中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1546–1557
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体计算与自然语言处理专刊 • 论文

基于句法和语义特征的疾病名称识别

何云琪, 刘苏文, 钱龙华*, 周国栋

苏州大学计算机科学与技术学院, 苏州 215006
* 通信作者. E-mail: qianlonghua@suda.edu.cn

收稿日期: 2018–08–10; 接受日期: 2018–09–19; 网络出版日期: 2018–11–14

国家自然科学基金重点项目 (批准号: 2017YFB1002101) 和国家自然科学基金项目 (批准号: 61373096) 资助

摘要 生物医学实体识别 (如基因/蛋白质、化学物和疾病等) 是生物医学文本挖掘的基础, 它对生物医学实体关系的抽取和生物医学知识库的建立等方面都有着重要的研究意义. 针对目前的疾病名称识别中存在的问题, 本文提出了一系列新的句法特征和语义特征来提高疾病名称识别的性能, 其中句法特征包括组块和依存信息, 语义特征包括疾病名称的缩写信息、字典信息和疾病概念之间的上下位关系等. 在 NCBI 疾病语料库上的实验表明, 结合一系列句法和语义特征的 CRF 模型可以显著提高疾病实体识别的性能, 取得了目前该语料库上的最高 $F1$ 值 85.3%.

关键词 疾病名称识别, 条件随机场, 句法特征, 语义特征

1 引言

疾病实体识别任务是指从生物医学文献中自动找出相应的疾病名称, 作为生物医学文本挖掘的第一步, 它对生物医学知识库的构建、新药研制、药物发现与安全监督有着重要的研究意义. 随着生物文献数量的爆炸式增长以及新的疾病的发现, 人工从医学文献中检索疾病名称, 由于其成本高、时间长而难以满足要求, 因此, 如何可靠地从生物医学文献中自动识别出疾病名称是当前亟待解决的首要问题之一.

疾病实体识别可以看作是命名实体识别在生物医学领域的应用, 通用领域的命名实体识别任务是从新闻文本中识别出人名、地名和机构名等实体, 传统的机器学习方法往往采用 CRF 序列标注模型来实现, 其特征包括词、词干、词缀和词形等词汇特征以及实体字典等语义特征. 由于语言的歧义性, 有时依靠词汇本身难于辨别出实体, 如 “Apple” 可能指水果也可能指苹果公司, 这就需要通过结合上下文来理解; 另一方面, 新的命名实体不断涌现出来, 如新的人名、机构名等, 因此借助于外部资源的实体识别方法必须不断扩充其实体字典以满足新实体的需求.

引用格式: 何云琪, 刘苏文, 钱龙华, 等. 基于句法和语义特征的疾病名称识别. 中国科学: 信息科学, 2018, 48: 1546–1557, doi: 10.1360/N112018-00210
He Y Q, Liu S W, Qian L H, et al. Disease name recognition based on syntactic and semantic features (in Chinese). Sci Sin Inform, 2018, 48: 1546–1557, doi: 10.1360/N112018-00210

与通用领域实体识别相似, 目前的疾病名称识别方法主要有基于词典和规则的方法、基于传统机器学习的方法和基于深度学习的方法等. 基于词典和规则方法虽然准确性高, 但灵活性较低, 不能适应不同领域的要求. 传统机器学习方法的关键问题是如何从文本中提取各种有效的词法、句法和语义特征, 然后利用序列标注模型进行疾病名称的识别. 深度学习方法重点在于如何在词向量的基础上利用各种神经网络模型进行疾病名称识别. 在疾病名称识别方面, 深度学习方法目前还没有超过传统的机器学习方法.

