**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。可以看出，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，可以为生物领域的研究者们提供最前沿的研究动态。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（binary classification）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（multi-classification）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，我们可以它认为是一个分类器，传统的分类器像感知器（perceptron），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。当然也可以认为是一系列的规则或者模板，通过这些规则和模板可以用来判断生物实体之间是否存在关系。因此，的不同选择会使得系统有所差异，具体的方法将在下个章节详细描述。

**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>”之间的数字(3)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可以通过下面的公式2来计算

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。

1. 基于语法规则匹配的方式

从上面的描述可以看出，基于实体共现的方法大多只是利用语料库的统计信息，然而更深层次，比如句子层面的语法，句法信息并没有被这些方法包括进去，这使得这些方法对于数据过于敏感，很难准确刻画数据的内部特征。此外利用实体共现的方法大多只能判断两个实体之间是否存在关系，而很难判断实体之间存在什么样的关系，因此深层次的分析变得尤为重要。以句子为例，句子的语法信息大多能够反映一句话的内部特性，因此，结合语法和规则进行关系抽取是一个很自然的选择。尤其是随着自然语言处理技术（NLP）的迅速发展，各种各样的语法分析工具也不断出现，比如传统NLP领域的Stanford parser，生物领域内的Genia等等，这些句法分析工具为分析文本提供了便利，这个章节描述的基于句法规则的方法大多是基于这些语法分析工具的结果进行的。

最早利用句法分析和规则来进行关系抽取是由Yakushiji等人提出的。他们首先利用大规模通用的语法构建了一个语法分析器，利用这个语法分析器，可以根据动词（因为大多数的实体之间的作用关系都是通过动词体现出来的，所有这里面主要根据动词来分析），将各种各样的句子表达转化成一个标准的结构，他们称这样的结构为“论元结构” （argument structure），在“论元结构”的基础上，利用少量的规则和模板，就可以完成生物实体关系抽取工作。图2展示该系统的一个流程。



图2：利用“论元结构”以及规则来进行生物实体关系抽取

通过分析，我们可以看出，利用“论元结构”的一个最大优势在于减少语言多样性对于结果的影响。比方说对于蛋白质实体P1和P2，它们之间的表达关系可以通过下面的表达方式来体现：

* P1 activates P2 （简单的主动表达）
* P2 is activated by P1 （一般的被动表达）
* P1 activating P2 （动名词形式的表达）
* Activation of P2 by P1 或者P2 activation by P1（名词化形式的表达）

可以看出，两个实体之间的表达方式是多样的，如果使用句子层面的模板匹配的方式，则需要定义至少5个模板来匹配上面的表达。但是利用“论元结构”的方式，只需要根据动词“activate”（或者activate的变形activation，activates，activated等）以及对应的生物实体对P1和P2，利用少量的规则，就可以分析出关系。换句话说，语法层面的规则是更深级别的（high level），相对于单纯句子层面的规则，语法层面的规则更加具有归纳性，能够反映出数据的内部特性。

Fundel 等人提出了一种利用解析依赖树（dependency parser tree）来进行关系抽取的方法。他们首先利用Stanford parser对句子进行语法的依赖分析，通过生成的语法树（如图3所示），加上少量定义的规则，在LLL-challenge 2005标准训练数据集上取得了0.75的准确率（测试集上0.72的准确率）。



图3：对于句子“The **sigmaB**-dependent promoter drives expression of **yvyD** under stress conditions and after glucose starvation whereas a **sigmaH**-dependent promoter is responsible for **yvyD** transcription”句法依赖分析的结果。图中所有的生物实体都用红色进行了标记，椭圆中每个单词后面的数字代表着一个单词在句子中的位置，深灰的矩形方框代表着依赖关系，每个椭圆形则代表着一个短语块，每个箭头则代表着两个单词之间存在的句法依赖关系，句法分析是在短语块的基础上进行的，其中黄色箭头部分代表着一对实体之间的依赖关系路径。

Fundel等人通过分析文本，针对生物领域实体关系的特殊性，首先定义了三种实体相互作用的关系类型，它们分别是：

* P1-relation-P2 （比如：P1 activates P2），也就是P1通过动词作用P2;
* Relation-of-P1-by-P2 （比如：activation of P1 by P2），P2通过介词作用于P1;
* Relation-between-P1-and-P2 （比如：Interaction between P1 and P2），P1和P2通过between相互作用。

