软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(11):2725−2746 [doi: 10.13328/j.cnki.jos.004880]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

中文电子病历命名实体和实体关系语料库构建*

杨锦锋[1], 关 毅[1], 何 彬[1], 曲春燕[1], 于秋滨[2], 刘雅欣[3], 赵永杰[4]

[1](哈尔滨工业大学 语言技术研究中心 网络智能研究室,黑龙江 哈尔滨 150001)

[2](哈尔滨医科大学 附属第二医院 病案室,黑龙江 哈尔滨 150086)

[3](哈尔滨医科大学 附属第二医院 呼吸内科,黑龙江 哈尔滨 150086)

[4](哈尔滨医科大学 附属第四医院 神经内科,黑龙江 哈尔滨 150001)

通讯作者: 关毅, E-mail: guanyi@hit.edu.cn

摘 要: 电子病历是由医务人员撰写的面向患者个体描述医疗活动的记录,蕴含了大量的医疗知识和患者的健康信息.电子病历命名实体识别和实体关系抽取等信息抽取研究对于临床决策支持、循证医学实践和个性化医疗服务等具有重要意义,而电子病历命名实体和实体关系标注语料库的构建是首当其冲的.在调研了国内外电子病历命名实体和实体关系标注语料库构建的基础上,结合中文电子病历的特点,提出适合中文电子病历的命名实体和实体关系的标注体系,在医生的指导和参与下,制定了命名实体和实体关系的详细标注规范,构建了标注体系完整、规模较大且一致性较高的标注语料库.语料库包含病历文本 992 份,命名实体标注一致性达到 0.922,实体关系一致性达到 0.895.为中文电子病历信息抽取后续研究打下了坚实的基础.

关键词: 中文电子病历;命名实体;实体关系;标注规范;标注语料构建

中图法分类号: TP391

中文引用格式: 杨锦锋,关毅,何彬,曲春燕,于秋滨,刘雅欣,赵永杰.中文电子病历命名实体和实体关系语料库构建.软件学报, 2016,27(11):2725−2746. http://www.jos.org.cn/1000-9825/4880.htm

英文引用格式: Yang JF, Guan Y, He B, Qu CY, Yu QB, Liu YX, Zhao YJ. Corpus construction for named entities and entity relations on Chinese electronic medical records. Ruan Jian Xue Bao/Journal of Software, 2016,27(11): 2725−2746 (in Chinese). http://www.jos.org.cn/1000-9825/4880.htm

Corpus Construction for Named Entities and Entity Relations on Chinese Electronic Medical Records

YANG Jin-Feng[1], GUAN Yi[1], HE Bin[1], QU Chun-Yan[1], YU Qiu-Bin[2], LIU Ya-Xin[3], ZHAO Yong-Jie[4]

[1](Web Intelligence Laboratory, Language Technology Research Center, Harbin Institute of Technology, Harbin 150001, China)

[2](Medical Record Room, the 2nd Affiliated Hospital of Harbin Medical University, Harbin 150086, China)

[3](Respiratory Department, the 2nd Affiliated Hospital of Harbin Medical University, Harbin 150086, China)

[4](Neurology Department, the 4th Affiliated Hospital of Harbin Medical University, Harbin 150001, China)

Abstract: An electronic medical record (EMR) is a patient's individual medical record written by health care providers and stored in digital format in which much medical knowledge and information about patient's personal health conditions are kept. The construction of annotated corpus for named entities and entity relations on EMR is a primary and fundamental task for information extraction which plays important role in clinical decision support, practice of evidence-based medicine, and other medical applications. Based on survey of current research on corpus construction for named entities and entity relations on EMR, this research proposes an annotation scheme for named entities and entity relations on Chinese electronic medical records (CEMR) according to characteristics of the records. Under the supervision of physicians, a complete and detailed annotation specification on CEMR is formulated, and an annotated corpus with high agreement is constructed. The corpus comprises 992 medical text documents, and inter-annotator agreement (IAA) of named entity

* 收稿时间: 2014-12-03; 修改时间: 2015-06-24; 采用时间: 2015-08-03; jos 在线出版时间: 2016-03-22
CNKI 网络优先出版: 2016-03-22 13:23:21, http://www.cnki.net/kcms/detail/11.2560.TP.20160322.1323.001.html

annotations and entity relation annotations attain 0.922 and 0.895, respectively. The work presented in this paper builds substantial foundations for the subsequent research on information extraction in CEMR.

Key words: Chinese electronic medical record; named entity; entity relation; annotation specification; annotated corpus construction

健康是人们最宝贵的财富,随着经济的发展,人们对自己的健康和社会所能提供的医疗服务越来越关注.目前,有限的医疗资源和医疗服务水平不能满足人们日益增长的需求,不利于医患关系的改善.为缓解这种矛盾,我国于 2009 年颁布的《关于深化医药卫生体制改革的意见》就已明确提出要建立实用共享的医药卫生信息系统,对医疗的每一个环节的信息技术应用都提出了更高的要求,重点建立医院电子病历管理系统和居民健康档案,旨在实现统一高效、互联互通的医疗服务信息平台.患者的电子病历贯穿医疗活动的始终,是医疗信息系统的核心数据.

电子病历(electronic medical record,简称 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],是由医务人员撰写的面向患者个体描述医疗活动的记录.为了规范电子病历系统的实施,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 1 所示),即可了解电子病历数据量的庞大.电子病历由医务专业人员撰写,不仅是具有法律效力的医疗活动证据,而且包含大量的专业医疗知识.通过分析电子病历能够挖掘出这些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或概念,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].近年来,在电子病历文本上应用自然语言处理、信息抽取等技术服务于临床决策支持的研究倍受关注[4].这个过程分为两个不同的阶段:自然语言处理研究主要关注病历文本的预处理,包括句子边界识别、词性标注、句法分析等;信息抽取以自然语言处理研究为基础,主要关注病历文本中各类表达医疗知识的命名实体或医疗概念的识别和关系抽取[5].

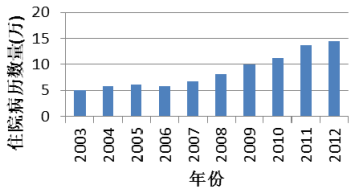


Fig.1 Statistics of in-patient records from medical record room of the 2nd affiliated hospital of Harbin medical university

图 1 哈尔滨医科大学附属第二医院病案室住院病历统计数据

海量的电子病历数据堪称医疗领域的大数据,是一座知识的宝库,蕴含了大量的医疗知识和患者的健康信息.电子病历数据不应只是封存在病案室里,而应得到有效利用.如何利用电子病历数据支持生物医学研究和临床研究,是医学信息学(medical informatics)和转化医学(translational medicine)的重要研究内容[6].医学信息学可简单地定义为系统地处理有关药品和临床治疗的信息、数据和知识的新兴学科[7],其两个重要分支——临床信息学(clinical informatics)和用户健康信息学(consumer health informatics),都与电子病历信息抽取密切相关.临床信息学主要研究利用信息技术实现临床决策支持(clinical decision support),改善临床治疗效果[8],电子病历是其重要的基础数据.临床信息学的应用领域主要是基于信息技术的循证医学(evidence-based medicine)[9]和电子病历系统的智能支持.病历电子化使得大规模病历的自动分析成为可能,由于电子病历记录了患者的疾病和症状、治疗过程和治疗效果,这些信息是重要的临床证据,自动抽取这些信息能够更加高效、精确地收集证据辅

助决策,促进循证医学这种数据驱动的医疗方法[10,11].电子病历已经成为和生物医学文献同等重要的循证医学实践的源数据.尽管电子病历系统提升了医生的工作效率,但仍然成为医生工作的负担,尤其表现在书写病程记录上,这也影响到了电子病历数据的质量[12−14].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[15,16].为了促进和规范电子病历系统智能支持的实施,美国和欧洲推出了电子病历系统分级实施模型,中国也于 2010 年推出电子病历系统功能应用水平分级评价方法及标准[17].卓越的临床智能支持是电子病历系统分级的主要依据,而临床智能支持的研究与实现必须立足于已有电子病历数据和生物医学文献的信息抽取和知识挖掘.随着医学信息学的发展和医疗信息化的普及,患者历次就诊的电子病历可聚集起来生成终身个人健康记录(personal health record)[18],一个典型案例可参见文献[19].通过分析个人健康记录,可以抽取患者个性化的健康知识,进而为患者个人需求、偏好建立模型并整合到医疗信息系统中,实现个性化医疗服务[20],这是用户健康信息学研究的主要内容之一[21,22].另外,基础医学研究和临床治疗之间的转化医学研究[23]也离不开对电子病历的分析处理.这方面的代表性工作主要体现在 I2B2(informatics for integrating biology and the bedside)[24]历年组织的与电子病历信息抽取相关的评测.I2B2 从 2006 年开始组织了一系列面向病历信息抽取的评测,并发布了共享语料集[25−31],这些评测任务和数据集使得临床研究者能够在现成的数据集上展开研究.以命名实体识别和实体关系抽取为主要研究内容的电子病历信息抽取研究引起了广大研究者的重视,该研究在英文病历上已经全面展开,而在中文病历上的研究却刚刚起步.

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研究大多关注于住院病历,并且只限于文本数据的挖掘.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(半结构或无结构)数据,如病程记录和出院小结等.这种文本信息方便表达概念以及事件等,是临床治疗过程的主要记录形式.结构化的数据处理起来相对容易,因而这些自由文本是电子病历命名实体识别和实体关系抽取的主要研究对象.当前,大多数命名实体识别和实体关系抽取方法是基于统计机器学习方法,并且在开放领域已经趋于成熟.电子病历文本具有半结构化特点和鲜明的子语言[32]特点,文献[33,34]分别对英文病历和中文病历的文本特点进行了总结.由于病历文本的特殊性以及统计机器学习方法的固有局限性,开放领域的研究成果很难应用于病历文本上.因而,展开电子病历命名实体识别和实体关系抽取研究首当其冲的就是构建标注语料库.正如 Roberts[35]所指出的,构建标注语料库有 3 个方面的主要原因:1) 标注体系清晰地界定了抽取任务的目标;2) 标注语料用于评价抽取系统的性能;3) 标注语料用于开发抽取系统(比如训练机器学习模型).因此,构建高质量的标注语料库对电子病历命名实体识别和实体关系抽取至关重要.然而,中文电子病历信息抽取研究领域还没有一个标注完整、规模较大、开放共享的命名实体和实体关系标注语料库.

