融合浅层句法分析的蛋白质互作用信息抽取方法*

钱伟中, 王 娟, 傅 翀, 秦志光
(电子科技大学 计算机科学与工程学院, 成都 610054)

摘 要: 针对传统基于机器学习方法在蛋白质互作用信息抽取中的缺陷, 提出融合浅层句法分析的信息抽取方法, 该方法将候选的句子进行浅层句法分析, 包括对短语切分、同位语分析、并列结构分析、句子切分的处理。经过该步骤, 句子被划分为多个单独的语法单元。然后, 对每个语法单元采用基于最大熵的分类方法进行蛋白质互作用信息抽取。该方法在 BC-PPI语料库中获得了 62.1%的 F_1 性能。比较实验结果表明, 该方法能有效减少误判和漏判, 提高信息抽取的性能。

关键词: 蛋白质互作用; 信息抽取; 浅层句法分析; 最大熵
中图分类号: TP391 文献标志码: A 文章编号: 1001-3695(2011)03-0972-04
doi: 10.3969/j.issn.1001-3695.2011.03.051

Protein-protein interaction extraction method using shallow parsing

QIAN Wei-zhong, WANG Juan, FU Chong, QIN Zhi-guang
(School of Computer Science & Engineering, University of Electronic Science & Technology of China, Chengdu 610054, China)

Abstract: In order to solve problems of protein-protein interaction extraction based on traditional machine learning methods, this paper proposed an information extraction method using shallow parsing. This method first processed candidate sentences by shallow parsing including phrase chunking, appositive parsing, coordinative parsing and sentence splitting. After this step, divided sentences into multiple individual grammar units. Secondly, extracted protein-protein interactions from each unit using maximum entropy classification method. Tested in the BC-PPI corpus, this method achieved F_1 value of 62.1%. Comparative experiments show the method decreases false positives and false negatives efficiently and improves performances of information extraction.

Key words: protein-protein interaction(PPI); information extraction; shallow parsing; maximum entropy

0 引言

随着生物医学研究的快速发展, 公开发表的生物医学学术论文和研究成果, 以及由此带来的潜在的生物医学知识, 正以越来越快的速度增长着。截至目前, 全球最大的生物医学文献数据库 Medline收录的论文总数已超过 2 000万篇, 并且还以每年 50余万篇的速度增长。但是, 人们对于生物文献中的生物医学数据的分析和处理的速度却远远落后于数据的增长。对于生物医学研究者而言, 人工方式从海量的生物医学数据库中提取感兴趣的研究信息是非常困难且效率低下, 由此生物医学数据的信息抽取方法应运而生。其中, 蛋白质互作用(PPI)是生物医学信息抽取研究的重要内容。基于生物和遗传方法发现的蛋白质互作用对广泛存在于生物医学文献中, 采用自动化方法, 将方便研究者迅速从文献中抽取蛋白质互作用信息, 对推进疾病预防、生物医学过程研究具有重要意义。目前, 为便于蛋白质互作用对信息抽取的研究, 一些生物医学关系抽取语料库和数据库相继被建立, 比较著名的有 BIND(biomolecular interaction network database), HPRD(human protein reference database)和 MINT(molecular interaction database)等。

抽取蛋白质互作用对的任务目标是从生物文献中抽取 PPI对, 抽取过程一般分为三个步骤:

a)从文献中识别蛋白质实体;

b)提取 PPI对;

c)交叉实验评估系统的性能。

抽取 PPI对的方法性能主要受如下三个因素的影响:

a)蛋白质实体提取方法的性能。缺乏命名标准及名称的易混淆性导致实体识别性能远低于新闻文献中实体的识别性能。

b)上下文特征的选择。PPI对以自然语言描述的形式存在于文献中, 如何提取 PPI描述文本的特征将极大地影响抽取性能。

c)复杂句子结构对性能的影响。生物文献中存在大量的复杂句子结构, 如同位语从句、定语从句、并列句等, PPI对中的蛋白质名称和互作用词分散于不同的语法体中, 增大了抽取的难度。

目前, 大量的 PPI抽取方法被提出, 主要分为三类:

a)基于自然语言处理的方法[1]。其一般过程为定义语法模式, 使用部分或全分析策略描述句子结构, 并且应用语言学技术从句子中提取句法或语义信息。然而, 该类方法需要大量的语义资源, 同时抽取性能较低。

收稿日期: 2010-08-23; 修回日期: 2010-10-08 基金项目: 国家高科技发展规划项目(2006AA01Z411)

