第 34 卷 第 17 期 计 算 机 工 程 2008 年 9 月
Vol.34 No.17 Computer Engineering September 2008
·博士论文· 文章编号：1000—3428(2008)17—0021—02 文献标识码：A 中图分类号：TP391.4

基于编辑距离和多种后处理的生物实体名识别

杨志豪，林鸿飞，李彦鹏
(大连理工大学计算机科学与工程系，大连 116024)

摘 要：基于编辑距离和多种后处理的生物医学文献实体名识别方法通过“全称缩写对识别算法”扩充词典，利用编辑距离算法提高识别召回率。在后处理阶段，使用前后缀词扩展、POS 扩展、合并邻近实体及利用上下文线索等方法进一步提高性能。实验结果表明，使用该方法即使利用内部词典也可以获得较好的识别效果。

关键词：文本挖掘；实体识别；编辑距离；条件随机域

Bio-entity Name Recognition Based on Edit Distance and Multiple Postprocessing

YANG Zhi-hao, LIN Hong-fei, LI Yan-peng
(Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024)

【Abstract】A bio-entity name recognition approach using edit distance and multiple postprocessing methods is presented, which expands dictionary via the abbreviation definitions identifying algorithm and improves the recall rate through the edit distance algorithm. The post-processing methods of improving the performance including first-keywords and post-keywords expansion, POS(Part of Speech) expansion, merge of adjacent entity names and the exploitation of the context cues are discussed. Experimental results show that with the above methods even an internal dictionary based system can achieve a fairly good performance.

【Key words】 text mining; entity recognition; edit distance; Conditional Random Fields(CRF)

1 概述

当前，生物医学文献呈指数级增长，增加了进行有效研究的复杂度，另一方面，这些海量资源又为文本挖掘技术提供了用武之地。利用文本挖掘技术从生物医学文献中抽取基因、蛋白质、疾病间的联系和相互作用有着非常重要的意义。而在文本中识别基因、蛋白质等生物实体是成功抽取的前提。

生物文献中实体的命名很不规范，可有多种拼写形式，像“N-acetylcysteine”，“N-acetyl-cysteine”和“NAcetyl Cysteine”都是指同一生物实体；缩写大量使用，也很不规范，如“TCF”可以是“Tcell Factor”和“Tissue Culture Fluid”的缩写。因此，生物实体命名识别是当前研究的难点和热点。目前最好系统的综合分类率也不超过 80%，这与可以应用的水平还有较大的差距。

生物医学实体识别的方法主要有基于词典、基于规则和基于机器学习的方法。基于机器学习方法的优势在于可以判别生物实体数据库中未包含的实体。当前已有一些学者使用各种机器方法进行生物实体识别，包括 SVM[1], HMM[2], CRFs[3]。然而机器学习方法不能提供识别实体的标识信息(如 GenBank ID 和 SwissProt ID)，即它们能辨别出是实体，但不能辨别出是哪种实体。而这些标识信息是与其他数据源(如医学数据库)进行信息融合不可或缺的。同样，基于规则的系统也无法提供识别实体的标识信息。因此，提高基于词典的实体识别方法的性能有重要的研究意义

基于词典的方法通常使用字符串完全匹配算法。但生物实体存在大量的变体名，使用完全匹配算法会导致极低的召回率。本文介绍一种基于编辑距离和多种后处理的生物实体名识别方法。

2 方法描述

2.1 词典的构造和扩充

本文实验语料是 NLPBA2004 测评数据集。包括 2 000 篇 MEDLINE 摘要的训练集和 404 篇测试集。要求识别出 protein, DNA, RNA, cell type 和 cell line 五类实体。使用的词典是利用训练集构造的内部词典。处理噪音后，该词典包含 17 726 条记录。

