doi:10.3772/j.issn.1002-0470.2015.02.005

中文电子病历命名实体标注语料库构建①

曲春燕②* 关 毅③* 杨锦锋* 赵永杰** 刘雅欣***

(*哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

(**哈尔滨医科大学附属第四医院 哈尔滨 150001)

(***哈尔滨医科大学附属第二医院 哈尔滨 150001)

摘 要 针对中文电子病历命名实体语料标注空白的现状,研究了中文电子病历命名实体标注语料库的构建。参考2010年美国国家集成生物与临床信息学研究中心(I2B2)给出的电子病历命名实体类型及修饰类型的定义,在专业医生的指导下制定了详尽的中文电子病历标注规范;通过对大量中文电子病历的分析,提出了一套完整的中文电子病历命名实体标注方案,而且采用预标注和正式标注的方法,建立了一定规模的中文电子病历命名实体标注语料库,其标注语料的一致性达到了92%以上。该工作对中文电子病历的命名实体识别及信息抽取研究提供了可靠的数据支持,对医疗知识挖掘也有重要意义。

关键词 中文电子病历(CEMR),命名实体,标注语料库,标注规范,标注一致性(IAA)

0 引 言

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]。电子病历包含了关于病人健康信息的全面、专业、准确的描述,是极其宝贵的医疗知识资源[2]。电子病历命名实体标注语料库的构建对医疗领域知识挖掘研究具有重要意义。从2006年至今,美国国家集成生物与临床信息学研究中心(Informatics for Integrating Biology and the Bedside,I2B2)先后组织了7次电子病历信息抽取研究评测并发布了9个语料库。Pestian等[3]构建了基于自动疾病编码任务的语料库。2011年以来,文本检索会议(text retrieval conference,TREC)连续组织了两次电子病历检索评测[4,5]。国外基于电子病历的相关共享任务的开展及语料库的构建,大大推动了医疗领域电子病历信息抽取的研究。而在国内,尽管各医院信息系统已初具规模,但是由于统一标准和规范的缺乏等原因[6],导致中文电子病历共享语料库仍处于空白状态。因此,中文电子病历命名实体及实体关系标注语料库构建的研究,对国内医疗领域的信息抽取研究有着填补空白的重大意义。

目前,语料标注模式主要有3种[7]:传统标注模式、众包标注模式及团体标注模式。传统标注模式指在标注规范的指导下,训练标注人员进行标注,同时不断修订规范。它的缺点是较费时费力,不仅需要领域专家参与标注规范的制定,还要花费大量的时间训练标注者。众包标注模式则是利用在线用户对同一数据进行标注,以投票的方式获取高质量的标注。它能够以较低的成本取得标注结果,但仅适用于简单的且不要求较多领域知识的标注任务。与前两种标注模式相比,团体标注能够以较小的成本快速地完成标注任务,即使专业性很强的语料,也可不用专家参与,但对标注团体的规模有一定的要

① 国家自然科学基金(60975077)资助项目。

② 女,1990年生,硕士生;研究方向:自然语言处理;E-mail:xxstar0509@163.com

③ 通讯作者,E-mail:guanyi@hit.edu.cn

(收稿日期:2014-08-11)

求。I2B2 2009 的评测即采用了团体标注的模式构建评价语料[8],实验表明该标注模式比传统标注更有优势。I2B2 2010 的评测任务[9]中,首次给出了较完整的电子病历命名实体类型及修饰类型的定义。其参照一体化医学语言系统(UMLS)定义的语义类型[10],将命名实体分为 3 类:医疗问题、检查及治疗,并给出了 6 种实体修饰类型:当前的、不存在的、非患者本人的、有条件的、可能的及待证实的。这是目前为止对电子病历命名实体较系统的分类。但是,由于中英文语言及病历书写的差异,英文电子病历命名实体的定义并不能直接被中文电子病历采用。