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.这些研究将为临床决策支持、循证医学实践和疾病监控等提供支持,从而提高医疗服务质量.电子病历命名实体和实体关系标注语料库的构建将为这些研究打下坚实的基础.本文对国内外已有的电子病历命名实体和实体关系标注语料库构建工作进行了细致的调研,并指出其不足之处;在此基础上提出适合中文电子病历的命名实体和实体关系的标注体系,并制定了详细的标注规范;在住院医生的指导下,选取电子病历中的出院小结和首次病程记录作为标注对象,构建了迄今为止规模最大的命名实体和实体关系标注语料库;基于当前已完成的工作,规划了今后的工作计划,并展望了未来的研究方向;文章最后对本文工作进行了总结.

1 相关工作

英文电子病历命名实体和实体关系标注语料库构建工作起步较早,其标注体系、标注方法和一致性评价(Inter-annotator agreement,简称 IAA)对中文电子病历语料库的构建是非常重要的参考.病历中的医疗问题(也就是疾病和症状)是首先受到关注的信息,Meystre 等人[36]于 2006 年构建了涉及 80 种常见的医疗问题的命名实体标注语料,医疗问题是否发生在患者身上也是很重要的信息,因此,该语料库针对每一个医疗问题同时标注了修饰信息,即,当前发生(present)还是当前没有发生(absent).该语料库包含 160 份文档,文档类型包括出院小结、病

程记录等10个类型,每份文档同时由两位医生标注,不一致的标注由第3位医生解决.由于该语料只标注了医疗问题,并且限定在指定范围内,修饰信息的分类比较粗糙,规模也不大,因此实用性受到了限制.著名医疗机构梅奥诊所为实现临床文本分析和知识抽取系统(clinical text analysis and knowledge extraction system,简称cTAKES)[37]于2008年构建了包含160份文档的命名实体语料,文档类型为门诊记录、出院小结、住院记录等3种,语料由医疗专家标注,一致性采用F值评价,只标注了疾病,疾病的界定参考美国国家医学图书馆建立集成的医疗知识库UMLS(unified medical language system)[38]所定义的10个语义类型,并首次对修饰信息进行了细致的分类.该语料用于识别病历中的疾病并映射疾病编码,没有涉及其他类型的命名实体.在临床电子化科学体系CLEF(clinical e-science framework)项目中,为了构建临床重要信息抽取系统,Roberts等人[35]于2009年随机抽取了临床记录文本、射线报告、病理报告各50份,标注了命名实体、修饰、实体关系(比如检查发现了结果,药物作用于症状)以及时间信息,首次把实体类型扩展到6类,并首次涉及到关系的标注和时间信息的标注,标注工作由医学专家完成,一致性评价采用计算相同标注的百分比.该语料在标注体系上进行了大胆的尝试,极大地启发了后继的语料构建工作.2010年,I2B2组织了概念抽取和概念关系抽取评测任务[29],并发布了包含871份文档的标注语料,其标注体系与Roberts等人的标注体系有相似的设计,但明显比Roberts等人的设计要合理.电子病历中病程记录的撰写方式基本采用Weed[39]于1968年提出的面向问题的组织方式(problem- oriented medical record,简称POMR).这种记录方式以医疗问题为中心,检查和治疗都围绕医疗问题而展开,有助于对医疗问题的治疗情况和进展进行跟踪和分析.因而,I2B2 2010语料的实体类型为医疗问题(medical problem)、检查(test)和治疗(treatment),这3类实体的界定参考UMLS定义的语义类型;实体间关系主要是3大类,即医疗问题和检查的关系、医疗问题和治疗的关系、医疗问题和医疗问题的关系;医疗问题的修饰类型也进行了仔细的划分,一共分为6类;标注工作由医疗专家完成,一致性评价采用F值.该语料库的标注体系是目前最合理的,因为充分考虑了临床实践过程和病历文本的组织方式.但是该语料把疾病和症状合并为医疗问题,没有作区分,而在医疗实践中,疾病和症状是要区别对待的.事实上,I2B2 2010的组织者Uzuner[3]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.电子病历命名实体识别和实体关系抽取不是一个孤立的任务,还依赖于对病历文本的词法句法分析等基础研究.鉴于此,Daniel 等人[40]于2013年构建了一个包含词性、短语结构句法树、谓词-论元结构、命名实体的多层标注语料库,以句子为标注单位,一共包含13 091句,一致性采用F值评价.值得注意的是,该语料在实体的标注体系中把疾病和症状分开为两种不同的类型.该语料最大的优点是在相同的数据上实施多层标注,便于训练联合模型,但由于是在语句级上标注,病历文本的半结构化信息丢失,这类半结构化信息对命名实体识别有着重要的提示作用.病历文本是医疗专业人员撰写的记录,非专业人员理解起来存在较大障碍.针对这种问题,CLEF(conference and labs of the evaluation forum)健康评估实验室于2013年发起了病例信息抽取和相关信息检索的评测[41],旨在帮助非专业人员看懂病历,评测在第1个任务中发布了只标注医疗问题的命名实体语料库ShARe(shared annotated resources),该语料继续用于2014年评测的第2个任务,同时也用于2014年SemEval第7个任务[42].另外,Mizuki等人[43]为了在日文病历上展开命名实体识别研究,构建了包含50份文档的语料,这些文档只标注了主诉和诊断,由于难以获得真实的电子病历,这50份病历文档全部是由医生虚构的.英文病历上的命名实体语料构建工作取得了丰硕的成果,其代表性的语料库就是I2B2在2010年构建的语料,分类体系最完整,规模最大,而且公开给其他研究者,组织共享评测任务.与I2B2从临床医生的需求角度抽取医疗信息不同,CLEF 健康评估实验室致力于帮助普通用户理解病历这种专业文本展开信息抽取研究,构建的标注语料ShARe已用于3次共享评测任务.这些研究内容以及公开的语料库极大地启发了中文电子病历命名实体语料库的构建.

中文电子病历名命名实体识别研究起步较晚,目前都没有公开,典型的研究有4个.Lei等人[44]于2013年构建了包含800份文档的命名实体语料库,400份入院记录,400分出院小结,电子病历来源于北京协和医院,命名实体的分类借鉴2010年I2B2的实体分类,把治疗细分为药物(medication)和过程(procedure),标注工作由两名医生完成,同时标注了语料中40份文档,用于一致性评价.中文电子病历文本的分词影响着后续的信息抽取研究,为了训练分词和命名实体识别的联合模型,Xu等人[45]于2013年构建了包含336份出院小结并且同时标注分词

和命名实体的语料,实体的分类与 Lei 等人的分类体系基本相同,同时增加了人体部位(anatomy)这个类型,这类实体在临床中是很重要的信息.中医病历的信息抽取研究也逐渐受到中医临床研究者的重视,中医病历文本中医的症状、综合症、复方制剂、中草药等也是关键信息,这些是中医病历命名实体识别要抽取的关键信息[46].为了展开中医病历命名实体研究,Wang 等人[47]于 2014 年构建了标注症状名的语料,包含 11 613 条主诉,标注工作由医生完成.该语料标注的文本和实体类型比较单一.对于特定疾病的病历文本,已有的实体分类体系都显得过于粗糙,一些重要信息得不到有效区分.为了识别肿瘤相关的信息,Wang 等人[48]构建的肿瘤病历命名实体标注语料做了有益的尝试.该语料主要包含 115 份肿瘤患者的手术记录,病历来自复旦大学附属中山医院,从临床医生需求的角度定义了 12 种实体类型,语料标注由 3 位医生完成,2 位医生标注,1 位医生处理不一致标注,最终标注了 961 个实体.尽管标注的语料规模较小,但是对特定疾病病历信息抽取研究提供了较好的借鉴,主要是两条:一是针对特定疾病对要抽取的信息进行细致的划分,二是根据医生的需求设计标注体系.类似地,Coden 等人[49]针对英文癌症病历信息抽取构建了实体标注语料.针对特定疾病的信息抽取研究除了肿瘤病历以外,基于心血管疾病病历的信息抽取和风险预测研究也是研究者关注的热点[50],这类研究也需要根据心血管疾病的特点对抽取的信息进行细分.国内其他学者在电子病历命名实体识别研究中也尝试构建了一些小规模的标注语料,如文献[51].表 1 总结了国内外已构建的一些重要的电子病历命名实体和实体关系标注语料,并从语料规模、是否标注实体、是否标注修饰、是否标注关系这几个方面进行了比较.表 1 中,“√”表示已标注,“–”表示未标注,完整标注实体、修饰和实体关系的语料只有两个.经过比较,可以总结出中文病历的命名实体标注语料存在 3 点不足:(1) 标注的文本类型比较单一;(2) 命名实体的分类体系不够完整;(3) 没有标注医疗问题的修饰信息和实体关系.因为这些不足,再加上这些语料没有公开,已构建的这些语料库很难被其他研究者使用.因此,构建一个分类体系完整、文本类型和科室覆盖面大的、适合中文病历的命名实体和实体关系标注语料库势在必行.

Table 1 Summary of existing annotated corpuses for named entities and entity relations on electronic medical records

表 1 已构建的电子病历命名实体和实体关系标注语料总结

作者	年份	语言	规模	实体	修饰	实体关系
Meystre[36]	2006	英文	160(文档)	√	√	–
Savova(cTAKES)[37]	2008	英文	160(文档)	√	√	–
Roberts[35]	2009	英文	150(文档)	√	√	√
Uzuner(I2B2)[29]	2010	英文	871(文档)	√	√	√
Daniel[40]	2013	英文	13,091(句)	√	√	–
Mizuki[43]	2013	日文	50(文档)	√	√	–
Jianbo Lei[44]	2013	中文	800(文档)	√	–	–
Yan Xu[45]	2013	中文	336(文档)	√	–	–
Yaqiang Wang[47]	2014	中文	11613(句)	√	–	–
Hui Wang[48]	2014	中文	115(文档)	√	–	–

语料构建的核心工作是制定规范和依据规范标注.文献[52]总结了 3 种语料标注模式:(1) 传统的领域专家标注,这种标注模式适合于特定领域的语料标注,能够保证标注的质量,但是标注成本高,周期长;(2) 众包标注,这种标注模式能够以较低的成本标注较大规模的语料,但是只限于简单的标注任务,并且标注过程也需要精心的设计,最有代表性的众包标注平台是 Amazon 的 Mechanical Turk[53],2012 年,SemEval 第 2 个任务关系相似度计算[54]使用的语料就是在 Mechanical Turk 上构建的;(3) 团体标注,这种标注模式构建语料的过程类似于信息检索评价集的构建,组织者首先提供小批量的标注数据作为参照,从众多提交的结果中自动筛选部分结果再有人工确认解决不一致标注,该方法成功应用于 2009 年 I2B2 药品信息抽取评测的语料构建[55],在不依赖于专家的情况下,能构建出高质量的语料,对组织方式和团体的配合与支持要求很高.电子病历文本是医疗专家撰写的专业文档,比较这 3 种标注模式,领域专家标注模式仍然最适合电子病历语料的标注[52].事实上,已有的中英文命名实体标注语料的标注工作几乎全部由医疗专家完成,医疗专家甚至还参与到规范的制定和完善当中,由于涉及较多医疗领域知识,标注规范的制定不是一蹴而就的,需要根据标注结果不断地修改,比如 Roberts 等人[35]在语料构建过程中,采取逐轮迭代的方法不断完善规范.文献[52]总结了一些电子病历语料构建的经验,虽然标注

过程以医生为主导,但是自然语言处理研究者可以在规范制定、标注流程、标注培训、一致性评价、辅助标注工具开发等方面尽可能多地与医疗专家配合,这些经验为我们的中文电子病历命名实体和实体关系语料构建工作提供了宝贵的借鉴.

