高技术通讯 2014 年 第24 卷 第6 期:609 ~615

doi:10.3772/j.issn.1002-0470.2014.06.009

面向中文电子病历的词法语料标注研究①

蒋志鹏② 赵芳芳 关 毅③ 杨锦锋

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 针对中文电子病历(CEMR)标注语料匮乏,目前面向中文电子病历的分词和词性标注研究仍处于空白阶段的实际情况,从中文电子病历语料的构建出发,提出了从数据预处理到语料标注的整体方案,获得了较高的标注一致性,为进行更大规模更高质量的病历语料标注工作提供了指导。通过实验量化中文电子病历与开放领域语料、英文电子病历语料的词法统计差异,系统地分析了通用标注模型在中文电子病历中的错误分布,为进行适用于中文电子病历分析的自然语言处理(NLP)技术研究奠定了基础。

关键词 中文电子病历(CEMR),词性标注,标注一致性,语料差异,错误分析

0 引 言

电子病历(electric 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,即能实现存储、管理、传输和重现的医疗记录[1]。电子病历中的非结构化数据的规模远大于结构化数据,非结构化数据蕴藏着丰富的医疗知识,但处理起来会更加困难,成为计算机自动分析的障碍。电子病历是智慧医疗的物质基础,其自动分析依赖于大量的自然语言处理(natural language processing, NLP)技术。目前以统计机器学习为代表的分词、词性标注模型在开放领域已获得了较高的精度,其中中文分词准确率已达到 98% 以上,英文词性标注准确率也达到了 97%,接近人工标注水平。在进行跨领域标注时,通用标注模型在限定领域中的表现与语言的屈折形态丰富性密切相关。比如德文具有非常丰富的屈折形态,能够较好地辅助词性预测,在德国开放语料上训练 TnT 词性标注器,标注医疗文本的准确率接近 97%[2]。英文的屈折形态弱于德文,宾州树库(Penn treebank,PTB)上训练的 TnT 词性标注器,在电子病历上的标注准确率仅为 89.79%[3]。中文几乎没有屈折形态,所以直接应用通用模型标注中文电子病历(Chinese EMR, CEMR)的表现应该更差,然而还没有任何面向中文电子病历的词法分析研究。文献[4]证明了引入领域标注语料能够更好地解决跨领域标注问题。国外早在 2003 年就开展了生物医学文献的标注工作,并形成了著名的 GENIA 语料库[5],该语料库的词性标注规范源于 PTB,只是针对生物医学文献的特殊性进行了调整。文献[6]对医学词表进行了泛化以扩充词性标注集,文献[3,7]则完全沿用 PTB 的标注集,只是针对电子病历中的符号、药名等标注方法做了特别说明。由于中英文词法存在较大差异,宾州中文树库(Penn Chinese treebank,PCTB)的词类数比英文少了近 1/3,所以无法直接应用英文生物医学的标注规范,另一方面,没有任何可以借鉴的中文生物医学标注语料。

本文以 PCTB 的标注规范为基础进行了中文电子病历的词法语料标注研究。在实际标注过程中,以迭代的方式不断调整规范,首次基于电子病历语料构建了国内生物医学领域的分词、词性标注语料,并获得了较高的标注一致性和准确性。基于已构建的标注语料,系统地分析了通用标注模型在不同病历部分中的错误分布,通过实验验证了基于英文电子病历的一系列假设,并总结了中文电子病历新的特点。

① 国家自然科学基金(60975077)资助项目。

② 男,1985 年生,博士生;研究方向:自然语言处理;E-mail:xyf-3456@163.com

③ 通讯作者, E-mail: guanyi@hit.edu.cn

(收稿日期:2013-09-30)

1 中文电子病历的特点

1.1 数据采集

本文所使用的全部电子病历均来自哈尔滨某三级甲等医院，目前已经获取 5000 份完整的神经内科病历，每份病历包括出院小结（死亡小结），以及住院期间所有的病程记录。由于病案室中的电子病历以只读的图片格式存储（例如 tif，jpg 等），不能直接应用自然语言处理（NLP）技术。另外，电子病历以半结构化的方式组织内容，包括入院日期、出院日期、门诊收治诊断、临床初步诊断、临床确定诊断、入院时情况、治疗经过、出院时情况、治疗效果、出院医嘱 10 个部分，各部分之间存在明显的段落分割，如图 1 所示。

