基于支持向量机的生物医学文献蛋白质关系抽取

杨志豪[1],洪 莉[2],林鸿飞[1],李彦鹏[1]
(1.大连理工大学 电子与信息工程学院,辽宁 大连 116024; 2.朝阳师范高等专科学校 数学计算机系,辽宁 朝阳 122000)

摘 要:从生物医学文献中抽取蛋白质(基因)交互作用关系对蛋白质知识网络的建立、蛋白质关系的预测以及新药的研制等均具有重要的意义.提出了一种基于支持向量机(SVM)的蛋白质(基因)交互作用关系抽取方法.该方法除了选取词项特征、关键词特征、实体距离特征、链接特征外,还利用链接语法分析方法可以获得较高准确率的特性,引入链接语法分析方法抽取结果特征.实验结果表明,该方法的召回率性能与使用同一测试语料的其他系统相比具有明显的优势,综合分类率 F指标也高于其他系统.

关键词:关系抽取;链接语法;支持向量机
中图分类号: TP391 **文献标识码**: A **文章编号**: 1673-4785(2008)04-0361-09

Extraction of information on protein-protein interaction from biomedical literatures using an SVM

YANG Zhi-hao[1], HONG Li[2], LIN Hong-fei[1], LI Yan-peng[1]
(1. College of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China; 2. Department of Mathematics and Computer, Chaoyang Teachers College, Chaoyang 122000, China)

Abstract: Automated extraction of protein-protein interaction information from biomedical literature is helpful when building a protein knowledge network, predicting protein functions and designing new drugs. This paper presents a method for protein-protein interaction extraction from biomedical literature using a support vector machine (SVM). In this method, besides common index parameters such as word features, keyword features, entity distance features and link path features, a link grammar extraction feature is used to improve precision when identifying protein-protein interactions. Experimental results indicated that the recall rate and the F-score of this method are much higher than that of other extraction systems for the same dataset.

Keywords: interaction extraction; link grammar; support vector machine (SVM)

随着高通量生物技术的发展,生物医学的实验手段和研究方法均发生了巨大的变革,领域内实验数据的"指数性"增长,给数据的存储与传输,数据的处理、理解与应用带来一系列问题,来自数学、化学、药学、统计学和计算机科学等领域专家给予了广泛关注,并取得了大量成果.生物医学文献作为成果展示和学术交流的主要方式之一,其数目之大,增长速度之快远远超过了其他学科领域,采用文本挖掘技术从这座宝库快速有效地提取生物医学知识的需求十分迫切.其中,从生物医学文献中抽取蛋白质(基因)相互作用关系可以帮助建立蛋白质知识网络、预测蛋白质关系以及辅助新药的研制,因此具有重要的研究意义.

1 相关研究

当前进行蛋白质(基因)交互关系抽取主要有 3种方法:基于自然语言处理的系统、基于模式匹配的系统和基于机器学习与统计的方法.

基于自然语言处理的系统通过分析语法结构进行关系抽取,依据它们的分析策略将它们分为浅层分析系统和深层分析系统. Pustejovsky等人使用浅

收稿日期: 2008-05-07.
基金项目:国家自然科学基金资助项目(60373095, 60673039);国家"863"高科技计划资助项目(2006AA01Z151).
通信作者:杨志豪. E-mail: Yangzh@dlut.edu.cn.

层分析方法从生物医学文献的摘要中抽取蛋白质抑制关系[1],但只得到了 57%的召回率. Leroy等人提出了一个浅层分析器,用于从文献中抽取名词短语间的类属关系,精确率达到了 90%[2]. 与浅层分析相比,深层分析方法也有很多人使用过. Park等人提出了一个基于可组合的分类语法的深层分析器,该分析器首先定位目标动词,然后使用双向增量分析技术扫描该动词的左部和右部以获得语法成分[3]. 该系统的召回率和精确率分别是 48%和 80%. 另一个深层分析器利用词典分析程序和上下文无关文法抽取蛋白质和基因的交互关系,得到的召回率为 63.9%,精确率为 70.2%[4]. 另外,Davulcu等人提出了一个基于链接语法分析器(link grammar)的关系抽取系统 IntEx[5],使用链接语法分析器将复杂句划分为简单句,又将简单句划分为更具体的句子成分,然后从这些成分中选择满足条件的部分进行关系抽取,他们的召回率是 26.94%,精确率是 65.66%. 浅层分析系统只限于把句子解析成较小的单元,而不揭示单元之间的句法关系. 对于句子中简单的实体关系能获得较好的性能,但对于复杂句子中多个实体间的关系抽取则性能较差. 深层分析系统着眼于充分分析整个句子的语法特点,从而最大限度地揭示句子所反映的主题内容,能获得更高的准确性,但需要更高的计算能力和时间复杂性.