除了与通用领域内存在的歧义性和新名词等共同难点外, 在生物医学领域内的疾病名称识别还具有自己的特点. 疾病名称的定义非常宽泛并且变体较多, "疾病" 定义为 "在一定病因作用下自稳调节紊乱而发生的异常生命活动过程, 表现为症状、体征和行为的异常"[1)]. 这使得疾病名称在医学文献中常常以症状、体征、行为等形式出现, 并常以其上位词来代替. 例如, 在子句 "cause serious illness and early death" 中, "early death" 一词为疾病名称, 但它在 MEDIC 数据库中对应的实体是其上位词 "death", 如何准确识别该类疾病的边界是疾病名称识别的一个难点. 另一方面, 由于疾病命名方式较为复杂, 因此常用简称或缩写来代替, 从而出现疾病名称变体较多的现象, 例如 "HIV" 表示 "Human Immunodeficiency Virus"、"CT" 表示 "Copper Toxicosis". 这种缩写形式本身携带的语义信息量较少, 如何识别疾病的缩写是疾病名称识别的另一个难点.

针对上述问题, 本文在序列标注模型的基础上提出了一系列新的句法和语义特征, 包括句法层面的组块和依存信息, 语义层面的疾病缩写、字典以及上下文层次关系等信息, 这些特征能够有效地捕获疾病名称的句法结构和语义信息, 从而显著提高疾病名称识别的性能.

2 相关工作

生物医学领域的疾病名称识别可以粗略地分成 3 大类: 基于词典和规则的方法、基于传统机器学习的方法和基于深度学习的方法.

基于词典 [1,2] 的方法使用生物医学方面的词典来匹配文献中的生物医学实体名称, 其性能取决于使用的匹配算法和词典的规模. 基于规则 [3,4] 的方法使用正则表达式来匹配生物文献中的实体名称, 这种模式匹配对领域知识要求较高. 由于疾病名称的多样化和新的疾病名称的出现使得词典和规则方法不能有效覆盖大多数的疾病名称, 所以上述方法往往召回率较低.

基于传统机器学习 [5,6] 的方法是目前解决疾病名称识别的主流技术, 常用的机器学习模型较多, 有支持向量机模型 (support vector machine, SVM) [7]、隐 Markov 模型 (hidden Markov model, HMM) [8]、最大熵模型 (maximum entropy model, MEM) [9]、条件随机场 (conditional random field, CRF) [10] 等, 其中 CRF 模型被广泛用于实体识别任务. Leaman 等 [11] 提出的 BANNER 系统, 使用 CRF 模型结合一系列词法特征识别通用的生物医学实体, 经过调整之后可以识别多种生物医学实体. Lu 等 [12] 应用此模型构建 DNorm 系统, 在 NCBI [13] 疾病数据集上的疾病名称识别任务中获得了 79.8% 的 $F1$ 值. Lu 等 [14] 还提出采用联合学习的模型来实现疾病实体识别和规范化系统 TaggerOne, 其中在 NCBI 疾病数据集上的疾病实体识别任务上取得了 82.9% 的 $F1$ 值. Lou 等 [15] 使用状态机的方式实现联合学习模型, 在该语料上取得 82.1% 的 $F1$ 值.

基于深度学习 [16~18] 的方法在自然语言处理任务上获得了广泛的应用, Sahu 等 [19] 首次提出使用深度学习的方法在 NCBI 疾病数据集上识别疾病名称, 他们采用 RNN (recurrent neural network) [20]

1) https://baike.baidu.com/item/疾病.

表 1 疾病名称识别的词法特征

Table 1 Lexical features for disease name recognition

Word	POS	Stem	Lemma	Affix	Context
Rheumatic	JJ	rheumat	Rheumatic	R, Rh, Rhe, c, ic, tic	for, Diseases
Diseases	NN	diseas	Disease	D, Di, Dis, s, es, ses	Rheumatic, database

中的双向 LSTM [21] 模型来捕获词序列语义信息, 并利用 CNN (convolution neural network) [22] 来捕获每个词中的字符序列语义信息, 实验结果的 $F1$ 值为 79.1%. Dang 等 [23] 提出的 D3NER 系统, 使用双向 LSTM-CRF 模型来实现序列化标注任务, 输入包含词、词性以及缩写等特征, 在 NCBI 语料库的实验结果 $F1$ 值为 84.68%.