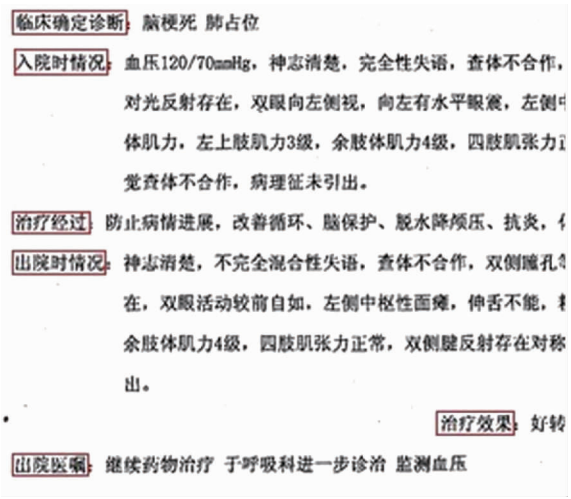

图 1 原始出院小结局部样例

1.2 词法语料标注

分词规范的总则不仅是对整个规范的概述，而且能够为解决规范中未出现问题提供指导。一些规范依靠无法量化的标准判断分词单位，例如，文献[8]将“结合紧密、使用稳定”作为判定标准，文献[9]将该标准与词典收入情况相结合；另一些规范则采用多维度的判定方式，例如，PCTB 规范[10]以共现频率、内部结构复杂性、界限词素等作为分词标准。电子病历中存在大量的专业术语、缩略词及其组合形式，例如神清、自觉、主诉等，上述标准无法较好地覆盖这些语言现象。

分词细则中，PCTB 分词规范仍不能覆盖电子病历中的所有歧义问题，尤其是组合词切分，其中，形容词与形容词的组合在诊断依据、病例特点部分比较常见，例如“左/上/肢”、“深浅/感觉”，PCTB 分词规范并没有给出区别方案；名词与名词的组合，例如血尿、脑实质，PCTB 规范中仅规定了修饰关系的切分标准。

词性标注方面，PCTB 规范[11]同样具有不同程度的不完备性，以下三类问题较为突出：

（1）中文电子病历经常使用特殊符号作为一种缩略词，例如“肌力 4 + 级”中“ + ”表示“强”，应标为形容词（VA），但“ + ”甚至没有在 PCTB 语料库中出现过，只能被标注为标点符号（PU）。

（2）中文电子病历经常以“动补短语作宾语”的形式描述病症，而类似语法模式在 PCTB 中几乎不会出现，导致动词（VV）与名词（NN）的歧义频现，例如“伴有视物模糊”中的“视物”。从描述症状角度看，“视物”可以标注为 NN，从短语结构看，“视物模糊”又是动补短语，可以标注为 VV。

（3）名词修饰语（JJ）的标注歧义在“查体”部分经常出现，例如“左侧肢体麻木”和“左侧中枢性面瘫”中的“左侧”。在 PCTB 规范中“左侧”可能词性为名词（NN）、定位词（LC）以及名词修饰语（JJ），其中 LC 可以根据是否为介词论元或直接修饰动词短语（子句）来判定，NN 和 JJ 则需要判断该词是否成为中心词，但在上例中“左侧”都不是中心词，无法达到消歧的目的。

2 中文电子病历的词法语料标注方案

2.1 数据预处理

针对中文电子病历数据采集的特点，本文首先提出了数据预处理流程，如图 2 所示。该流程既体现了中文电子病历采集过程的特殊性，又包含了面向不同用途的随机抽样方案。

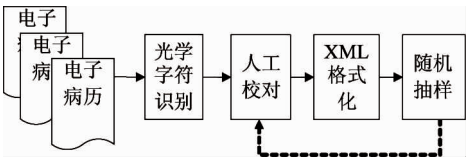

图 2 数据预处理流程

光学字符识别阶段，我们直接使用开源引擎 Tesseract 完成识别工作。经过初步人工校对后，将电子病历处理成可扩展标记语言（extensible markup language，XML）文档，如图 3 所示。XML 能够提供清晰的标签层次结构，并容易通过图形化工具展示，便于后续信息抽取、句法分析等标注工作的开展，并