作者简介: 钱伟中(1976-), 男, 江苏无锡人, 博士研究生, 主要研究方向为面向生物领域的文本挖掘(qwz_617@163.com); 王娟(1982-), 女, 博士研究生, 主要研究方向为数据挖掘; 傅翀(1976-), 男, 博士研究生, 主要研究方向为信息安全; 秦志光(1956-), 男, 教授, 博士, 主要研究方向为信息安全.

b)基于模式匹配的方法[2]。其一般过程为通过人工定义或自动学习的方法, 从语料库中提取蛋白质互作用文本模式。该类方法对规模较小的语料库易出现过拟合的现象, 为了提高抽取的准确性, 需要创建大量的文本模式, 这样会降低系统性能, 而且, 对于文本模式不匹配的未知蛋白质对, 无法正确抽取信息。

c)基于机器学习的方法[3], 采用机器学习的方法, 如支持向量机、最大熵、决策树等, 其一般过程分为两个阶段: 学习阶段对语料库进行训练, 从样例数据集合中统计出相关特征和参数, 建立识别模型; 抽取阶段对模型输入测试文本集, 输出抽取结果。该类方法采用丰富的上下文特征, 不依赖于具体句子模式, 具有较强的泛化性能。然而, 由于忽略句法结构分析, 会导致出现以下问题:

(a)对于同位语和并列结构中出现的蛋白质名称, 无法进行有效的互作用关系抽取, 出现漏抽取(false negative)结果。

(b)文本中的各个蛋白质实体, 大量存在位于不同的句法单元中的情况, 存在实体之间文本距离较远的情况, 单独采用基于词频特征的机器学习方法, 将会出现大量的误抽取(false positive)结果。

本文的主要贡献如下: a)提出浅层句法分析的方法, 能够对英文句子进行快速、有效的句法分析, 将待抽取句子划分成独立的语法单位; b)针对传统的基于机器学习的蛋白质互作用抽取方法的缺点, 提出一种融合浅层句法分析的信息抽取方法, 经过浅层句法分析, 句子被切分为多个单独的语法单元, 将每个语法单元作为机器学习训练的基本单位, 进行蛋白质互作用信息抽取。通过实验, 以最大熵为基本学习方法, 与传统的机器学习方法抽取性能进行比较表明, 该方法能有效提高蛋白质互作用抽取性能。本文采用的抽取模型同样适用于其他机器学习方法, 并可以扩展到其他信息抽取应用中, 如新闻文本关系抽取、问答系统等。

1　生物文本浅层句法分析

1.1　分析过程

本文主要针对英文生物文献进行蛋白质互作用关系抽取, 英文生物文献中广泛存在同位语、并列结构及复杂句, 需要针对这些问题, 提出浅层句法分析方法。

首先, 给出蛋白质互作用对和候选句的定义。

定义 1　蛋白质互作用对是一个三元组(protein1, iword, protein2)。其中, protein1和 protein2为两个蛋白质名, iword为两者的交互词。

定义 2　候选句是语料库中包含至少 1个蛋白质互作用对的句子。

浅层句法分析过程如下:

a)输入候选句;

b)对句子进行词性分析(part of speech);

c)对句子的词性分析结果进行浅层句法分析, 划定短语边界;

d)对句子结构进行同位语规则分析;

e)对句子结构进行并列结构分析;

f)对句子结构进行从句结构分析。

对步骤 a), 首先对句子进行蛋白质名称识别和交互词识别, 并计算蛋白质互作用对个数, 保留内含多于 1个 PPI的句子作为候选句。对步骤 b)的词性分析, 采用 MedPost[4]作为词性分析器。对步骤 c), 使用 Ramshaw等人[5]提出的基于转换的浅层句法分析器, 该分析器对基本的名词短语(noun chunk, NP)识别的准确率(precision)和召回率(recall)达到 92%。

1.2　同位语分析

同位语结构中包含两个语法实体: a)下位词(hyponym), 为同位语的中心部分; b)上位词(hypernym), 为同位语的修饰部分, 用于对中心部分进行说明。生物文献中存在大量的同位语结构, 其中, 对于包含蛋白质名的同位语结构, 存在上位词和下位词均为蛋白质名的情况。

通过对文献文本的分析, 发现同位语结构可以用以下两类模式代表。

模式 1　[NC1], [NC2]

模式 2　[NC1] |, |[keywords] [NC2]

其中, NC1和 NC2为名词短语, keywords为关键词, 表示满足同位语结构出现频率较大的词语, 本文取 keywords= {"such as" "including"}。