3 CRF 模型

条件随机场模型被广泛应用于生物医学实体识别任务, 它是一种判别式无向图模型, 对多个特征函数在给定观测值后的条件概率进行建模, 假设 $X = \{x_1, x_2, \ldots, x_n\}$ 为观测序列, $Y = \{y_1, y_2, \ldots, y_n\}$ 为相应的标记序列. 条件概率 $P(y|x)$ 的定义如下所示:

$$P(y|x) = \frac{1}{Z} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right), \tag{1}$$

其中

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right), \tag{2}$$

其中 t_k 是定义在边上的特征函数, 称为转移特征, s_l 是定义在结点上的特征函数, 称为状态特征. t_k 和 s_l 都依赖于位置, 是局部特征函数, 取值为 1 或 0. λ_k 和 μ_l 是对应的权值, $Z(x)$ 是规范化因子.

条件随机场完全由特征函数和对应的权值决定. 当使用条件随机场模型来实现命名实体识别任务时, 关键问题是如何根据具体情况定义大量有效的特征函数. 本文根据疾病实体名称识别的特点设计了一系列词法、句法和语义特征等, 旨在提高其识别性能.

4 CRF 特征

4.1 词法特征

与其他疾病识别系统 [11,12] 一样, 本文使用了疾病名称识别中的基本词法特征集合, 包括词汇特征 (word)、词性特征 (POS)、词干特征 (stem)、词原型特征 (lemma)、词缀特征 (affix)(1~3)、上下文特征 (context) 等. 在句子 “Twins with AS were identified from the Royal National Hospital for Rheumatic Diseases database” 中疾病名称 “Rheumatic Diseases” 的词法特征如表 1 所示.

以单词 “Rheumatic” 为例, 词性为 “JJ” (形容词), 其词干为 “rheumat”, 词原型为 “Rheumatic”, 词缀表示词的前缀和后缀 (长度分别为 1~3), 上下文特征窗口为 2, 即考虑前一个单词 “for” 与后一个单词 “Diseases”.

表 2 组块特征

Table 2 Chunk features

Feature name	Feature type	Feature meaning
Chunk	SBIEO	Chunking tag of the current word

表 3 依存特征

Table 3 Dependency features

Feature name	Feature type	Feature meaning
Dependency	String	Headword of the current word

4.2 句法特征

生物医学文献中存在大量结构相对复杂的短语和句式, 造成疾病名称识别中的边界错误. 在大多数情况下, 疾病名称表现为名词或名词短语, 如 “skin tumors”, 因此将其短语结构作为特征将有助于识别疾病名称的边界; 另一方面, 对于复杂短语结构的疾病名称, 其内部单词之间也存在依存关系, 如复杂实体 “male and female breast cancer” 中 “male” 和 “female” 之间存在并列关系且它们都依赖于 “cancer”, 这种依存关系表明了疾病名称的内部结构, 也有利于改善边界识别问题.

为了提取句法特征, 本文首先对句子进行组块分析和句法分析, 得到组块结构和词汇之间的依存关系, 然后在此基础上构造句法特征集, 具体特征包括:

组块特征. 组块, 也称短语块, 是指一个句子中不包含其他短语块的最大短语, 组块除名词组块 (NP) 外, 还包括动词组块 (VP), 如 “will require” 及介词组块 (PP), 如 “such as”. 在获得组块特征后, 本文根据 SBIE 机制[2)]给短语块中的每个单词分别赋予不同的特征标记, 如表 2 所示.

依存特征. 依存关系是指句子中各个单词之间存在的语法关系, 如主语、谓语和宾语等, 依存关系能有效地捕获句子中单词之间的长距离依存关系. 本文首先利用依存句法分析器得到词汇之间的依存关系, 然后将一个单词所依赖的单词作为依存特征加入. 初步实验表明, 在依存特征中, 当前词所依赖的中心词对疾病名称识别的作用最大, 特别地, 当中心词为复合词时, 取其最右面的部分, 如 “5-phosphatase” 中 “phosphatase”. 依存特征表示形式如表 3 所示.

如图 1 所示, 表 1 所提及句子的依存分析树表示了各单词间的依存关系, 如 Rheumatic 和 database 存在名词修饰关系, 即 NMOD (Rheumatic, database).

