文章编号：1003-0077(2008)03-0089-10

生物医学文本挖掘技术的研究与进展

王浩畅，赵铁军

（哈尔滨工业大学 教育部—微软语言语音重点实验室，黑龙江 哈尔滨，150001）

摘 要：生物医学研究是二十一世纪最受关注的研究领域之一，该领域发表了巨量的研究论文，已经达到年平均60万篇以上。如何在规模巨大的研究文献中有效地获取相关知识，是该领域研究者所面临的挑战。作为生物信息学分支之一的生物医学文本挖掘技术就是一项高效自动地获取相关知识的新探索，近年来取得了较大进展。这篇综述介绍了生物医学文本挖掘的主要研究方法和成果，即基于机器学习方法的生物医学命名实体识别、缩写词和同义词的识别、命名实体关系抽取，以及相关资源建设、相关评测会议和学术会议等。此外还简要介绍了国内研究现状，最后对该领域近期发展作了展望。

关键词：计算机应用；中文信息处理；生物信息学；文本挖掘；信息抽取；机器学习
中图分类号：TP391　　文献标识码：A

Research and Development of Biomedical Text Mining

WANG Hao-chang, ZHAO Tie-jun

(MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: 21st century is the era of biology and there are more than 6 hundred thousand academic papers published annually in this field. The challenge to researchers is how to automatically and effectively acquire relevant knowledge from huge size of biomedical literature. To address this issue, the biomedical text mining has become a new branch of bioinformatics and made great progress.. This survey introduces main approaches and relevant achievements in this research, including machine learning methods to named entity recognition, abbreviation and synonym recognition, relation extraction, as well as relevant resource constructions, international evaluations and academic gatherings.. Some domestic researches are briefly described and, finally, prospective developments in the near future are anticipated.

Keyword: computer application; Chinese information processing; bioinformatics; text mining; information extraction; machine learning

1 引言

当前，生物医学领域的研究正在飞速发展，大量的生物医学知识以非结构化的形式存在于各种形式的文本文件中。国际上生物医学领域的权威数据库MEDLINE（Medical Literature Analysis and Retrieval System Online）的文献总数目前已达到1 600万篇，近年来年均发表文献超过60万篇。如何才能有效地利用这些文本中所蕴含的生物医学知识无疑对分析海量的生物医学数据是非常重要的。常用方法是通过关键词在MEDLINE中或者互联网上进行检索，但是这只能从大量文档集合中找到与用户需求相关的文件列表，而不能从文本中直接获取用户感兴趣的事实信息。因此，提供从大规模生物医学文献中自动获取相关知识的有效工具是一项迫在眉睫的任务。

文本挖掘技术在文本知识自动获取中起到了重

收稿日期：2007-05-28　定稿日期：2007-12-03
基金项目：国家863计划项目(2006AA010108, 2006AA01Z150)

作者简介：王浩畅(1974—)，女，博士生，研究方向为生物信息智能计算，自然语言处理，信息抽取；赵铁军(1962—)，男，教授、博导，主要研究领域为自然语言处理和人工智能。

要作用。文本挖掘通常包括信息检索、信息抽取、数据挖掘三个步骤。其中信息检索(Information Retrieval, IR)用于识别相关文本,信息抽取(Information Extraction, IE)用于识别实体、关系、事件等信息,数据挖掘(Data Mining, DM)则从结构化信息中识别出相互间的关联[1,2]。生物医学文本挖掘的研究重点主要由信息抽取和数据挖掘两方面的研究组成。具体来说,包括生物医学领域命名实体识别、同义词和缩写词识别、关系抽取、利用推理进行关系抽取的假设生成、文本分类以及上述工作的集成框架等[2,3]。该领域研究的主要方法是通用的机器学习方法、领域知识、面向任务的前处理和后处理技术的相互结合。

文本挖掘在生物医学领域中的应用,可以提高生物医学信息建设和管理的效率。生物医学数据库的建设是最早推动生物医学文本挖掘的动力。通过信息抽取技术可以建设以疾病诊断、药物设计为目的的专用蛋白质作用关系数据库。例如建设特定疾病如乳腺癌、老年痴呆症的蛋白质作用关系相关数据库。通过数据库描述的蛋白质作用网络,将极大地有利于疾病诊断、药物设计,促进相关生物医学研究的进展。近年来文本挖掘技术在生物医学领域中的应用多是通过挖掘文本发现生物学规律,例如基因、蛋白质及其相互作用的关系,进而对大型生物医学数据库进行自动注释。例如:现有研究成果已经可以对蛋白质数据库加注功能关键词,并利用这项功能发现大分子间的相互作用关系。使用标准词汇对实验数据统一标注,架起了生物医学文献与生物医学实验数据的桥梁。借助生物医学文本挖掘技术进行数据标注的方法,广泛应用在功能基因组学数据上。经过人手工核对,正确的标注信息将赋予实验数据,有效的文献信息也将作为标注依据链接到实验数据。

生物医学文本挖掘的更大意义在于可以通过对文本分析研究帮助人们发现在文本中隐含的知识,从文献中挖掘出来实验假设和实验建议,以便生物学家验证得到新的科学发现,从而提高人们对生物医学现象的认识。例如,运用分子生物学文献的信息抽取技术来分析海量的生物医学数据,可以帮助分子生物医学专业人员理解分子生物学实验数据,研究分析实验结果。