在生物文献中出现的同位语中, 包含蛋白质名称的同位语结构对蛋白质互作用对抽取性能直接相关, 而其他则关系较弱。为了减少同位语分析数量, 优化系统性能, 规定进行分析的同位语结构需满足如下基本条件: NC1与 NC2具有相同的单复数形式; NC1与 NC2均为包含蛋白质名称实体名称的短语。

对满足基本条件的同位语结构, 分析处理方法如表 1所示。

表 1　同位语处理方法

模式	处理
模式 1	if firsttag(NC1) = DT or CD then (hypernym, hyponym) = (NC1, NC2) else (hypernym, hyponym) = (NC2, NC1)
模式 2	(hypernym, hyponym) = (NC1, NC2)

函数 firsttag()功能为取短语中首词词性。分析同位语结构后, 在抽取蛋白质互作用对时, 去掉包含同位语结构中上位词的蛋白质互作用对, 保留下位词所对应的蛋白质互作用对。

1.3　并列结构分析

定义 3　并列结构是多个短语由连词组成的语法体, 各组成短语在语法成分上是相同的。

对名词并列结构可以用以下有限状态自动机识别。并列结构识别的有限自动机如图 1所示。

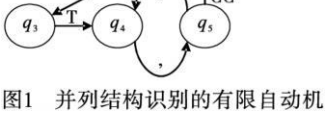

图1　并列结构识别的有限自动机

其中: q_0为起始状态, S为终止状态, 状态之间的有向连接线代表状态迁移, 状态线匹配的标志为集合 S= {T, CC, ′,′}中的元素。其中, T为任意的名词短语, CC为连词。如果待匹配的名词短语序列具有一条从起始状态至终止状态的连接, 则表明该序列满足名词并列结构条件。

为了缩小分析范围,与同位语分析类似,对识别出的名词并列结构,提出如下的限制条件:

a)名词并列结构中的每个名词短语,具有相同的单复数形式;

b)至少存在一个识别出的蛋白质名称。

在生物医学文献中,位于名词并列结构中的蛋白质名称具有相同的语法属性,因此,可以将名词并列结构中的蛋白质名称集合作为整体,进行信息抽取。

对名词并列结构的信息抽取,采用两阶段处理:

a)封闭。设名词并列结构的短语集合为 $S=\{n_1, n_2, n_3, \cdots, n_k\}$。其中,$n_i\ (i=1, \cdots, k)$为单个蛋白质名称,则令 $N=n_1$,作为整体的代表名称进行互作用信息抽取,抽取过程提取单词序列为整体代表名称中的单词序列。

b)开放。如整体名称 N 代表集合 $S=\{n_1, n_2, n_3, \cdots, n_k\}$,且存在 1蛋白质互作用对 (N, iword, X),则该蛋白质互作用对分解为 $\{(n_1, \text{iword}, X), (n_2, \text{iword}, X), \cdots, (n_k, \text{iword}, X)\}$。

经过名词并列结构的处理,可以有效避免对并列结构中的蛋白质作用对抽取的错误或遗漏,提高抽取性能。

1.4 从句结构分析

生物文献中常见蛋白质互作用信息分布于不同的从句中,传统的机器学习方法没有判定从句边界,也没有考虑代词指代消解问题[6],抽取结果中出现分布于不同从句中的蛋白质对,降低了抽取精度。

从句结构分析主要分为对宾语从句和定语从句的分析,从句分析的主要任务是对从句边界进行定位,同时,对于定语从句,考虑指代消解问题。

英文文献中的从句主要可分为两类,即有限定词(如 that, which等)作为前置词的从句与无限定词从句。对于无限定词从句,需要采用复杂的语义分析确定句子边界,且识别率较低。本文主要针对有限定词的从句识别,提出启发式规则。

从句结构分析可分为两个阶段:

a)判别从句类型,在本文,判断从句属于宾语从句还是定语从句。

b)对从句进行定界和指代消解处理。

步骤 a)根据前述指定的从句识别范围,宾语从句判定方法如下:

单词 that前一词词性为动词或介词,且 that后至句子末尾的单词序列至少包含一个动词。

1)定语从句判定

前置词 that或 which前一词词性为名词,且前置词后至句子末尾的单词序列至少包含一个动词。

2)对宾语从句处理

搜索句子单词序列中的单词 that,如果包含多个 that,搜索是否包含连词 CC,将每个 that及后一个 that之间的单词序列作为一个整体(不包括末尾标点符号),判断是否满足并列结构,如果满足,将每一个 that序列作为一个单独的抽取句子。