表 4 列出该句中部分词的句法特征, 包括组块特征和依存特征 (总称为 SYN), 例如:

- 以词 “identified” 为例, 其组块特征 “E-VP” 表示 “identified” 是动词短语 “were identified” 中的最后一个词;
- 以词 “from” 为例, 其依存特征为 “identify” 表示 “from” 的中心词是 “identify”.

4.3 语义特征

疾病实体名称中存在大量的缩写形式以及使用下位词来代替实体名称的现象, 这种缩写形式往往携带的语义信息量较少且歧义性较强, 因而难于识别, 因此本文提出用缩写特征集来丰富疾病缩写的语义信息, 旨在提高疾病缩写的识别性能; 另一方面, 疾病实体之间存在丰富的上下位关系, 如 MeSH

2) SBIE 分别表示单词构成的疾病名称, 多词疾病名称中的第一个词、中间词和最后一个词.

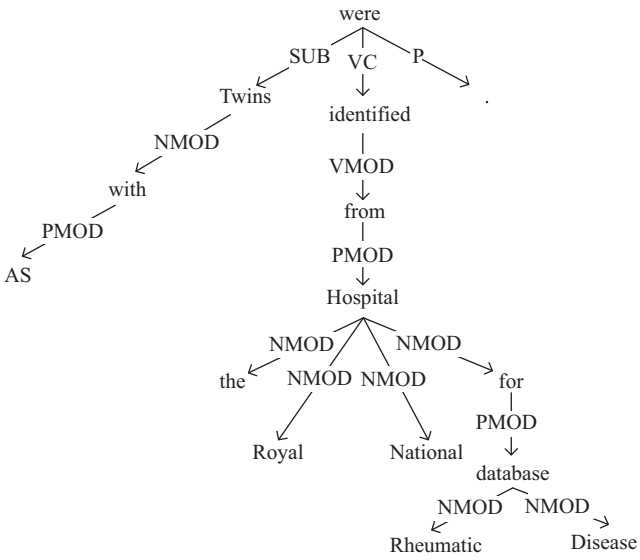


图 1　依存分析树

Figure 1　Dependency parse tree

表 4　句法特征集

Table 4　Syntactic feature set

Word	Chunk feature	Dependency feature
Twins	S-NP	be
with	S-PP	Twin
AS	S-NP	with
were	B-VP	be
identified	E-VP	be
from	S-PP	identify

表 5　缩写特征集

Table 5　Abbreviation feature set

Feature name	Feature type	Feature meaning
IS_ABB	Binary	Whether the current word is an abbreviation
IS_SF	Binary	Whether the short form of the abbreviation is the name of a disease
IS_LF	Binary	Whether the long form of the abbreviation is the name of a disease
Headword	String	Headword of the long form in the abbreviation

树中的分层体系, 这些语义分类编码特征有助于识别特定的或具有上下位关系的疾病名称. 构造缩写特征集和语义编码特征集的具体步骤如下:

缩写特征集 (ABB). 在生物医学文献中经常对复杂的疾病名称进行缩写, 如用缩写 “APC” 来代替疾病 “Adenomatous Polyposis Coli”, 前者称为疾病名称的短形式, 后者称为长形式, 疾病的短形式和长形式构成了缩写对. 缩写特征集合 (ABB) 包含 4 个子特征, 如表 5 所示.

在获得 “IS_SF” 和 “IS_LF” 特征时, 通过字符串匹配来判断短形式和长形式是否出现在 MEDIC

表 6 语义编码特征集

Table 6 The feature set of semantic code

Feature name	Feature type	Feature meaning
MEDIC	SBIEO	Whether the current word sequence appears in MEDIC
MeSH	SBIE+NONE	Classification code of the current word sequence in MeSH tree

表 7 语义特征集

Table 7 Semantic feature set

Word	Abbreviation features	Semantic code features	Tag
Twins	0, 0, 0, None	O, S-M01.438.873	O
with	0, 0, 0, None	O, NONE	O
AS	1, 1, 1, spondylitis	S_Disease, NONE	S_Disease
were	0, 0, 0, None	O, NONE	O
identified	0, 0, 0, None	O, NONE	O
from	0, 0, 0, None	O, NONE	O
the	0, 0, 0, None	O, NONE	O
Royal	0, 0, 0, None	O, NONE	O
National	0, 0, 0, None	O, NONE	O
Hospital	0, 0, 0, None	O, NONE	O
for	0, 0, 0, None	O, NONE	O
Rheumatic	0, 0, 0, None	B_Disease, B-C17.300.775	B_Disease
Disease	0, 0, 0, None	E_Disease, E-C17.300.775	E_Disease
database	0, 0, 0, None	O, S-V02.300	O

