电子病历实体识别

目录

一、课题背景

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, 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，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。电子病历由医务专业人员撰写，不仅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。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。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。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。

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。 电子病历数据不应只是封存在病案室里，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。 医学信息 学可简单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， 其两个重要分支，临床信 息学(Clinical

Informatics)、用户健康信息学(Consumer Health Informatics)，都与电子病历信息抽取密切相关。 杨锦锋 等: 中文电子病历命名实体和实体关系标注体系及语料库构建 3 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，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)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。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。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。为了促进和规范电子病历系统智能支持的实施,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。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)，一个典型案例。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。另外，基础医学研究和临床治疗之间 的转化医学研究，也离不开对电子病历的分析处理。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特点。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之上。 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。 如 Roberts 所指出的，构建标注语料库有三个方面的主要原因: 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。

二、中文电子病历命名实体和实体关系标注体系建立

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 疗活动主要涉及四类重要信息: 检查、症状、疾病和治疗。 这四类信息在 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的之间关系。 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: 第一类实体是疾病, 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。 第二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”。 第三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。 第四类实体是检查手段, 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。 第五类实体是治疗手段, 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六类实体关系的抽取: 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 4 所示, 圆圈表示五类命名实 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。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, 医生可以预先快速的浏览病人的信息, 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、治疗在这五类。 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的遵循实体间不重叠、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)这三个原则。

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”)

.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。 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 8 Journal of Software 软件学报 2016,27 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)。 自诉症状指的是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”) 异常检查结果指的是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)

治疗指的是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。 治疗指的是能够治疗疾病或者缓解症状的医疗 概念，这是和检查最大的不同。 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七个，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 义这个七个修饰。 杨锦锋 等：中文电子病历命名实体和实体关系标注体系及语料库构建 9 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五个修饰：(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 要有三类，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。

三、实体标注细节

通过分析电子病历，医生针对患者的诊疗活动可以概括为：通过患者自述（自诉症状）和检查结果（检查项目）发现疾病的表现（症状），给出诊断结论（疾病），并基于诊断结论，给出治疗措施（治疗方案）。这个过程可以看出，医疗活动主要涉及四类重要信息：症状、疾病、检查和治疗，涉及的具体描述如下：

1）疾病：泛指导致患者处于非健康状态的原因，比如：诊断、病史。

2）疾病诊断分类：疾病诊断相关分组，比如“高血压，极高危组”中的“极高危组”。

3）症状：泛指疾病导致的不适和显示表达的检验检查结果，分为：自诉症状和体征（异常检验检查结果）。

4）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。

5）治疗手段：泛指为了治愈疾病、缓解或改善症状而给予患者的药物、手术和措施等。

3.1 疾病 DIS,DISEASE

疾病必须是能够治疗的，其语义范围包括：疾病或者综合征、受伤或中毒、先天性畸形、病毒细菌、病理功能、细胞或分子功能障碍、获得性异常、解剖异常、肿瘤进程、精神或行为障碍等。

1。1。1 疾病诊断分型 DT, **DISEASE TYPE**

疾病的具体分类，表示疾病的进展程度，疾病诊断分类一般出现在诊断里。如：

1）**失代偿期** DT

2）**III 期** DT

3）**II 型** DT

3.2 症状

症状是能够被改善或治愈的，并且能够被否定词修饰，为疾病的表现。包括患者向医生陈述的不适感觉(症状)和医生观察到的（体征）或者检查结果，如：

3.2.1 患者向医生陈述的不适感觉(症状) SYM,SYMPTOM

1）疼痛时伴有右下肢活动受限。（“疼痛 “ 、” 右下肢活动受限”）；

2）伴活动后心慌气短。（“心慌”、“气短”）

3.2.2 医生观察到的（体征）ST

1）双肺听诊可闻及少量痰鸣音。（“痰鸣音”）

2）自带胸片示左下肺症病变。（“左下肺症病变”）

3）双肺听诊无著征。（“著征”）

3.3 检查 TES,TEST

检查是为了发现、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项目，包括：化验过程，诊断过程等。如：

1）头 **CT** 显示脑实质内高密度灶。（“CT”）

2）血压最高达到 180/130mmHg。（“血压”）

3）双肺听诊无著征。（“听诊”）

4）自带胸片示左下肺症病变。（“胸片”）

3.4 治疗

治疗是能够治疗疾病或者缓解症状而施加给患者的手段，包括手术、药品、措施等。本

标注语义类型包括:药物、手术。如：

3.4.1 药品 DRU,DRUG

1）奥扎格雪、脑蛋白水解物等静点（药物“奥扎格雪”和“脑蛋白水解物”）。

3.4.2 手术 SUR,SURGERY

1）4 年前行胆囊切除术。（手术“胆囊切除术”）

2）鼻内镜下行双筛、双上颌窦。（手术“鼻内镜”）

3.4.3 措施(非手术，非药品的治疗) PRE,precaution

3.5 实体修饰词标注

3.5.1 否认词(AT,,absent)标注：

各瓣膜区未闻及病理性杂音

全腹无压痛、反跳痛及肌紧张

3.5.2 条件词(CL,conditional)标注：

在某种条件具备的情况下才发生的词。

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**。

再如：**吃红薯后**血糖升高

3.5.3 既往信息词（PT,past）

明确表示患者过去有过的治疗史或疾病症状，比如：

有多年心脏病**史**。

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。

3.5.4 时间标注统一标为 TE

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。

3.5.5！！可能性词:

不确定当前会发生，需要进一步的证据确认的词。如：

不排除缺血性疾病。/右肺中下叶考虑创伤性湿肺

待证实词：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

3.5.6 程度词标注（AM,AMOUNT），非量化的数量描述词，如大小、多少、程度（明显等）等

双肺听诊可闻及**少量**痰鸣音。

3.5.7 解剖位置

器官（**REG，REGION**）
部位词（**ORG,ORGEN**）

3.5.8 频率词 （FW,Frequency Word）

患者走路不稳，偶有头晕。时有胸闷气短。

反复胸闷，憋气，持续时间长短不等。

标注格式：

突发 AM

头晕 SYM

伴 O

恶心 SYM

呕吐 SYM

3 小时 TE

四、难点

由于基于传统正则匹配或者机器学习的方法提取的特征有限，对于大量书写用词不一的非结构化电子病历实体识别带来准确率和召回率不高的情况。电子病历具有噪音大，特殊符号，书写错误等情况存在，因此需要采用深度学习模型来对其进行实体识别。因此难点总结如下：

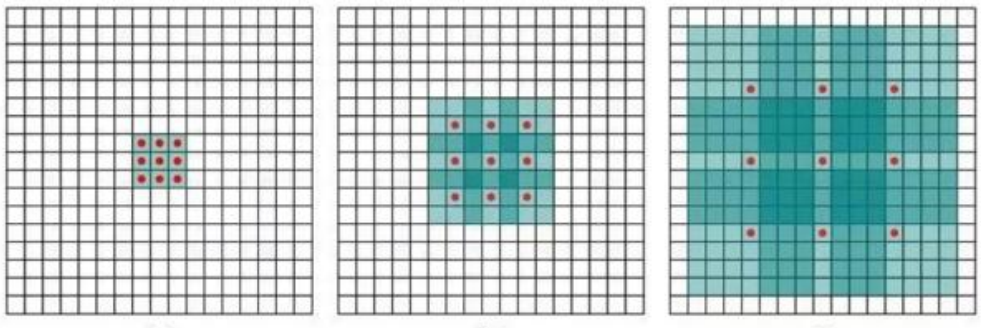
1、 如何生产深度学习可用的标签数据，给大量的医疗文本打上实体类别的标签

2、 传统的基于规则和机器学习算法的实体识别准确率不高，如何提升

3、 部分医疗实体名字特别长，如何获取更长的上下文依赖，让识别准确率更高

4、 类别不均衡，有些类别的词特别少，有些很多

5、 未登录词如何处理



1. 评价

本评测采用精确率（Precision）、召回率（Recall）以及 F1-Measure 作为评价指标。

命名实体识别介绍

命名实体识别（Named Entity Recognition，简称 NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。

命名实体识别作用

命名实体识别是信息提取、问答系统、句法分析、机器翻译、面向 Semantic Web 的元数据标注等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。一般来说，命名实体识别的任务就是识别出待处理文本中三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。

命名实体识别过程组成

通常包括两部分：（1）实体边界识别；（2） 确定实体类别（人名、地名、机构名或其他）。英语中的命名实体具有比较明显的形式标志（即实体中的每个词的第一个字母要大写），所以实体边界识别相对容易，任务的重点是确定实体的类别。和英语相比，汉语命名实体识别任务更加复杂，而且相对于实体类别标注子任务，实体边界的识别更加困难。

命名实体识别难点

（1）汉语文本没有类似英文文本中空格之类的显式标示词的边界标示符，命名实体识别的第一步就是确定词的边界，即分词；（2）汉语分词和命名实体识别互相影响；（3）除了英语中定义的实体，外国人名译名和地名译名是存在于汉语中的两类特殊实体类型；（4）现代汉语文本，尤其是网络汉语文本，常出现中英文交替使用，这时汉语命名实体识别的任务还包括识别其中的英文命名实体；（5）不同的命名实体具有不同的内部特征，不可能用一个统一的模型来刻画所有的实体内部特征。