基于模式匹配的系统比基于自然语言处理的系统要简单得多,它们根据预先定义好的模式和匹配规则将标注好词性的序列与结构信息进行匹配. 人们开发了许多系统,用于自动模式获取和关系抽取. 其中大部分需要特殊的训练资源,比如标注了领域特定标签的文本. Ono等人提出了一个基于模式的系统,该系统使用简单词的人工编码规则和标注了词性的模式从生物医学文献的摘要中抽取特殊种类的蛋白质交互关系[6],得到了较高的召回率和精确率,分别为 82.5%和 94.3%. Huang等人提出了一种从语料中自动获取模式的方法和一个基于动态规划的匹配算法[7],精确率和召回率大约都在 80%. David等人开发了一个名为 BioRAT[8]的系统,该系统使用了一个信息抽取引擎和一个模板设计工具来进行关系抽取,得到的召回率为 20.31%,精确率为 55.07%. 基于模式匹配系统性能依赖于模式的数量和质量,难以处理较复杂的句子,而且无法抽取跨句子的实体关系.

基于机器学习和统计的方法较前 2种方法的优势在于不需要付出繁重的努力去定义规则或语法,它自动提取实体交互模式而不需要人的参与. 目前已经出现了许多基于蛋白质称名共现的机器学习与统计的方法[9-13].

其中根据挖掘单位(如摘要、句子等)又分为不同的类型. Andrade等[9]和 Marcottle等[13]的方法在摘要集合中提取蛋白质交互关系. 前者将一组相关文档与一组随机选取的文档对比来提取领域知识(如基因功能和交互);后者检索到可能包含蛋白质交互关系的文档; Craven等[10]最早开发了基于机器学习的句子级蛋白质交互关系抽取系统. 使用贝叶斯分类器,对于一个包含 2个实体名的句子,返回它们存在交互关系的概率. 后续的研究者使用了包括隐马尔科夫模型、支撑向量机的机器学习方法来判别包含蛋白质交互关系的句子;还有的方法研究句子中一对实体存在交互关系的概率. Stapley等[11]使用固定的基因名列表,借助共现方法在 Medline记录中构建每个基因对的相似性矩阵来检测它们的关系. Jenssen等[12]采用类似的方法发现了 DNA array实验中的人类基因聚类间的关系.

简单的统计方法(如基于蛋白质名称共现的方法)不能准确地描述蛋白质之间的关系,因此会导致较高的抽取错误率;而复杂的统计模型为了获得准确的模型参数需要大量的训练集,在实际应用中通常是难以得到的.

以上 3种蛋白质(基因)交互关系抽取的方法都有各自的优缺点,实际上许多系统都采用混合的方法以获得更好的性能.

当前研究存在的一个问题是:大部分抽取系统采用的是自己制作的语料,缺乏统一的性能评价标准. 2001年 Blaschke和 Valencia推荐使用 DIP数据库,并将其作为评测生物关系抽取系统的标准库[14]. DIP是 1999年由 UCLA的 D. Elsenberg实验室建立,它的目标是成为一个蛋白质–蛋白质相互作用的数据库,把关于蛋白质相互作用的多样的实验信息整合成一个容易进行查询的专一数据库. DIP数据库中的每条记录都定义了一对相互作用的蛋白质,并且提供了描述这些交互作用的文档. 到 2004年,DIP数据库已有超过 18 500个蛋白质交互关系[15]. 研究人员可以首先使用自己的系统进行关系抽取,然后将抽取结果与 DIP数据库中的记录进

行比较.这样做可以使评测结果更加令人信服. 2004年 David等人使用 DIP数据库的一个子集 (392条记录)进行了蛋白质相互作用关系抽取测试.他们的 BioRAT系统得到的召回率为 20.31%,精确率为 55.07%. 2005年 Davulcu等人开发的 IntEx系统,使用链接语法分析器在同一子集上得到 26.94%的召回率和 65.66%的精确率.

笔者曾提出了一个基于链接语法分析的蛋白质(基因)交互作用关系的抽取方法[16].该方法使用条件随机域 (conditional random fields, CRF)与上下文线索结合的生物实体识别方法,再通过链接语法分析划分语法成分,从语法成分及其合理组合中抽取蛋白质 (基因)交互作用关系.使用与 BioRAT和 IntEx系统相同的 DIP语料进行测试,实验结果表明该方法的召回率以及综合分类率 F指标都高于 BioRAT和 IntEx系统.类似链接语法分析器的深层分析系统着眼于充分分析整个句子的语法特点,从而最大限度地揭示句子所反映的主题内容,能获得更高的准确性,但需要较高的计算能力和时间复杂性,而且召回率较低.基于机器学习和统计的方法优势在于不需要付出繁重的努力去定义规则或语法,它自动提取实体交互模式而不需要人的参与,通常可以获得较高的召回率,而在大多数情况下,这是生物医学研究者更倾向于得到的.