2 中文电子病历命名实体和实体关系标注体系的建立

通过分析电子病历,医生针对患者的诊疗活动可以概括为:通过检查手段(做什么检查)发现疾病的表现(什么症状),给出诊断结论(什么疾病),并基于诊断结论,给出治疗措施(如何治疗).从这个过程可以看出,医疗活动主要涉及 4 类重要信息:检查、症状、疾病和治疗.这 4 类信息在 UMLS 中也具有明确对应的语义类型定义.中文病历中对患者症状和检查结果的描述占有相当大的比重,因此在中文电子病历命名实体识别研究中,有必要把疾病和症状分开,并且定义疾病和症状之间的关系.中文电子病历命名实体识别主要研究以下几类实体的识别:

- 第 1 类实体是疾病,泛指导致患者处于非健康状态的原因(不包括不良生活习惯),或者医生根据患者的身体状况做出的诊断.疾病是可以治愈或改善的.
- 第 2 类实体是疾病诊断分类,一般紧跟一个具体的疾病,是疾病的一个具体分类,比如“高血压,极高危组”中的“极高危组”.
- 第 3 类实体是疾病的表现,在本研究中称为症状,泛指疾病导致的不适或异常感觉和显式表达的异常检查结果.虽然这两类症状都是疾病的表现,但又明显不同,因此症状细分为两个子类:自诉症状和异常检查结果.
- 第 4 类实体是检查手段,在本研究中简称为检查,泛指为了得到更多的由疾病导致的异常表现以支持诊断而采取的检查设备、检查程序、检查项目等.
- 第 5 类实体是治疗手段,在本研究中简称为治疗,泛指为了治愈疾病、缓解或者改善症状而给予患者的药物、手术等.

另外,医生在描述患者的疾病和症状时,通常都表达出不同的确定程度,这是诊断过程中的重要信息,比如肯定发生的、肯定不发生的(否认的)、可能发生的等等.这些信息在本规范中称为疾病和症状的修饰信息.患者曾经历过的治疗信息或者明确否认的既往治疗史也是临床诊断的重要信息,因此,针对治疗类实体,也要识别修饰信息.修饰信息的识别是电子病历命名实体识别研究独有的任务.

中文电子病历实体关系抽取研究主要关注这 6 类实体关系的抽取:治疗和疾病之间的关系,比如治疗施加于疾病;治疗和症状之间的关系,比如为缓解症状而施加的治疗;检查和疾病之间的关系,比如检查证实疾病;检查和症状之间的关系,比如检查发现症状;疾病和症状之间的关系,比如疾病导致症状;疾病和疾病诊断分类之间的关系,该关系表示疾病的进展程度.实体及实体之间的关系如图 2 所示,圆圈表示 5 类命名实体,连接两个圆圈之间的箭头表示两类命名实体之间的关系,箭头的方向表示实体关系的方向.

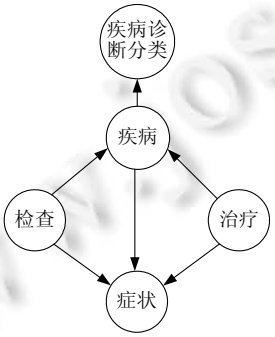


Fig.2 Medical entity types and relations between each other

图 2 医疗实体分类及实体间关系

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,医生可以预先快速地浏览病人的信息,后续再关注特定的细节.下面详细描述中文电子病历中的命名实体、实体修饰、实体关系的定义和分类.

2.1 命名实体分类

如前所述,命名实体的类型有疾病、疾病诊断分类、症状、检查、治疗这 5 类.借鉴 I2B2 对概念类型的定义方法,本研究使用 UMLS 语义类型界定每一类实体涵盖的范围,涉及到的语义类型也参照 I2B2 选用的语义类型.采用语义类型来确定实体类型的范围可看成是采用医疗领域的惯例对实体类型进行细分,使规范更具可操作性.本研究所定义的命名实体遵循实体间不重叠、不嵌套、实体内不含有表示停顿的标点符号(比如逗号、句号、顿号等)这 3 个原则.

2.1.1 疾病(disease)

在本研究里,疾病是个宽泛的概念.导致患者处于非健康状态的原因或者医生对患者做出的诊断统称为疾病,其对应的 UMLS 语义类型有:疾病或者综合征(disease or syndrome)、受伤或中毒(injury or poisoning)、先天性畸形(congenital abnormality)、病毒细菌(virus/bacterium)、病理功能(pathologic function)、细胞或分子功能障碍(cell or molecular dysfunction)、获得性异常(acquired abnormality)、解剖异常(anatomic abnormality)、肿瘤进程(neoplastic process)、精神或行为障碍(mental or behavioral dysfunction)等.疾病必须是能够被治疗的,并且能够被否定词修饰.否定词是指病历中描述病史时经常使用的一些表示否定的词,比如“否认高血压”和“无心脏病史”中的“否认”、“无”.几个典型的疾病类实体如下:

1) 老年女患,否认高血压、糖尿病史(“高血压”和“糖尿病史”).
2) 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”).
3) 交通意外致头部外伤(“头部外伤”).
4) 查 EB 病毒,巨细胞病毒(“EB 病毒”和“巨细胞病毒”).

2.1.2 疾病诊断分类(disease type)

在诊断里,通常出现对某个诊断疾病的分类信息,如 II 型、极高危组.这类信息不是疾病名,而是对疾病的一个具体分类,表示疾病的进展程度,因此引入疾病诊断分类这类实体,这类实体通常出现在诊断里,并且一般紧跟一个具体的疾病.比如:

1) 肝硬化 失代偿期(“失代偿期”).
2) 多发性骨髓瘤 轻链型 III 期 A(“轻链型”和“III 期 A”).
3) 糖尿病 II 型(“II 型”).

2.1.3 症状(symptom)

在本研究里,症状区别于临床医疗上的症状概念,泛指由疾病导致的不适表现或者异常表现、显式表达的异常检查结果,其对应的 UMLS 语义类型主要是症状或体征(symptom or sign).症状是能够被治疗手段改善或治愈的,并且能够被否定词修饰.在本研究里,症状作为疾病的表现,可分患者的自诉症状和医生的检查结果.患者的自诉症状取决于患者的感受,有较大的主观性;而医生(通过检查设备)检查到的结果是比较客观的发现.因此,在临床诊断上这两类疾病的表现作为诊断依据的重要性是不一样的;而且在病历中,这两类信息也是分开记录的.因此,根据医生的建议,我们把症状进一步分为自诉症状和异常检查结果两个子类,相当于把语义类型“症状或体征”细分为症状(symptom)和体征(sign).

自诉症状是指患者自己向医生陈述(或是别人代述)的不适感觉或异常感觉,比如下面的例子:

1) 疼痛时伴有右下肢活动受限(“疼痛”和“右下肢活动受限”).
2) 伴活动后心慌气短(“心慌”和“气短”).
3) 伴出汗、 乏力、恶心、略感气短(“出汗”、“乏力”和“气短”).

异常检查结果是指医生观察到的或者通过检查程序或设备检查到的发生于患者的异常变化以及异常检查结果,并且显式地表明是异常的,比如下面的例子:

4) 因肌酐高做腹膜透析时也有恶心呕吐(“肌酐高”).

5) 双肺听诊可闻及少量痰鸣音(“痰鸣音”).

6) 自带胸片示左下肺炎症病变(“左下肺炎症病变”).

2.1.4 检查(test)

检查是指为了发现、证实疾病或症状,找到更多关于疾病或症状的信息而施加给患者的检查过程、仪器等,也包括检查项目,对应的 UMLS 语义类型有:化验过程(laboratory procedure)、诊断过程(diagnostic procedure)等.检查只是为了寻找更多跟疾病或症状相关的信息,并不能治疗疾病或者缓解症状,比如下面的例子:

1) 头 CT 显示脑实质内高密度灶(“头 CT”是辅助检查).

2) 血压最高达到 180/130mmHg(“血压”是检查项目).

3) 心肺听诊无著征(“心肺听诊”是专科检查).

2.1.5 治疗(treatment)

治疗是指为了解决疾病或者缓解症状而施加给患者的治疗程序、干预措施、给予药品,其对应的 UMLS 语义类型有:药物(pharmacologic substance)、治疗或预防过程(therapeutic or preventive procedure)、药物输送设备(drug delivery device)、医疗设备(medical device)、类固醇(steroid)、生物医学或牙科材料(biomedical or dental material)、抗生素(antibiotic)、临床药物(clinical drug)等.治疗是指能够治疗疾病或者缓解症状的医疗概念,这是与检查最大的不同,比如下面的例子:

1) 奥扎格雪、脑蛋白水解物等静点(“奥扎格雪”和“脑蛋白水解物”是药物).

2) 4 年前行胆囊切除术(“胆囊切除术”是治疗过程).

3) 鼻内镜下行双筛、双上颌窦(“鼻内镜”是治疗设备).