本文在 2010 年 I2B2[9]给出的电子病历概念识别任务定义的基础上,在专业医生的指导下,首次提出了中文电子病历命名实体标注规范①。通过分析中文电子病历特点,采用预标注和正式标注的方法,以迭代的方式不断修订规范,同时采取多种措施控制标注质量,构建了首个中文电子病历命名实体标注语料库,达到了较高的一致性。

1 中文电子病历命名实体标注语料库的构建

1.1 标注规范

中文电子病历命名实体标注规范的制定,为命名实体标注语料的构建提供了指导。该规范的制定主要参照 I2B2 2010 评测数据构建的两个标注规范——概念标注规范②和修饰标注规范③,结合中文电子病历的特点[11],对命名实体类型及修饰类型给出了定义。标注过程中主要遵循三个原则:不重叠、不嵌套、不含有起分隔作用的标点符号。该规范共包括约 13000 字,并包含超过 300 个的正例与反例,对于一些特殊情况也给出了详细的处理方案。

该规范共定义了 5 种实体类型,分别为疾病、疾病诊断分类、症状、检查和治疗。根据中文电子病历的特点,本文将症状细分为两类:自诉症状、异常检查结果。该规范使用 UMLS 语义类型界定每一类实体涵盖的范围,但不局限于 UMLS 中的概念[10]。

(1) 疾病(disease):泛指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。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)等。例如**糖尿病**;**冠心病**。

(2) 疾病诊断分类(diseasetype):指对疾病的一个具体的分类,这类实体通常出现在诊断中,且一般紧跟一个具体的疾病。例如高血压病 3 级**极高危组**;急性白血病 AML-**M2**。

(3) 症状(symptom):泛指由疾病导致的不适表现、异常表现或者显式表达的异常检查结果。其对应的 UMLS 语义类型有症状或体征(sign or symptom)、精神或行为障碍(mental or behavioral dysfunction)、异常的检查结果(abnormal test results)等。症状是能够被治愈或者改善的。根据中文电子病历的特点,又将症状细分为两类:自诉症状和异常检查结果。

(i) 自诉症状(complaint symptom):指患者自己向医生陈述(或别人代述)的不适感觉或者异常感觉。表示症状严重程度的修饰成分也包括在症状里。自诉症状主要包括不适感觉和异常的精神或行为状态,例如无**肢体活动障碍**及**抽搐**发作;**反应迟钝**。

(ii) 异常检查结果(test result):指的是医生观察到的或通过检查程序、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。异常检查结果通常包括异常体征、异常检查结果,例如**低血压**;**脑实质内高密度灶**。

(4) 检查(test):指的是为了发现、否认、证实疾病或者症状和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及仪器检查项目等。对应的

① http://wi.hit.edu.cn/dev/YuLiao/NER.pdf
② https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf
③ https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf

UMLS 语义类型有化验过程(laboratory procedure)、诊断过程(diagnostic procedure)等。与治疗类似,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,它阐述了为了找到疾病或症状所采用的方法,例如**胸 X 光**;**血常规**。

(5) 治疗(treatment):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、干预措施或者给予的药品。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。治疗通常包括药物名称、治疗过程、医疗设备等,例如**奥扎格雪**、**脑蛋白水解物**等静点;**改善脑循环**。

针对疾病和症状,从是否发生于患者本人以及是否发生于患者本人的确定程度,该标注规范定义了疾病和症状的 7 种修饰类型:从是否发生于患者本人的维度,则是否认、非患者本人;从是否发生于患者本人的确定程度,则是当前的、有条件的、可能的、待证实的、偶有的。每个疾病或症状只有一个修饰。

(1) 当前的(present):指当前肯定发生的不适症状或疾病,包括已经确定的疾病或正在遭受的症状、检查结果等。通常情况下,如果症状或者疾病是本次患者就诊的问题或者是在就诊时发现的,就应该标注为当前的。例如行走时**步态欠灵活稳定**;自诉有**冠心病史**。