因此,本文提出了一种基于支持向量机 (support vector machines, SVM)的蛋白质交互作用关系抽取方法.该方法通过适当特征的选取 (包括词项特征、关键词特征、实体距离特征、链接特征以及链接语法分析特征),利用 SVM分类器判断句子中每对蛋白质 (基因)是否存在相互作用关系.实验证明这种方法可以获得比基于自然语言处理和基于模板方法更高的召回率.

2　方法描述

使用上述方法,实现了一个蛋白质相互作用关系的抽取系统 BioPISVMExtractor.该系统使用 IEPA语料作为训练语料[17],使用 DIP语料作为测试语料.系统首先对 DIP语料进行指代消解,然后进行实体识别,之后对语料句子中的每个蛋白质对进行特征提取,并使用 SVM分类器进行二值分类,即分为存在相互作用关系的蛋白质对和不存在相互作用关系的蛋白质对,从而抽取出蛋白质间的相互作用关系. BioPISVMExtractor系统的框架如图 1所示.

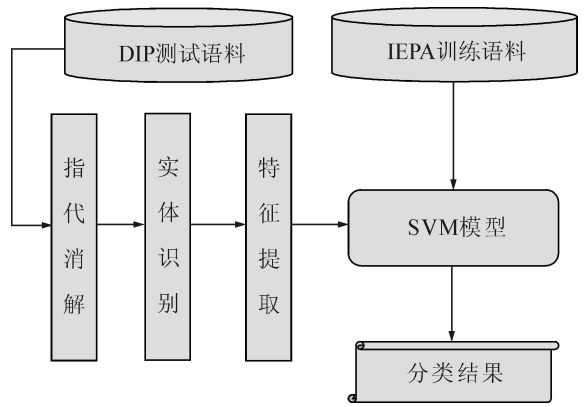


图 1　BioPISVMExtractor系统框架

Fig 1　System framework of BioPISVMExtractor

2.1　指代消解

指代消解是自然语言处理的重要内容,在信息抽取系统中,指代消解是一个关键问题[18].生物医学文本中的基因交互关系经常通过与实体相关的代词来表述.因此,信息抽取方法必须考虑代词的消解.

所设计系统的指代消解部分目前处理的是第三人称代词和反身代词,因为第一人称和第二人称代词经常都被用于指代文本的作者,与关系抽取中所涉及到的实体几乎没有关系.指代消解部分首先使用 GENIA Tagger进行词性标注[19],将文本中的单数名词、名词短语和复数名词、名词短语标注出来,然后使用与该代词最为接近的并且单复数相吻合的名词或名词短语来消解该代词.

2.2　实体识别

实体识别的目的是在生物医学文本中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.进行蛋白质相互作用关系的抽取,第 1步要做的就是进行生物实体识别.笔者曾提出过的基于条件随机域 (CRF)与上下文线索结合的生物实体识别方法在 JNLPBA2004数据集上,可以达到 75.04%的 F值,在 BioCreative 2004的测试集上,可以达到 83.71%的 F值[20].在 BioPISVMExtractor系统中,也使用了该方法.

2.3　SVM模型

近年来,支持向量机 (SVM)的研究在广泛开展.支持向量机是 V. Vapnik等根据统计学习理论 (statistical learning theory, SLT)提出的一种新的机器学习方法,该方法能较好地解决小样本、非线性、高维数和局部极小点等实际问题[21-23],已成为机器学习界的研

究热点之一,并成功地应用于分类、函数逼近和时间序列预测等方面[24-26].

SVM是从线性可分情况下的最优分类面发展而来的,所谓最优分类面就是要求分类面不但能将两类正确分开,而且使分类间隔最大.分类线性方程为 $x \cdot w + b = 0$,其中,w为分类面的法线,b决定相对于原点的位置.可以对它们进行归一化,使得对线性可分的样本集 (x_i, y_i), $i = 1, \cdots, n$, $x_i \in R^d$, $y_i \in \{+1, -1\}$满足:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, \ i = 1, \cdots, n \quad (1)$$

此时分类间隔等于 $2/\|w\|$,使间隔最大等价于使 $\|w\|^2$ 最小.满足式(1)且使 $\frac{1}{2}\|w\|^2$ 最小的分类面就是最优分类面.