数据库中; 在获得缩写对中长形式的中心词时, 直接选择长形式序列的最右边的词作为中心词. 虽然这样做在少数情况下会导致中心词提取错误, 但可以省去句法分析所带来的开销和错误.

语义编码集. 字典编码特征和分类编码特征通常对命名实体识别具有促进作用, 前者是指词序列是否出现在实体字典中, 后者是指词序列在 MeSH 树中的结点编码. MeSH 中的所有概念都按照层次体系进行分类组织, 因而每个概念都是 MeSH 树中的一个结点, 每个结点都有一个表示其层次关系的分类编码, 如 “C17” 表示 “Skin and Connective Tissue Diseases”. 语义编码特征集中包含两个子特征, 即字典编码和分类编码, 用以体现疾病的字典形式和上下位层次关系, 如表 6 所示, 两个子特征的获取方法如下:

- 在获得 MEDIC 特征时, 本文使用最长匹配的原则在句子中查找 MEDIC 中的疾病名称, 如果找到一个匹配, 则根据 SBIE 机制给匹配的各个词分别赋予不同的标记, 否则标记为 O.
- 在获得 MeSH 特征时, 同样使用最长匹配原则在句子中查找 MeSH 中的疾病名称, 如果找到一个匹配, 则根据 SBIE 机制给匹配的各个词分别赋予不同的树结点编码标记, 否则标记为 NONE. 初步实验表明, 使用第 1 层至第 3 层的分类编码信息能够获得最佳性能.

表 7 列出了表 1 所提及的句子中的所有单词的语义特征, 包括缩写特征集和语义编码集. 例如:

- 以词 “AS” 为例, 其缩写特征集 “1, 1, 1, spondylitis” 表示 “AS” 是缩写形式, 其短形式和长形式均为疾病名称, 且长形式的中心词为 “spondylitis”.

表 8 NCBI 数据集

Table 8 NCBI dataset

Statistics	Train set	Dev. set	Test set
Abstracts	593	100	100
Sentences	5818	958	1080
Entities	5145	787	960

• 以词 “Rheumatic” 为例, 其语义编码集 “B_Disease, B-C17.300.775” 表示 “Rheumatic” 在 MEDIC 数据库中表现为疾病名称的第 1 个单词且在 MeSH 树中出现于结点 “C17 (Skin and Connective Tissue Diseases)” 下的子结点 “300 (Connective Tissue Diseases)” 下的子结点 “775 (Rheumatic Diseases)” 中. 这些特征能够发现部分疾病之间的上下位关系, 改善疾病名称识别边界错误.

5 实验

5.1 实验语料及评估标准

本文使用 NCBI 疾病语料库 [13], 该语料库总共包含 793 篇 PubMed 摘要. 它由两名专家标注疾病实体, 因而大大提高了标注的正确性. NCBI 数据集包含训练集、开发集和测试集, 3 个数据集的分布情况如表 8 所示.

实验评估采用标准的精确度 (precision, P)、召回率 (recall, R) 和 $F1$ 值等性能指标, 其计算方式如下:

$$\begin{aligned} P &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ R &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ F1 &= \frac{2PR}{P+R}, \end{aligned} \tag{3}$$

其中 TP 表示识别出的实体中正确的数量; FP 表示识别出的实体中不正确的数量; FN 表示没有识别出的实体数量.

5.2 语料预处理

首先对语料进行句法分析, 即将语料分句之后使用 GDep [24] 工具得到句法树, 用于构造句法特征集. 其次, 因为生物文献中存在大量的名称缩写, 所以本文将语料经过缩写识别工具 Ab3P [25] 识别缩写对, 得到缩写对的短形式和长形式, 用于缩写特征集的构造. 然后对句子进行符号化, 即在标点符号处分开, 最后使用 NLTK 工具包中的 POS Tag, PorterStemmer 和 Lemmatizer 分别得到每个单词的词性特征、词干特征以及词原型特征.