词典的大小对识别效果影响很大，为此利用全称缩写词对识别算法对词典进行了扩充。生物医学文献中的许多全称缩写词对，如“Toll-like receptor 2 (TLR2) ”和“NF-Y-Associated Factors(YAFs)”，可用来扩充词典：首先使用识别算法识别测试集中的全称缩写词对，共得到 654 个词对。为了提高准确率，使用了条件随机域(Conditional Random Fields, CRF)模型[4]进行过滤。

将 CRF 应用于命名实体识别中，则 o 表示句子的单词序列，s 表示相应的状态序列，标注的过程就是根据已知的单词序列推断出最有可能的状态序列，即 $P(s|o)$的最大值。本文实验使用了一阶线性 CRF，见式(1)：

$$P(s|o)=\frac{1}{Z}\exp(\sum_i\sum_k\lambda_k f_k(s_{i-1},s_i,o,i)) \quad (1)$$

基金项目：国家自然科学基金资助项目(60373095, 60673039)；国家“863”计划基金资助项目(2006AA01Z151)

作者简介：杨志豪(1973－)，男，讲师、博士研究生，主研方向：文本挖掘；林鸿飞，教授、博士、博士生导师；李彦鹏，硕士研究生

收稿日期：2007-09-15 **E-mail：**yangzh@dlut.edu.cn

其中， $f_k(s_{i-1}, s_i, o, i)$ 是二值特征函数，表明当前句子中第 i 个位置上是否具有第 k 个特征。λ_k 是特征的权重，通过训练得到。训练结束后各参数的值均已知，然后通过动态规划维特比(Viterbi)算法求得最优路径，使得 $P(s|o)$值最大。

CRF 这样基于特征的统计模型将问题归结为特征的选择。选取的特征包括单词本身、构词特征、词缀特征、词形特征、特征联合、词性标记特征、关键词特征、边界词特征共 9 类。

使用条件随机域模型过滤后，得到的 450 个全称缩写词对被扩充到词典里。

2.2 编辑距离算法

发现实体名候选词的最简单方法是基于词典使用字符串完全匹配算法。但生物实体存在大量的变体名，使用完全匹配算法会导致极低的召回率。为了解决实体名变体的问题，采用一种近似字符串匹配算法——编辑距离算法[5]。编辑距离是 2 个字符串通过插入、删除、改写字符等编辑操作而变为相同字符串所需要的最小操作数。编辑距离的计算通过一个二元数组 C 实现，数组元素的计算公式如下：

$$C_{i,0} = i \tag{2}$$

$$C_{0,j} = j \tag{3}$$

$$C_{i,j} = \text{if } (x_i = y_j) \text{ then } C_{i-1,j-1} \text{ else } 1 + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}) \tag{4}$$

编辑距离算法计算示例见表 1，“ EGR-1 ”和“ EGR 1 ”间的距离是 1。

表 1 编辑距离算法示例

		E	G	R		1		i	n	c	l	u	d	e
	0	1	2	3	0	1	0	1	2	3	4	5	6	7
E	1	0	1	2	1	1	1	1	2	3	4	5	6	7
G	2	1	0	1	2	2	2	2	2	3	4	5	6	7
R	3	2	1	0	3	2	3	3	3	3	4	5	6	7
-	4	3	2	1	3	2	3	4	4	4	4	5	6	7
1	5	4	3	2	2	1	2	3	4	5	5	5	6	7

为识别出给定文本中的实体名候选词，系统对词典中的每个实体名进行以上的规范化编辑距离计算，当发现文本中的某一字符串的规范化编辑距离值小于阈值时，会被当作实体名候选词。

2.3 后处理

本文采用前后缀词扩展、POS(Part of Speech)扩展、合并邻近实体及利用上下文线索等后处理方法进一步提高性能。

(1)前后缀词扩展

部分匹配错误是实体识别的主要错误之一，原因是实体名只有部分在词典中存在。例如实体名“ human interleukin-2 gene ”，如果只有“ interleukin-2 gene ”包含在词典中，就会导致部分匹配错误，在这里称“ human ”这类的词为前缀词。又如实体名“ CD4 gene ”，如果只有“ CD4 ”包含在词典中，也会导致部分匹配错误，在这里称“ gene ”类词为后缀词。为此，根据训练集构建了每个类别的高频前缀词表和后缀词表。当识别到一个候选实体名时，如果它的前缀词或后缀词属于高频前缀词表和后缀词表，则其前缀词或后缀词将和该实体名一起标注为一个实体。