利用 Lagrange优化方法可以把上述最优分类面问题转化为其对偶问题[28],即在约束条件:

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \quad (2)$$

和

$$\alpha_i \geq 0, \ i = 1, \cdots, n \quad (3)$$

下对 α_i 求解:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_j \cdot x_j) \quad (4)$$

的最大值.式(4)中,α_i 为与每个样本对应的 Lagrange乘子.这是一个不等式约束下二次函数寻优的问题,存在惟一解.容易证明,解中将只有一部分(通常是少部分)α_i 不为零,对应的样本就是支持向量.解上述问题后得到的最优分类函数为

$$f(x) = \mathrm{sgn}\{(w \cdot x) + b\} = \mathrm{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\}. \quad (5)$$

式中的求和实际上只对支持向量进行.α_i^* 是非零 loagrang乘子,b^* 是最优分类阈值,可以用任意一个支持向量(满足式(1)中的等号)求得,或通过 2类中任意一对支持向量取中值求得.

在线性不可分的情况下,可以在式(1)条件中增加一个松弛项 $\varepsilon_i \geq 0$,成为

$$y_i[(w \cdot x_i) + b] - 1 + \varepsilon_i \geq 0, \ i = 1, \cdots, n, \quad (6)$$

将目标改为求 $(w, \varepsilon) = \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{n} \varepsilon_i)$ 最小,即折衷考虑最少错分样本和最大分类间隔,得到广义最优分类面.其中,$C > 0$是一个常数,它控制对错分样本惩罚的程度.广义最优分类面的对偶问题与线性可分情况下几乎完全相同.只是条件(3)变成了条件(7):

$$0 \leq \alpha_i \leq C, \ i = 1, \cdots, n \quad (7)$$

对非线性问题,可以通过非线性变换转化为某个高维空间中的线性问题,在变换空间上求最优分类面.这种变换可能比较复杂,因此这种思路在一般情况下不易实现.但是注意到,在上面的对偶问题中,不论是式(4)还是表达的寻优函数式(5)都只涉及训练样本之间的内积运算 $(x_i \cdot x_j)$,在高维空间实际上只需进行内积运算,而这种内积运算是可以用原空间中的函数实现的,甚至没有必要知道变换的形式.根据泛函的有关理论,只要一种核函数 $K(x_i, x_j)$满足 Mercer条件,它就对应某一变换空间中的内积.因此,在求解最优分类面中采用适当的内积函数 $K(x_i, x_j)$就可以实现某一非线性变化后的线性分类,而计算复杂度却没有增加,此时目标函数变为

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j). \quad (8)$$

而相应的分类函数也变为如式(9)所示:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*). \quad (9)$$

2.4 特征选取

使用 SVM分类器进行蛋白质相互作用关系抽取的核心工作是特征项的选取.选取特征项的好坏将直接影响到分类的精度.为了使蛋白质相互作用关系抽取系统达到较高的精度,采用了多种特征,包括特征词项特征、关键词特征、实体距离特征、链接特征以及链接语法分析特征.

2.4.1 词项特征

在本文系统中使用了 3种词项特征.它们分别是包含在 2个蛋白质名中的词项、2个蛋白质名之间的词项以及 2个蛋白质名周围的词项.

2.4.1.1 包含在 2个蛋白质名中的词项

顾名思义,这些特征包括出现在 2个蛋白质名中的所有词.因为一个蛋白质名可以是一个词,也可以是多个词,所以相应的特征也是包含一个词或者多个词.例如句子 *A*:"We show here that recombinant **bovine prion protein** strongly interacts with the catalytic alpha/alpha' subunits of **protein kinase**".

句子中用黑体标注的词便是蛋白质名.这样,在特征向量中它们的特征值就分别被表示为 p1_bo-

vine、p1_prion、p1_protein以及 p2_protein、p2_kinase.

2.4.1.2　2个蛋白质名之间的词项

这些特征包括位于 2个蛋白质名之间的所有单词.如果 2个蛋白质名之间没有单词出现,那么这个特征就被设置为空 (NULL).

对于上面例句 *A*中的句子,位于 2个蛋白质名之间的单词串是“strongly interacts with the catalytic alpha/alpha' subunits of”,那么在特征向量中它们的特征值就被表示为 b_strongly、b_interacts、b_with、b_the、b_catalytic、b_alpha/alpha'、b_subunits和 b_of

2.4.1.3　2个蛋白质名周围的词项

这些特征由 2部分组成:一部分是第 1个蛋白质名左边的 *n*个词项;另 1部分是第 2个蛋白质名右边的 *n*个词项.这里,*n*是需要考虑的蛋白质名周围的词项个数,在本文系统中 *n*值被设为 3.与 2个蛋白质名之间的词项特征相似,如果在第 1个蛋白质名左边没有词项,那么这个特征就被设置为空 (NULL);同理,如果在第 2个蛋白质名右边没有词项出现,那么这个特征也被设置为空 (NULL).这里不考虑这些词出现的顺序.