2.2 实体修饰分类

实体的修饰也叫断言,修饰信息反映了实体与患者的关系,该关系体现在两个方面:是否发生于患者本人、发生于患者本人的确定程度.从这个角度理解,修饰信息体现了电子病历中医疗知识的患者个性化特点.修饰信息对于正确理解病历至关重要,因此,电子病历命名实体识别同时要研究实体修饰的识别,也就是对识别出来的实体(包括疾病、症状和治疗)在预定义的修饰类型上进行分类.在本研究中,我们定义了疾病、症状以及治疗的修饰,并举出示例(实例中的斜体字表示对应修饰的重要指示词).

2.2.1 疾病和症状的修饰

疾病和症状的修饰一共有 7 个,分别是否认(absent)、非患者本人(family)、当前的(present)、有条件的(conditional)、可能的(possible)、待证实的(hypothetical)、偶有的(occasional).我们从实体与患者关系角度来定义这个 7 个修饰.

- 在是否发生患者本人这个方面有两个修饰.
  - (1) 否认:患者主动否认、或肯定不发生于患者身上,比如:
    各瓣膜区**未闻及**病理性杂音;
    全腹**无**压痛、反跳痛及肌紧张;
    腹壁静脉曲张:**无**.
  - (2) 非患者本人:发生于患者家属,该种修饰可能和“否认”重叠.若发生此种情况,选择否认.比如:
    其**父母**均患有糖尿病.
- 在发生于患者本人的确定程度这个方面有 5 个修饰.
  - (3) 当前的:肯定发生或正在发生于患者本人的疾病和症状,比如:
    头晕、呕吐**伴**右下肢无力;
    自诉**有**冠心病史;
    头 CT **示**:双侧多发腔梗.
  - (4) 有条件的:当前不一定发生,在某种条件具备的情况下才发生.比如:
    该患者于入院前 3 个月开始出现阵发性胸闷、心慌,常于**饮酒后出现**.

(5) 可能的:不确定当前会发生,需要进一步的证据才能确定.比如:
不排除缺血性疾病;右肺中下叶**考虑**创伤性湿肺;
临床**初步诊断**:脑梗死、高血压病、糖尿病.

(6) 待证实的:当前不会发生,但预期会发生.比如:
手术一周后**会有**局部瘙痒;
多在**皮疹出现后** 1~4 周左右出现血尿和(或)蛋白尿.

(7) 偶有的:指症状或者疾病当前不经常出现,或者出现的频率较低.比如:
病程中患者走路不稳,**偶有**头晕;
大便**偶有**一过性发白;
时有胸闷气短.

2.2.2 治疗的修饰

患者是否既往经历过某种治疗对临床诊断有重要参考作用,尤其是手术类治疗手段.治疗的修饰信息主要有 3 类:既往的(history)、否认的(absent)、当前的(present).

(1) 既往的:明确表示是患者过去经过的治疗史,比如:
有多次输血**史**;
18 年前剖宫产手术;
后自行间断口服拜糖平及二甲双胍 **8 天**;
胃溃疡穿孔切除术**史**.

(2) 否认的:对既往治疗史的否认,比如:
未接种疫苗;
否认人流术史.

(3) 当前的:表示治疗是患者当前正在经历的或者即将要经历的,这类治疗通常出现在首次病程记录里的治疗计划和医嘱以及出院小结里的治疗经过里面.比如:
改善脑循环;
保护脑组织;
营养神经;
抗炎、化痰.

2.3 实体关系分类

实体关系抽取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.如图 2 所示,本研究将实体关系分为 6 大类,这 6 类关系只限于一个句子范围内实体之间的关系,跨句子范围的关系不作考虑.

在中文电子病历中,同一类别的实体经常一起出现,这些同类别实体间通常具有并列关系,比如并发、伴随、配合施治等关系.而且,这些同类别实体与其他实体存在着相同的关系.要把这些关系拆分成单个实体间关系要么缺乏明确的一一对应关系而无法拆分,要么特别琐碎.比如下面的例子:

2012 年 6 月出现咳嗽,咳黄痰伴有血丝,在家自行口服消癌平中药抗肿瘤及静点青霉素抗炎治疗,上诉症状未缓解.

在这个例子中,治疗类实体是“消癌平”、“青霉素”,症状类实体是“咳嗽”、“咳黄痰”、“血丝”.这段描述没有明确指出哪个治疗是针对哪个症状的,因此我们很难判断出其一一对应关系.另一方面,建立单个实体间的关系也比较繁琐.这个例子中症状间的伴随关系有 3 个,治疗间的配合关系有 1 个,治疗和症状间的关系有 6(=2×3)个,一共是 10 个关系,繁琐程度可见一斑.为了解决这个问题,我们引入实体组这一概念,每一类实体都有对应的实体组.我们把出现在一个句子中的同时满足以下两个条件的同一类实体构成一个实体组.

1) 同时性;

2) 与句内的其他实体或实体组具有相同的关系.

上述条件中的第 1 个条件是实体组的内部约束,同时性是指两个相同类型的实体在一次医疗活动期间同时发生,反映了实体间的并列关系.对疾病(症状)说是并发或者伴随的关系,对治疗(检查)来说是配合或者协同的关系.第 2 个条件是实体组的外部约束,反映了不同类型实体间的关系,比如治疗和症状的关系,这类关系是本研究关注的重点.

引入实体组之后,实体关系抽取任务在实现上可考虑采取两种不同的思路:一是先识别实体组,而后识别两个实体组的关系类型;二是先识别两个实体间的关系,然后根据实体关系类型合并实体得到实体组,进而得到实体组之间的关系.

实体组的引入弱化了单个实体间的关系,从而解决关系模糊和繁琐的问题.引入实体组来定义关系其实也符合了临床诊疗的习惯.临床上,在根据患者的症状得出患者当前患有的疾病时,医生并不是只根据某一个症状来判断患者的疾病,而是根据一组症状来综合判断出患者的疾病;而且在治疗的时候,也是多种治疗手段配合一起使用.定义了实体组(下文中用大括号“{}”表示实体组),下面给出实体关系的分类.

2.3.1 治疗和疾病的关系

基于治疗对疾病产生效果,治疗和疾病的关系分为 5 种.

(1) 治疗改善了疾病(TrID),表示治疗改善或者治愈了疾病.比如:
高血压病史 20 年,平素口服波依定,代文控制在 130/90mmHg 左右(治疗组{“波依定”,“代文”}改善了“高血压”).

(2) 治疗恶化了疾病(TrWD),表示治疗没有改善也没有治愈疾病,或者恶化了疾病.比如:
糖尿病皮下注射胰岛素诺和灵 30R 控制,血糖控制不佳(治疗“胰岛素诺和灵 30R”恶化了疾病“糖尿病”).

(3) 治疗导致了疾病(TrCD),表示治疗不是针对该疾病的,而是导致了该疾病.比如:
2008 年行输尿管镜下手术,自称术后肾脏破裂、出血(治疗“输尿管镜下手术”导致了疾病组{“肾脏破裂”,“出血”}).

(4) 治疗施加于疾病(TrAD),表示治疗是施加于该疾病的,但是结果没有提及.比如:
既往糖尿病史 20 余年,皮下注射胰岛素控制,血糖控制不详(治疗“胰岛素”施加于“糖尿病”).

(5) 因为疾病而没有采取治疗(TrNAD),因为该种疾病而不采取治疗或中断治疗,且该种疾病不是该治疗导致的.比如:
如果无手术禁忌症,行骨折切开复位内固定术治疗(因为“手术禁忌症”而不采取“骨折切开复位内固定术治疗”).

2.3.2 治疗和症状的关系

类似于治疗和疾病的关系分类,基于治疗对症状产生的效果,治疗和症状的关系分为 4 类:

(1) 治疗改善了症状(TrIS),表示治疗改善或消除了症状.比如:
患者诉规律服用钙剂等治疗后,后背部疼痛显著缓解,余无不适(治疗“钙剂”改善了症状“后背部疼痛”).

(2) 治疗恶化了症状(TrWS),表示治疗没有缓解也没有治愈症状,或者恶化了症状.比如:
2012 年 6 月出现咳嗽,咳黄痰伴有血丝,在家自行口服消癌平中药抗肿瘤及静点青霉素抗炎治疗,上诉症状未缓解(治疗组{“消癌平”,“青霉素”}恶化了症状组{“咳嗽”,“黄痰”,“血丝”}).

(3) 治疗导致了症状(TrCS),表示治疗不是针对该症状的,而是导致了该症状.比如:
后继续口服复方造矾丸治疗,血小板计数明显上升(治疗“复方造矾丸治疗”导致了症状“血小板计数明显上升”).

(4) 治疗施加于症状(TrAS),表示治疗是施加于症状的,但是结果没有提及.比如:
颜面部伤口多处擦伤,伤口疼痛出血,伤口敷料包扎(治疗“伤口敷料包扎”施加于症状组{“伤口疼

痛”,“出血”}).

(5) 因为症状而没有采取治疗(TrNAS):因为该种症状而不采取治疗或中断治疗,且该种症状不是该治疗导致的.比如:

2 周前在当地医院复查血常规及肝功时发现转氨酶高,停用达那唑,出现咳嗽,无咳痰(因为“转氨酶高”而不采取“达那唑”).

2.3.3 检查和疾病的关系

(1) 检查证实了疾病(TeRD),表示疾病是通过检查确诊的.比如:

头 MRI 示:腔隙性脑梗死(检查“头 MRI”证实了疾病“腔隙性脑梗死”).

(2) 为了证实疾病而采取检查(TeCD),表示为了证实疾病而采取某种检查手段,但结果未知.比如:

患者病情尚不除外脑炎,建议腰穿,患者家属拒绝该项检查(为了证实疾病“脑炎”而采取检查“腰穿”).

2.3.4 检查和症状的关系

类似于检查和疾病的关系,检查和症状的关系有 2 类:

(1) 检查发现了症状(TeRS),表示某种症状视同检查发现的.比如:

辅助检查:心脏彩超示:左房稍大,主动脉弹性减低,左室顺应性减低(检查“心脏彩超”发现了症状组{“左房稍大”,“主动脉弹性减低”,“左室顺应性减低”}).

(2) 因为症状而采取检查(TeAS),根据患者的症状而采取的检查项目或检查手段.比如:

4 年前再次出现发热,鼻出血,当地查血常规示血小板少(因为症状组{“发热”,“鼻出血”}而采取检查“血常规”;检查“血常规”发现了症状“血小板少”).

