文章编号：2095-2163(2019)02-0130-06 中图分类号：TP391.1 文献标志码：A

基于字级别条件随机场的医学实体识别

何 彬，关 毅
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：开放域分词器对临床文本进行分词处理时，其性能受到了临床文本独特的子语言特性的极大限制，给后续医学实体边界识别造成了不少错误累积。本文针对该问题构建了面向临床文本的分词器，将该分词器用于医学实体识别模型的词特征提取来减少医学实体边界错误，还构建了字级别的条件随机场模型用于识别医学实体，避免了分词给实体边界识别造成的错误累积问题。

关键词：医学实体识别；条件随机场；临床分词器

Character-based CRF for Medical Entity Recognition

HE Bin, GUAN Yi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The unique sub-language characteristics of clinical text greatly limit the performance of the open-domain word segmenter, and this causes a lot of error accumulation for subsequent medical entity boundary recognition. Therefore, a word segmenter dedicated to clinical text is constructed for this problem. This clinical word segmenter is used to extract word features for the medical entity recognition model to reduce medical entity boundary errors. Besides, a character-based CRF model is built to identify medical entities, which avoids the error accumulation problem caused by word segmentation.

【Key words】 Medical Entity Recognition; Conditional Random Fields; clinical word segmenter

0 引 言

医学实体作为临床文本中医学知识的主要载体之一，其识别效果对从临床文本中抽取医学知识尤为关键。目前的研究通常采取监督学习方法在一定数量的标注数据集上训练统计机器学习模型，不过这类模型在特征提取过程中对于已有的自然语言处理工具包比较依赖。然而，在中文临床文本上，由于公开的数据集尤为有限，极大地限制了医学实体识别模型中所使用特征的提取效果。例如，Xu 等人[1]发现开放域文本上训练的分词器在临床文本的分词上效果不佳，词边界的划分直接影响了实体边界的准确度，这说明了医学实体识别模型对于特定的分词器的依赖性较强。因此，构建临床文本上训练的分词器对于医学实体识别模型的特征提取十分关键。此外，医学术语的表述多样性总会给分词器带来一些词划分错误，这些错误会直接传递给实体边界识别造成错误累积，故研究字级别的医学实体识别模型是避免分词带来的错误累积问题的一种解决方案。本研究首先构建词级别的医学实体识别模型，基于该模型上对比了开放域文本和临床文本上训练的分词器对于识别性能的影响，同时也探索了字级别的医学实体识别模型的性能。

在本文中，研究的主要贡献有：

（1）构建了临床文本上的分词器用于医学实体识别模型的词特征提取，并与开放域分词器进行比较，得出分词器对于医学实体识别模型性能的影响。

（2）构建了字级别条件随机场模型用于医学实体识别。

（3）实验结果表明，临床文本上训练的分词器对于医学实体识别性能有显著提升，字级别条件随机场模型避免了词切分对医学实体边界识别带来的错误累积问题，并且提升了模型性能。

1 方法概述

条件随机场模型[2]是命名实体识别任务中最常用的模型，该模型是在给定输入序列 $x=(x_1,\cdots,x_n)$ 的条件下，求条件概率 $p(y\mid x)$ 最大的输出序列 $y=(y_1,\cdots,y_n)$。线性链条件随机场模型可以形式化为：

$$p(y\mid x)=\frac{1}{Z(x)}\exp\Big(\sum_{i,k}\lambda_k\,t_k(y_{i-1},y_i,x,i)+$$

作者简介：何 彬(1989-)，男，博士研究生，主要研究方向：自然语言处理、信息抽取；关 毅(1970-)，男，博士，教授，博士生导师，主要研究方向：自然语言处理、健康信息学、认知语言学等。

收稿日期：2018-12-09

$$\sum_{i,l} \mu_l s_l(y_i, x, i)), \quad (1)$$

其中，

$$Z(x) = \sum_{y \in Y} \exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)). \quad (2)$$

此处，t_k 和 s_l 是特征函数；λ_k 和 μ_l 为特征对应的权重；$Z(x)$ 为归一化因子。

基于统计机器学习的模型通常可以划分为数据预处理、特征提取、模型编码和解码模块。本研究采用条件随机场模型来进行医学实体识别，其流程如图 1 所示。这里将对该模型的各个功能模块展开如下研究描述。