对于上面的例句 *A*,2个蛋白质名周围的词项包括:第 1个蛋白质名左边的 3个词项“here that recombinant”;第 2个蛋白质名右边的 3个词项“. .”.那么在特征向量中它们的特征值就分别被表示为 l_here、l_that、l_recombine以及 r_.

2.4.2　交互词特征

这里所说的交互词,指的是表示 2个蛋白质名之间交互作用关系的交互动词 (interactor),如例句 *A*中的“interact”就是表明句中 2个蛋白质间关系的交互词.构造的关互词表中包含了大约 500个交互词.

在本文系统中,如果有交互词位于 2个蛋白质名之间或者位于 2个蛋白质名周围,那么这个交互词就被加入到交互词特征中.如果句中出现多个包含在交互词表中的关交互词,那么系统会选取句中的第 1个出现的交互词.如果句中没有关键词出现,那么这个特征就被设置为空 (NULL).

对于例句 *A*中的句子,查找交互词表,找到的关键词为“interacts”,那么在特征向量中它的特征值就被表示为 k_interacts

2.4.3　实体距离特征

距离较近的实体存在交互关系的可能性较大,系统因此引入实体距离特征.如果两实体间距离小于等于 3个单词,则相应的特征值就被表示为“DISLessThree”;如果两实体间距离大于 3个单词而小于等于 6个单词,则相应的特征值就被表示为“DISBetweenThreeSix”;如果两实体间距离大于 6个单词而小于等于 9个单词,则相应的特征值就被表示为“DISBetweenSixNine”;如果两实体间距离大于 9个单词而小于等于 12个单词,则相应的特征值就被表示为“DISBetweenNineTwelve”;两实体间距离大于 12个单词,则相应的特征值就被表示为“DISMoreTwelve”.

2.4.4　链接特征

系统中对链接特征的提取用到的是链接语法分析器.链接语法 (link grammar)是 D. Sleator和 D. Temperley[28]于 1991年提出的.它便于语言工程的实现,是计算语言学中引人注目的一种新的语法理论.一部链接语法就是一个单词的集合,其中每个单词后面记录着各自的链接要求.这些链接要求可以通过一系列链接子表达式指定.一个由单词组成的串,如果在单词之间存在满足下列条件的链的话 (或者说能够在单词之间画出一些链,并且这些链满足下面的条件),就说这个单词串是链接语法所定义的语言中的句子.首先这些链满足了其中所有单词的链接要求,其次满足下面 4条元规则:1)平面性,这些链之间互相不交叉;2)连通性,这些链足以把所有的单词链接在一起;3)顺序性,公式中较左边的链接子必须和距离单词较近的单词链接;反之,公式中较右边的链接子必须和距离单词较远的单词链接;4)排他性,一对单词之间同时不能有 2条链链接.

简单的链接表达式由链接子、二元操作符 &和 or以及圆括号组成.每个链接子由名字和后缀 2部分组成.后缀有 2个,分别是 +和 -. +和 -表示链接的方向,+表示向右链接,-表示向左链接.单词串中某个单词如果有一个向右的链接子,例如 X+,而另一个单词有一个向左的链接子 X-,那么这 2个链接子就相互匹配,这两个单词之间就可以画一条 X链.同时可以说,链接子 X+或 X-得到了满足或说链接满足了链接子 X+或 X-.在链接子被满足的基础上,可以定义:1)公式 X & Y要被满足,则链接必须同时满足链接子 X和 Y;2)公式 X or Y要被满足,则链接必须满足链接子 X和 Y中的一个.

图 2显示了链接语法分析器应用于例句“Bovine PRION protein as a modulator of protein KINASE CK2 is described.”上的效果.

在识别出语料中的每个句子中包含的命名实体后,使用链接语法分析器提取句子中 2个命名实体之间存在链接路径.如果相应的链接路径能够被提取出来,那么相应的特征值就被表示为“Link_YES”;否则,就被表示成“Link_NO”.

对于例句 A中的句子,经过链接语法分析器分析后得到的链接路径为:“bovine. a- > A- > protein. n - > Ss - > interacts v - > MVp - > with - > Jp - > subunits[!]. n - > Mp - > of - > Jp - > kinase[?]. n < - AN < - protein. n”.那么相应的特征就被表示为“Link_YES”.