3)定语从句处理

将定语从句前置词 that替换为前一名词短语,其中主语为一名词短语或名词短语的并列结构,谓语为一动词短语或动词短语的并列,将主语与谓语末尾词之间的序列作为一个单独的抽取句子,并从原句子中删除该句子。

2 融合浅层句法分析的信息抽取方法

基于机器学习的信息抽取方法具有较强的泛化性能,在生物文本信息抽取中得到了广泛的应用,由于生物蛋白质名称分布于不同的语法结构中,造成识别错误率和遗漏率较高,识别精度较低。本文提出的融合浅层句法分析的信息抽取方法主要思想是在进行信息抽取前,将句子进行语法分析,去掉冗余信息,缩小搜索范围,以提高文本抽取性能。融合浅层句法分析的信息抽取框架如图 2所示。

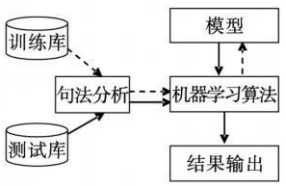

图2　融合浅层句法分析的信息抽取过程

图 2中,虚线表示训练过程,实线表示测试过程,方法的训练和测试两个阶段在预处理过程中加入了第 2章中提出的句法分析方法,句法分析方法使用句法规则,对候选句划定短语边界,提取同位语,并列结构分析提取对等语法结构,从句结构分析划分从句边界,从而将复杂句划分为多个简单句集合。

语料库划分为训练库和测试库,其中的句子需要事先经过蛋白质名称识别过程,以返回文本中蛋白质名称及单词在文本中的序号。浅层句法分析所输出的简单句集合进一步分解为候选蛋白质对集合。蛋白质对集合中的元素需满足,即组成蛋白质对的蛋白质名称位于句法分析输出的同一个简单句中。

可定义为 $S=\{(p_{k_1}, i_{k_1}, p_{k_2}, i_{k_k}), k\in N\}$。其中:$N$ 为元素总数;p_{k_1}、p_{k_2}为蛋白质对中的蛋白质名称;i_{k_1}、i_{k_2}分别为 p_{k_1}、p_{k_2}在该简单句中的单词索引号。

机器学习算法以候选蛋白质对为抽取单位,进行实际的信息抽取任务。在生物文献蛋白质互作用对抽取任务中,蛋白质互作用信息抽取是一个二元分类任务。任意的机器学习算法,如支持向量机、最大熵模型等,都能用于蛋白质互作用对的识别。与使用分析树或依赖关系树的方法对比,浅层词法特征被证明对蛋白质互作用对的抽取性能有更大的影响[7],因此,使用从简单句中抽取的词法特征作为机器学习算法的输入。

本文采用的词法特征如表 2所示。

表 2　词法特征

特征名	描述
Pr	蛋白质名称中的词
Bw	蛋白质对之间的词
Wf	第一个蛋白质名称之前三个词
Wa	第二个蛋白质名称之后三个词
K	交互词及位置

交互词由交互词表定义,如 interaction, cleave等。交互词位置属于下述三种情况之一:

a)位于第一个蛋白质名之前;

b)位于两个蛋白质名称之间;

c)位于第一个蛋白质名之后。

蛋白质互作用对抽取过程如下:

a)预处理。蛋白质名称识别,使用句法规则,对训练库和测试库中的句子进行句法分析,输出简单句组成的集合。使用蛋白质识别返回的结果,输出候选蛋白质对集合。

b)训练。以候选蛋白质对集合为单位,抽取词法特征,并对分类进行标注,对正例标注为"+",反例标注为"-"。应用机器学习算法学习、输出目标模型。

c)抽取。以训练阶段得到的模型,对测试库中的候选蛋白质对进行分类,将得到的概率最大的分类作为抽取结果。

3　实验结果

为了验证方法的有效性,将本文所提出的方法应用到 BC-PPI语料库[8]中,并以最大熵方法为学习算法。

BC-PPI语料库包含 1 000个句子,每个句子中的蛋白质名称及互作用词都已事先进行了人工标注,因此,可以不考虑蛋白质名称识别问题对抽取性能的影响。语料库中总共包括 1 426个蛋白质对,其中,255个包含互作用信息,为正例,其余不含互作用信息,为反例。

为了验证不同的句法分析组件对实验结果的影响,将最大熵分类算法作为基本的方法,然后,将同位语规则分析、并列结构分析及从句结构分析与基本方法的组合进行对比实验。实验过程采用 10折 1方式,即进行 10次实验,每次实验中,将语料库中的 PPI对随机划分为训练集和测试集,其中,训练集和测试集的 PPI数目比为 9∶1。同时,将实验所得结果与同一数据集上其他方法进行比较。实验性能的度量用精度(precision)、召回率(recall)及 F_1 值表示。其中,F_1 为精度与召回率的综合性能度量,计算式为

$$F_1=\frac{2\times \text{precision}\times \text{recall}}{\text{precision}+\text{recall}} \tag{1}$$