针对不同部分进行单独分析。

随机抽样阶段,我们分别以篇章为单位和句子为单位随机抽取小规模病历样本,记为 G 语料和 E 语料。借鉴 I2B2 评测(基于英文临床数据的自然语言处理技术评测)将整篇电子病历作为评测集的做法,为了保证电子病历的完整性,能够更好地为信息抽取服务。我们选择 G 语料作为最终开发的标准语料,抽样时唯一的约束条件是"首次病程记录和出院小结不属于同一患者"。E 语料作为开发标注规范的评价语料,应尽可能覆盖各种分词、词性情况。

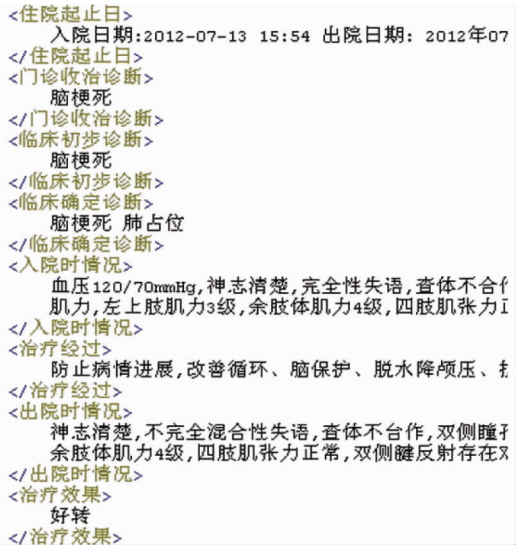

图 3　XML 格式化的出院小结局部样例

2.2　标注规范开发流程

借鉴英文生物医学语料标注经验,本文提出了适用于中文电子病历的标注规范开发方案,如图 4 所示。由于中英文词性标注集存在较大差异,我们将开放领域公认的 PCTB 标注规范[10,11]作为基础规范。在规范评价阶段,标注人数一般会控制在 2 至 4 人[3,7,12],我们选择两位具有语言学背景的研究生作为标注者,相比文献[3]选择医学专家作为标注者,节省了培训通用标注知识的时间,缩短了标注周期。每次迭代开发都从 E 语料中随机抽取 100 句电子病历,两名标注者分别按照标注规范校对自动标注结果,依次评价标注一致性及准确性,最后根据标注歧义修订规范,以进行下一次迭代。考虑到人机互助能够大幅提升标注速度,我们选择附带自定义词典功能的 ICTCLAS 分词系统作为分词工具,每次迭代后通过更新词典提高下次自动分词的准确率。

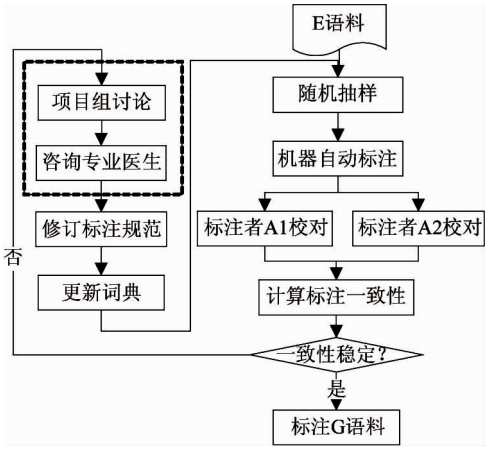

图 4　标注规范开发方案

图 4 中虚线框是整个方案的核心部分,即项目组讨论和咨询专业医生阶段,主要目的是通过解决标注分歧,为修订标注规范提供依据,并生成用于评价一致性的标准语料。一致性判定方法如 3.1 节所述,当内部标注一致性(inter annotator agreement, IAA)、标注准确率(precision)及分词 F1 值在 5 次迭代后均保持在一个较高水平,我们就认为标注一致性达到稳定,该规范能够用于标注 G 语料。

2.3　标注规范修订

在规范修订过程中,本文针对电子病历语料特有的语言现象,结合不同的中文标注规范,补充、细化 PCTB 规范中未覆盖的歧义说明,提出了适用于中文电子病历的分词和词性标注规范,3.1 节实验结果表明,新的规范能够更好地解决电子病历中的标注歧义问题,并最大程度地匹配 PCTB 的语料库。

为了方便后续讨论,下面给出了 3 个核心概念:

定义 1(组合性) 如果某词能够进一步被切分为两个子词的组合,并且切分后的各子词均具有词性意义,则该词具有组合性。

定义 2(替换性) 当某词具有组合性,组合中的子词被其他词替换后,各部分词性均保持不变,并且新的组合仍可能出现在电子病历中,则该词具有替换性。

定义 3(还原性) 如果某词能够还原为完整语义形态,则该词具有还原性。

由于电子病历中整体为名词或者不具组合性的术语通常不切分,例如糖尿病、巩膜,所以我们在总则中重点解决存在较多歧义的非名词性术语切分问题。切分方案如图 5 所示,以术语"抗凝"为例,首先判断该词具有组合性和还原性,能够预切分为"抗"和"凝",并分别还原成动词"阻止"及"凝固",当"凝固"替换成"发炎"时,两部分仍为动词保持不

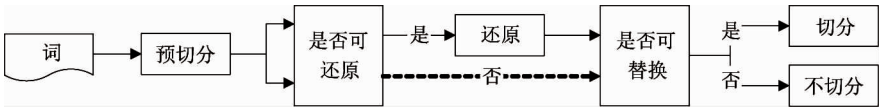

图 5 非名词性术语切分方案

变,且新词“抗炎”同样会出现在电子病历中,所以该词具有替换性,需要进行切分。

在分词总则基础上,本文结合每次迭代的错误分析结果,对 PCTB 分词规范细则进行了修订。对于 1.2 节中组合词切分问题,我们规定当形容词并列修饰后缀时切分,否则不切分;关于名词和名词组合,增加了并列关系的切分标准,并且进一步细化了不同字数修饰关系的切分标准。

针对 1.2 节中词性标注的三类歧义,本文给出如下解决方案:根据上下文标注特殊符号,例如“肌力 4 + 级”中“ + ”表示“强”,应标为形容词(VA),“头 MRI + MRA 示”中“ + ”则表示“和”,标为并列连词(CC);保留词语的本质属性,例如“视物模糊”是动补短语,则统一标注为动词;先补充语法成分再进行标注,例如“左侧肢体麻木”是正常语法结构,“左侧”做 JJ 直接修饰“肢体”,而“左侧中枢性面瘫”实际上省略了谓语“存在”,补充完整后“左侧”应为名词主语,标为 NN。

为了统计词性分布,辅助信息抽取的需要,我们对 PCTB 词性标注集进行了一定的扩展,增加了名词的下属子类药品名(NM)和疾病名(ND),对其他 PCTB 中鲜有的语言现象做了特别说明。另一方面,为了匹配 PCTB 语料库,还保留了直接应用 PCTB 标注集的病历语料,用于评价通用模型的标注效果。

本文提出的分词规范属于针对中文电子病历的简则,整体分为总则、特殊词、组合词三大部分,其中特殊词和组合词根据是否具有组合性划分,共包括 26 项分词规则,均结合了中文电子病历中的实际例子进行说明。词性标注规范保留了 PCTB 规范 30 个词类,增加了名词下属的疾病名和药品名两个子类,并规定短语式疾病名中,相邻名词均标为疾病名。具体歧义说明为 29 项,其中部分说明针对中文电子病历的特点进行了调整。

3 实验与分析

3.1 标注质量控制

保证较高的内部标注一致性(IAA)和标注准确率(precision)是获得高质量语料的关键。双重标注是常用的标注质量控制手段[3,7,12],每次从 E 语料中随机抽取小规模病历样本进行自动标注,标注者 A1 和 A2 分别独立地按规范校对,通过计算 IAA、Precision 及 F1 值对各自标注质量进行评价,即图 4 中一致性判定过程。IAA 及 Precision 计算公式如下:

$$\text{IAA} = \frac{\text{一致标签个数}}{\text{所有标签个数}} \times 100\% \quad (1)$$

$$\text{Precision} = \frac{\text{正确标签个数}}{\text{所有标签个数}} \times 100\% \quad (2)$$

其中 IAA 度量的是两标注者标注一致的结果所占比例,Precision 度量的是某一标注者标注正确结果所占比例,F1 值则是中文分词评测中常用的评价指标,正确标注语料通过图 4 中项目组讨论及咨询专业医生获得。

为了评价规范修订对标注效果的影响,本文对比了规范修订前后主要词性歧义项的分布,表 1 给出了第 1 次和第 3 次标注结果中数量最多的前 5 对歧义项。可以看出,直接使用 PCTB 词性标注规范,名词与动词、名词修饰语的歧义问题占较大比例,经过 2.3 节修订之后,数量分别下降了 47% 和 37%,尽管歧义项分布在一定程度上会受到标注者机械错误的影响,但是结合表 2 中标注准确率整体上升的趋势,仍能证明本文的修订方案对于解决中文电子病历的歧义问题的有效性。