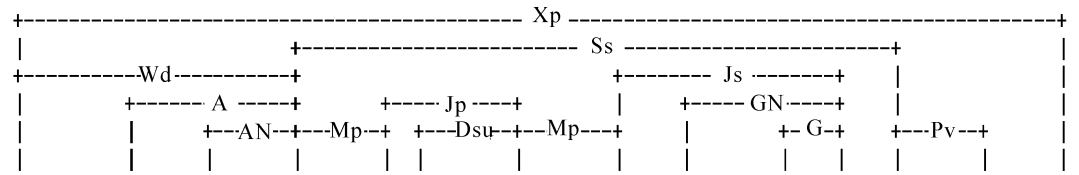


图 2 一个生物医学文献中的句子的链接语法结果

Fig. 2 Results of a sentence from biomedical literature after link grammar parsing

2.4.5 链接语法分析特征

通过链接语法分析划分语法成分抽取蛋白质交互作用关系[16],能获得较高的准确率(55.41%).因此将该方法识别的结果作为特征引入,可能会提高 SVM分类器的准确率.如果句子中的一对蛋白质用链接语法分析方法被提取出来,那么相应的特征值就被表示为“LinkExtracted_YES”,否则,就被表示成“LinkExtracted_NO”.

对于例句 A“We show here that recombinant bovine prion protein strongly interacts with the catalytic alpha/alpha' subunits of protein kinase”,抽取的特征如表 1所示.

表 1 例句 A的特征选取

Table 1 Feature choice of example sentence A

特征名	特征值
第 1个蛋白质名	p1_bovine, p1_prion, p1_protein
第 2个蛋白质名	p2_protein, p2_kinase
2个蛋白质名之间的词项	b_strongly, b_interacts, b_with, b_the
左边的词项	l_here, l_that, l_recombine
右边的词项	r_
实体距离特征	DistBetweenSixNine
交互词特征	k_interacts
链接特征	Link_YES
链接语法分析特征	LinkExtracted_YES

3 实验与讨论

3.1 实验语料

系统所采用的训练语料是 IEPA(interaction extraction performance assessment)语料.该语料是由美国爱荷华州立大学的 J. DING和 D. BERLEANT等人构建的[18].它包含 303篇 Medline摘要,这些摘要是使用 10个查询串对 PubMed进行查询得到的结果,其中每个查询串都包含由“AND”连接词连接的 2个生物医学名词,它们是生物医学研究者根据文本挖掘系统用户的兴趣来制定的.这些由查询串查询到的摘要包含 336个正例(蛋白质与蛋白质之间存在相互作用关系)和 308个负例(蛋白质与蛋白质之间不存在相互作用关系).在 IEPA语料中所有的蛋白质名都已经被正确地标注,以使语料更适合关系抽取的使用.系统所采用的测试语料来源于交互蛋白质数据库(DIP),共包含 392条正确关系记录. IEPA语料和 DIP语料都是来自于包含蛋白质与蛋白质交互信息的 MEDLINE摘要,属于同质的数据,适合分别作为训练语料和测试语料.

3.2 实验结果

由于未得到这 229篇 PubMed文献的正文,本文只使用它们的摘要进行了关系抽取测试,然后对比 DIP 392条记录对结果进行了人工评测.表 2列出了引入不同特征对性能的影响.可以看到只使用词项特征获得的召回率和准确率都比较低,但随着更多特征的引进,召回率和准确率都得到提高.在链接语法分析特征引入后,召回率虽然略有下降(从 71.2%降到 70.4%),但准确率提高较多(从 37.8%提高到 43.6%),也得到了更好的综合分类率 F指数(从 49.4%提高到 53.8%).

表 2　引入特征对性能的影响

Table 2　Performance of using different features

特征组合	特征类型					召回率/%	准确率/%	F值/%
	词项特征	实体距离特征	关键词特征	链接特征	链接语法分析特征			
组合 1	*					63.5	23.4	34.2
组合 2	*	*				64.2	28.5	39.5
组合 3	*	*	*			70.7	33.2	45.2
组合 4	*	*	*	*		71.2	37.8	49.4
组合 5	*	*	*	*	*	70.4	43.6	53.8

注:标记"*"的特征项表示被引入的特征.

将结果与基于链接语法分析的 BioPIExtractor[16]、IntEx系统和 BioRAT系统的结果对比,结果如表 3和表 4所示. BioPISVMExtractor系统的召回率(70.4%)要明显高于 BioPIExtractor(39.80%)、IntEx(26.94%)和 BioRAT(20.31%)系统,这表明基于支持向量机方法能获得比语法分析方法更高的召回率. 在大多数情况下,生物医学研究者更倾向于获得更高的召回率,在这一点上 BioPISVMExtractor系统具有明显的优势,在准确率上,BioPISVMExtractor系统低于其他 3个系统.