(2) 否认(absent):指的是症状或疾病的否定,是肯定不发生于患者本人的,同时还包括以前的症状或疾病经过治疗后不再发生的情况。常见的否定词有未及、未见、未触及、未诉、未闻及、否认、不伴、无等。例如:各瓣膜区未闻及**病理性杂音**;不伴**意识障碍**。

(3) 非患者本人的(family):指的是患者亲属患有的疾病或出现的症状(有些疾病为家族病)。例如:其父母均患有**糖尿病**。

(4) 有条件的(conditional):指当前不一定发生,在特定条件下才会发生的疾病或症状。例如:长期饮酒会引发**酒精肝**;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(possible):指的是可能发生的症状或者根据当前症状做出的可能的疾病诊断。例如:**糖尿病**待除外;既往的**心律失常**。

(6) 待证实的(hypothetical):指的是疾病或者症状当前不会发生,但预期以后会发生。例如: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(occasional):指的是当前不经常出现的症状或者疾病。例如:病程中患者走路不稳,偶有**头晕**;大便偶有**一过性发白**。

治疗的修饰信息主要有 3 类:既往的、否认的、当前的。每个治疗只有一个修饰。

(1) 既往的(history):明确表示是患者过去经历过的治疗史,或者病历里描述的患者近期已经经历过的治疗。例如:**髌骨骨折手术史**;后自行间断口服**拜糖平**及**二甲双胍** 8 天。

(2) 否认的(absent):一般是否认既往的治疗史。例如:未**接种疫苗**;否认**人流术史**。

(3) 当前的(present):指患者当前经历的治疗或者即将要经历的治疗。这类治疗一般都是本次治疗提出的,例如:**保护脑组织**;**营养神经**。

该规范将英文电子病历标注规范定义的医疗问题拆分为疾病和症状,同时又将症状细分为自诉症状和异常检查结果,并增加了疾病诊断分类这一类型。在实体类型修饰的定义上,不仅在原有的对疾病和症状定义的修饰类型的基础上增加了一个修饰类型"偶有的",同时新增了对治疗的修饰,包括既往的、否认的、当前的。实体类型及修饰类型的增加不仅保留了更多的医疗知识,让实体的信息更加丰富,有利于后续实体抽取规则的制定,同时也使得实体关系更加清晰、明了,为实体关系的建立打下了基础。

1.2 标注过程

本文所使用的电子病历均来源于哈尔滨医科大学附属第二医院,共包含来自普通外科、心血管内科、血液内科等 35 个大科室、87 个小科室的 144230

份电子病历。考虑到不同科室语言特点及记录方式的不同，我们从中随机选取 3825 份不同科室的病历文本进行研究，其中 992 份用于命名实体语料库的构建，具体科室分布如图 1 所示。

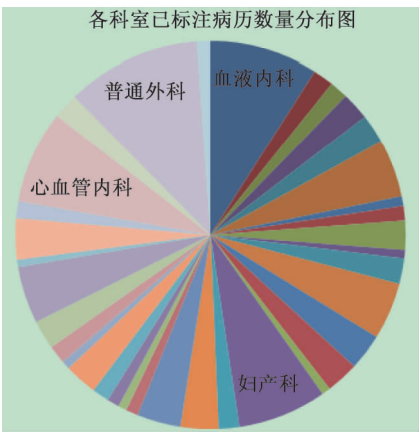

图 1　各科室已标注病历数量分布图

目前电子病历的文本信息主要包括病程记录、手术记录、护理记录、医嘱、出院小结、死亡小结等。为了尽可能涵盖患者的医疗信息，本文选取首次病程记录和出院小结这两部分作为重点研究对象。其中，首次病程记录详细描述了患者的自诉情况、检查项目及结果、医生初步诊断及治疗措施。而出院小结则主要描述医生给出的诊断结论、出入院时的情况、治疗过程及效果，以及出院医嘱。