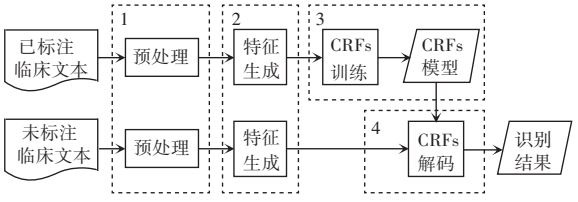

图 1　CRF 模型流程

Fig. 1　Flow diagram of the CRF model

1.1　数据预处理

本研究使用的医学实体标注语料库是以文档为标注单元的数据集，首先就要对该数据集进行必要的预处理操作。涉及的流程步骤可分述如下。

(1)对文本进行句子切分。

(2)针对文本中出现的中英文标点使用不统一的情况进行标点替换。

1.2　基于词的模型特征提取

中文文本没有空格对词来做出划分，在模型训练前需要利用分词器对句子进行词切分。研究中利用斯坦福分词器(Stanford Word Segmenter)[3-4]对临床文本进行分词处理，并基于得到的词序列提取特征。首先，研究利用斯坦福词性标注器(Stanford POS Tagger)[5-6]生成句子的词性标记来提取词性特征，接着参照文献[7]提取词的拼写特征。拼写特征由词内的字的字符类型组合而成，同时在文献[7]中使用的大写字母(X)、小写字母(x)和数字(D)的基础上增加了汉字(C)、符号(S)和其他(O)的字符类型。研究还从网络上爬取了医学术语来构建医学术语字典(数据源有好大夫在线(http://www.haodf.com./ask/)、万方医学网(http://lczl.med.wanfangdata.com.cn/)、国家食品药品监督管理总局(http://samr.cfda.gov.cn/WS01/CL0001/)等)，术语规模见表 1。接下来，将通过判断词是否出现在术语字典中或是术语的组成部分来提取模型的字典特征。表 2 对词级别医学实体识别模型的特征进行了举例说明，详情见表 2。同时，由于临床文本中许多句子长度过长，故在本研究中未引入句法特征。

表 1　医学术语字典规模

Tab. 1　The scale of the medical term dictionary

实体类型	数量
治疗	69 812
检查	5 779
疾病	31 051
症状	17 765
其它	317 142

表 2　词级别医学实体识别模型的特征

Tab. 2　Features for the word-based medical entity recognition model

特征类别	特征	特征值举例*
词汇特征	词本身	高血压
	词性标记	NN
拼写特征	词内的字的字符类型组合	C
字典特征	词是否出现在“治疗”术语字典中	0
	词是否出现在“检查”术语字典中	0
	词是否出现在“疾病”术语字典中	1
	词是否出现在“症状”术语字典中	0

注：* 表示样例“既往 有 高血压 病史 2 年，无 糖尿病史，否认 肝炎”在当前词为“高血压”时的特征值。

1.3　面向临床文本的分词器构建

考虑到开放域文本上训练的分词器在临床文本上分词效果不佳的现状[1]，文中基于宾州中文树库的分词规范[8]构建了中文临床文本上的分词规范[9]，在 138 份临床记录上标注了分词信息，其中包含 2 612 个句子，47 426 个词。研究在该数据集上训练了基于条件随机场的分词器，使用的特征有 c_i，c_{i-1}，c_{i+1}，$c_{i-2}\ c_{i-1}$，$c_{i-1}\ c_i$，$c_i\ c_{i+1}$，$c_{i+1}\ c_{i+2}$，其中 c 表示字，i 表示当前字的索引。而后用该分词器替换第 1.2 节中的分词器进行临床文本的词划分，在保持其它特征不变的情况下重新构建了医学实体识别模型。

1.4　基于字的模型特征提取

虽然本研究在医学实体识别模型中引入了临床文本上训练的分词器，但是分词器总会产生一些词切分错误，给后续实体识别任务造成错误累积。因此，研究中也进行了基于字的医学实体识别构建，将分词信息作为特征的形式加入到模型中，避免词切分直接造成实体边界识别错误。这里对基于字的医学实体识别的字序列和标记序列进行了举例说明，详情参见表 3。

表 3 基于字的医学实体识别的字序列和标记序列距离

Tab. 3 An example of the character sequence and label sequence for the character-based medical entity recognition model

字	既	往	有	高	血	压	病	史	2	年
标记	O	O	O	B-D	I-D	I-D	I-D	E-D	O	O

在表 3 中,“O”是实体外部标记,“B-D”是实体开始标记,“I-D”是实体内部标记,“E-D”是实体结束标记,“D”表示疾病。

表 2 中的特征在适配到字序列上即可得到字级别医学实体识别模型的特征。在分析时,则使用双向最大匹配算法对句子进行医学术语匹配,根据匹配结果来生成句中每个字对应的字典匹配标记。此后,就对模型中的特征进行了举例说明,具体见表 4。