2.3.5 疾病和症状的关系

疾病和症状的关系只有一种,就是疾病导致了症状(DIS),这种关系也表示疾病的确诊是由于症状发生了.比如:

胆囊炎,提示:右肾增大伴其内密度减低疾病“胆囊炎”导致了症状组{“右肾增大”,“密度减低”});

患者于 20 多天前无诱因出现口角不自主流涎,并有轻度咳嗽、咳痰,遂至医院就诊,行头颅 CT 检查提示“脑出血”并住院治疗(疾病“脑出血”导致了症状组{“口角不自主流涎”,“轻度咳嗽”,“咳痰”}).

2.3.6 疾病和疾病诊断分类的关系

该类关系只有一种,数量较少,表示疾病诊断分类是疾病的一个具体类别(DADt).比如:

肝炎后肝硬化 乙型 丙型(疾病“肝炎后肝硬化”的类别是{“乙型”,“丙型”});

肝硬化 失代偿期(疾病“肝硬化”的类别是“失代偿期”).

2.4 和已有的标注体系对比

本研究的标注体系借鉴了 2010 年 I2B2 语料库的标注体系,并基于中文电子病历文本的特点而设计的.对比已有的中英文病历文本上命名实体和实体关系的标注体系,本研究的标注体系主要有 3 个创新之处:

(1) 把医疗问题拆分为疾病和症状,并且进一步把症状细分为自诉症状和异常检查结果两个子类.

(2) 既往经历过的治疗措施影响着当前治疗方案,因此,给治疗类实体也标注了修饰信息.

(3) 考虑到临床诊疗活动的习惯以及中文病历中文本表述特点,引入实体组,并把实体关系定义为实体组之间的关系.这一点是与已有标注体系最大的不同,也是对信息抽取研究中实体关系的进一步扩展.

3 中文电子病历命名实体和实体关系标注规范的制定和语料构建

基于上一节我们提出的命名实体和实体关系的标注体系,我们制定了完整的命名实体和实体关系标注规范[56],开发了标注工具[57],对 992 份真实病历文本进行了命名实体和实体关系标注实践,整个过程历时一年半.在标注语料构建队伍中,有两名医生深度参与:一位是哈尔滨医科大学附属第二医院呼吸内科住院医生(医学博士),另一位是哈尔滨医科大学附属第四医院神经内科住院医生(医学硕士).医生的工作主要是辅助我们完善标注规范,并标注全部语料的命名实体.根据表 1 所列出的一些重要语料的规模数据来看,本文所构建语料的规模

是最大的.下面从 4 个方面介绍我们的语料构建工作.

3.1 病历文本数据准备

哈尔滨医科大学附属第二医院较早实施电子病历系统,到 2011 年底,该医院各科室已全面实施电子病历系统.对比最新国家颁布的电子病历规范,该医院的电子病历基本符合规范,所以本研究用于构建语料的电子病历来源于该医院病案室 2012 年住院病历.电子病历数据的形式主要有表格、自由文本、图像这 3 种,自由文本形式的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疗活动过程;主诉、现病史和病历小结的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.因此,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,本研究使用的病历文本主要选择首次病程记录和出院小结.所选病历涵盖该医院全部 35 个科室,子科室分为 87 个,为了保证语料能够均匀覆盖到所有科室,病历抽样是按照子科室平均抽样的.各科室抽取病历的数量分布图如图 3 所示,图中每一种颜色对应一个科室,我们只标出了病历数量最大的 4 个科室.

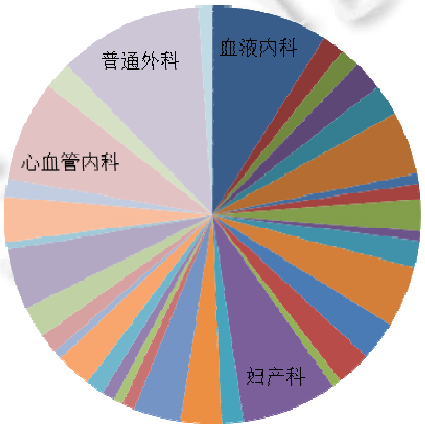
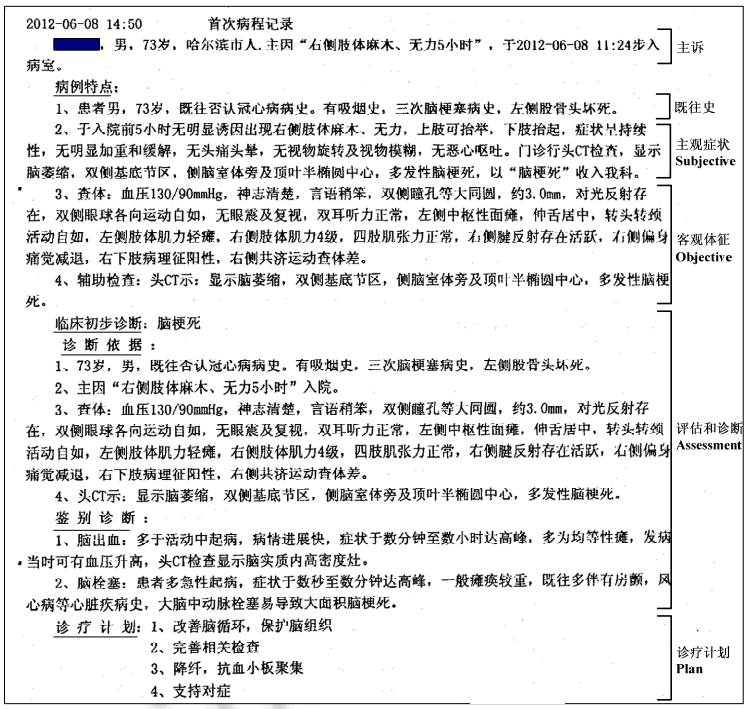


Fig.3 Percentages of annotated medical records from different departments

图 3 各科室标注病历的数量分布图

抽取出的病历文本包含患者和医生的隐私信息,所以删除隐私信息是第 1 步要做的工作.隐私信息主要包括患者的姓名、证件号码、家庭住址、工作单位,医生的隐私主要是姓名,文本中出现的医疗机构名称也是隐私信息.我们通过规则匹配和人工检查的方式删除隐私信息.

出院小结和首次病程记录是半结构化文本数据,每一部分文本有独立的描述功能,并且有一个标题说明该文本描述的内容,比如“病例特点”、“临床初步诊断”、“诊疗计划”等.这种半结构化信息可以很好地指导命名实体的识别,特别是实体类型的判断,并且在标注时也能起到提示作用.图 4 和图 5 的病历文本样例展示了文本的半结构化特点.病程记录基本上按照面向医疗问题的方式(POMR)组织内容,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做出诊断,最后给出相应的诊疗计划.这种组织可以从图 5 所示的首次病程记录看出,并且也印证了把症状分为主观的自诉症状和客观的检查结果是合理的.为了在以后的命名实体识别和实体关系抽取研究中能够方便地提取文本的半结构化信息,我们把病历文本格式化为 XML 格式,标题表示为 XML 的节点.


2012-06-08 14:50 首次病程记录

[redacted]，男，73岁，哈尔滨市人.主因“右侧肢体麻木、无力5小时”，于2012-06-08 11:24步入病室。 主诉

病例特点：

1、患者男，73岁，既往否认冠心病病史。有吸烟史，三次脑梗塞病史，左侧股骨头坏死。 既往史

2、于入院前5小时无明显诱因出现右侧肢体麻木、无力，上肢可抬举，下肢抬起，症状呈持续性，无明显加重和缓解，无头痛头晕，无视物旋转及视物模糊，无恶心呕吐。门诊行头CT检查，显示脑萎缩，双侧基底节区，侧脑室体旁及顶叶半椭圆中心，多发性脑梗死，以“脑梗死”收入我科。 主观症状 Subjective

3、查体：血压130/90mmHg，神志清楚，言语稍笨，双侧瞳孔等大同圆，约3.0mm，对光反射存在，双侧眼球各向运动自如，无眼震及复视，双耳听力正常，左侧中枢性面瘫，伸舌居中，转头转颈活动自如，左侧肢体肌力轻瘫，右侧肢体肌力4级，四肢肌张力正常，右侧腱反射存在活跃，右侧偏身痛觉减退，右下肢病理征阳性，右侧共济运动查体差。

4、辅助检查：头CT示：显示脑萎缩，双侧基底节区，侧脑室体旁及顶叶半椭圆中心，多发性脑梗死。 客观体征 Objective

临床初步诊断：脑梗死

诊 断 依 据 ：

1、73岁，男，既往否认冠心病病史。有吸烟史，三次脑梗塞病史，左侧股骨头坏死。

2、主因“右侧肢体麻木、无力5小时”入院。

3、查体：血压130/90mmHg，神志清楚，言语稍笨，双侧瞳孔等大同圆，约3.0mm，对光反射存在，双侧眼球各向运动自如，无眼震及复视，双耳听力正常，左侧中枢性面瘫，伸舌居中，转头转颈活动自如，左侧肢体肌力轻瘫，右侧肢体肌力4级，四肢肌张力正常，右侧腱反射存在活跃，右侧偏身痛觉减退，右下肢病理征阳性，右侧共济运动查体差。

4、头CT示：显示脑萎缩，双侧基底节区，侧脑室体旁及顶叶半椭圆中心，多发性脑梗死。

鉴 别 诊 断 ：

1、脑出血：多于活动中起病，病情进展快，症状于数分钟至数小时达高峰，多为均等性瘫，发病当时可有血压升高，头CT检查显示脑实质内高密度灶。

2、脑栓塞：患者多急性起病，症状于数秒至数分钟达高峰，一般瘫痪较重，既往多伴有房颤，风心病等心脏疾病史，大脑中动脉栓塞易导致大面积脑梗死。 评估和诊断 Assessment