从召回率与准确率的综合分类率 F指数来看,BioPISVMExtractor系统的综合分类率 F达到了53.8%,高于同一测试语料上的其他系统:BioPIExtractor系统(46.33%)、IntEx系统(38.20%)和 BioRAT系统(29.68%).

表 3　BioPISVMExtractor召回率与 BioPIExtractor、IntEx和 BioRAT的比较

Table 3　Recall comparison among BioPISVMExtractor, BioPIExtractor, IntEx and BioRAT

结果	BioPISVMExtractor		BioPIExtractor		IntEx		BioRAT	
	个数	百分率/%	个数	百分率/%	个数	百分率/%	个数	百分率/%
召回	276	70.40	156	39.80	142	26.94	79	20.31
未召回	116	19.60	236	60.20	385	73.06	310	79.69
总数	392	100.00	392	100.00	527	100.00	389	100.00

表 4　BioPISVMExtractor准确率及综合分类率与 BioPIExtractor、IntEx和 BioRAT的比较

Table 4　Precision comparison among BioPISVMExtractor, BioPIExtractor, IntEx and BioRAT

结果	BioPISVMExtractor		BioPIExtractor		IntEx		BioRAT	
	个数	百分率/%	个数	百分率/%	个数	百分率/%	个数	百分率/%
正确	672	43.60	543	55.41	262	65.66	239	55.07
不正确	869	56.40	437	44.59	137	34.34	195	44.93
总数	1 541	100.00	980	100.00	399	100.00	434	100.00
F值		53.80		46.33		38.20		29.68

3.3　错误分析与讨论

从生物医学文本中进行蛋白质(基因)相互作用关系的抽取受限于自然语言语法、语义的复杂性,要想取得较高的性能,也是极具挑战性的任务. 对关系抽取的错误原因进行了分析.

需要指出的是,由于 DIP数据库中的记录包含的关系来自 229篇 PubMed文献的摘要和正文,而实验语料只是 229篇 PubMed文献的摘要,所以不可避免地会影响召回率的性能. 这种情况占召回率错误的大多数,如果不考虑只包含在正文中的实体关系,BioPISVMExtractor系统在该测试集上的召回率会进一步提高.

在关系抽取各个阶段中,指代消解、命名实体识别、链接分析、特征提取等处理环节,每个环节都可能导致错误的发生. 其中,指代消解部分的错误是因为自然语言文本中指代关系的复杂性. 命名实体识别部分的错误是因为生物实体命名很不规范,生物实体命名识别是当前研究的一个难点和热点,当前

最好的实体识别系统的 F指数也不超过 80%.这部分错误所占的比例相对较大.此外,链接语法分析器本身也会产生链接提取错误.由于训练语料较少,特征较稀疏所导致的关系抽取错误也占了一定比例.

4 结束语

本文提出的一种基于支持向量机的蛋白质交互作用关系抽取方法,该方法通过适当特征的选取,利用 SVM分类器判断句子中每对蛋白质(基因)是否存在相互作用关系.实验结果表明该方法取得的召回率和综合分类率优于同一测试语料上其他系统,尤其在召回率方面,该方法的效果明显高于其他系统,能较好地满足生物医学研究者的要求.下一步的工作,将考察引入其他特征对抽取性能的影响.此外,会对语料文献的正文和摘要同时进行关系抽取,进一步考察抽取效果.

参考文献:

[1] PUSTEJOVSKY J, CASTANO, ZHANG J. Robust relational parsing over biomedical literature: extracting inhibit relations[C]// Proceedings of the Seventh Pacific Symposium on Bio-Computing [S.l.], 2002: 362-373.

[2] LEROY G, CHEN H, MARTINEZ J D. A shallow parser based on closed-class words to capture relations in biomedical text[J]. Journal of Biomedical Informatics, 2003, 36(3): 145-158.

[3] PARK J C, KIM H S, KIM J J. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar[C]// Proceedings of the Pacific Symposium on Bio-Computing Hawaii, USA, 2001: 396-407.

[4] TEMKIN J M, GILDER M R. Extraction of protein interaction information from unstructured text using a context-free grammar[J]. Bioinformatics, 2003, 19: 2046-2053.

[5] AHMED S T, CHINDAMBARAM D, DAVULCU H, et al. IntEx: a syntactic role driven protein-protein interaction extractor for bio-medical text[C]// Proceeding of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics Detroit, Michigan, USA, 2005: 54-61.

[6] ONO T, HISHIGAKI H, TANIGAMI A, et al. Automatic extraction of information on protein-protein interactions from the biological literature[J]. Bioinformatics, 2001, 17(2): 155-161.

