第 34 卷 第 7 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.7 Computer Engineering April 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)07—0203—02 文献标识码：A 中图分类号：TP391.43

条件随机域与上下文线索结合的生物实体识别

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：介绍一个用于在生物医学文献中识别基因、蛋白质等生物实体的识别方法。该方法基于条件随机域方法，选取适当特征进行实体识别，利用上下文线索进一步提高识别性能。实验结果表明上下文线索的引入使识别性能在条件随机域方法基础上提高了近 3%，从而获得了较好的最终识别效果。

关键词：文本挖掘；生物实体识别；条件随机域；上下文线索

Bio-entity Recognition Based on Combination of Conditional Random Fields and Contextual Cues

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

【Abstract】 A method is proposed for recognizing the bio-entities (protein, gene, etc.) in biomedical literatures. This method is CRFs-based and selects the appropriate features to recognize the bio-entities. Then the exploitation of the contextual cues is studied to further improve performance. Experimental results show the introduction of contextual cues achieves a performance improvement of nearly 3 percent in F-score, therefore achieving a fairly good overall performance.

【Key words】 text mining; bio-entity recognition; Conditional Random Fields(CRFs); contextual cue

1 概述

当前，对基因、蛋白质等实体的研究是生命科学的研究重点，从医学文献中发现基因、蛋白质分子、疾病间的联系和相互作用有着非常重要的意义。要得到基因、蛋白质以及疾病等之间的联系，必须首先在文本中识别基因、蛋白质等生物实体，即生物实体识别。

生物实体识别的目的是在生物及医学领域对专业词汇加以确认和分类，这类实体包括基因、蛋白质、DNA 和 RNA 等。生物文献中实体命名很不规范，如可能有多种拼写形式，像“N-acetylcysteine”,“N-acetyl-cysteine”和“NAcetylCysteine”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“T Cell Factor”和“Tissue Culture Fluid”的缩写。因此，生物实体命名识别是当前研究的一个难点和热点。在 JNLPBA2004 任务测评中，最好的系统达到 72.6%的综合分类率；在 BioCreative 2004 task 1A 测评中最好的系统获得 74.3%的综合分类率，这与可以应用的水平还有较大的差距。

目前的生物医学实体识别的方法主要有基于字典、基于规则和基于机器学习的方法。基于词典的方法简单实用，但受限于词典的规模和质量；基于规则的方法的优点是：规则可以按照需求灵活地加以定义和扩展。但是产生相应规则需要花费大量的时间，并且要有专家参与；机器学习方法的优势在于它们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]等。

由于对于训练集规模和质量以及特征值选取的依赖性，机器学习方法的识别性能仍有改进的空间。本文介绍了一个基于 CRFs 的生物实体识别系统，并研究了利用上下文线索来提高性能的方法。

2 方法描述

本文的方法包含 2 个阶段：CRFs 识别阶段和利用上下文线索进行性能提高的后处理阶段。

2.1 CRFs 识别

条件随机域(Conditional Random Fields, CRF)[4]是计算具有无向图G结构的随机变量集合S在给定随机变量集合O下的条件概率$P(s|o)$。

将 CRF 应用于命名实体识别中，则 O 表示一个句子的单词序列，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，有式(1)：

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

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，并且取决于当前状态s_i和前一个状态s_{i-1}；λ_k 是特征的权重，通过训练得到。设训练集为 $D=\{<o_1,s_1>,\cdots,<o_j,s_j>,\cdots,<o_n,s_n>\}$，训练的过程是最大化$D$中所有样本出现的条件概率的似然函数，定义为式(2)：

$$LL(D)=\sum_j \log_N(P(s_j \mid o_j))-\sum_k \frac{\lambda_k^2}{2\sigma^2} \quad (2)$$

其中的第 2 项是高斯平滑因子，用于处理训练数据的稀疏问

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，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：2007-09-05 **E-mail：**hflin@dlut.edu.cn

题。训练结束后各参数的值均已知，然后通过动态规划韦特比(Viterbi)算法求得最优路径，使得 $P(s|o)$值最大。

CRFs 这样基于特征的统计模型将问题归结为特征的选择。本实验选取的特征共分为 9 类：

(1)单词本身(F1)：将所有的单词都转化成小写字母，一定程度减少了特征的维数，与其他特征结合可以弥补大小写信息的丢失。

(2)构词特征(F2)：包括首字母大写，是否包含横线，是否是数字，希腊字母等对识别未登录词有用的信息。

(3)词缀特征(F3)：对每个单词都取了 3 个和 4 个字符的前缀和后缀。

(4)词形特征(F4)：将大写字母替换成 A，小写字母替换成 a，数字替换成 0，特殊符号替换成 x。 如：c-fos gene, c-jun gene, c-myb gene 的第一个词都有相同的词形 axaaa。