根据这些关系类型，在句法分析树的基础上，他们制定了相应的规则来匹配上面的关系类型。对于第一种通过动词作用的类型，可以利用生物实体之间的依赖关系路径来提取关系。如图3所示，对于实体sigmaB和yvyD，它们之间的依赖关系路径为“The sigmaB-dependent promoter---nsubj---drives---dobj---expression---of---yvyD”。可以看出这个路径中包含关键词drives，同时包含依赖关系nsubj，因为nsubj这个依赖关系通常表示主语和谓语的关系，因此这时候则判定sigmaB和yvyD之间存在依赖关系。对于第二个和第三个通过介词作用的类型，他们同样制定了一些规则来进行关系抽取，跟第一个关系类型不同的是，第一种类型主要考虑动词的作用，因此抽取的是关于动词的路径，第2,3类型考虑的是介词，因此主要抽取实体和介词之间的路径。如果检测到实体之间通过介词“of，by，to，for，in”连接到关键词，则认为存在关系。具体的例子可以参考图4。

跟基于“论元结构”的方法相比，Fundel 等人的方法可以认为是对句子更深层次的分析，这里面涉及到更多自然语言处理方面的应用，包括短语块的抽取，句法分析，句法依赖分析等等。这种深层次的分析更加有利于针对特定的问题来制定特定的策略，比如在Fundel等人的方法中，规则已经被限制到只有3条，跟在句子层面的模板比较，数量上已经大大减少，但是却取得了很好的效果。但基于句法分析的系统同样存在缺陷，其中一个最大的问题在于系统的稳定性完全取决于句法或者依赖关系分析器的稳定性，虽然传统的语法分析器已经可以达到很高的准确率（比如Stanford parser的准确率在标准的测试数据集上已经超过0.90），但是考虑到生物领域的文本具有一定的特性，比如生物领域的特有名词相对较多，句子相对较长，而且针对生物领域句子句法分析的研究较少，这导致一般的句法分析器的性能很难得到保证，这种由句法分析造成的错误的传播，会影响整个生物实体关系抽取的性能。这也是目前基于语法规则的方法面临的一个挑战。

C:\Users\Administrator\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Relex中的介词规则.tif

图4：实体之间通过介词或者between这种方式连接的句法分析结果。黄色部分代表着一对实体通过介词作用的路径。可以看出在左边图中实体TNF和IL-1通过介词of连接到关键词stimulatory，因此判定这两者之间存在关系；在右边图中，实体TNF-alpha和IFN-gamma通过介词between连接到关键词antagonism，因此判断两者之间存在关系。

1. 基于统计机器学习算法的生物关系实体抽取的研究

大多数基于统计学习算法的生物实体关系抽取系统都把关系抽取当作分类问题，所以问题的关键在于如何构建特征以及选择合适的分类器。对于分类器而言，支持向量机（SVM）由于基于结构风险最小化理论，在各项有监督学习的任务上展现出强大的优势，因此SVM也在生物实体关系抽取系统中被广泛的应用，并且体现出良好的性能。因此，在基于SVM分类器的基础上，特征的选择对于实体关系的抽取显得尤其重要。此外由于SVM中核函数的存在，可以将低维空间的数据映射到高维空间，使得低维线性不可分的数据在高维空间是可分的，核函数的选择对于SVM分类问题至关重要，因此本章节也重点介绍了如何利用抽取到的特征来构建不同的核函数，从而实现自动关系抽取的相关工作。

总的来说，核函数衡量的是两个输入样本之间的相似性，为了描述问题的方便，在本文中，我们使用来衡量输入样本和之间的相似性，其中可以认为是某一种核函数。

3.1 n-gram核函数

n-gram作为自然语言中最基本的特征，被广泛的应用到各种任务中。Setre等人提出可以根据两个生物实体的位置关系，可以将n-gram特征分成三个方面，即位于第一个实体前面，两个实体之间，第二个实体后面三个特征集合。例如对于句子Activation of P2 by P1 has been confirmed，可以构造如表2所示的unigram，bigram，tri-gram特征。这些特征通过one-hot的方式，可以表示成一个特征向量，n-gram核函数通过计算特征向量的相似性从而实现自动关系抽取。