每个医疗实体共包含起止位置、实体类型、实体修饰类型三部分信息。起止位置由选定文本的开头字符和结尾字符位于整个文本的位置组成。实体类型指的是规范提出的五大医疗实体类型。实体修饰类型与实体类型对应，根据不同类型实体的特点赋予不同的修饰类型，其中疾病诊断分类和检查没有修饰类型。疾病和症状的修饰主要包括 7 种，治疗则主要有三种修饰类型。

实体标注形式如下①：

C = “实体” P = 实体开始位置:实体结束位置 T = “实体类型” A = “实体修饰类型”

例如：

于我科行 **VAD 方案**六个周期，**化疗副反应**轻。

C = VAD 方案 P = 260:265 T = treatment A = history

C = 化疗副反应 P = 271:276 T = complaintsymptom A = present

为了便于实体的标注，我们自行开发了实体标注工具及实体比较工具②。如图 2 所示，为了便于区分，不同类型的实体以不同的颜色显示。对于待定的标注，通过选择不确定标记，便于后续的讨论与修改。

图 2　中文电子病历实体标注工具

① https://www.i2b2.org/NLP/Relations/assets/Annotation%20File%20Formatting.pdf

② https://github.com/yangjinfeng/emrproject

图 3 为实体比较工具,比较工具不仅可以进行基本的实体标注,还可以对比两个标注者标注的内容,通过用不同颜色区分两个标注者的标注,使得不一致标注更加醒目、直观,便于标注者的修改以及问题的记录。

图 3　中文电子病历实体比较工具

考虑到电子病历的领域性较强,以及人员的限制,我们采用了传统的标注方案,采取标注人员标注为主,规范制定人员从旁指导的模式,遇到疑难问题经讨论后达成一致,以此来不断完善规范。标注团队主要包括两名住院医生,一名是呼吸内科的医学博士,一名是神经内科的医学硕士,她们在工作中有书写电子病历的经历,具有丰富的医疗知识和临床经验;规范制定团队主要包括与自然语言处理领域相关的一名博士和一名硕士。

整个标注过程分为四轮,其中前三轮是预标注部分,第四轮是正式标注部分。预标注旨在培训医生,在熟悉规范的同时给予专业的指导,便于规范的完善及问题的修正。经过三轮预标注,两个标注者的标注一致性达到标准,规范趋于稳定,开始正式标注。

预标注一共包含 150 份病历文本,每轮标注由 50 份病历文本组成。每轮的 50 份病历文本由两个标注者分别独立标注,简称为 A1 和 A2。如图 4 所示,两个标注者具有完全相同的标注任务,使用完全相同的标注工具,并给予完全相同的标注规范。在每个标注者单独完成 50 份病历的标注之后,通过一致性评价得到所有不一致集。通过标注团队与规范制定团队的集体讨论,解决所有不一致问题,并根据出现的问题,充实样例并完善规范,用于指导下一次的标注。

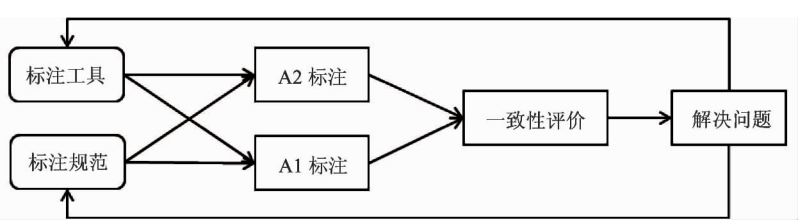

图 4　预标注过程

第四轮正式标注一共包含 992 份病历文本,其中包括预标注的 150 份病历文本。正式标注由两名标注者共同完成。为保证标注进度及质量,正式标注采取如下措施:

(1)及时反馈。标注者针对无法确认的问题,通过不确定标记或提问的方式及时提出,避免相同问题重复出现。

(2)质量监测。保证两个标注者的病历中有 50 份重复的病历文本,通过 F 值评价检验两个标注者的标注质量。

(3)抽样检查。对于两个标注者提交的病历,规范制定者会抽取 20% 进行检查,及时纠正错误。

1.3 标注一致性(inter-annotator agreement,IAA)