(5)特征联合(F5)：将相邻位置的特征进行联合，得出新的特征，有助于识别长距离词。本实验选择窗口的大小为(-1,+1)。

(6~7)词性标记特征(F6)和短语切分标记特征(F7)：本实验使用GENIA Tagger[5]对训练语料和测试语料进行标注，得到相应的词性标记和短语切分标记作为特征。

(8)关键词特征(F8)：本实验统计了训练集中出现 20 次以上的 1-gram 和 2-gram 的关键词，将这些词是否出现作为特征。

(9)边界词特征(F9)：从结果的统计中发现，相当多的错误都是发生在边界。本实验统计了训练集中出现 5 次以上的边界词作为特征。

2.2 上下文线索的利用

CRFs 的识别性能依赖于训练集的数量和质量以及特征值的选取。笔者发现，可以利用上下文线索进一步提高识别的性能。本文使用了 2 类这样的上下文线索：括号对和启发式语法结构。

2.2.1 括号对

“…实体名 1(实体名 2)”是生物医学文献中常见的语言现象，括号中的实体名 2 常常是实体名 1 的进一步说明。这种现象又可以分为 2 类：全称缩写词对和非全称缩写词对。

(1)全称缩写对

在生物医学文献中存在许多全称缩写词对，如“Glucocorticoid Receptors(GR)”，可以利用这些全称缩写词对来识别实体及其所属类别。通过全称缩写词对识别算法来提取测试集中的全称缩写词对，然后对 CRFs 模型的识别结果进行调整。表 1、表 2 分别显示了 2 个调整示例。

表 1 利用全称缩写词对调整示例 1

调整前		调整后	
NF-Y-associated	B-protein	NF-Y-associated	B-protein
factors	I-protein	factors	I-protein
(	O	(	O
YAFs	O	YAFs	B-protein
)	O	)	O

表 2 利用全称缩写词对调整示例 2

调整前		调整后	
posttransplant	O	posttransplant	O
lymphoproliferative	O	lymphoproliferative	O
disorders	O	disorders	O
(	O	(	O
PTLDs	B-protein	PTLDs	O
)	O	)	O

在表 1 中， 调整前“NF-Y-associated factors”被识别为一个基因，而它的缩写形式“YAFs”并未被识别。由于两者之间存在全称缩写关系，从而可以推断“YAFs”也应识别为一个基因。而在表 2 中，调整前缩写形式“PTLDs”被识别为一个基因，但从其全称词“posttransplant lymphoproliferative disorders”可以推断出(其后缀词“disorders”指的是一种疾病)“PTLDs”应被标为“O”(非实体)。为此，根据训练集构建了每个类别的高频后缀词表，来判断一个全称词是否属于某类。

(2)非全称缩写对

除了全称缩写词对，还存在许多非全称缩写词对的情况。括号中的实体名常常是括号外实体名的进一步说明，如“B-protein lymphoid Specific octamer binding protein (OTF-2B)”。对这些非全称缩写词对也进行类似全称缩写词对的处理。

2.2.2 启发式语法结构

在生物医学文献中，还有一些启发式语法结构提示生物实体的存在及其类别。示例如表 3 所示。

表 3 启发式语法结构示例

编号	示例
1	…two discrete complexes, NFX1.1 and NFX1.2
2	TF-1 cells, an erythroleukemia cell line…
3	Egr2 and Egr3 are NFAT target genes.
4	…target genes such as IRF1 and c-fos
5	D609 is a strong electrolyte…

示例 1~示例 4 展示了能帮助识别生物实体及其类别的的语法结构。例如在示例 1 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。在示例 5 中，可以推断“D609”应被标注为“O”，因为“D609”是一个“electrolyte”而“electrolyte”不属于任何一个类别的后缀词。

3 实验及讨论

3.1 语料

本文实验所用语料是 NLPBA2004 中使用的 GENIA3.02 数据集。训练集包含 2 000 篇 MEDLINE 摘要。测试集是 404 篇摘要。JNLPBA2004 测评要求识别出 protein, DNA, RNA, cell type 和 cell line 5 类实体。

3.2 CRFs 模型识别结果

本文的 CRFs 模型得到了 71.87%的综合分类率，其中 protein 和 cell type 类别的识别效果较好(73.43%和 73.98%)，而 cell line 类别的识别效果最差，只有 58.35%。

3.3 利用上下文线索提高性能

2 种上下文线索：括号对(bracket pairs (BP))和启发式语法结构(heuristic syntax structure (HSS))对性能的提高如表 4 所示。其中第 1 行 baseline 是初始的 CRFs 模型识别性能；第 2 行是利用括号对线索得到的性能，比 baseline 提高了近 1%(从 71.87%提高到 72.84%)；第 3 行是进一步利用启发式语法结构线索得到的性能，提高了近 3%(从 71.87%提高到 74.80%)。

表 4 上下文线索对性能的提高 (%)

	召回率	准确率	综合分类率
baseline	72.86	70.90	71.87
BP	74.32	71.41	72.84
HSS	76.96	72.76	74.80

（下转第 208 页）