诊 疗 计 划：1、改善脑循环，保护脑组织

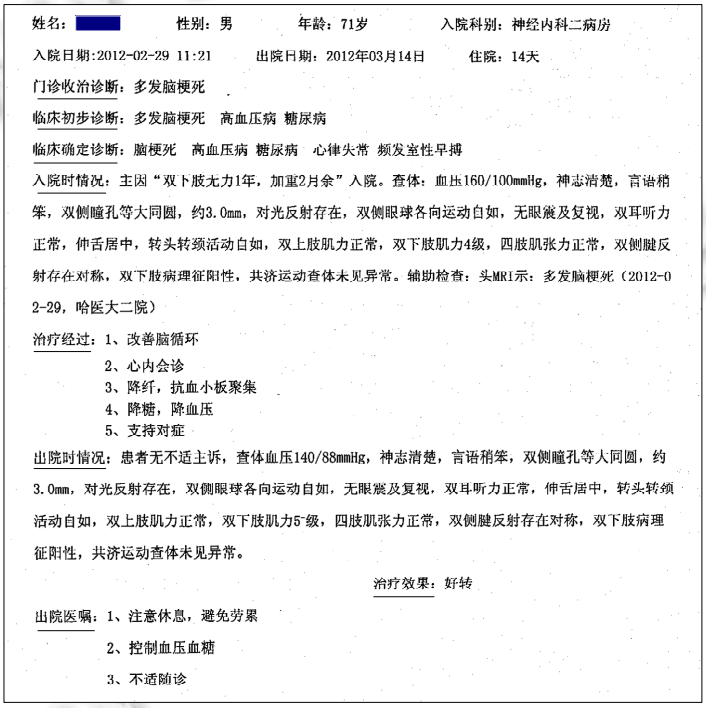
2、完善相关检查

3、降纤，抗血小板聚集

4、支持对症 诊疗计划 Plan

Fig.4 A first progress note from the 2nd affiliated hospital of Harbin medical university

图 4 哈尔滨医科大学第二附属医院首次病程记录


姓名：[redacted] 性别：男 年龄：71岁 入院科别：神经内科二病房

入院日期:2012-02-29 11:21 出院日期：2012年03月14日 住院：14天

门诊收治诊断：多发脑梗死

临床初步诊断：多发脑梗死 高血压病 糖尿病

临床确定诊断：脑梗死 高血压病 糖尿病 心律失常 频发室性早搏

入院时情况：主因“双下肢无力1年，加重2月余”入院。查体：血压160/100mmHg，神志清楚，言语稍笨，双侧瞳孔等大同圆，约3.0mm，对光反射存在，双侧眼球各向运动自如，无眼震及复视，双耳听力正常，伸舌居中，转头转颈活动自如，双上肢肌力正常，双下肢肌力4级，四肢肌张力正常，双侧腱反射存在对称，双下肢病理征阳性，共济运动查体未见异常。辅助检查：头MRI示：多发脑梗死（2012-02-29，哈医大二院）

治疗经过：1、改善脑循环

2、心内会诊

3、降纤，抗血小板聚集

4、降糖，降血压

5、支持对症

出院时情况：患者无不适主诉，查体血压140/88mmHg，神志清楚，言语稍笨，双侧瞳孔等大同圆，约3.0mm，对光反射存在，双侧眼球各向运动自如，无眼震及复视，双耳听力正常，伸舌居中，转头转颈活动自如，双上肢肌力正常，双下肢肌力5-级，四肢肌张力正常，双侧腱反射存在对称，双下肢病理征阳性，共济运动查体未见异常。

治疗效果：好转

出院医嘱：1、注意休息，避免劳累

2、控制血压血糖

3、不适随诊

Fig.5 A discharge summary from the 2nd affiliated hospital of Harbin medical university

图 5 哈尔滨医科大学第二附属医院出院小结

3.2 规范制定和标注过程

电子病历命名实体和实体关系标注规范的制定难度较大,不仅涉及专业的医疗知识,而且涉及到对医疗实体的定义和分类.基于此,我们深入分析了 2010 年 I2B2 的规范和中文电子病历文本的特点,制定出初步规范,然后采用多轮迭代的模式进行规范的修订和标注工作.整个过程明显分为 3 个阶段,如图 6 所示,多轮迭代主要是在第 2 阶段.

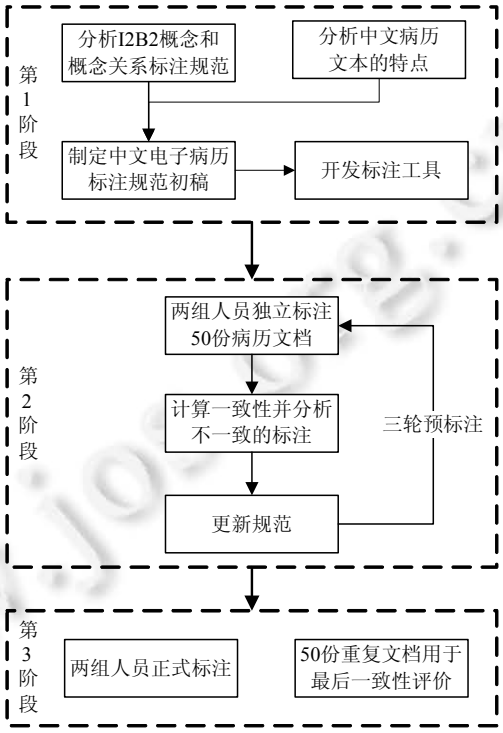


Fig.6 Flow diagram of corpus construction practice

图 6 语料构建实施流程图

第 1 个阶段,深入分析 I2B2 概念标注和关系标注规范,对比英文电子病历和中文电子病历的差别,总结中文电子病历文本的特点,并与临床医生多次交流.在医生的指导下分析了大量电子病历,确立了中文电子病历中实体和实体关系的分类框架,制定了标注规范的初稿,开发了标注工具,并进行初步的标注.这个阶段我们的主要成果是标注规范初稿和标注工具.在标注规范的制定过程中,我们从以下 3 个方面来保证规范的可操作性:

1) 针对每一类实体的标注,规范首先给出总体标注原则,该原则指出每类实体对应的 UMLS 语义类型所必须具有的基本特征,不符合基本特征的就不能被标注为该类实体.比如,疾病类实体必须是医生的诊断结论,必须是能够被治疗的,是可以被“否定”修饰的.
2) 辅以大量的正确标注示例和错误标注示例,从两方面给出标注指导.
3) 针对一些容易混淆的问题,规范还特别给出标注提示.

第 2 个阶段,采用迭代的模式进行预标注,目的是完善规范和培训标注人员,流程如下:

1) 随机抽取 50 份涵盖各科室的病历,两组标注人员独立标注.
2) 评价标注结果的一致性,分析每一个不一致的标注.
3) 针对不一致的标注展开讨论,并对规范进行完善.
4) 基于新的规范展开下一轮的预标注,流程转至 1).

经过 3 轮预标注,规范趋于稳定,标注人员得到了充分的训练,标注一致性达到较高水平.

第 3 个阶段,正式展开标注,标注过程中,我们采取以下 3 个方面的措施来保证标注质量:

1) 标注工具提供标记功能,对不确定的标注暂时打上标记,经过讨论后再确定标注.

2) 标注人员定期提交标注结果,其他人员对标注结果进行抽样检查,把认为可能与规范相冲突的标注拿出来讨论.抽样的方式是人工任意选择至少三分之一的标注结果.

3) 在分配给标注人员的病历中,仍然保持有 50 份重复的病历,这些重复的病历用作最后的一致性评价.

命名实体标注先于实体关系标注,遵循相同的三阶段过程.命名实体标注工作涉及太多的医疗知识,因此标注人员由两名医生担任.关系标注相对简单,标注人员由 4 名研究室成员担任,医生负责审核规范,并参与讨论.最后给出标注结果的存储格式示例.

命名实体标注结果的存储格式如下:

C=颈静脉怒张 P=220:225 T=testresult A=absent

C 表示命名实体名称,P 表示命名实体在文档中的边界位置,T 表示命名实体的类型(示例中是异常检查结果),A 表示命名实体的修饰(示例中是否认).

实体关系标注结果的存储格式如下:

E={肺腺癌【183-186】disease;}||R=DIS||E={胸痛【172-174】complaintsymptom;胸闷【175-177】complaintsymptom;}

第 1 个 E 表示关系的第 1 个实体组(示例中是疾病组),R 表示关系类型(示例中是“疾病导致症状”),第 2 个 E 表示关系的第 2 个实体组(示例中是症状组).

3.3 标注语料一致性评价及分析

标注一致性一般可以用两种指标表示:Kappa 值[58]和 F 值[59].Kappa 值一般用于正例和负例的标注评价,比如情感极性分类的语料标注.而在实体识别语料标注中,未标注的文字只能视为负例,而且难以统计,在负例很多却难以统计的情况下,可以采用 F 值评价,并且这种情况下 F 值接近于 Kappa 值[59].英文病历命名实体识别语料标注一致性评价一般也采用 F 值评价[40,55,60],所以本研究中,语料标注的一致性用 F 值评价,具体做法是,把一个标注者($A1$)的标注结果视为标准答案,计算另一标注者($A2$)标注结果的精确度(P)和召回率(R),进而计算 F 值,计算公式如公式(1)~公式(3)所示.

$$P=\frac{A1\text{和}A2\text{一致的标注结果总数}}{A2\text{的标注总数}} \tag{1}$$

$$R=\frac{A1\text{和}A2\text{一致的标注结果总数}}{A1\text{的标注总数}} \tag{2}$$

$$F=\frac{2\times P\times R}{P+R} \tag{3}$$

计算命名实体标注一致性时,实体字符串、类型和修饰同时一致时才认为实体标注是一致的.为了比较这 3 种因素对标注一致性的影响,我们计算了实体字符串的一致性,并且在实体字符串一致的情况下,分别计算了实体类型的一致性和实体修饰的一致性.命名实体标注一致性以及最后的一致性评价结果见表 2.

Table 2 Agreement evaluation of named entity annotation

表 2 命名实体标注的一致性评价

次数	实体字符串	类型	修饰	实体字符串+类型修饰
第1轮	0.867	0.992	0.977	0.848
第2轮	0.939	0.990	0.980	0.920
第3轮	0.942	0.994	0.984	0.927
最终	0.942	0.989	0.980	**0.922**

从表 2 可以看出,在实体字符串标注一致的情况下,实体类型和实体修饰标注的一致性一直都比较高,在迭代过程中次数变化不大,而实体字符串标注一致性在迭代过程中逐渐提升.可见,对标注一致性影响比较大的是

实体字符串的起止位置判断.在规范的制定中,实体边界的确定规则也是我们感觉最难的地方.最终,实体标注一致性达到 0.922.

在计算实体关系一致性时,实体组内的实体和实体组之间的关系同时一致时,才认为关系的标注是一致的.为了比较这两种因素对实体关系标注一致性的影响,我们计算了实体组的标注一致性,并且在实体组标注一致的情况下,计算了关系的标注一致性.实体关系标注一致性以及最后的一致性评价结果见表 3.从表 3 可以看出:实体组标注对实体关系的一致性评价影响较大,并且实体组的标注一致性在迭代中逐步改善.