5.3 实验结果及分析

语义特征和句法特征对识别性能的贡献. 表 9 比较了句法特征集和语义特征集对识别性能的贡献, 其中基准系统 (baseline) 表示包含基本词法特征 (词特征、词性特征、词干特征、词原形特征以及

表 9 句法特征集和语义特征集对性能的贡献

Table 9 Performance contributions of syntactic and semantic feature sets

Features	P (%)	R (%)	$F1$ (%)
Baseline	86.16	78.44	82.12
+Chunk	86.10	79.38	82.60
+Dependency	87.05	79.79	83.26
+Abbreviation	87.35	81.52	84.19
+Semantic_codes	**87.43**	**83.33**	**85.33**

表 10 缩写特征 (ABB) 对性能的贡献

Table 10 Performance contributions of abbreviation features

Features	P (%)	R (%)	$F1$ (%)
Baseline+SYN	87.05	79.79	83.26
+IS_ABB	87.02	81.04	83.93
+IS_SF	86.27	81.15	83.63
+IS_LF	86.98	80.73	83.74
+Headword	**87.35**	**81.25**	**84.19**

词缀特征) 时的性能, 各个特征按照累加的方式逐一添加到特征集中, 每一列性能中的最高值用粗体表示. 从表中可以看出:

• 组块特征略微提升了 $F1$ 指数, 且主要是通过提升召回率实现的 (∼0.9%), 说明短语块信息具有一定的帮助作用. 例如, 复合实体 "skin tumors" 的组块信息表明该字串是一个名词短语, 从而帮助系统识别出整个短语为疾病名称.

• 依存特征主要通过提升准确率 (∼1%) 来提升 $F1$ 指数, 说明依存关系能够排除部分假正例. 分析表明, 依存特征主要通过改善实体边界错误来提升性能, 如 "pineal and retinal tumors" 中 "pineal" 和 "tumors" 之间的依存关系使整个短语识别为疾病, 而不仅仅是把 "retinal tumors" 识别为疾病名称; 另外, 依存关系有助于捕获句内长距离依存关系, 如句子 "CUG-BP was found to bind to the human cardiac troponin T (cTNT) pre-messenger RNA and regulate its alternative splicing." 中 "cardiac troponin T" 为蛋白质, 该词的依存特征 "RNA" 排除了该词作为疾病名称的可能性.

• 缩写特征集明显提升了 $F1$ 指数, 增加幅度达 (∼1%), 且性能提升主要体现在召回率的提高 (∼1.5%), 这说明缩写特征集可以让 CRF 模型学习到更多的疾病缩写形式. 例如疾病缩写对 "Cowden disease" 和 "CD", 缩写特征集有助于它被识别为疾病.

• 语义编码特征显著提升了 $F1$ 指数 (∼1.1%), 同样也是主要通过提高召回率来实现的 (∼2.1%), 这说明词典形式和上下位关系同样有助于从文本中找出更多的疾病名称. 例如词语 "Rheumatic Diseases" 的语义编码特征依次为 "B_Disease, B-C17.300.775" 和 "E_Disease, E-C17.300.775", 这些特征使得该名称被识别为疾病实体.

缩写子特征对识别性能的贡献. 表 10 比较了各个缩写子特征对识别性能的贡献, 其中第 1 行的性能是基于基本特征加上句法特征集 (Baseline+SYN), 各个子特征按照累加的方式逐一加到特征集中, 每一列性能中的最高值用粗体表示. 从表中可以看出, 除了 IS_SF, 其他子特征都有助于性能提高, 不过去掉该特征后会降低总体性能, 因此本文选择保留该子特征.