表 1 规范修订前后主要词性歧义项分布

PCTB 词性标注规范			修订后的词性标注规范		
歧义项		数量	歧义项		数量
NN	VV	89	NN	VV	47
JJ	LC	72	NN	M	15
M	LC	64	JJ	NN	9
NN	VA	63	NN	VA	9
JJ	NN	37	ND	NN	6

表 2 为前 3 次迭代中分词和词性标注的准确率及一致性,其中第 1 次迭代直接使用 PCTB 的标注规范,之后依次迭代修订。可以看出,受标注经验及规范适用性影响,第 1 次分词和词性标注的各项指

标均比较低,经过初次大规模的修订之后,第 2 次迭代各项指标有较大幅度提升,一方面证明了规范修订的有效性,另一方面也说明具有语言学背景的标注者能够较快地掌握新的标注知识。第 3 次词性标注时,我们尝试扩展了词性标注集,各项指标并没有较大浮动,究其原因,一方面疾病名和药品名具有较高的辨识度,另一方面可能这两类词语所占比例并不大,本文将在 3.2 节进一步验证该假设。从前 5 次迭代结果的整体趋势看,各项指标已经稳定在较高水平,证明能够开展高质量的语料标注工作。

表 2　前 3 次迭代分词和词性标注准确率及一致性

迭代次数	中文分词			词性标注		
	A1 F1 值(%)	A2 F1 值(%)	IAA(%)	A1 Precision(%)	A2 Precision(%)	IAA(%)
1	96.76	92.27	96.53	96.68	88.53	89.25
2	95.51	96.94	97.89	97.36	97.81	95.18
3	98.49	96.47	98.25	97.80	97.60	95.60

3.2　词法语料分析

本文最终构建了 70 份经过中文分词、词性标注的电子病历语料,包括 30 份首次病程记录及 40 份出院小结。整个语料共包含 1094 个句子,28430 个词,平均每句 25.99 个词,与开放语料(PCTB)27.09 的句长十分接近。从表 3 可以看出,两种语料的词性分布没有显著差异,使用频繁的名词(NN)、标点(PU)、动词(VV)比例较接近,由于电子病历的去隐私化处理,导致专有名词(NR)在电子病历中并不多见,而更多数词(CD)、形容词(VA)的出现也表明中文电子病历的词法特征与英文电子病历相近。

表 3　词性分布比较

中文电子病历		开放领域语料	
词性	比例(%)	词性	比例(%)
NN	27.41	NN	27.21
PU	21.69	PU	15.29
VV	13.15	VV	13.87
CD	8.07	AD	7.22
VA	5.9	NR	6.15

为了进一步分析中文电子病历与开放语料的词法分布差异,我们对电子病历不同部分的句长进行了统计,发现电子病历各部分句长相差较大,出院小结平均 18.02 词/句,其中词数最少的“治疗效果”部分仅为 1 词/句,而首次病程记录平均37.29 词/句,最长的“病例特点”部分达到 67.67 词/句,这也是导致中文电子病历平均句长接近开放语料的原因。另一方面,我们统计了 5000 份电子病历的标点符号分布,如表4 所示。逗号数量超过了标点总数的 1/2,大量描述检查、病征的语句都以逗号分隔,以致中文电子病历中短语频繁出现,而非英文中的短句。所以,“英文电子病历包含大量短句”[3] 的假设并不适用于中文电子病历,需要结合不同病历部分的特点具体分析,另外,这种特殊的语法现象也对句法分析中常用的标点符号分割造成更多的干扰,需要探索更为有效的长句分割方式。

表 4　标点符号分布

标点符号	逗号	句号	顿号	冒号	引号	其他
比例(%)	53.8	14.67	13.75	5.73	3.59	8.5

在单一词性分布差异不明显的情况下,本文引入逐点互信息(pointwise mutual information,PMI)进行对比分析,计算两个相邻词性 t_1 和 t_2 的 PMI 的公式为

$$PMI(t_1,t_2) = \log_2(p(t_1,t_2)/(p(t_1)p(t_2))) \quad (3)$$

当 $PMI(t_1,t_2)$ 为正时表示 t_1,t_2 组合的可能性较大,为负时表示 t_1,t_2 拆分的可能性较大,越接近 0 则彼此越独立。表 5 为中文电子病历与开放领域的平均 PMI 比较,可以看出,中文电子病历平均 PMI 正值高于开放领域语料,甚至比英文电子病历的差距[13] 更明显,说明中文电子病历文法存在更强的规律性,基于规则的 NLP 技术能够更好地发挥作用。