表 4 字级别医学实体识别模型的特征

Tab. 4 Features for the character-based medical entity recognition model

特征类别	特征	特征值举例*
词汇特征	字本身	高
	字的分词标记	B
	字的词性标记	B-NN
拼写特征	字的字符类型	C
字典特征	字对应的字典匹配标记	B-D

注:* 表示样例“既往 有 高血压 病史 2 年 , 无 糖尿病史 , 否认肝炎”在当前字为“高”时的特征值。

2 实验与分析

2.1 实验数据

研究选取文献[10]中构建的医学实体标注语料库作为实验数据。并且以文档为单位对每个科室的数据进行随机等比划分,得到了包含 521 份文档的训练集和 471 份文档的测试集。本数据集中包含治疗、检查、疾病和症状四大类医学实体,其类别数量分布见表 5。

表 5 医学实体数据集类别数量分布

Tab. 5 The distribution of entity types in the medical entity dataset

类型	训练集	测试集
治疗	2 730	2 488
检查	3 454	3 270
疾病	4 246	4 074
症状	9 755	9 034
总计	20 185	18 866

2.2 实验设置

在本研究中,重点选取 BIESO 的标记策略给数据打标签。同时运用准确率、召回率和 $F1$ 值来评价每一个实体类别的性能,并用所有实体类别上的微平均 $F1$ 值来评估模型的整体性能,研究推得其数学计算公式如下:

$$Precision = True\ Positive/(True\ Positive + False\ Positive); \quad (3)$$

$$Recall = True\ Positive/(True\ Positive + False\ Negative); \quad (4)$$

$$F1 = 2 * Precision * Recall/(Precision + Recall). \quad (5)$$

其中, *True Positive* 表示系统输出中边界和类型都与标准数据相匹配的实体数量;*False Positive* 表示系统输出中未出现在标准数据中的实体数量;*False Negative* 表示系统输出中未召回的在标准数据中的实体数量。

研究中采取了标准的数据划分进行实验,并有针对性地设计了下列 3 种模型在该数据集上做出比较。这 3 种模型的功能设计可描述如下。

(1)词级别的条件随机场模型(CRF):该模型以词序列作为输入,使用斯坦福分词器对实验数据进行分词处理。

(2)引入临床文本上训练的分词器的 CRF 模型(CRF+CWS):研究使用临床文本的分词语料上训练的分词器替换 CRF 模型中的分词器来进行词划分,其它特征保持不变。

(3)字级别的条件随机场模型(Char-CRF):该模型以字序列作为输入,模型中的分词特征通过临床文本上训练的分词器提取。

2.3 实验结果与分析

2.3.1 系统性能

各个医学实体识别模型的综合性能对比见表 6。表 6 中,Lex 表示词汇特征;Ort 表示拼写特征;Dic 表示字典特征;CWS 表示临床分词器。从表 6 中可以看出,基于字的条件随机场模型取得了最好的识别结果,其 $F1$ 值达到了 91.00%。该模型比使用等同特征集合的词级别的医学实体识别模型在 $F1$ 值上提升了 1.02 个百分点,说明了将分词信息以特征的形式加入模型训练不仅避免了词切分对医学实体边界识别带来的错误累积问题,还对医学实体识别性能有所提升。

表 6 医学实体识别模型性能比较

Tab. 6 Performance comparison of medical entity recognition models

模型	准确率	召回率	$F1$ 值
CRF (Lex)	87.30	80.97	84.01
CRF (Lex)+CWS	91.18	87.89	89.50
CRF (Lex+Ort)+CWS	91.18	88.17	89.65
CRF (Lex+Ort+Dic)+CWS	91.42	88.58	89.98
Char-CRF	**92.18**	**89.86**	**91.00**

在只使用词汇特征的模型上，引入临床文本上训练的分词器对医学实体识别效果有大幅的提升，模型 *F*1 值提升了 5.49 个百分点，这说明了临床文本具备的独特的文本特点使得开放域分词器直接作用在该文本类型上时会出现领域适应问题，不同的文本特点降低了开放域分词器的分词精度。为减少对后续任务造成的错误累积，训练面向临床文本的分词器是有必要的。