生物医学文本挖掘是生物信息学研究的分支之一,是生物学研究中不可缺少的环节,它汇集着具有不同专业背景研究者的共同努力,推动和促进了生物医学的发展,对实现疾病的辅助诊断、预防和治疗,新药的辅助发现等起到了重要的作用,为人类对生命的探索做出了重要贡献。生物医学为文本挖掘技术提供了大量的验证数据,对文本挖掘技术起到了反推动作用。这是一种跨学科性研究,涉及到自然语言处理、机器学习、生物信息学等方面的技术,非常具有挑战性。目前,该研究领域吸引了来自计算语言学、生物信息学、机器学习等方面研究者的广泛关注,本文侧重介绍生物医学命名实体识别、缩写词和同义词识别、生物医学实体关系抽取、建立相关资源以及技术评测等。

2 命名实体识别

生物医学文本挖掘的基本任务之一是生物医学命名实体识别(Biomedical Named Entity Recognition, Biomedical NER),其目的是从生物医学文本集合中识别出指定类型的名称,如蛋白质、基因、核糖核酸、脱氧核糖核酸等。这是进一步抽取关系和其他潜在信息的关键步骤。

生物医学领域的命名实体具有如下特点:新的命名实体不断出现,目前并不存在一个完整的包含各种类型的生物医学领域命名实体的词典,所以简单的文本匹配算法已经失去了作用;很多生物医学命名实体都是多词短语,有些有前置修饰语,例如:activated B cell lines,有些名称很长,例如:47 kDa sterol regulatory element binding factor,这些特点给确定命名实体的边界带来了很大的困难;相同的词或者短语可以表示不同类别的生物医学命名实体,要依据上下文才能推断出来,例如:IL-2 既表示蛋白质名称,又表示 DNA 名称;很多生物医学命名实体拥有多个不同的书写形式,例如:N-acetyl-cysteine, N-acetylcysteine, NAcetylCysteine 等表示同一命名实体;很多生物医学命名实体是用"and"或者"or"连接的并列结构,它们共享同一个中心名词,例如:91 and 84 kDa proteins,这样的命名实体也很难正确识别;生物医学命名实体还存在着嵌套现象,例如:<PROTEIN><DNA> kappa 3</DNA> binding factor</PROTEIN>,因此还要解决候选命名实体的重叠问题;缩写词占有较高的比例,例如:IFN, TPA 等等。很多缩写词的形成是没有规律可言的,并且缩写词还具有高度的歧义性,一般情况下,扩展形式比缩写词形式有更多的证据确定它的类别,缩写词形式和它的扩展形式相比

更难分类。总之缩写词的识别很大程度上依赖于上下文，而不能依赖于现存的生物词典。因此，生物医学命名实体识别是富有挑战性的一项研究。

目前，生物医学命名实体识别的方法分为以下三类：基于启发式规则的方法，基于字典的方法和基于机器学习的方法。基于规则的方法需要耗费大量人力建立识别规则库，而基于字典的方法存在着名称冲突和覆盖率受限的不足。目前研究的重点主要是基于机器学习的方法。

机器学习方法是从样例数据集合中统计出相关特征和参数，以此建立识别模型。目前已经有很多机器学习方法应用到生物医学命名实体识别当中，如贝叶斯模型、隐马尔可夫模型(HMM)、支持向量机(SVM)、条件随机场(CRFs)、最大熵(ME)等[4]。基于机器学习的方法依赖于大量的标注语料，因此所面临的问题是如何获得廉价的大量训练数据。

支持向量机方法是一种比较有效的学习方法，已经成功应用到自然语言处理的多项任务中。Kazama 等应用支持向量机来识别生物医学命名实体并使用 GENIA 语料作为训练语料[5]。Lee 等提出了一种基于支持向量机和查找字典的两阶段生物医学命名实体识别的方法[6]，在第一阶段，使用 SVM 分类器识别命名实体并且用简单的字典查找作为后期处理来校正由 SVM 模型识别带来的错误；在第二阶段，把识别后的命名实体用 SVM 划分成语义类。该方法把任务划分成以上两个子任务，能够针对每一个任务选择更相关的特征，选择更为合适的分类方法，减轻了不平衡的类分配问题所产生的影响，提高了整体任务识别的精确率。AbGene 系统是比较成功的生物医学命名实体识别系统之一[7]，曾被多个研究者作为命名实体识别组件用于关系抽取研究当中。该系统使用 7 000 个手工标注命名实体类别的句子作为贝叶斯模型的训练语料，并采用手工统计规则作为后处理，同时使用命名实体所在的上下文来帮助校正识别错误。该系统达到了 85.7%的精确率和 66.7%的召回率。Chang 等设计的 GAPSCORE 系统[8] 考虑到单词的出现次数、词形和上下文并以此为句子中每个词分配一个得分，然后使用基于词形和上下文等特征来训练 N-gram 模型，具有高分的单词更可能是基因和蛋白质名称。Zhou 等人使用基于丰富特征集合的方法训练隐马尔可夫模型，他们在 GENIA 语料上获得了 66.5%的精确率和 66.6%的召回率[9]。Yi-Feng Lin 等[10] 使用基于特征的最大熵模型并结合后处理过程，在分类为 23 个实体类别的 genia 语料上获得了 72.9%的精确率和 71.1%的召回率。Tzong-han Tsai 等[11] 使用条件随机域模型结合丰富的特征集合和后处理过程在 BIONLP2004 测试语料上获得了 69.1%的精确率和 71.3%的召回率。

