1): 129−140

99 de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at I2B2 2010. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 557−562

100 Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC′08). Marrakech, Morocco: European Language Resources Association, 2008. 28−30

101 Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 2007, **15**(1): 14−24

102 Uzuner O. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 2009, **16**(4): 561−570

103 Aronow D B, Fangfang F, Croft W B. Ad hoc classification of radiology reports. *Journal of the American Medical Informatics Association*, 1999, **6**(5): 393−411

104 Goryachev S, Sordo M, Zeng Q T, Ngo L. Implementation and Evaluation of Four Different Methods of Negation Detection, Technical Report, Decision Systems Group, Harvard Medical School, 2006

105 Mutalik P G, Deshpande A, Nadkarni P M. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 2001, **8**(6): 598−609

106 Sohn S, Wu S, Chute C G. Dependency parser-based negation detection in clinical narratives. In: Proceedings of the 2012 AMIA Summits on Translational Science. San Francisco, USA: AMIA, 2012. 1−8

107 Harkema H, Dowling J N, Thornblade T, Chapman W W. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 2009, **42**(5): 839−851

108 Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association*, 2009, **16**(1): 109−115

109 Demner-Fushman D, Apostolova E, Islamaj D R, Lang F M, Neveol A, Shooshan S E, Aronson A R. NLM′s system description for the fourth I2B2/VA challenge. In: Proceedings of the 2010 I2B2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: I2B2, 2010

110 Grouin C, Abacha A B, Bernhard D. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. In: Proceedings of the 2010 I2B2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: I2B2, 2010

111 Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, Hirschman L. MITRE system for clinical assertion status classification. *Journal of the American Medical Informatics Association*, **18**(5): 563−567

112 Frunza O, Inkpen D. Extraction of disease-treatment semantic relations from biomedical sentences. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 91−98

113 Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 594−600

114 Stone P J, Dunphy D C, Smith M S, Ogilvie D M. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press, 1966

115 Ryan R J. Groundtruth Budgeting: A Novel Approach to Semi-Supervised Relation Extraction of Medical Language [Master dissertation], Massachusetts Institute of Technology, USA, 2011

116 Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. In: Proceedings of the 2008 AMIA Annual Symposium, 2008. 783−787

117 Chen E S, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 2008, **15**(1): 87−98

118 Roberts A, Gaizauskas R, Hepple M. Extracting clinical relationships from patient narratives. In: Proceedings of the 2008 Workshop on Current Trends in Biomedical Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. 10−18

119 Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson′s legacy. *Biomedical Digital Libraries*, 2006, **3**(1): 2

120 Cameron D, Bodenreider O, Yalamanchili H, Danh T, Vallabhaneni S, Thirunarayan K, Sheth A P, Rindflesch T C. A graph-based recovery and decomposition of Swanson′s hypothesis using semantic predications. *Journal of Biomedical Informatics*, 2013, **46**(2): 238−251

121 Chapman W W, Nadkarni P M, Hirschman L, D′Avolio D W, Savova G K, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 2011, **18**(5): 540−543

122 Pestian J P, Brew C, Matykiewicz P, Hovermale D J, Johnson N, Cohen K B. A shared task involving multi-label classification of clinical free text. In: Proceedings of the 2007 Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. 97−104

123 Pestian J P, Matykiewicz P, Linn-Gust M. What′s in a note: construction of a suicide note corpus. *Biomedical Informatics Insights*, 2012, **5**: 1−6

124 Voorhees E, Tong R. Overview of the TREC 2012 medical records track. In: Proceedings of the 21st Text REtrieval Conference. Gaithersburg, MD: National Institute for Standards and Technology, 2008. http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf

125 Jiang Zhi-Peng, Zhao Fang-Fang, Guan Yi, Yang Jin-Feng. Research on Chinese electronic medical record oriented lexical corpus annotation. *High Technology Letters*, 2014, **24**(6): 609−615
(蒋志鹏, 赵芳芳, 关毅, 杨锦锋. 面向中文电子病历的词法语料标注研究. 高技术通讯, 2014, **24**(6): 609−615)