(2)POS 扩展

许多生物实体名是描述性的，而且较长。词性(POS)标注可以帮助确定实体名的边界。本实验使用 GENIA Tagger 对测试语料进行 POS 标注，它是一个应用于生物医学领域的性能较好的词性标注和浅层语法分析工具。

(3)合并邻近实体

正确合并相邻的实体可以提高识别的性能。通过分析训练集语料，总结了以下合并规则：

1)合并“ and ”和“ or ”连接的相邻已标注实体

如果由“ and ”和“ or ”连接的 2 个实体共享相同的修饰词，则应该将它们合并成一个实体，如表 2 所示。

2)合并由“ and ”和“ or ”连接的相邻标注实体与未标注实体。

在某些情况下，由“ and ”和“ or ”连接的一个标注实体和一个未标注实体也应被合并，如表 3 所示。

表 2 合并“ and ”和“ or ”连接的相邻已标注实体

Before		After	
FOG	B-protein	FOG	B-protein
and	O	and	I-protein
GATA	B-protein	GATA	I-protein
proteins	I-protein	proteins	I-protein

表 3 合并“ and ”和“ or ”连接的相邻标注与未标注实体

Before		After	
Toll-like	B-protein	Toll-like	B-protein
receptors	I-protein	receptors	I-protein
2	I-protein	2	I-protein
and	O	and	I-protein
4	O	4	I-protein

(4)利用上下文线索

在生物医学文献中，有一些上下文线索结构提示生物实体的存在及其类别。如“ …two discrete complexes, NFX1.1 and NFX1.2 ”、“ TF-1 cells, an erythroleukemia cell line… ”、“ Egr2 and Egr3 are NFAT target genes ”都包含这样的上下文线索结构。例如在第 1 个句子中，可以推断“ NFX1.1 ”和“ NFX1.2 ”属于 protein 类别，因为它们都是“ complexes ”而“ complexes ”是 protein 类别的高频后缀词。

3 实验结果

(1)利用编辑距离算法提高召回率

通过实验验证了编辑距离算法对召回率的提高。结果如表 4 所示，最左列是规范化编辑距离阈值。随着阈值增加，准确率下降，召回率增加。最好的综合分类率 F(68.48%)是在阈值为 0.02 时取得的，远高于基于字符串完全匹配的 baseline。

表 4 采用编辑距离算法的召回率提高

Threshold	Recall/(%)	Precision/(%)	F-score/(%)
0.00	64.01	68.21	66.04
0.01	67.44	68.30	67.87
0.02	68.81	68.16	68.48
0.04	70.74	63.04	66.67
0.06	71.05	61.28	65.80
0.08	71.44	59.36	64.84
baseline	52.60	43.60	47.70

(2)利用词典扩充和后处理提高性能

使用全称缩写词对词典进行了扩充，使用前后缀词扩展、POS 扩展、合并邻近实体和上下文线索等后处理方法提高识别性能。它们对性能的提高如表 5 所示。

表 5 采用词典扩充和后处理的性能提高 (%)

	Recall	Precision	F-score
baseline	54.91	52.51	53.68
全称缩写词	55.88	54.19	55.02
前后缀词扩展	63.65	62.20	62.92
POS 扩展	64.04	62.51	63.26
合并邻近实体	68.44	67.75	68.09
上下文线索	68.81	68.16	68.48

其中，baseline 是未使用词典扩充和后处理的字符串近似匹配结果。前后缀词扩展与合并邻近实体贡献最大，综合分类率分别提高了 7.9%和 4.8%。

（下转第 25 页）