表 5　平均逐点互信息比较

	中文电子病历	开放领域语料
平均 PMI 正值	1.68	1.4
平均 PMI 负值	-1.81	-2.22

在整体分析的基础上,我们训练了开放领域达到最优水平的 Stanford 词性标注器,分别在中文电

子病历不同部分进行自动标注。图 6 和图 7 分别为首次病程记录和出院小结不同部分的词性标注准确率及未登录词(out of vocabulary,OOV)率分布,从准确率曲线可以看出,一方面,中文电子病历整体标注准确率较英文差,仅为82.35%,另一方面,不同病历部分的标注准确率存在显著差异,最低的"治疗效果"部分仅为 25%,而最高的"治疗经过"部分可达 93.98%,相当于开放领域的标注水平。进一步对比两图,尽管这两种电子病历均来自同一科室,可是曲线走势却并不相同,首次病程记录符合传统词性标注准确率与 OOV 率的负相关假设,相关系数达 -0.99,说明未登录词对首次病程记录词性标注效果影响较大,而出院小结相关系数仅为 -0.23,与英文出院小结 -0.94的相关系数[3]截然不同。

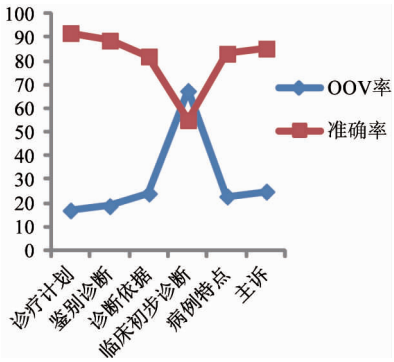

图 6 首次病程记录各部分标注效果

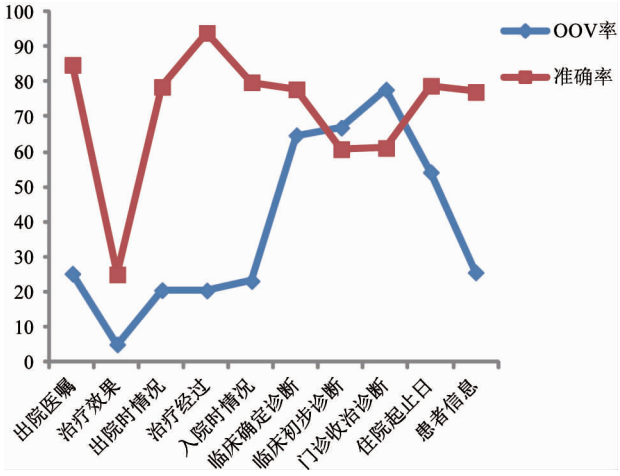

图 7 出院小结各部分标注效果

通过错误分析我们发现准确率较低的"临床初步诊断"、"门诊收治诊断"等部分,不仅有较高的 OOV 率,而且内容多为疾病名称及其修饰语的罗列,缺少词性预测所需的上下文环境。内容组织方式是差异产生的主要原因,首次病程记录以描述病征、检查等叙述性文字为主,导致标注准确率与 OOV 率呈明显的负相关,而中文出院小结相比英文,包含更多的结构化文本,为词性标注设置了新的障碍。针对此类问题,需要与构词法、医学词表等资源相结合,单独设计词性标注、分词模型。另外,准确率较高的"诊疗计划"、"治疗经过"等部分仍然存在一些共性,例如大多属于叙述性短句,与开放领域文法拟合较好,类似地,"入院时情况"、"病例特点"等部分则通常以并列长句的形式描述症状、检查。基于上述假设,在进行适用于中文电子病历的 NLP 技术研究时,不仅要针对同种病历的不同部分展开,还要考虑不同病历相似部分的融合,以达到最佳效果。

最后,为了分析医学术语对词性标注的影响,本文在 2.3 节特别标注了药品名和疾病名,然而,在标准语料中,这两类词仅占未登录词的 7%,错误率也仅占未登录词的 12%,虽然对词性标注结果影响较小,但也说明中文电子病历相比英文还不够完善,特别是出院小结,英文出院小结包含了中文所没有的"入院用药"、"出院用药"等重要信息。受中文电子病历发展现状的限制,进行深入的知识挖掘需要结合不同种类的电子病历,从而带来了更多的困难。