近两年来，生物医学领域命名实体识别的研究不断扩展和深入。一是命名实体识别扩展到新的语义类型，如临床术语[12]、化学名词语义类[13] 等。二是各种新方法的应用，如自动构建训练语料的 bootstrapping 方法[14]，多分类器结果的重新排序(reranking)方法[15] 等。此外还有嵌套命名实体识别[16, 17]。

目前性能最好的生物医学领域 NER 系统的 F 测度已经达到 80%以上，但与通用领域 NER 结果(90%以上)还存在一定差距[18]，还需要研究人员的进一步努力。

3　缩写词和同义词的识别

很多生物医学命名实体存在多个名称和缩写形式，因此必须有效地识别这些同义词和缩写词，目前大部分研究工作都集中在未登录的基因名同义词和命名实体缩写词的识别上。

抽取生物医学命名实体缩写词及其全称形式，所用方法依赖于全称和缩写词的接近程度。一般而言，全称或者缩写词通常在括号里，因此，识别缩写词被简化为寻找最佳的缩写词和对应全称的对齐过程，这样的对齐过程在很大程度上依赖于上下文。

大部分缩写词的识别方法属于以下三种方法之一：首字母匹配法、首字母和其他字母匹配法、特定模式匹配法。首字母匹配法最简单，即匹配缩写词每一个字母和周围文本中若干词的首字母。第二种方法是放宽条件，即允许匹配首字母之外的其他字母，这种方法一般使用启发式规则进行识别。第三种方法是识别那些后面还添加一定模式的缩写词，这也需要手工建立一些规则。

Liu 和 Friedman 在大量 MEDLINE 文本中统计缩写词和全称的搭配，以此作为规则来检测缩写词与全称的配对，取得了 96.3%的精确率和 88.5%的召回率[19]。在应用手工规则识别缩写词和全称的研究中，Yu 等获得了 95%的精确率和 70%的召回率[20]，Schwartz 和 Hearst 在 1 000 篇 MEDLINE 摘要的集合上识别与酵母有关的缩写词，获得了 96%的精确率和 82%的召回率[21]。Chang 使

用缩写词特征训练逻辑回归模型,并且用这些特征评价缩写词的候选全称形式,在 Medstract 语料上获得了 80%的精确率和 83%的召回率[22]。就目前识别精度来看,在单篇文章中自动识别生物医学缩写词和相应全称的问题已经基本解决,上述识别系统都取得了较高的精确率和召回率。今后的研究将把缩写词识别与其他文本挖掘任务结合,并应用到实际的生物医学文本挖掘系统当中。

同义词识别是建立一个能自动更新的同义词词表的基础,具有重要的应用价值。虽然从在线数据库中能获得基因名称的同义词列表,但这些数据库中多数为基因的正式名称,因此相对于文献中的实际基因名称,其数据并不完整。为了建立出现在文献中有代表性的基因和蛋白质名称同义词列表,需要从生物医学文本中自动抽取基因和蛋白质名称同义词。

Yu 等人结合了 AbGene 基因命名实体识别系统,采用统计方法、基于支持向量机的分类器、基于自动生成模式和手工生成规则等算法相结合,同义词识别的召回率为 80%,精确率为 9%[23]。Cohen 采用自动模式抽取方法对 MEDLINE 摘要进行同义词抽取,通过分析同义词共现网络结构选取最佳同义词模式,获得的精确率为 23%,召回率为 21%[24],该系统可以根据文本中出现的词间的明确逻辑关系来推断它们是否为同义词,与没有类似推断的系统相比,召回率提高了 10%。

基因和蛋白质名称的同义词抽取研究结果的精度普遍还较低,因此更具挑战性。目前,一种新的基因蛋白质名称的标准化工作正在开展,其研究步骤是首先进行基因和蛋白质名称的识别,然后再进行基因名称的规范化(Gene Name Normalization)[25]。此外,使用 Ontology 方法用于同义词识别也是最新的研究趋势[26]。

4 关系抽取

生物医学文本中关系抽取的目的是从多个给定类型的命名实体如基因、蛋白质和药物名称等当中检测是否存在预先指定类型的关系,如蛋白质之间的抑制关系,实体之间的从属关系等。大多数生物医学命名实体关系抽取系统主要抽取特定命名实体之间的二元关系,即两类命名实体之间的关系。

生物医学文本中的关系抽取还存在相当的困难,主要原因包括:文本中陈述同一事实有多种不同的陈述方式;文本中并不仅仅是简单的语法类型;文本中包括很多未登录命名实体;关系信息存在于多个句子之中;存在很多不能抽取出任何关系信息的句子。

目前生物医学领域命名实体关系的抽取主要使用了以下方法:共现方法、关键词方法、机器学习方法和自然语言处理方法。

共现方法认为离得越近的命名实体越可能相关,越经常一起出现的命名实体越可能相关。PubGene 系统使用共现方法建立了一个包含基因和基因交互关系的数据库[27],实验结果达到了 60%的精确率和 51%的召回率。当仅考虑出现在 5 篇或 5 篇以上文章中的基因对关系时,精确率上升到 72%。还有研究者在同一个短语中[28]或者同一个句子中[29]查找共现的基因对。Ding 等做了一项全面的量化研究实验[30],发现用共现方法识别关系在同一摘要中得到的精确率为 57%,召回率为 100%;而在同一句子中精确率为 64%,召回率为 85%;在同一短语中精确率为 74%,召回率为 62%。