在使用词汇特征的模型中先后增加了拼写特征和字典特征后，模型的 *F*1 值分别提升了 0.15 和 0.33个百分点，验证了词内的字符类型表达与医学术语字典对于医学实体识别性能上的增益作用。

2.3.2　各类医学实体的识别效果

针对每一种医学实体类型，研究也比较了上述各个医学实体识别模型的识别性能，相应比较结果见表 7。表 7 中，Lex 表示词汇特征；Ort 表示拼写特征；Dic 表示字典特征；CWS 表示临床分词器；表 7 中，括号里的数值是与前一模型对比的差值。

表 7　各类医学实体的识别性能比较

Tab. 7　Performance comparison on each medical entity type

模型	*F*1 值			
	治疗	检查	疾病	症状
CRF（Lex）	82.64	84.78	86.00	83.23
CRF（Lex）+CWS	84.39（+1.75）	90.18（+5.39）	88.93（+2.93）	90.88（+7.64）
CRF（Lex+Ort）+CWS	84.68（+0.29）	89.89（−0.29）	89.30（+0.37）	91.05（+0.17）
CRF（Lex+Ort+Dic）+CWS	85.42（+0.74）	89.97（+0.08）	89.85（+0.55）	91.26（+0.21）
Char-CRF	**87.19**（+1.77）	**90.66**（+0.69）	**90.90**（+1.05）	**92.21**（+0.95）

研究发现，临床分词器对于所有医学实体类别都有识别性能提升，尤其是检查和症状实体，分别在 *F*1 值上提升了 5.39 和 7.64 个百分点。通过对比斯坦福分词器和临床分词器对于医学实体边界的切分结果，研究中对分词结果与测试集标准中实体边界冲突的数量进行了统计，统计后结果见表 8。从表 8 中可以看出，与斯坦福分词器相比，临床分词器可以大量减少分词结果与实体边界的冲突，其中症状实体减少得最多，检查实体次之，这也解释了在加入临床分词器后对于各类医学实体性能提升的幅度。

表 8　分词结果与测试集标准中实体边界的冲突

Tab. 8　Conflict between word segmentation results and entity boundaries in the test set

分词器	治疗	检查	疾病	症状
Stanford Word Segmenter	115	247	165	858
Clinical Word Segmenter	54（−61）	44（−203）	41（−124）	108（−750）

在模型中加入字典特征后，各类医学实体的识别效果都有所提升。结合表 1 中对于各类医学术语字典规模的统计值可以发现，加入字典特征后的模型提升幅度与字典规模成正比。因此，收集更丰富的医学术语资源是提高医学实体识别的一种可行的措施。

3　相关工作

医学实体识别任务通常采取 2 种方法，分别是：基于字典和规则的方法和基于机器学习的方法。对该部分内容可依次给出探讨论述如下。

在医学领域，大量的医学术语给该领域的实体识别研究增加了难度，故医学术语的术语表在该任务中发挥了巨大的作用，这些术语表通常可称为受控词表（controlled vocabulary）。目前，在医学术语上收录最全的是 UMLS 超级叙词表，该表中汇集了 100 多种受控词表，而跻身其间的最著名的受控词表有 ICD−10[11]、MeSH[12]、SNOMED CT[13] 等。基于字典模式来识别医学实体的典型系统是梅奥诊所的 cTAKES（clinical Text Analysis and Knowledge Extraction System）[14]，该系统利用浅层句法来识别所有的名词短语作为字典的查找窗口，而后利用 UMLS 中的字典子集来识别疾病、治疗过程、解剖学部位、药品实体。

基于字典和规则的方法虽然在识别准确率上有一定的可靠性，但是同一个医学术语的多种表达方式以及术语表对所有医学术语表述覆盖不全的问题，则使得该方法在识别的泛化能力上要逊色于基于机器学习的方法。i2b2（Informatics for Integrating Biology & the Bedside）在 2010 年首次对临床文本上的医学实体进行系统分类，并参照 UMLS 中的语义类型把医学实体划分为医疗问题、检查和治疗三类[15]。这 3 种医学实体一定程度上体现出临床文本的面向问题的组织方式[16]，检查是为了揭示或证实医疗问题，治疗是为了治愈或缓解医疗问题。在

2010 i2b2/VA 评测中,大多数参与者采用基于统计学习模型的方法,有一些参与者还选取基于规则的方法,将其用于研发所涉及的预处理或后处理。de Bruijn 等人[17]采用半马尔科夫模型对词序列进行医学实体识别,并选用 4 种标记(outside,problem,treatment,test)替代 BIO 标记体系给词序列打标签。研究中借助半监督学习方法提取的词级聚类特征,同时还利用 cTAKES 和 UMLS 来抽取词级特征。在实验过程中,通过采取自训练方式的半监督学习方法对未标注数据加以利用,训练出的模型在 2010 i2b2/VA 评测上取得了最佳效果。

