**介绍**

**背景部分**

随着人类基因组计划的完成，人类对于生命科学的理解进一步加深，并由此衍生出了大量的研究方向，而系统生物学作为其中的一门新兴学科，获得了研究者的广泛关注。跟传统的生物学相比，系统生物学不仅仅只局限于单个蛋白质，单个基因对于生物机体产生的影响，而是通过研究蛋白质和蛋白质，基因和基因，蛋白质和基因，蛋白质和疾病之间的相互作用和相互关系，利用计算机和数学建模，通过计算的方式来预测细胞，器官甚至整个生物机体的系统表现。由此可见，生物实体之间的基本关系作为系统生物学最重要的基础部分，对于更深层次去理解机体的表现具有重要的意义。

21世纪以来，随着生物技术的迅猛发展，相关的生物文献也以一种爆炸式的方式在增长，这些生物文献大多是以电子版的方式存储在生物文献的数据库中，全世界的科研人员可以通过检索这些数据库，从而获得最新的研究成果。以最出名的Medline数据库为例子，在过去的十年时间里，它收录的文档总数翻了一倍，目前该数据库已经涵盖了70多个国家地区的7000多种期刊，收录接近2300万的生物文献，基本上涵盖了整个临床医学和生命科学的范围。值得注意的是，这些文献大多是以纯文本的，非结构化的方式存储在数据库中，这些海量的数据加上非结构化的存储方式，使得高效的检索成为技术难题。所以如何利用自然语言处理和数据挖掘的方式去分析这些海量文本，从原始文本中提取有用的信息成为了一个研究热点。

PubMed作为最大的，免费的生物文献搜索引擎，基本上涵盖了Medline，PMC，OLDMEDLINE，record in process等几个主流的生物文献的数据库，进一步的，它能提供各种生物医学论文的搜索以及摘要展示。通过收集用户的搜索记录，PubMed能够提供比较友好的检索体验。以关键检索词“breast cancer”为例，图1展示了PubMed的检索结果。从图中可以看出，在键入关键词“breast cancer”以后，搜索引擎能够联想出很多关于breast cancer的主题，比如breast cancer的治愈，breast cancer的风险等等。但是更深层次的主题，比如想要获取breast cancer相关的基因，如果没有丰富的生物医学相关的知识，很难直接从搜索引擎获得这些基因的信息。而生物实体关系抽取技术的研究，为提高搜索引擎的搜索精确性提供了一个很好的技术支持，如果关系抽取系统已经抽取了所有与breast cancer相关的实体，那么搜索引擎技术就可以从这些实体中选择合适的内容，为用户提供更加精确的搜索结果，因此，高质量的，精确的，自动的生物关系抽取系统是有必要的。



图1：基于关键词“breast cancer”的PubMed的检索结果

**生物实体的关系抽取**

生物实体关系抽取任务旨在从生命科学文献中准确有效地发现高质量的生物实体之间的关系。从这个定义可以看出，这里面存在两个概念：实体和关系。实体在不同的领域具有不同的定义，在非生物领域，实体可以认为是一个人名（name），一个地名（location），一个时间点（time）等等，而在生物领域，实体特指生物学上的专有个体名称，比如基因（gene），蛋白质（protein），疾病（disease）或者药物（drug）。可以看出，如果想要进行生物实体关系的抽取，第一步需要发现识别出这些生物实体。已经有大量的工作研究如何从纯文本中抽取这些命名实体，其中最简单的方法可以通过词典匹配的方式来抽取实体，但是这类方法需要人工构建实体字典；目前研究最广泛，效果最好的命名实体识别系统大多基于条件随机场（CRF），通过标记少量的数据，利用模型训练的方式，可以实现新的命名实体的识别。定义中的另一个概念则是关系，对于生物实体的关系抽取系统而言，在本文中，我们把关系定义为实体之间的相互作用。从表1可以看出一些典型的实体关系。

表1：生物实体之间典型的关系

|  |  |  |  |
| --- | --- | --- | --- |
| 实体类别 | 关系关键词 | 英文关键词 | 例子 |
| 基因之间 | 依赖 | depend | GeneA is depended on GeneB |
| 蛋白质之间 | 绑定 | bind | ProteinA binds ProteinB |
| 药物与药物 | 交互 | interaction | The interaction of DrugA and DrugB has been established |

从上面的描述可以看出，实体关系抽取问题可以转化成实体关系的分类问题，如果只是判断实体之间是否存在关系，则可以将问题简化成一个二分类问题；如果需要判断实体之间存在什么样的关系，那问题则可以转化成一个多分类的问题。从数学的角度来说，生物实体关系抽取任务可以描述如下：给定一个句子，其中和是命名实体，可以定义映射函数如公式1所示（这里面为了问题描述的方便，仅仅考虑二分类，因为多分类问题一般都可以转化成二分类问题）：

（1）

可以看作是从句子中提取到的特征，而映射函数则决定了两个实体之间是否存在关系，可以认为是一个分类器，传统的分类器像支持向量机（SVM），贝叶斯分类器等都可以用来当作。所以如果存在大量的标记数据，那我们就可以训练这个分类器，从而实现自动的实体关系抽取。

**生物实体关系抽取工作的研究现状**

生物实体关系抽取方面的现有工作总的来说可以分成3个大的方面。它们分别是：基于实体共现（Co-occurrence）的方法，基于规则的方法（Rule-based）以及基于机器学习（Machine Learning）的方法。本章节重点介绍这三方面的现有工作以及相应的效果。

1. 基于实体共现的方法

大多数基于实体共现的系统都是从句子层面上去考虑一对实体对是否存在关系（基于实体共现paper），为了判断两个实体之间是否存在关系，这类系统通常采用模板匹配的方式来实现自动的生物实体关系的抽取。第一个用来抽取蛋白质之间关系的系统是由Blaschke等人提出的，他们手工定义了一系列的模板用来匹配一对蛋白质实体对，这些定义的模板大多是以一系列的单词数字以及特殊符号组成，一个示例的模板如下所示：“Interaction of (3) <P> (3) with (3) <P>”，其中<P>代表一个蛋白质实体，单词”of”和特殊符号”<P>”代表这两个单词在一个句子中相隔的最大距离。通过检测实体的共现，利用模板匹配的方式，可以实现生物实体关系的自动抽取。

上述提出的方法还是基于句子的层面，然而大多数时候，我们需要处理文档层面的，基于以上的原因，Bunescu等人利用实体对共现的统计特性来检测一对实体是否存在关系。像点估计（PMI），卡方估计（chi-square），对数似然比（LLR）这类估计的方法都可以用来检测两个实体的出现在一起是否是偶然。以PMI估计为例子，两个实体之间的PMI

其中代表着两个生物实体同时出现的概率，它可以利用同时出现的次数来近似的估计，同理和作为两个实体各自出现的概率，同样也可以利用和各自出现的次数和来近似的估计。Bunescu指出PMI这种估计方式可以用来估计两个实体之间是否存在潜在的关系。

综上所述，基于实体共现的方法相对比较单一，实现起来比较简单，可以看出一个好的模板库会显著地影响系统的性能，另一方面，由于语言本身的多样性，预定义的模板可能很难完全的概况各种生物实体存在依赖关系的情况，这导致这些系统会取得很好的准确率，但是召回率会相对较低，如果要实现跨领域的关系抽取系统，需要重新定义模板，代价比较大。此外，利用各种估计的方法检测两个实体之间是否存在关系的系统，它们的性能严重依赖于语料库的分布情况，随着不同的语料库，会使得最终的结果产生震荡。