为了识别关系的类型,识别算法必须检验相关的信息。一种简单的推断方法是识别那些可以区分特定类型关系的关键词或者短语,这就是关键词方法,其具体应用是使用词模式。在此方法中,研究者给出了一些生物医学命名实体模式和区分特定类型关系的常用词。这些模式通常比较简单,不需要更多的词性信息或者复杂的语义信息,如< protein A> <action><protein B>,这里的< action> 是由 14 个词及其变体组成的词表[31];Ono 等的方法中则使用了 20 个模式[28]。

在基于机器学习方法的关系识别中,把句子中的关系共现表示成向量空间模型,然后使用分类器给句子中可能存在的关系打分。Eskin 和 Agichtein 使用 SVM 算法和基因序列 kernel 来预测蛋白质在细胞质中的位置,其性能达到 87%的精确率和 71%的召回率;而预测蛋白质在过氧化物酶体中的位置,其精确率为 44%,召回率为 21%[32]。Juan Xiao 等[33]使用基于特征的最大熵模型识别蛋白质的交互作用关系(Protein-Protein Interaction, PPI)获得了 88.0%的精确率和 93.9%的召回率。Ameet Soni[34]使用条件随机域模型识别 PPI 并和基于规则的系统作了对比,实验证明基于 CRFs 的系统比基于规则的系统识别性能有很大的提高。

用于关系抽取的自然处理方法一般要使用领域 Ontology 和句法结构分析。简单的方法可以只考

虑词性，如在识别蛋白质和蛋白质的关系中，句子中的蛋白质名称都必须是名词。Thomas 等仅使用词性作为是否存在关系的评分标准[35]。句法分析器是生物医学文本中进行关系抽取的有利工具。如使用浅层句法分析器(Shallow Parser)确定已知动词的主语和宾语[36]，使用完全句法分析器(Full Parser)确定句子中所有组成部分的关系[37]。Park 等使用句法分析器，使关系抽取结果达到了 80%的精确率和 48%的召回率[38]。Zhongmin Shi 等[39]使用统计句法分析技术同时识别生物医学命名实体及其间的功能关系，通过使用有噪音标注数据的半指导学习方法获得了 83.2%的 F 测度值。

Stephens 等提出了使用向量空间模型从文本中识别基因对关系及其共现强度的方法，使用了 TF-IDF 计算公式和用户定义的阈值来挖掘命名实体之间的关系[40]。文献[41]中提出一种无指导的关系抽取方法，该方法使用了类似于互联网页面重要性评价 HITS 算法的思想，称为基于图的交互增强方法。

5　语料库建设和领域本体知识库

统计机器学习方法需要大量的已标注文本数据作为学习器的训练语料，所以，生物医学文献语料库的标注成为相关研究的基础。生物医学文本标注的内容主要包括命名实体、命名实体关系。目前国际上可以公开获取的生物医学文本挖掘的标注语料库有：GENIA 语料库[42]、GENETAG 语料库(也是 BioCreAtIve Task 1A 的评测语料)[43]、Medstract 语料库[44]、Yapex 语料库[45]、Protein Design Group (PDG)语料库[46]和 University of Wisconsin 语料库[47]等。表 1 中列出了每个语料的发行时间，语料内容的切分单位(以句子或摘要为单位)，语料的大小(以词为单位)。表 2 中列出了各个语料可以应用的文本挖掘任务[48]。

表 1　生物医学文本挖掘语料库基本数据

语料名称	发行时间	单　位	语料大小(词数)
PDG	1999	Sentences	10 291
Wisconsin	1999	Sentences	1 529 731
GENIA	1999	Abstracts	432 560
Medstract	2001	Abstracts	49 138
Yapex	2002	Abstracts	45 143
GENETAG	2004	Sentences	342 574

表 2　语料库应用任务

语料名称	分句	分词	词性	命名实体识别	关系抽取	缩写词识别	指代消解
PDG				√	√		
Wisconsin				√	√		
GENIA	√	√	√	√			
Medstract				√		√	√
Yapex				√			
GENETAG				√			

GENIA 语料库是标注规模最大、语义分类最多、应用最广泛的标注语料库。该语料库标注包括词切分、句子切分、词性标注。语料中标注了关于人类血细胞转录因子领域的基因和基因产物命名实体，由 2 000 篇 MEDLINE 摘要组成，共有 18 545 个句子，39 373 个命名实体，36 个语义类。它也是 JNLPBA 语料库的母语料库。需要指出的是：PDG 和 Wisconsin 语料库中只列出所包含的命名实体，但没有指出所在文本中的位置，无法实现正确的评价，因此较难应用于一般的命名实体识别任务。

Medstract是这些语料中唯一有指代消解标注的，并且给出了缩写词的扩展形式。

上述语料库的原始语料皆出自国际权威的生物医学数据库MEDLINE，信息检索和文本挖掘研究主要集中在该数据库上。MEDLINE中生物医学文献数量目前已超过1 600万篇文献，其中超过300万篇文献是近5年内出版的。美国国立医学图书馆的Entrez-PubMed提供了免费的MEDLINE检索服务，是世界最著名和使用最广泛的MEDLINE网上检索系统，于1995年7月推出，已经成为科研人员获取医学文献信息的首选。PubMed提供了主题词检索和自由词检索。