Table 3 Agreement evaluation of entity relation annotation

表 3 实体关系标注的一致性评价

次数	实体组	关系	实体组+关系
第1轮	0.929	0.978	0.808
第2轮	0.932	0.942	0.879
第3轮	0.951	0.984	0.910
最终	0.945	0.959	0.895

从表 2 和表 3 一致性评价结果看:一致性逐渐递增,最终一致性评价结果和第 3 轮评价结果几乎持平,说明标注人员对规范的理解趋于一致.事实上,到第 3 轮预标注的时候,规范已基本稳定,这表明通过迭代的方式修订规范是有效的.表 2 和表 3 的标注一致性评价结果均显示,最终评价结果略低于第 3 轮预标注的评价结果.这是一个值得探讨的现象.用来评价最终一致性的病历是在正式标注时故意掺杂进去的,而且并未告知标注者.采取这种方式是为了检验标注者在没有监督的情况下,长时间进行标注工作是否会出现懈怠的情况.最终评价结果令人满意,表明标注者的工作是严谨的.虽然比第 3 轮评价结果略低,但这是可以接受的,因为标注者是人,不是机器,肯定会或多或少受到一些主观因素的影响.从这个意义上说,最终一致性才真正地反映了语料的一致性.因此我们认为,这种标注一致性评价方式是必要的.文献[61]指出,当标注一致性达到 0.8 时,即可认为语料的一致性是可信赖的.从最终一致性结果看,我们构建的语料库在一致性上是可靠的.

我们对最终标注完成的命名实体数量和实体关系数量进行了统计,见表 4 和表 5.对比已有的中英文命名实体语料库,我们构建的语料库标注体系最完整,规模最大.

Table 4 Statistics of entity types and entity counts of the corpus

表 4 语料库中实体类型与数量统计表

实体类型	数量
疾病	8 327
自诉症状	6 859
异常检查结果	12 092
检查	6 989
治疗	5 243
疾病诊断分类	223
总计	39 733

Table 5 Statistics of relation types and relation counts of the corpus

表 5 语料库中实体关系类型与数量统计表

关系大类	关系小类	数量	总和
治疗与疾病的关系	TrID	89	407
	TrWD	41	
	TrCD	4	
	TrAD	270	
	TrNAD	3	
治疗与症状的关系	TrIS	136	489
	TrWS	101	
	TrCS	85	
	TrAS	164	
	TrNAS	3	

Table 5 Statistics of relation types and relation counts of the corpus (Continued)

表 5 语料库中实体关系类型与数量统计表(续)

关系大类	关系小类	数量	总和
检查与疾病的关系	TeRD TeCD	416 2	418
检查与症状的关系	TeRS TeAS	827 216	1043
疾病和症状的关系	DIS	578	578
疾病和疾病诊断分类的关系	DADt	200	200

3.4 我们的经验

标注语料库在自然语言处理研究中的重要性不言而喻,但是构建语料库的难度也是众所周知的,涉及较多领域知识的特定领域语料构建更是难上加难.电子病历文本是个最特殊的特定领域文本,非专业人员理解起来尚有困难,所以我们的语料构建工作必须依赖于医生,而且是临床医生.我们的语料构建工作始于最初的真实病历获取,整个过程历时一年半,遭遇过不少挫折,也有过柳暗花明,最终圆满地完成了这个任务.一些较好的经验总结如下:

1) 充分调研国内外已有的电子病历命名实体语料构建工作,主要是标注体系、标注规范、标注方法和一致性评价方法,对比不同工作的优缺点.
2) 在医生的指导下阅读大量病历文本,分析病历文本的组织方式,并向医生了解医生在临床工作中书写病历的方式.
3) 选定一个现有的标注体系作为参照,结合中文病历的特点和临床医生的实际需求制定我们自己的标注体系.在我们的标注体系中,把症状细分为自诉症状和异常检查结果、给治疗类实体增加修饰信息等,都是考虑了临床医生的需求,而把实体关系定义为实体组的关系则主要是考虑了中文病历文本的特点.
4) 制定的规范要有可操作性,我们从定义、标注原则、正例标注、反例标注、特别提示这几个方面撰写规范,并且尝试性地标注了大量病历文本以验证规范的合理性,并充实规范中的正例和反例.规范初稿完成后,和医生一起逐条讨论并确认.
5) 多轮预标注的两个目的是训练医生和逐步完善规范,在展开预标注之前,形成完善的预标注方案,做好数据的准备工作.
6) 正式标注时,我们采取的 3 个方案最大限度地保证了语料标注的可靠性:拿不准的不勉强,讨论后做决定;抽样检查,发现可能有误的标注就拿出来讨论;保持一定数量的语料得到双份标注,做最终一致性评价.
7) 标注过程中,积极听取医生的意见,不断调整标注工具,提高标注效率.
8) 寻找合作医生考虑的几个因素:是临床医生,对我们的研究有兴趣,愿意抽出时间,工作认真负责.
9) 虽然医生是专家,但不能完全以医生为主导,医生可能因为执业习惯,对我们的一些标注方案有异议,我们是调研了现有的中英文病历命名实体的标注体系才形成的标注方案,认为不应修改的就坚持了我们的方案.比如医生对我们的实体类型中的检查(test)和症状的子类异常检查结果(test result)的标注方案不认同,原因是在临床工作中,他们所指的检查既包括检查手段又包括检查结果,在我们的研究中,这两类实体是要严格区分的.

4 后续工作及展望

我们已经开发完成了目前规模最大、科室覆盖面最广、分类最完备中文电子病历命名实体和实体关系标注语料库;在同样的病历文本上,研究室已完成了分词和词性标注语料[62]、短语句法树标注语料[63]的构建.基于这些语料库,我们的后续工作主要是两个方面:一是研究医疗知识的自动抽取方法,二是探索用户个性化健康知

识的表示和应用.

如前所述,构建标注语料库的目的是为了研究命名实体和实体关系自动识别的机器学习模型以及对模型进行评价,从而实现对大规模电子病历自动信息抽取.后续工作将以此为重点,并逐步展开时间信息抽取、共指消解、隐含关系发现、实体链接等信息抽取任务,最终实现中文电子病历医疗知识的抽取和整合.

电子病历是面向患者的个性化健康记录,从电子病历中抽取的医疗知识和健康信息与患者密切相关.因此,后续工作的另一个方面就是研究面向患者的个性化健康知识表示.从用户建模的角度,这种个性化健康知识表示也称为用户健康模型(consumer health model).围绕用户健康模型,研究将进一步扩展为用户健康模型的更新和维护、基于用户健康模型的推理(特别是引入大规模知识库的概率逻辑推理[64,65])、基于用户健康模型的预测和推荐.这些研究在临床信息学研究中可用于个性化诊断,在用户健康信息学研究中可用于用户个性化健康知识的推送和健康状况预测等.

5 结束语

本文主要总结了我们在中文电子病历命名实体和实体关系标注语料库构建方面的工作,主要体现在 3 个方面:

(1) 结合中文病历的特点,建立命名实体和实体关系标注体系;

(2) 制定命名实体和实体关系的标注规范;

(3) 构建命名实体和实体关系标注语料库.

我们的病历来源于正规权威医院,研究队伍中有两名住院医生深度参与,最后标注的语料具有较高的一致性,这些因素共同保证了我们构建语料库的质量.与其他电子病历命名实体和实体关系语料库构建相比,本文充分考虑了中文电子病历特点,具有 3 个主要的创新:

1) 把症状细分为自诉症状和异常检查结果;

2) 给治疗类实体标注修饰信息;

3) 引入实体组并把实体关系定义为实体组之间的关系.

我们已完成的工作构筑了中文电子病历命名实体识别和实体关系抽取研究的标准平台,正式启动了中文电子病历信息抽取研究的进程,开创了中文电子病历智能处理的新局面.我们相信,在全民关注医疗和健康的当下,我们的工作会吸引越来越多的研究者投入到中文电子病历信息抽取研究中来,为提高医疗服务质量,改善医患关系,从信息技术应用的角度贡献我们的智慧.

致谢 感谢黑龙江省人民医院皮肤科主任医师尤海燕和哈尔滨医科大学附属第二医院神经内科住院医师李明杰在我们前期调研阶段给予的指导.

References:

[1] Ministry of Health of the People's Republic of China. The basic specifications of electronic medical records (trial). 2013 (in Chinese). http://www.gov.cn/gzdt/att/att/site1/20100304/001e3741a2cc0cf99ded01.doc

[2] Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: Reflections on EMRs and future pediatric clinical research. Academic Pediatrics, 2011,11(4):280−287. [doi: 10.1016/j.acap.2011.02.007]

[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. [doi: 10.1016/j.artmed.2010.05.006]

[4] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 2009,42(5):760−772. [doi: 10.1016/j.jbi.2009.08.007]

[5] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. Journal of the American Medical Informatics Association, 2011,18(5):544−551. [doi: 10.1136/amiajnl-2011-000464]

[6] Prokosch HU, Ganslandt T. Perspectives for medical informatics: Reusing the electronic medical record for clinical research. Methods of Information in Medicine, 2009,48(1):38−44. [doi: 10.3414/ME9132]

[7] Greenes RA, Shortliffe EH. Medical informatics: An emerging academic discipline and institutional priority. JAMA: the Journal of the American Medical Association, 1990, 263(8):1114−1120. [doi: 10.1001/jama.1990.03440080092030]

[8] Gardner RM, Overhage JM, Steen EB, Munger BS, Holmes JH, Williamson JJ, Detmer DE. Core content for the subspecialty of clinical informatics. Journal of the American Medical Informatics Association, 2009,16(2):153−157. [doi: 10.1197/jamia. M3045]

[9] Sackett DL. Evidence-Based medicine. Seminars in Perinatology, 1997,21(1):3−5. [doi: 10.1016/S0146-0005(97)80013-4]

[10] Frankovich J, Longhurst CA, Sutherland SM. Evidence-Based medicine in the EMR era. The New England Journal of Medicine, 2011,365(19):1758−1759. [doi: 10.1056/NEJMp1108726]

[11] Fowler SA, Yaeger LH, Yu F, Doerhoff D, Schoening P, Kelly B. Electronic health record: Integrating evidence-based information at the point of clinical decision making. Journal of the Medical Library Association, 2014,102(1):52−55. [doi: 10.3163/1536-5050. 102.1.010]

[12] Miller RH, Sim I. Physicians' use of electronic medical records: Barriers and solutions. Health Affairs (Project Hope), 2004,23(2): 116−126. [doi: 10.1377/hlthaff.23.2.116]

[13] O'Donnell HC, Kaushal R, Barrón Y, Callahan MA, Adelman RD, Siegler EL. Physicians' attitudes towards copy and pasting in electronic note writing. Journal of General Internal Medicine, 2009,24(1):63−68. [doi: 10.1007/s11606-008-0843-2]

[14] Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. In: Proc. of the AMIA Annual Symp. Bethesda: American Medical Informatics Association, 2003. 269−273.