4 结 论

本文提出了适用于中文电子病历的分词及词性标注方案,给出了开放领域规范未覆盖的歧义消解办法,最终构建了 1094 个句子的测试语料。这是国内生物医学领域语料标注研究的初步探索,此研究证明,以语言学研究生为主,医生为辅的方式,结合迭代规范修订,能够更快地完成高质量电子病历语料标注工作。通过统计中文电子病历与开放领域语料的词性及句长分布,证明中文电子病历具有与英文相似的词法特征,而各部分句长存在显著差异,导致平均句长接近开放领域语料,验证了英文电子病历的短句假设并不适用于中文病程记录。逐点互信息的对比结果说明中文电子病历文法具有较英文更强的规律性,基于规则的 NLP 技术能够更好地发挥作用。最后,本文利用开放领域训练的 Stanford 词性标注器对中文电子病历各部分进行自动标注,验证了跨领域标注效果与语言屈折形态的相关性。基于各部分的错误分析结果,本文强调进行面向中文电子病历的 NLP 研究时,不仅要考虑同种病历不同部分的差异性,还要兼顾不同病历相似部分的一致

性,融合相似内容进行协同训练。

许多后续工作仍有待开展:深度上,探索句法分析、信息检索等高级 NLP 任务的标注方案;广度上,通过标注分析不同医疗机构、不同科室的电子病历,获得更具普适性的结论;方法上,探索以主动学习为代表的半自动标注模型,开展大规模病历标注工作。

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). 北京: 中华人民共和国卫生部, 2010

[2] Wermter J, Hahn U. Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. *Studies in health technology and informatics*, 2004, 107(Pt 1): 560

[3] Pakhomov S V, Coden A, Chute C G. Developing a corpus of clinical notes manually annotated for part-of-speech. *International journal of medical informatics*, 2006, 75(6): 418-429

[4] Liu K, Chapman W, Hwa R, et al. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *Journal of the American Medical Informatics Association*, 2007, 14(5): 641-650

[5] Kim J D, Ohta T, Tateisi Y, et al. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003, 19(suppl 1): i180-i182

[6] Smith L, Rindflesch T, Wilbur W J. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, 2004, 20(14): 2320-2321

[7] Savova G K, Masanz J J, Ogren P V, et al. 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

[8] 中国国家标准化管理委员会. GB/T 13715-92,信息处理用现代汉语分词规范. 北京: 中国标准出版社, 1992

[9] 俞士汶, 段慧明, 朱学锋等. 北京大学现代汉语语料库基本加工规范. 中文信息学报, 2002, 16(5): 49-64

[10] Xia F. The Segmentation Guidelines for the Penn Chinese Treebank Project. Pennsylvania: University of Pennsylvania, 2000

[11] Xia F. The Part-of-speech Guidelines for the Penn Chinese Treebank Project. Pennsylvania: University of Pennsylvania, 2000

[12] Albright D, Lanfranchi A, Fredriksen A, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. http://jamia.bmj.com/content/early/2013/01/24/amiajnl-2012-001317.full, 2013

[13] Campbell D A, Johnson S B. Comparing syntactic complexity in medical and non-medical corpora. *Proceedings of the AMIA Symposium*, 2001: 90-94

Research on Chinese electronic medical record oriented lexical corpus annotation

Jiang Zhipeng, Zhao Fangfang, Guan Yi, Yang Jinfeng

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

Abstract

Considering that the research on Chinese word segmentation and part-of-speech (POS) tagging for Chinese electronic medical record (CEMR) is currently at a blank stage because of the lack of annotated corpus on CEMR, a complete scheme for data preprocessing to corpus annotation was proposed starting from corpus construction on CEMR so as to obtain a higher annotation consistency, and to build corpus with larger scale and higher quality on CEMR. Furthermore, the statistical lexical differences between CEMR, open-domain corpus and English electronic health record were quantified, and the systematic error analysis was performed on a POS tagging model trained on open-domain corpus. The work lays the foundation for the research on natural language processing (NLP) technologies for CEMR.

Key words: Chinese electronic medical record (CEMR), part-of-speech tagging, annotation consistency, statistical lexical differences, error analysis