MeSH(Medical Subject Headings)是美国国家医学图书馆(NLM)用以分析生物医学期刊文献等资源的主题内容的控制语汇表，也是NLM出版的MEDLINE数据库主题检索的索引词典。MeSH由22 995个主标题(Descriptors, main headings)组成，分为15个层次。MeSH主标题层级结构安排的目的是为信息检索提供服务。生物信息学中最具有权威性的本体论是基因本体论(Gene Ontology, GO)，由基因本体论协会建立。其目标是建立一套结构化的、精确定义的、通用的控制性词汇，使其在任何生物体内都能描述基因和基因的产物所表现的角色。GO构建了3个相对独立的本体，即生物过程(Biological Process)、分子功能(Molecular Function)和细胞成分(Cell)，它们是基因和基因产物的所有属性。

6 评测会议和相关学术会议

文本信息处理技术的评价通常包含两个部分：标准的评测数据集和评价标准。标准评测数据集一般由领域专家通过手工标注相关文本来获得，这样的数据集通常称为金标准(Golden Standard)。其中比较流行的金标准语料库是GENIA语料库。将自动识别结果与标准数据集相比较，就可以评价某个文本挖掘技术目前所达到的水平。

生物医学文本挖掘评价标准与通常的文本挖掘评价标准类似，也是由精确率(Precision)，召回率(Recall)和F测度(F-score)来评价的。

近年来出现了很多公开评测生物医学文本挖掘算法的国际会议，对本领域研究的发展起到了重要推动作用。表3列出了当前国际上主要的评测会议。

表3 国际上生物医学文本挖掘算法评测会议

会议名称	主办单位	召开时间	评测任务	参加单位
TREC Genomics Track[49~51]	美国国家技术标准局(National Institute of Standards and Technology)	2003年开始，每年举办1次	分子生物学领域的文本检索和分类	2004年29个，2005年41个/我国大陆有复旦大学、清华大学、大连大学等参加
JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and its Applications)/BioNLP 2004[52] BioNLP 2007[53]	2004: National Institute of Informatics, Japan; University of Tokyo 2007: 美国Cincinnati大学等4个单位	2004/2007年	生物医学命名实体识别/临床医学文本多标记分类	2004年有8个参赛系统/2007年有50个参赛系统
BioCreAtIve (Critical Assessment of Information Extraction systems in Biology)[54]	西班牙国家癌症研究中心CNIO、美国MITRE公司、美国生物技术信息中心NCBI等5个机构	2004年，2006年	识别文本中的基因和蛋白质名称/用GeneOntology codes注释蛋白质，识别出蛋白质的功能	2004年10个国家27个单位
KDD(Knowledge Discovery and Data Mining)挑战杯[55]	ACM知识发现与数据挖掘特别兴趣组SIGKDD	2001, 2002, 2004, 2006年	识别基因功能；预测基因对信号传输路径的影响等	2006年18个国家68个单位参加，我国中科院获一个子任务的第二名

在近年来举行的竞赛和评测中，最有影响力的是TREC Genomics Track[49~51]，该评测由美国国家技术标准局(National Institute of Standards and Technology)支持。Genomics Track从2003年开始，以后每年一次，评测任务主要为分子生物学领域的文本检索和分类。2004年有29个研究组参加，2005年有41个研究组参加。我国大陆有复旦大学、清华大学、大连大学等几家单位先后参加这两届

评测[51]。

JNLPBA/BioNLP 2004（Joint Workshop on Natural Language Processing in Biomedicine and its Applications）评测是与国际计算语言学会议同时召开的研讨会，其主要评测任务是生物医学命名实体识别，共有八个参赛系统参加评测[52]。BioNLP 2007 的评测任务是临床医学文本多标记分类，共有 50 个参赛系统参加评测[53]。BioCreAtIve（Critical Assessment of Information Extraction systems in Biology）也是一个重要的生物医学文本挖掘评测会议，由西班牙国家癌症研究中心 CNIO、美国 MITRE 公司、美国生物技术信息中心 NCBI 等 5 个机构负责组织。该评测包括两个任务：其一是识别文本中的基因和蛋白质名称，除了识别命名实体外，各参评系统还要识别出这些命名实体的同义词；其二是用 GeneOntology codes 注释蛋白质，识别出蛋白质的功能。目前该评测已经举行了两届，2004 年评测包括多种文本挖掘任务，共有 10 个国家的 27 个研究组参加了此次评测[51]。2006 年评测的总结会议在 ACL2007 上进行。

KDD（Knowledge Discovery and Data Mining）挑战杯是一个公开评测数据挖掘算法的竞赛。尽管 KDD 的传统任务既和文本无关也和生物医学领域应用无关，但是从 2002 年已经开始了生物医学文本挖掘任务的评测，这也是最早的关于生物医学文本挖掘的评测。KDD 竞赛包括两部分：第一部分是识别基因功能；第二部分是预测基因对信号传输路径的影响。第二个任务可以作为一个统计分类问题来处理，其中涉及到基因功能信息、蛋白质定位以及蛋白质交互等。参赛者所建立的系统用来帮助 FlyBase 数据库管理[55]。

生物医学文本挖掘是一个跨学科的交叉领域，自然语言处理、生物信息学、机器学习领域都召开了关于这个主题的学术研讨会（Workshop），在自然语言处理领域已经发展成为一个相对独立的研究分支。表 4 给出了各领域中相关学术会议情况。2000 年以来，国际计算语言学界的两个主要学术会议 ACL（Annual Meeting of the Association for Computational Linguistics）和 COLING（International Conference on Computational Linguistics）的每届会议都有相关文章发表；从 2003 年起，每届会议均设有一个相关主题研讨会（Workshop）。