国内研究现状、发展动态

汉语的命名实体识别研究的难度相对英语较大,起步于 20 世纪 90 年代初。当时的任务主要是识别汉语中一些姓名、地名、组织机构等各种通用类别的词。1992 年，张俊盛和刘显仲提出用多语料库来辨识中文姓名[1]。此方法通过查语料字典，分别判断姓氏和名字的初始概率，然后构建一个规则约束集，判断姓名在规则约束集下的联合概率，最后通过动态规划法来求解约束条件下的最优解。1995 年，孙茂松，黄昌宁和高海燕等[2]用基于统计分析的规则法来识别汉语中的姓名。作者统计和分析的单名和双明的概率分布，以及汉字的成词规律及边界等规律，启发式的构建了一些规则集来进行姓名识别。在随机抽取的 300 句新闻语料库中，识别的准确率在 70.06%，召回率 99.77%。2001 年，季姮和罗振声把姓名反比概率（INF）模型引入了中文姓名的识别中[3]。此方法不需要自动分词，避免由分词不准带来的识别错误。作者通过大量语料分析构建了候选姓名表，在此基础上，分析句子中的上下文，位置依存关系，邻接链和特殊性来综合判断并输出句子中识别的真实姓名。2006 年，俞鸿魁，张华平和刘群等提出了用层叠隐马尔可夫模型去识别中文中的命名实体[4]。该

方法首先采用了隐马尔科夫模型识别普通的无歧义的人名、地名和机构名。对于复杂嵌套的命名实体，再次采用高层的隐马尔科夫模型进行识别，通过层叠的隐马尔科夫模型的融合，该方法在大规模封闭语料库测试中，人名、地名和机构名识别的 F1 值分别是 92.55%，94.53%，86.51%，达到了较好的效果。2016 年，孙晓，孙重远和任福继把命名实体识别技术应用于生物医学文献获取实体的任务当中[5]。此方法与传统的条件随机场有一定不同，它的层数相对较深，可达到更高的精度。通过增量式学习策略，选择更优的词和词性特征集，通过多种特征组合分类效果对比，来确定最优的条件随机场分类器的特征组合。最后通过 SVM 模型，CRFs-Margin 模型和 Genia Tagger 模型对识别结果进行投票，把多数分类器的投票的识别实体用规则库进行过滤，修正潜在的边界错误。2017 年，侯伟涛和姬东鸿提出一种基于深度学习的双向 LSTM 和多层感知器的集成实体识别模型[6]，用于医疗事件的识别。通过双向的 LSTM 捕捉文本中上下文隐含的语义特征，再把生成的特征向量放入全连接的多层感知器，判断实体的类别属性。2017 年张帆和王敏,提出一种深度学习命名实体识别模型[7]。该方法把实体识别看做分类问题，使用了词向量作为输入的特征，通过词向量特征代替了传统的人工提取的手工特征，再把词向量放入一个多层的神经网络，最后通过一个 SoftMax 分类器预测实体的类别。

国外研究现状、发展动态

命名实体识别是任务在 1996 年由 MUC-6 会议（the sixth in a series of Message Understanding Conferences[8]）提出。MUC-6 命名实体识别的有三个子任务组成，分别是实体名（Entity Name）：人名、地名和机构名称；时间表达式（Temporal Expressions）：时间、日期和持续时间；数字表达式（Number Expressions）：金钱，百分比和衡量度。MUC-6 主张从非机构化文本从抽取实体信息后结构化文本，自提出以来，每年都有大量的算法被提出。2002 年，J Su 提出用隐马尔科夫模型来做命名实体识别和基于隐马尔科夫模型的数据块标注方式[9]，在 MUC-6 和 MUC-7 英文实体识别的任务上 F1 准确率分别是 96.6%和 94.1%。2015 年，CND Santos 和 V Guimarães[10]利用单词级和字符级的词向量特征作为深度学习模型的输入特征，通过把两种特征进行串联，送入神经网络进行实体识别，达到较好的效果。2015 年，JPC Chiu 和 E Nichols[11]提出一种整合递归神经网络 LSTM 和卷积神经网络 CNN 的命名实体识别的混合新模型，同时把自定义字典匹配模式融入模型中，词方法在 CoNLL-2003 和 OntoNotes 公开数据集上，实体识别的 F1 值分别为 91.62%和 86.28%,获得 2015 年报道的最好效果。2015 年 Li L，Jin L 和 Jiang Z 等[12]提出一种拓展版的递归神经网络专门针对生物信息如基因和蛋白质的识别任务。和传统的神经网络提取语言学特征不同，此模型采用蛋白质先前节点的预测信息，生物的主题信息和聚类信息作为特征的输入，结果显示采用 DNN 神经网络比条件随机场分类器效果要好。为了更好的提高化合物和药品的识别准确率，

2016 年，Lample G， Ballesteros M 和 Subramanian S 等[17]提出了用两个神经网络的新框架进行实体识别。此框架其中一个是双向的 LSTM 神经网络加上条件随机场分类器，另外一个是模型是基于过渡转移方式标注实体。此方法新颖之处在于采用监督学习下的字向量和非监督学习下的词向量作为模型的训练特征。此模型在四种语言的公开数据集命名

实体识别任务上达到了当时最好的效果。为了应对较小标注样本无法很好的训练出较好的实体识别模型，文献[18]提出利用迁移学习的方法来解决小样本问题。具体而言，通过在大量训练样本下训练出一个模型，然后学习源实体和目标实体的关系，最后用在样本环境下训练好的模型初始化条件随机场分类器参数，在此基础上微调特定领域的实体识别模型。为了验证不同词向量特征对实体识别准确率的影响。