[15] Wilcox L, Lu J, Lai J, Feiner S, Jordan D. ActiveNotes: Computer-Assisted creation of patient progress notes. In: Proc. of the 27th Int'l Conf. on Extended Abstracts on Human Factors in Computing Systems. New York: ACM Press, 2009. 3323−3328. [doi: 10. 1145/1520340.1520480]

[16] Wilcox L, Lu J, Lai J, Feiner S, Jordan D. Physician-Driven management of patient progress notes in an intensive care unit. In: Proc. of the 28th Int'l Conf. on Human Factors in Computing Systems. New York: ACM Press, 2010. 1879. [doi: 10.1145/1753326. 1753609]

[17] Ministry of Health of the People's Republic of China. Measurement and standard of the level of electronic medical records (EMR) capabilities (trial). 2010 (in Chinese). http://www.moh.gov.cn/publicfiles/business/cmsresources/mohyzs/cmsrsdocument/doc13271. doc

[18] Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: A scoping review. Journal of the American Medical Informatics Association, 2011,18(4):515−522. [doi: 10.1136/amiajnl-2011-000105]

[19] Barbarito F, Pinciroli F, Barone A, Pizzo F, Ranza R, Mason J, Mazzola L, Bonacina S, Marceglia S. Implementing the lifelong personal health record in a regionalised health information system: The case of Lombardy, Italy. Computers in Biology and Medicine, 2015,59(C):164−174. [doi: 10.1016/j.compbiomed.2013.10.021]

[20] Wiesner M, Pfeifer D. Health recommender systems: Concepts, requirements, technical basics and challenges. Int'l Journal of Environmental Research and Public Health, 2014,11(3):2580−2607. [doi: 10.3390/ijerph110302580]

[21] Eysenbach G. Recent advances: Consumer health informatics. BMJ, 2000,320(7251):1713−1716. [doi: 10.1136/bmj.320.7251. 1713]

[22] Alpay L, Verhoef J, Xie B, Te'eni D, Zwetsloot-Schonk JHM. Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding. Biomedical Informatics Insights, 2009,2(1):1−10.

[23] Lehmann CU, Altuwaijri MM, Li YC, Ball MJ, Haux R. Translational research in medical informatics or from theory to practice: A call for an applied informatics journal. Methods of Information in Medicine, 2008,47(1):1−3.

[24] i2b2. https://www.i2b2.org/

[25] 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. [doi: 10.1197/jamia.M2444]

[26] 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. [doi: 10.1197/jamia.M2408]

[27] Uzuner O. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association, 2009, 16(4):561−570. [doi: 10.1197/jamia.M3115]

[28] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association, 2010,17(5):514−518. [doi: 10.1136/jamia.2010.003947]

[29] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 2011,18(5):552−556. [doi: 10.1136/amiajnl-2011-000203]

[30] Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. 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. [doi: 10.1136/amiajnl-2011-000784]

[31] 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. [doi: 10.1136/amiajnl-2013-001628]

[32] 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. [doi: 10.1016/S1532-0464(03)00012-1]

[33] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. 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.

[34] Yang JF, Yu QB, Guan Y, Jiang ZP. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. Acta Automatica Sinica, 2014,40(8):1537−1562 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.01537]

[35] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, Setzer A. Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics, 2009,42(5):950−966. [doi: 10.1016/j.jbi.2008.12.013]

[36] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. Journal of Biomedical Informatics, 2006,39(6):589−599. [doi: 10.1016/j.jbi.2005.11.004]

[37] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 2010,17(5):507−513. [doi: 10.1136/jamia.2009.001560]

[38] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 2004, 32(Database Issue):D267−D270. [doi: 10.1093/nar/gkh061]

[39] Weed LL. Medical records that guide and teach. New England Journal of Medicine, 1968,278(12):593−600. [doi: 10.1056/NEJM196803212781204]

[40] Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, Choi JD, Dligach D, Nielsen RD, Martin J, Ward W, Palmer M, Savova GK. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association, 2013,20(5):922−930. [doi: 10.1136/amiajnl-2012-001317]

[41] Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. Journal of the American Medical Informatics Association, 2015,22(1):143−154. [doi: 10.1136/amiajnl-2013-002544]

[42] Analysis of clinical text. http://alt.qcri.org/semeval2014/task7/

[43] Mizuki M, Yoshinobu K, Tomoko O, Mai M, Aramaki E. Overview of the NTCIR-10 MedNLP task. In: Proc. of the NTCIR-10. 2013.

[44] Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association, 2014,21(5):808−814. [doi: 10.1136/amiajnl-2013-002381]

[45] Xu Y, Wang Y, Liu T, Liu J, Fan Y, Qian Y, Tsujii J, Chang EI. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. Journal of the American Medical Informatics Association, 2014,21(e1):84−92. [doi: 10.1136/amiajnl-2013-001806]

[46] Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: A survey. Journal of Biomedical Informatics, 2010,43(4):650−660. [doi: 10.1016/j.jbi.2010.01.002]

[47] Wang Y, Yu Z, Chen L, Chen Y, Liu Y, Hu X, Jiang Y. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. Journal of Biomedical Informatics, 2014,47:91−104. [doi: 10.1016/j.jbi.2013.09.008]

[48] Wang H, Zhang W, Zeng Q, Li Z, Feng K, Liu L. Extracting important information from Chinese operation notes with natural language processing methods. Journal of Biomedical Informatics, 2014,48(C):130−136. [doi: 10.1016/j.jbi.2013.12.017]

[49] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, De Groen PC. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. Journal of Biomedical Informatics, 2009,42(5):937−949. [doi: 10.1016/j.jbi.2008.12.005]

[50] Lloyd-Jones DM. Cardiovascular risk prediction: Basic concepts, current status, and future directions. Circulation, 2010,121(15): 1768−1777. [doi: 10.1161/CIRCULATIONAHA.109.849166]

[51] Ye F, Chen YY, Zhou GG, Li HM, Li Y. Intelligent recognition of named entity in electronic medical records. Chinese Journal of Biomedical Engineering, 2011,30(2):256−262 (in Chinese with English abstract). [doi: 10.3969/j.issn.0258-8021.2011.02.014]

[52] Xia F, Yetisgen-Yildiz M. Clinical corpus annotation: Challenges and strategies. In: Proc. of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) of the Int'l Conf. on Language Resources and Evaluation (LREC). 2012. 32−39.

[53] Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2008. 254−263.

[54] Measuring degrees of relational similarity. http://www.cs.york.ac.uk/semeval-2012/task2/

[55] Uzuner Ö, 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. [doi: 10.1136/jamia.2010.004200]

[56] Yang JF, Qu CY, He B. Annotation specification for named entities and entity relations on Chinese electronic medical records. Harbin Institute of Technology, 2004 (in Chinese). http://wi.hit.edu.cn/dev/YuLiao/NER.pdf

[57] Annotation tool. https://github.com/yangjinfeng/emrproject

[58] Carletta J. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 1996,22(2):249−254.

[59] Hripcsak G, Rothschild AS. Agreement, the *f*-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association, 2005,12(3):296−298. [doi: 10.1197/jamia.M1733]

[60] Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In: Proc. of the 12th World Congress on Health (Medical) Informatics. Marrakech: European Language Resources Association (ELRA), 2008. 2325−2330.

[61] Artstein R, Poesio M. Inter-Coder agreement for computational linguistics. Computational Linguistics, 2008,34(4):555−596. [doi: 10.1162/coli.07-034-R2]

[62] Jiang ZP, Zhao FF, Guan Y, Yang JF. Research on Chinese electronic medical record oriented lexical corpus annotation. High Technology Letters, 2014,24(6):609−615 (in Chinese with English abstract). [doi: 10.3772/j.issn.1002-0470.2014.06.009]

[63] Jiang Z, Zhao F, Guan Y. Developing a linguistically annotated corpus of Chinese electronic medical record. In: Proc. of the IEEE Int'l Conf. on Bioinformatics and Biomedicine (BIBM). Belfast: IEEE, 2014. [doi: 10.1109/BIBM.2014.6999174]

[64] Wang WY, Mazaitis K, Lao N, Mitchell TM, Cohen WW. Efficient inference and learning in a large knowledge base. Machine Learning, 2015,100(1):101−126. [doi: 10.1007/s10994-015-5488-x]

[65] Lao N, Mitchell T, Cohen WW. Random walk inference and learning in a large scale knowledge base. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011. 529−539.

附中文参考文献:

[1] 中华人民共和国卫生部.电子病历基本规范(试行).2013. http://www.gov.cn/gzdt/att/att/site1/20100304/001e3741a2cc0cf99ded01.doc

[17] 中华人民共和国卫生部.电子病历系统功能应用水平分级评价方法及标准(试行).2013. http://www.moh.gov.cn/publicfiles/business/cmsresources/mohyzs/cmsrsdocument/doc13271.doc

[34] 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]

[51] 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. [doi: 10.3969/j.issn.0258-8021.2011.02.014]

[56] 杨锦锋,曲春燕,何彬.中文电子病历命名实体和实体关系标注规范.哈尔滨工业大学,2004. http://wi.hit.edu.cn/dev/YuLiao/NER.pdf

[62] 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. [doi: 10.3772/j.issn.1002-0470.2014.06.009]

杨锦锋(1978－),男,湖北麻城人,博士,CCF专业会员,主要研究领域为自然语言处理,信息抽取,机器学习.

关毅(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为智能信息检索,网络挖掘,自然语言处理,认知语言学.

何彬(1989－),男,博士生,主要研究领域为自然语言处理,信息抽取.

曲春燕(1990－),女,硕士生,主要研究领域为自然语言处理,信息抽取.

于秋滨(1966－),女,副主任医师,主要研究领域为电子病案的数据挖掘.

刘雅欣(1986－),女,博士,医师,主要研究领域为基因多态性与急性肺损伤的相关性.

赵永杰(1985－)女,医师,主要研究领域为血管性痴呆.