表 4　各领域中与生物医学文本挖掘相关的国际会议

会议名称和主题	召开时间	所属领域及主会
Natural Language Processing in Biomedicine	2003 年	自然语言处理；ACL（Annual Meeting of the Association for Computational Linguistics）
BioNLP（Natural language processing of biology text）：Natural Language Processing in Biomedicine and its Applications	2004 年	自然语言处理；COLING（International Conference on Computational Linguistics）
Linking Biological Literature，Ontologies and Databases：Mining Biological Semantics	2005 年	自然语言处理；ACL
BioNLP' 06：Linking Natural Language Processing and Biology：Towards deeper biological literature analysis	2006 年	自然语言处理；北美 ACL
BioNLP-2007：Biological，translational，and clinical language processing	2007 年	自然语言处理；ACL
Pacific Symposium on Biocomputing（PSB）	1996 年开始，每年一届	生物信息学
PSB-2007：New Frontiers in Biomedical Text Mining	2007 年	生物信息学
Intelligent Systems in Molecular Biology（ISMB）	1993 年开始，每年一届	计算生物学
LLL05（Learning Language in Logic）；Challenge task：Extracting Relations from Bio-medical Texts	2005 年	机器学习；ICML（International Conference on Machine Learning）

生物信息学领域与生物医学文本挖掘相关的学术会议主要有从 1996 年开始每年一月在夏威夷举行的 Pacific Symposium on Biocomputing（PSB）会议和始于 1993 年 Intelligent Systems in Molecular Biology（ISMB）年度会议。从PSB2000 开始，该会议几乎每届都把文本挖掘作为会议主题之一；PSB2007 则提出了“New Frontiers in Biomedical Text Mining”的主题。和 PSB 差不多同时，ISMB

也在每届会议发表了这方面的文章,并且近几年把文本挖掘和信息抽取列为会议主题之一。国际生物信息学杂志 Bioinformatics 近年来开辟了数据和文本挖掘专栏,每期均有此类文章发表。

国际机器学习研究领域也对生物医学文本挖掘表现了很大兴趣,2005 年国际机器学习大会 ICML (International Conference on Machine Learning)的一个 Workshop 是 LLL05 (Learning Language in Logic),其主题为:Challenge task:Extracting Relations from Bio-medical Texts。会议为关系抽取提供了训练和测试语料、评测程序,有 5 个国家的 6 个研究小组参加了评测。

7 国内相关研究

目前国内在生物医学文本挖掘领域的研究相对还比较少,主要有清华大学和哈尔滨工业大学,均取得了一定成果。清华大学研究者在蛋白质关系抽取方面做了深入研究,其主要工作包括:基于动态规划算法的模式匹配方法,用于抽取蛋白质交互作用关系,取得了 80%的召回率和精确率[56];在此基础上采用最小描述长度原理进行模式优化,进一步提高了抽取精度[57]。他们还将模式匹配与浅层句法分析结合起来,通过句法和语义约束,很好地识别了生物医学文本中的同位和并列句,将原模式匹配方法的精确率和 F 测度提高了 7%[58]。哈工大研究人员主要致力于生物医学命名实体识别和关系的识别的研究,先后尝试了多种机器学习方法。先后应用 SVM 算法[59]、Generalized Winnow、CRF 等方法[60]进行命名实体识别,在实现中选择了丰富特征并结合后处理过程,在相同测试集上取得了优于国际同类研究的结果。目前,他们在综合多种统计学习方法进行多分类器融合的研究上取得了一定的成果,进一步提高了生物医学命名实体识别的精确率和召回率。在关系识别的研究上主要应用基于特征的机器学习方法并取得了一定的成果。

8 结论与展望

生物医学文本挖掘在近年来的发展中已经取得了很多成果,但现有成果离真正实用还有一定距离。相关资源已经建立,各种评测会议相继召开,有力地推动了本领域研究的发展。在今后 5 年甚至更长一段时间内,本领域未来研究的主要趋势包括:

(1)采用新的机器学习方法或多方法融合进一步提高多类型命名实体识别的精确率和召回率,为关系抽取等研究打下坚实基础;(2)生物医学文本中各类关系包括复杂关系、关系产生的条件等的挖掘;(3)文本挖掘结果的可视化,如关系可视化网络等;(4)充分利用自然语言处理技术,对生物医学文本进行句法分析和语义分析等,为相关知识挖掘提供更好的基础;(5)探索在全文(Full Text)而不仅是文摘数据上进行文本挖掘,以期获得更多、更精确的生物医学知识;(6)探索将文本分类、多文档综述和自动问答等技术应用于生物医学文本数据库,实现更加智能化的文本检索和挖掘系统;(7)建立更大规模的多级标注资源和标准测试数据集合,进一步提高国际同行间研究的可比性。

参考文献:

[1] Cohen, A. M., W. R. Hersh. A survey of current work in biomedical text mining, [J]. Briefings in Bioinformatics, 2005, 6(1): 57-71.

[2] Wang, Sammy. Application of Data and Text Mining to Bioinformatics [EB/OL]. http://cs.uga.edu/~zhiming/datamining/TM.ppt.

[3] Ananiadou, Sophia, Kell, D. B. Tsujii, Jun-ichi. Text mining and its potential applications in systems biology [J]. Trends in Biotechnology. 2006, 24(12): 571-579.

[4] Polajnar, T. Survey of Text Mining of Biomedical Corpora [EB/OL]. http://www.brc.dcs.gla.ac.uk/~tamara/surveyoftm.pdf.

[5] Kazama, Jun'ichi, Takaki Makino, et al. 2002. Tuning support vector machines for biomedical named entity recognition [A]. In: Proc. of ACL-02 Workshop on Natural Language Processing in the Biomedical Domain [C]. 2002. 1-8.