表 11 语义编码特征对性能的贡献

Table 11 Performance contributions of semantic features

Features	P (%)	R (%)	$F1$ (%)
Baseline+SYN+ABB	87.35	81.25	84.19
+MEDIC	87.23	83.23	85.18
+MeSH	**87.43**	**83.33**	**85.33**

表 12 本文系统与主流系统的性能比较

Table 12 Performance comparison with the state-of-the-art systems

Systems	Methods	P (%)	R (%)	$F1$ (%)
BANNER [11]	CRF	82.2	77.5	79.8
TaggerOne [14]	Joint Learning	85.1	80.8	82.9
Lou et al. [15]	Joint Learning	**90.7**	74.9	82.1
Sahu et al. [19]	Bi-LSTM	84.9	74.1	79.1
D3NER [23]	Bi-LSTM-CRF	85.0	**83.8**	84.4
Ours	CRF	87.4	83.3	**85.3**

语义编码子特征对识别性能的贡献. 表 11 比较了两个语义编码子特征对识别性能的贡献, 其中第 1 行的性能是基于基本特征加上词法特征集和缩写特征集 (即 Baseline+SYN+ABB), 两个语义编码子特征按照累加的方式逐一加到特征集中, 每一列性能中的最高值用粗体表示.

从表 11 中可以看出, MEDIC 特征明显提高了 $F1$ 指数 (∼1%), 主要是通过提高召回率实现的, 说明词典形式有助于从文本中找出更多的疾病名称, 而精确度轻微下降, 主要是因为词典匹配出现了一些假正例, 如 “disease” 等在不同语境中可能表示为疾病也可能表示为非疾病实体. MeSH 特征在 MEDIC 特征的基础上小幅度提高了 3 个性能指标.

与其他系统的性能比较. 表 12 列出了本文实验系统与当前系统的性能比较, 其中 BANNER [11] 采用的是结合通用词法特征的 CRF 模型; TaggerOne [14] 及 Lou 等 [15] 则采用联合学习方法来实现疾病名称识别和规范化, 不同的是, TaggerOne 使用隐 Markov 模型, Lou 等则使用状态机来实现实体识别和实体规范化的联合学习; Sahu 等 [19] 及 D3NER [23] 系统则采用神经网络的方法, 前者使用 Bi-LSTM 模型, 后者使用 Bi-LSTM-CRF. 本文采用 CRF 模型并且加入了句法和语义特征, 表 12 中每一列性能的最高值用粗体表示. 从表 12 中可以看出, 联合学习的方法优于单一的 CRF 模型, 其中 Lou 等的方法精确度较高; 深度学习的方式根据模型的不同, 其性能也不同, 其中, 我们可以看到 D3NER 系统使用 CRF 结合 Bi-LSTM 的性能优于单独使用 Bi-LSTM 的性能; 本文方法的 $P/R/F1$ 性能指标均高于当前大部分系统的性能指标, 尤其是 $F1$ 值获得了当前最高性能 85.3%, 这说明本文提出的句法特征和语义特征是非常有效的.

6 结论

本文在传统 CRF 模型的基础上, 提出了一系列的句法和语义特征, 如组块、依存、缩写和语义编码特征等, 有效提高了生物文献中疾病名称识别的性能, 其中句法特征可以更好地区分疾病名称的边界, 缩写特征可以让 CRF 模型学到更多的疾病名称, 而语义编码特征可以让模型学到 MEDIC 数据库

中的名称以及在 MeSH 树中的上下位层次信息, 进一步提高疾病名称识别性能.

在下一步的工作中, 我们将关注两个方面的问题, 一是在生物文献中存在很多复合实体, 例如 “VHL-positive and VHL-negative RCC” 分别代表 “VHL-positive RCC” 和 “VHL-negative RCC” 两种疾病, 解决方法是引入复合实体的识别过程. 二是鉴于深度学习在疾病识别中的性能还没有超过传统机器学习方法, 我们将探索不同的神经网络模型, 利用深度学习方法来进一步提高疾病实体识别的性能.