[7] HUANG M L, ZHU X Y, HAO Y, et al. Discovering patterns to extract protein-protein interactions from full texts[J]. Bioinformatics, 2004, 20(18): 3604-3612.

[8] DAVID C, BEMARD B, WILLIAM L, et al. BioRAT: extracting biological information from full-length papers[J]. Bioinformatics, 2004, 20(17): 3206-3213.

[9] ANDRADE M A, VALENCIA A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families[J]. Bioinformatic, 1998, 14(7): 600-607.

[10] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]// Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. Heidelberg, Germany, 1999: 77-86.

[11] STAPLEY B, BENOIT G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts[C]// Proceedings of the Pacific Symposium on Biocomputing [S.l.], 2000: 529-540.

[12] JENSSEN T K, 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.

[13] MARCOTTE E M, XENARIOS I, EISENBERG D, et al. Mining literature for protein-protein interactions[J]. Bioinformatics, 2001, 17(4): 359-363.

[14] BLASCHKE C, VALENCIA A. Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study[J]. Comparative and Functional Genomics, 2001(2): 196-206.

[15] LUKASZ S, CHRISTOPHER S M, ADAM J S, et al. The database of interacting proteins: 2004 update[J]. Nucleic Acids Research, 2004, 32(1): 449-451.

[16] YANG Zhihao, LIN Hongfei, WU Baodong. BioPPIExtractor: a protein-protein interaction extraction system for biomedical literature[J]. Expert Systems with Applications, 2007(12): 14-19.

[17] DING J, BERLEANT D, NETTETON D, et al. Mining MEDLINE: abstracts, sentences, or phrases? [C]// Proceedings of the Pacific Symposium on Biocomputing Hawaii, USA, 2002: 326-37.

[18] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报, 2002, 16(6): 9-17.
WANG Houfeng. Survey: computational models and technologies in anaphora resolution[J]. Journal of Chinese Information Processing, 2002, 16(6): 9-17.

[19] TSURUOKA Y, TATEISHI Y, KIM J D, et al. Developing a robust part-of-speech tagger for biomedical text [C]// Proceedings of Advances in Informatics-10th Panhellenic Conference on Informatics. Volos, Greece, 2005: 382-392.

[20] YANG Zhihao, LIN Hongfei, LI Yanpeng. Exploiting the contextual cues for bio-entity name recognition in biomedical literature [J]. Journal of Biomedical Informatics, 2008(1): 36-42.

[21] VAPNIK V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[22] 阎　辉,张学工,李衍达. 应用 SVM方法进行沉积微相识别 [J]. 物探化探计算技术, 2000, 22(2): 158-164.
YAN Hui, ZHANG Xuegong, LI Yanda. Support vector machine methods in pattern recognition of sedimentary facies [J]. Computing Techniques for Giophysical and Geochenical Exploration, 2000, 22(2): 158-164.

[23] 张学工. 关于统计学习理论与支持向量机 [J]. 自动化学报, 2000, 26(1): 32-42.
ZHANG Xuegong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42.

[24] 李　凯,郭子雪. 一种基于 SVM的函数模拟方法 [J]. 微机发展, 2001(3): 5-6.
LI Kai, GUO Zixue. A function simulation based on support vector machine [J]. Microcomputer Development, 2001(3): 5-6.

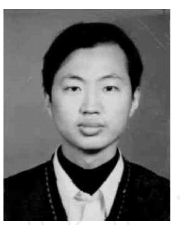
[25] 马云潜,张学工. 支持向量机函数拟合在分形插值中的应用 [J]. 清华大学学报, 2000, 40(3): 76-78.
MA Yunqian, ZHANG Xuegong. Application of support vector machines function regression in fractal interpolation [J]. Journal of Tsinghua University, 2000, 40(3): 76-78.

[26] MÜLLER K R, SMOLA A J, RATSCH G, et al. Predicting time series with support vector machines[C]// Proceedings of the 7th International Conference on Artificial Neural Networks. Lausanne, Switzerland, 1997.

[27] BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[28] SLEATOR D, TEMPERLEY D. Parsing English with a link grammar [C]// Proceedings of Third International Workshop on Parsing Technologies. Tilburg, Netherlands, 1993.

作者简介:



杨志豪,男,1973年生,讲师,主要研究方向为文本挖掘和中文信息处理,发表学术论文 20余篇.

洪　莉,女,1962年生,副教授,主要研究方向为智能信息处理.

林鸿飞,男,1962年生,教授,博士生导师,主要研究方向为搜索引擎、文本挖掘、情感计算、中文信息处理以及商业智能的研究. 主持 2项国家自然科学基金和 1项国家 863高科技计划研究项目. 发表学术论文百余篇.