[6] Lee, Ki-Joong, Young-Sook Hwang, et al. 2003. Two-Phase Biomedical NE Recognition based on SVMs [A]. In: Proc. of ACL-03 Workshop on Natural Language Processing in the Biomedical Domain [C]. 2003. 33-40.

[7] Tanabe, L., Wilbur, W. J. Tagging gene and protein names in biomedical text [J]. Bioinformatics, 18(8): 1124-1132.

[8] Chang J T, Schutze H, Altman R B. GAPSCORE: Finding gene and protein names one word at a time [J]. Bioinformatics, 2004, 20(2): 216-225.

[9] Zhou G, Zhang J, Su J, et al. Recognizing names in biomedical texts: A machine learning approach [J].

Bioinformatics, 2004, 20(7): 1178-1190.

[10] Yi-Feng Lin, Tzong-Han Tsai, Wen-Chi Chou, et al. A Maximum Entropy Approach to Biomedical Named Entity Recognition[A]. In: Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2004)[C]. 2004. 56-61.

[11] Tzong-han Tsai, Wen-Chi Chou, Shih-Hung Wu, et al. Integrating Linguistic Knowledge into a Conditional Random Field Framework to Identify Biomedical Named Entities[J]. Expert Systems with Applications. 2006, 30(1), 117-128.

[12] Elhadad, No emie; and Komal Sutaria (2007) Mining a lexicon of technical terms and lay equivalents[4]. BioNLP 2007: Biological, translational, and clinical language processing[C]. 49-56.

[13] Corbett, Peter; Colin Batchelor; and Simone Teufel (2007) Annotation of chemical named entities[A]. BioNLP 2007: Biological, translational, and clinical language processing[C]. 57-64.

[14] Andreas Vlachos, Caroline Gasperin. Bootstrapping and Evaluating Named Entity Recognition in the Biomedical Domain[A]. Proc. Of BioNLP-2006[C]. June, New York City: 2006. 138-145.

[15] Kazuhiro Yoshida, Jun'ichi Tsujii. Reranking for Biomedical Named-Entity Recognition[A]. Proc. Of BioNLP-2007[C]. Prague: 2007. 215-222.

[16] Baohua Gu, Recognizing Nested Named Entities in GENIA corpus[A]. Proc. Of BioNLP-2006[C]. New York City: 2006. 112-113.

[17] Beatrice Alex, Barry Haddow and Claire Grover. Recognising Nested Named Entities in Biomedical Text[A]. Proc. Of BioNLP-2007[C]. Prague: 2007. 65-72.

[18] Lorraine Tanabe, John Wilbur, A Priority Model for Named Entities[A], Proc. of BioNLP-2006[C]. New York City 2006. 33-40.

[19] Liu H, Friedman C. Mining terminological knowledge in large biomedical corpora[A]. Pac Symp Biocomput 2003[C]. 2003. 415-426.

[20] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles[J]. J Am Med Inform Assoc 2002, 9(3): 262-272.

[21] Schwartz, AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text[A]. Pac Symp Biocomput 2003[C]. 2003, 451-462.

[22] Chang JT, Schutze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE[J]. J Am Med Inform Assoc 2002, 9(6): 612-620.

[23] Yu. H, Agichtein. E. Extracting synonymous gene and protein terms from biological literature[J]. Bioinformatics, 2003, 19(Suppl 1. 1). i340-349.

[24] Cohen AM, Hersh WR, Dubay C, et al. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts[J]. BMC Bioinformatics, 2005, 6(103).

[25] Haw-ren Fang, etc. Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries[4]. Proc. Of BioNLP-2006[C]. New York City: 2006. 41-48.

[26] Alona Fyshe, Duane Szafron, Term Generalization and Synonym Resolution for Biological Abstracts: Using the Gene Ontology for Subcellular Localization Prediction[A]. Proc. Of BioNLP-2006[C]. New York City: 2006. 17-24.

[27] Jenssen TK, Laegreid A, Komorowski J, et al. A literature network of human genes for high throughput analysis of gene expression[J]. Nature Genetics, 2001, 28(1): 21-28.

[28] T Ono, H Hishigaki, A Tanigami, et al. Automated extraction of information on protein protein interactions from the biological literature[J]. Bioinformatics, 2001, 17(2): 155-161.

[29] Rindflesch TC, Tanabe L, JN Weinstein, et al. Extraction of drugs, genes and relations from the biomedical literature[A]. In Pacific Symposium on Biocomputing[C]. 2000. volume 5, 517-528.

[30] Ding J, D Berleant, D Nettleton, et al. Mining MEDLINE: abstracts, sentences, or phrases?[A]. Pacific Symposium on Biocomputing[C]. 2002. 326-337.

[31] Blaschke C, MA Andrade, C Ouzounis, et al. Automatic extraction of biological information from scientific text: protein-protein interactions[A]. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology[C]. 1999, 60-67.

[32] Eskin E, Agichtein E. Combining text mining and sequence analysis to discover protein functional regions[A]. Pac Symp Biocomput 2004[C]. 2004. 288-299.

[33] Juan Xiao, Jian Su, GuoDong Zhou, et al. Protein-protein interaction extraction: A supervised learning approach[A]. First International Symposium on Semantic Mining in Biomedicine (SMBM)[C]. 2005. 148-156.