表 3列出了实验结果。其中,方法 base为未加入句法分析的基准方法,本实验采用最大熵分类算法,$L_i(i=1,\cdots,3)$表示采用的句法分析方法,其中,L_1 为同位语分析方法,L_2 为并列结构分析方法,L_3 为从句结构分析方法,符号+连接各种方法的名称,表示各种方法的集成。

表 3　实验结果　　%

Method	Precision	Recall	F_1
Base	64.3	43.1	51.6
Base+ L_1	66.5	45.3	53.9
Base+ L_2	66.8	53.2	59.2
Base+ L_3	69.1	43.5	53.4
Base+ L_1+ L_2+ L_3	70.3	55.6	62.1

从表 3中可以看出,与基准方法比较,加入句法分析方法的各个组件均有效提升了抽取性能。其中,从句结构分析对精度提升最大,相较于基准方法精度提升了 4.8%,但对于召回率影响不大,这是因为通过从句结构分析,将位于不同从句的蛋白质对过滤掉,从而减少了误报(false positive),但是由于从句结构分析并没有对位于同一从句中的蛋白质对及相关文本进行变化,对于召回率影响较小,相较于基准方法只提升了 0.4%,并列结构分析对于召回率提升最大,相较于基准方法召回率提升了 10.1%。从语料库中分析得出,生物文献中广泛使用了并列结构,而并列结构分析能够有效减少对并列结构中作用关系的漏报(false negative),同位语分析方法对于精度和召回率的影响位于另外两种方法之间。加入三种句法分析方法后,总体性能 F_1 较基准方法提升了 10.5%,达到 62.1%。结果表明,综合使用三种句法分析方法能够有效提升生物文献中蛋白质互作用信息的抽取性能。

与另两种在同一数据集上的方法进行比较,Plake等人[8]提出的基于模式匹配的抽取方法使用了遗传算法进行模式学习,获得了 60%的精度和 46%的召回率,F_1 值为 52%;Leif使用 SVM方法[9]获得了 48%的精度和 58%的召回率,F_1 值为 52%。本文提出的方法均优于其他两种方法。

4　结束语

本文在传统的基于机器学习抽取方法的基础上,提出一种融合句法分析的信息抽取方法,并应用在生物文献中蛋白质互作用对的抽取任务中,在 BC-PPI语料库中获得了 F_1 值 62.1%的水平,该方法运用句法分析方法对待抽取文本进行预处理,通过提出的同位语分析、并列结构分析和从句结构分析方法,有效地降低了误报和漏报,提升了总体性能。

参考文献:

[1] YAKUSHIJI A, TATEISIY, MIYAO Y, *et al*. Event extraction from biomedical papers using a full parser[C]//Proc of the 6th Pacific Symposium on Biocomputing 2001: 408-419.

[2] HUANG Min-lie, ZHU Xiao-yan, YU Hao. Discovering patterns to extract protein-protein interactions from full biomedical texts[J]. Bioinformatics, 2004, 20(18): 3604-3612.

[3] MITSUMORI T, MURATA M, FUKUDA Y. Extracting protein-protein interaction information from biomedical text with SVM[J]. IEICE Trans on Information and Systems, 2006, E89-D(8): 2464-2466.

[4] SMITH L, RINDFLESCH T, WILBUR W J. MedPost: a part-of-speech tagger for biomedical text[J]. Bioinformatics, 2004, 20(14): 2320-2321.

[5] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[C]//Proc of ACL the 3rd Workshop on Very Large Corpora. 1995: 82-94.

[6] 孔芳,周国栋,朱巧明,等. 指代消解综述[J]. 计算机工程, 2010, 36(8): 33-36.

[7] XIAO Juan, SU Jian, ZHOU Guo-dong. Protein-protein interaction extraction: a supervised learning approach[C]//Proc of the 1st International Symposium on Semantic Mining in Biomedicine. 2005: 51-59.

[8] PLAKE C, HAKENBERG J, LESER U. Optimizing syntax-patterns for discovering protein-protein interactions[C]//Proc of ACM Symposium on Applied Computing. Santa Fe: ACM, 2005: 195-201.

[9] NIELSEH L A. Extracting protein-protein interactions using simple contextual features[C]//Proc of BioNLP Workshop. Morristown: Association for Computational Linguistics, 2006: 120-121.