在评测结束后,Tang 等人[18]在评测数据集上对比了使用相同特征情况下结构化支持向量机(Structural Support Vector Machine,SSVM)和条件随机场模型的性能,并在系统中引入了基于聚类的和分布式的词表示特征。实验表明,SSVM 模型比 CRF 模型具有更好的泛化能力,在医学实体识别上取得了更好的效果,而且 2 种词表示特征对模型效果提升也是有利的。Lv 等人[19]利用基于实例的迁移学习方法 TrAdaBoost 来探索从一个医疗机构的数据上训练的实体识别模型在应用到另一个医疗机构的数据上时的效果。在此次工作中,为了能够负迁移带来的影响,研究使用 Bagging 策略来进行权重更新。实验表明,该方法只需使用目标领域上少量的标注数据就能够达到更优的性能,因而对解决跨医疗机构临床文本的实体识别任务十分有效。

中文临床文本上的医学实体识别研究起步较晚,以叶枫等人[20]利用条件随机场模型在中文临床文本上识别疾病、临床症状、手术操作这 3 类医学实体的研究作为标识,随即开启了中文临床文本上的医学实体识别研究进程,不过该研究中定义的医学实体类型还未臻至全面。借鉴 2010 i2b2/VA 评测中的医学实体分类体系,Xu 等人[1]在 336 份出院小结上构建了分词和医学实体的语料库,增加了药品和解剖学部位这 2 类实体。研究中基于 CRF 模型,对比了独立模型、增量模型、标签联合模型和对偶分解联合模型,实验结果表明,对偶分解模型结合了分词与医学实体识别任务的相关关系,在医学实体识别任务上的效果达到了最好。与 Xu 等人的研究类似,Lei 等人[21]将治疗细分为药物和过程,在北京协和医院的临床文本上构建了 800 份医学实体语料库。研发时重点基于字袋、分词、词性、区段信息等特征在该语料库上对比了 CRF、SVM、ME 和 SSVM 模型的效果。实验结果表明分词和区段信息的特征组合取得了最佳的识别效果,SSVM 模型在该任务上的效果要优于其它 3 种统计机器学习模型。基于 Lei 等人构建的数据集,Wu 等人[22]利用深度神经网络训练医学实体识别模型,并基于中文临床文本训练词向量加入模型训练,实验结果表明该模型效果优于 CRF 模型,并且验证了中文临床文本上训练的词向量对模型性能有较大提升。

前述研究都是面向所有科室的临床文本进行医学实体识别,但对于特定疾病的信息抽取任务时,现有的实体分类体系就显得比较粗糙。为了抽取肿瘤相关信息,Wang 等人[23]在肝癌手术记录上探索了基于规则和基于 CRF 模型的抽取方法。但是该模型中利用的特征比较局限,并未引入更多的语言学特征,导致模型效果受限。

在中文临床文本上,中医的临床文本也是位列其中的一种数据类型。Wang 等人[24]在中医临床文本的主诉上研究症状实体识别方法。通过在标注单元、序列标注单元选取、标记集合上进行适配,研究对比了 HMM、MEMM 和 CRF 模型在该任务上的性能。实验表明,针对主诉内容采取的序列标注策略适配在该任务上是合适且有效的,CRF 模型的效果在该任务上要优越于 HMM 和 MEMM 模型。

4 结束语

本文面向临床文本的特点提出了不同的模型进行医学实体识别。临床文本上训练的分词器改善了词边界切分的准确度,有效地提升了模型识别性能。此外,字级别的条件随机场模型将分词结果以特征的方式引入到模型训练中,避免了词边界错误给医学实体识别带来的错误累积,并且进一步提升了模型性能。

参考文献

[1] XU Yan, WANG Yining, LIU Tianren, et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries [J]. Journal of the American Medical Informatics Association, 2014, 21(E2): e84-92.

[2] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling data [C] // Proceedings of 8th International Conference on Machine Learning (ICML). Bellevue, Washington, USA: Morgan Kaufmann,2001: 282-289.

[3] TSENG H, CHANG P, ANDREW G,et al. A conditional random field word segmenter [J]. Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. Jeju Island, Korea: ACL, 2005:1-4.

(下转第 142 页)