[34] Ameet Soni. Protein Interaction Extraction from Medline Abstracts Using Conditional Random Fields[EB/OL]. http://pages.cs.wisc.edu/~apirak/cs/cs838/soni_report.pdf, May 4, 2006.

[35] Thomas J, D Milward, C Ouzounis, et al. Automatic extraction of protein interactions from scientific abstracts[A]. In Pacific Symposium on Biocomputing[C]. 2000. volume 5, 541-552.

[36] Limsoon Wong. PIES, a protein interaction extraction system[A]. In Pacific Symposium on Biocomputing[C]. 2001. volume 6, 520-531.

[37] Yakushiji A, Yuka Tateisi, Yusuke Miyao, et al. Event extraction from biomedical papers using a full

parser[A]. In Pacific Symposium on Biocomputing [C]. 2001. volume 6, 408-419.

[38] Park J C, Hyun Sook Kim, Jung Jae Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar [A]. In: Pacific Symposium on Biocomputing [C]. 2001. volume 6, 396-407.

[39] Zhongmin Shi, Anoop Sarkar, Fred Popowich. Simultaneous Identification of Biomedical Named-Entity and Functional Relation Using Statistical Parsing Techniques [A]. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007) [C]. Rochester, NY: April 22-27, 2007. 161-164.

[40] M. Stephens, M. Palakal, S. Mukhopadhyay, et al. Detecting Gene Relations From Medline Abstracts [A], PSB 2001 [C]. 2001. 483-495.

[41] Amgad Madkour, *Kareem Darwish, etc. BioNoculars: Extracting Protein-Protein Interactions from Biomedical Text[A]. Proc. Of BioNLP-2007[C]. Prague: 2007. 89-96.

[42] N. Collier, H. Park, N. Ogata, et al. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers [A]. EACL 1999 [C]. 1999. 271-272.

[43] Tanabe, Lorraine, Natalie Xie, et al. GENETAG: a tagged corpus for gene/protein named entity recognition [J]. BMC Bioinformatics, 2005, 6(Suppl. 1): S3.

[44] James Pustejovsky, Jose Castano, Jason Zhang, et al. Medstract: creating large-scale information servers for biomedical libraries [A]. Proc. workshop on natural language processing in the biomedical domain [C]. Association for Computational Linguistics, 2002. 85-92.

[45] Brandeis University. Medstract Project -- Initial Annotion Corpora, 2001; http://scylla.cs.brandeis.edu/gold-standards.html, accessed June 24, 2003.

[46] Franzen K, Gunnar Eriksson, Fredrik Olsson, et al. Protein names and how to find them [J]. International Journal of Medical Informatics, 2002, 67(1-3): 49-61.

[47] Blaschke, Christian, Miguel A. Andrade, et al. Automatic extraction of biological information from scientific text: protein-protein interactions [A]. ISMB-99 [C]. AAAI Press, 1999. 60-67.

[48] Cohen K B, Lynne Fox, Philip V. Ogren, et al. Corpus design for biomedical natural language processing [A]. Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics [C]. 2005. 38-45.

[49] Hersh W, Bhuptiraju RT. TREC Genomics track overview [A]. The Twelfth Text Retrieval Conference (TREC 2003), National Institute of Standards and Technology [C]. 2003.

[50] Hersh W, Bhuptiraju RT, Ross L, et al. TREC 2004 Genomics track overview [A]. The Thirteenth Text Retrieval Conference (TREC 2004), National Institute of Standards and Technology [C]. 2004.

[51] Hersh W, Cohen A, Yang J, et al. TREC 2005 Genomics track overview [A]. The Fourteenth Text Retrieval Conference (TREC 2005), National Institute of Standards and Technology [C]. 2005.

[52] KIM Jin-Dong, OHTA Tomoko, TSURUOKA Yoshimasa, et al. Introduction to the Bio-Entity Recognition Task at JNLPBA [A]. In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004) [C]. Geneva, Switzerland: 2004. 70-75.

[53] John P. Pestian1, Christopher Brew, etc. A Shared Task Involving Multi-label Classification of Clinical Free Text[A]. Proc. Of BioNLP-2007[C]. 2007, Prague: 97-104.

[54] Hirschman L, Colosimo M, Morgan A, et al. Overview of BioCreAtIvE Task 1B: normalized gene lists [J]. BMC Bioinformatics, 2005, 6(Suppl. 1), S11.

[55] Yeh A, Hirschman L, Morgan A. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup [J]. Bioinformatics, 2003, 19(Suppl. 1), i331-i339.

[56] Huang M, Zhu X, Yu Hao, et al. Discovering patterns to extract protein-protein interactions from full texts [J]. Bioinformatics, 2004, 20(18): 3604-3612.

[57] Yu Hao, Zhu X, Huang M, et al. Discovering Patterns to Extract Protein-Protein Interactions from the Literature: Part II [J]. Bioinformatics, 2005, 21(15): 3294-3300.

[58] Huang M, Zhu X, Ming Li. A hybrid method for relation extraction from biomedical literature [J]. International Journal of Medical Informatics, 2006, 75(Issue 6): 443-455.

[59] 王浩畅, 赵铁军. 基于SVM的生物医学命名实体的识别[A]. 第十六届中国神经网络大会论文集, 哈尔滨工程大学学报[C]. 2006, 第27卷增刊: 570-574.

[60] 王浩畅等, 赵铁军, 刘延力等. 生物医学文本中命名实体识别的智能化方法[A]. 信息、知识、智能及其转换理论第一次高峰论坛会议论文集, 北京邮电大学学报[C]. 2006, 第29卷增刊: 54-58.