标注语料质量常用 IAA 评价[12,13],通常有 F 值和 kappa 值两种计算方法[14]。kappa 值常用于分类实验的一致性评价,如二分类问题,其中必须已知反例的数量。而在命名实体语料标注中,未标注的实体即为反例,其数量是无法统计的。当反例无法统计且数量很大时,期望一致性将趋近于 0,无法正确衡量标注的一致性,但此时的 F 值将趋近于 kappa 值[14],故可以采用 F 值的计算方法进行评价。同时,F 值也是常用的命名实体标注语料的评价方法[15,16],因此,本文也采取 F 值的方式进行 IAA 评价。

将一个标注者(如 A1)的标注视为标准,通过计算另外一个标注者(如 A2)的精度(Precision,P)和召回率(Recall, R)计算而得。其中,无论将哪位标注者的标注结果作为标准,都不影响 F 值。精度 P 和召回率 R 是广泛用于评价结果质量的两个度量值,而 F 值则是精度和召回率的调和平均值,F 值越高,则标注质量越高。其中,

$$P = \frac{\text{A1 和 A2 标注一致数}}{\text{A2 标注总数}} \quad (1)$$

$$R = \frac{\text{A1 和 A2 标注一致数}}{\text{A1 标注总数}} \quad (2)$$

则 F 值可以表示为

$$F = \frac{(1+\beta^2) \times R \times P}{(\beta^2 \times P) + R} \text{(这里一般取} \beta = 1) \quad (3)$$

2 构建结果

目前已经构建完成的中文电子病历命名实体标注语料库由来自正式标注的 992 份电子病历文本构成,四轮标注都采用标注一致性(IAA)评价。表 1 是对 A1 和 A2 两个标注者标注的实体的总结,其中包括每个标注者标注的实体数、两个标注者标注的总实体数及实体标注的 IAA 结果。可以看出,随着标注的深入,实体标注的 IAA 逐渐提升,从第一轮的86.7% 提升到了第三轮的94.2%。而正式标注的 IAA 与最后一轮预标注持平,保持在 94% 以上,表明经过三轮的预标注,标注规范已经趋于稳定,同时标注者对标注规范的理解已经达到了一致。

表 1 三轮预标注的实体数及实体标注一致性

轮数	标注者	实体数	实体总数	IAA
第一轮	A1	1690	2008	86.7%
	A2	1856		
第二轮	A1	1817	1917	93.9%
	A2	1798		
第三轮	A1	1967	2090	94.2%
	A2	1985		
第四轮	A1	1637	1737	94.2%
	A2	1645		

表 2 从类型、修饰、整体(实体 + 类型 + 修饰)三个方面分别计算了 IAA 结果。经过三轮预标注,类型和修饰的 IAA 提高不超过 1%,但整体的 IAA 提高了 7.9%。由此可见,实体的标注在整个标注过程中起到关键作用。随着实体识别程度的提升,整体标注的质量会不断提高。

表 2 预标注的类型、修饰及整体的标注一致性

	类型	修饰	实体 + 类型 + 修饰
第一轮	99.2%	97.7%	84.8%
第二轮	99.0%	98.0%	92.0%
第三轮	99.4%	98.4%	92.7%
第四轮	98.9%	98.0%	92.2%