参考文献

1 Song M, Yu H, Han W S. Developing a hybrid dictionary-based bio-entity recognition technique. BMC Med Inform Decis Mak, 2015, 15: S9

2 McCray A T, Srinivasan S, Browne A C. Lexical methods for managing variation in biomedical terminologies. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1994. 235

3 Bunescu R, Ge R, Kate R J, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med, 2005, 33: 139–155

4 Wang H, Zhao T, Tan H, et al. Biomedical Named Entity Recognition Based on Classifiers Ensemble. Int J Comput Sci Appl, 2008, 5: 1–11

5 Leaman R, Wei C H, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminf, 2015, 7: S3

6 Wei C H, Kao H Y, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int, 2015, 2015: 918710

7 Cortes C, Vapnik V. Support-vector networks. Machine Learn, 1995, 20: 273–297

8 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE, 1989, 77: 257–286

9 Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing. IRCS Technical Reports Series, 1997, 81: 1–14

10 Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, San Francisco, 2001. 282–289

11 Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Biocomputing, 2008, 13: 652–663

12 Leaman R, Doğan R I, Lu Z Y. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics, 2013, 29: 2909–2917

13 Doğan R I, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inf, 2014, 47: 1–10

14 Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics, 2016, 32: 2839–2846

15 Lou Y, Zhang Y, Qian T, et al. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics, 2017, 33: 2363–2371

16 Yao L, Liu H, Liu Y, et al. Biomedical named entity recognition based on deep neutral network. Corpus, 2015, 8: 279–288

17 Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2017, 1: 8

18 Xu K, Zhou Z, Hao T, et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In: Proceedings of International Conference on Advanced In-telligent Systems and Informatics. Berlin: Springer, 2017. 355–365

19 Sahu S K, Anand A. Recurrent neural network models for disease name recognition using domain invariant features. 2016. ArXiv: 1606.09371

20 Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. In: Proceedings of the 11th

Annual Conference of the International Speech Communication Association, 2010
21 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9: 1735–1780
22 Santos C D, Zadrozny B. Learning character-level representations for part-of-speech tag-ging. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014. 1818–1826
23 Dang T H, Le H Q, Nguyen T M, et al. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics, 2018, 1: 8
24 Miyao Y, Saetre R, Sagae K, et al. Task-oriented evaluation of syntactic parsers and their representations. In: Proceedings of ACL-08: HLT, 2008. 46–54
25 Sohn S, Comeau D C, Kim W, et al. Abbreviation definition identification based on automatic precision estimates. BMC Bioinf, 2008, 9: 402

Disease name recognition based on syntactic and semantic features

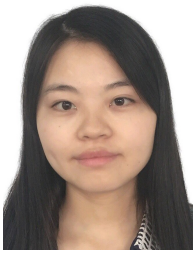
Yunqi HE, Suwen LIU, Longhua QIAN* & Guodong ZHOU

School of Computer Science and Technology, Soochow University, Suzhou 215006*, China*
* Corresponding author. E-mail: qianlonghua@suda.edu.cn

Abstract Biomedical entity recognition (such as genes, proteins, chemicals, diseases, etc.) is the foundation of biomedical text mining, which plays a significant role in extracting biomedical entity relations and constructing biomedical knowledge bases. To deal with existing issues of the current disease name recognition systems, this paper proposes a series of new syntactic and semantic features to improve disease name recognition. The syntactic features include chunk and dependency information, while the semantic features include the disease abbreviation form, its dictionary entry form, and hyponymy relationships between disease concepts. Experiments over the NCBI disease corpus show the CRF model, combined with these syntactic and semantic features, can significantly improve the state-of-the-art performance of disease entity recognition, achieving an F1 score of 85.3%.

Keywords disease name recognition, conditional random fields, syntactic features, semantic features

Yunqi HE was born in 1994. She is a graduate student at Soochow University with majors in computer science and technology. Her research interests include entity named recognition, entity linking, and natural language processing.



Suwen LIU was born in 1995. She is a graduate student at Soochow University with majors in computer science and technology. Her research interests include information extraction and natural language processing.

Longhua QIAN was born in 1966. He received his Ph.D. from Soochow University in 2009. He is currently a professor in the School of Computer Science and Technology, Soochow University. His research interests include natural language understanding, information extraction, and knowledge discovery.

Guodong ZHOU was born in 1967. He received his Ph.D. from the National University of Singapore, Singapore, in 1999. He is currently a professor in the School of Computer Science and Technology, Soochow University. His research interests include natural language understanding, information extraction, statistical machine translation, and machine learning.