126 Xia F. The Segmentation Guidelines for the Penn Chinese Treebank (3. 0). Technical Report IRCS-00-06, University of Pennsylvania, USA, 2000

127 Xia F. The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3. 0). Technical Report IRCS-00-06, University of Pennsylvania, USA, 2000

128 Xue N, Xia F. The Bracketing Guide-lines for Penn Chinese Treebank Project. Technical Report IRCS-00-06, University of Pennsylvania, USA, 2000

129 Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. *IEEE Transactions on Information Technology in Biomedicine*, 2002, **6**(2): 102−108

130 Slaughter L, Ruland C, Rotegard A K. Mapping cancer patients′ symptoms to UMLS concepts. In: Proceedings of the 2005 AMIA Annual Symposium, 2005. 699−703

131 Jimeno-Yepes A J, Aronson A R. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*, 2010, **11**(1): 569−580

132 Jonquet C, Shah N H, Youn C H, Callendar C, Storey M A, Musen M A. NCBO annotator: semantic annotation of biomedical data. In: Proceedings of the 8th International Semantic Web Conference. Washington, DC, USA, 2009. 171−172

133 Pedersen T, Pakhomov S, McInnes B, Liu Y. Measuring the similarity and relatedness of concepts in the medical domain. In: Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics. New York, USA: ACM, 2012. 879−880

134 Ruiz-Martinez J M, Valencia-Garcia R, Fernandez-Breis J T, Garcia-Sanchez T, Martinez-Bejar R. Ontology learning from biomedical natural language documents using UMLS. *Expert Systems with Applications*, 2011, **38**(10): 12365−12378

135 Rosse C, Mejino J. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 2003, **36**(6): 478−500

136 Pisanelli D M, Battaglia M, De Lazzari C. ROME: a reference ontology in medicine. In: Proceedings of the 2007 Conference on New Trends in Software Methodologies, Tools and Techniques. Amsterdam, The Netherlands: IOS Press, 2007. 485−493

137 Wang X, Thompson P, Tsujii J, Anani-adou S. Biomedical Chinese-English CLIR using an extended CMeSH resource to expand queries. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey: European Language Resources Association, 2012. 1148−1155

138 Shen Tong. The Chinesization and Formalization of Unified Medical Language System [Master dissertation], Harbin Insititute of Technology, China, 2013
(沈彤. 一体化医学语言系统的中文化和形式化表示研究 [硕士学位论文], 哈尔滨工业大学, 中国, 2013)

杨锦锋 哈尔滨工业大学博士研究生. 主要研究方向为自然语言处理, 电子病历信息抽取.
E-mail: yangjinfeng2010@gmail.com
(**YANG Jin-Feng** Ph. D. candidate at Harbin Institute of Technology. His research interest covers natural language processing and information extraction on electronic medical records.)

于秋滨 哈尔滨医科大学附属第二医院副主任医师. 主要研究方向为电子病案的数据挖掘.
E-mail: yuqiubin6695@163.com
(**YU Qiu-Bin** Deputy chief physician at the Second Affiliated Hospital of Harbin Medical University. Her research interest covers data mining on electronic medical records.)

关 毅 哈尔滨工业大学教授. 主要研究方向为智能信息检索, 网络挖掘, 自然语言处理, 认知语言学. 本文通信作者.
E-mail: guanyi@hit.edu.cn
(**GUAN Yi** Professor at Harbin Institute of Technology. His research interest covers intelligent information retrieval, web mining, natural language processing, and cognitive linguistics. Corresponding author of this paper.)

蒋志鹏 哈尔滨工业大学博士研究生. 主要研究方向为中文分词, 词性标注, 句法分析. E-mail: xyf-3456@163.com
(**JIANG Zhi-Peng** Ph. D. candidate at Harbin Institute of Technology. His research interest covers word segmentation, part-of-speech tagging, and syntactic analysis.)