从表 1 和表 2 可以看出,正式标注的 IAA 都达

到了92%以上。而当IAA超过0.8时,标注结果即可视为可靠的[17]。由此可见,本文的标注结果是真实可靠的。

3 结 论

本文主要提出了中文电子病历命名实体标注方案,并构建了规模为992份的中文电子病历命名实体标注语料库。标注语料的制定,离不开标注规范的指导。因此,本文在英文标注规范的基础上,细分了实体类型,同时对实体修饰类型进行了相应的扩充,制定了适用于中文电子病历的标注规范,开创了国内电子病历标注规范的先河。在标注方面,本文采取了以专业人员标注为主,规范制定人员从旁协助的模式,在完善规范的同时,也保证了标注语料的可靠性,证明该方法是医学领域语料标注的一种行之有效的方式。此外,为了控制标注语料的质量,本文在确保规范的准确性及标注人员的标注能力的基础上,采取了及时反馈、质量监测、抽样检查等多项措施,保证了标注结果的真实可靠。最后,通过对比实体、类型、修饰及整体的标注一致性可以看出,实体标注是整个标注工作的主要难题,也是未来工作的重中之重。与此同时,电子病历书写不规范产生的歧义,人工校对产生的人为错误,也给标注带来了一定程度的困难。

后续的工作将从提高数据质量及标注自动化两个方面展开。在提高数据质量方面,研发自动校对工具,提高校对效率和质量。在提高标注自动化方面,探索基于ActiveLearning的半自动语料标注,在降低标注成本的同时提高标注效率。

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). http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3585/201003/46174.htm:国家卫生计生委统计信息中心,2010

[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] Pestian J P, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Stroudsburg, USA, 2007. 97-104

[4] Voorhees E, Tong R. Overview of the TREC 2011 medical records track. In: Proceedings of the 20th Text REtrieval Conference Proceedings, Montgomery, USA, 2011

[5] Hersh W R, Voorhees E M. Overview of the TREC 2012 medical records track. In: Proceedings of the 21st Text REtrieval Conference Proceedings, Montgomery, USA, 2012

[6] 任彩玲. 电子病历遭遇三大障碍. 信息系统工程, 2008,(2): 28-30

[7] Xia F, Yetisgen-Yildiz M. Clinical corpus annotation: challenges and strategies. In: Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012

[8] Uzuner Ö, Solti I, Xia F, et al. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 2010, 17(5): 519-523

[9] Uzuner O, South B R, Shen S D S. 2010 i2b2 / VA challenge on concepts , assertions , and relations in clinical text. *Journal of the American Medical Informatics Association*, 2011,18(5):552-557

[10] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*,2004,32(Database issue):267-270

[11] 杨锦锋, 于秋滨, 关毅等. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014,40(8): 1537-1562

[12] Ogren P V, Savova G, Buntrock JD, et al. Building and evaluating annotated corpora for medical NLP systems. In: Proceedings of the American Medical Informatics Association, 2006 Annual Symposium, Washington, USA, 2006. 1050

[13] Roberts A, Gaizauskas R, Hepple M, et al. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 2009,42(5):950-966

[14] 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

[15] Albright D, Lanfranchi A, Fredriksen A, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 2013, 20(5): 922-930

[16] Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics, Amsterdam, USA, 2008. 2325-2330

[17] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 2008, 34(4): 555-596

The construction of annotated corpora of named entities for Chinese electronic medical records

Qu Chunyan*, Guan Yi*, Yang Jinfeng*, Zhao Yongjie**, Liu Yaxin***

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

(**The Fourth Hospital of Harbin Medical University, Harbin 150001)

(***The Second Hospital of Harbin Medical University, Harbin 150001)

Abstract

In view of the current blank in semantical annotation of named entities of Chinese electronic medical records (CEMRs), a study on construction of annotated corpora for CEMRs' named entities was conducted. By reference to the definitions of named entity type and modification type of electronic medical records given by the US Informatics for Integrating Biology and the Bedside (I2B2) in 2010, an annotation specification for CEMRs was developed under the guidance of professional doctors; Based on the analysis of a large number of CEMRs, a complete scheme for annotation of CEMRs' named denties was proposed, and a large-scale annotated corpus for named entities of CEMRs was established by using the methods of pre-annotating and formal annotating. Its annotation consistency is over 92%. This annotated corpora can provide reliable data for named entity recognition for CEMRs and information extraction research, and it is very useful for medical knowledge mining.

Key words: Chinese electronic medical record(CEMR), named entity, annotated corpora, annotation specification, inter-annotator agreement (IAA)