表2：根据实体的位置构建的unigram，bigram，tri-gram特征

|  |  |  |  |
| --- | --- | --- | --- |
| 特征类型 | unigram特征 | bigram特征 | Tri-gram特征 |
| 前向 | Activation, of | Activation of | - |
| 中间 | by | - | - |
| 后项 | has, been, confirmed | has been, been confirmed | has been confirmed |

* 1. 基于语法分析的核函数

从上面的叙述可以看出，基于n-gram的核函数提取到的信息还是停留在词（或者多个词）的层面，语法之间的信息并没有被考虑进去，而且由于n-gram信息特征太过于稀疏，这对系统的性能存在一定的影响。基于以上的原因，很多研究者根据语法树，语法依赖树等语法信息，提出了一系列基于语法分析的核函数。

Chao等人首先提出了卷积句法树核（convolution parse tree kernel），并通过对句法树的剪枝算法，在蛋白质和蛋白质任务上取得了很好的效果。卷积树核函数通过比较句法树包含生物实体的子树的相似性，从而判断两个输入样本的相似性。图5展示了如何计算两个句法树之间相似性。



图5：箭头右边代表着一个句法树的所有子树结构，图的上部分是短语”A Dog“的所有子树结构，图的下部分是”A Cat“的所有的子树结构，两者之间相同的子树结构都通过红色进行标记，因此可以计算两个树的卷积核

对于每对蛋白质，Chao抽取包含这对蛋白质的子树，利用一些规则对这些子树进行剪枝，然后利用上述的卷积句法核，通过修改SVM的核函数，从而实现蛋白质关系抽取。

Airola等人利用图的一些算法，提出了基于图的核函数（All-path graph kernel），他们首先对输入的句子进行句法依赖分析，对于句法依赖的结果，他们将其看作一个有向图。图中每个节点都是句子中的一个单词，而每个依赖关系都是有向图的一条边，并且给每条边分配一个权重，可以用来表示这个有向图，其中代表有向图节点，另外用代表着每个节点所代表的类别（比如这个节点是不是处于两个实体的中间，是否在第一个实体的前面等等）。那么可以用代表顶点和之间的权重，而则代表着顶点拥有类别。他们首先计算出所有的顶点相连接的情况，如下所示。

因为需要把自连接的情况排除掉，所以上面的邻接矩阵可以用下面的方式表示。

根据上面的最终的邻接矩阵以及，可以定义基于图的核函数为公式所示，其中，和则是两个输入的样本。

可以看出基于图的核函数主要衡量的是在每个类别下两个有向图之间的联系或者相似性，通过统计所有类别组合下的相似性，就可以计算出两个输入样本之间的相似性。同样的，Airola等人也利用SVM作为基本分类器，实现实体关系抽取。

3.3 混合核函数

通常来说单个核函数可以获得样本某个方面的信息，比如n-gram核函数可以获得句子词层次的信息，而基于句法分析的核函数则可以很好地抓住语法层次的信息。混合核函数则是考虑将这些核函数结合起来，从而能够体现各个层次的信息。

Miwa等人充分利用bag-of-word (bow)核，Subset tree (ST)核以及graph核，并且结合依赖分析（dependency）和深度依赖（deep dependency）分析，提出了以下的混合核函数。

可以看出Miwa等人提出的核函数是一种归一化的核，因为他们存在一个归一化因子，Li等人则提出了一种加权的混合核函数，他们定义混合函数如下：

其中和代表着不同的核函数，注意到这里面的是一个权重值，这个数值可以衡量某个核函数对于最终分类器的贡献。

总结来说，核函数的提出主要是为了衡量两个样本之间的相似性，不管是在词的层面还是语法的层面。混合核函数由于综合了多方面的信息，其性能一般优于单个核函数，这个已经在很多的工作中得到了证实，目前最先进的基于统计机器学习的系统大多基于混合核函数。

**基于神经网络的生物实体关系抽取技术**

在过去几年中，神经网络已经重新成为强大的机器学习模型，在诸如图像识别和语音的领域中取得了很多最先进的成果。最近，神经网络模型也开始应用于自然语言处理领域，并在句法分析，问答系统，语义分析等多个方面取得了巨大的突破。在非生物领域，已经有工作使用卷积神经网络（CNN）或者循环神经网络（RNN）来进行关系抽取。但是在生物领域，基于神经网络的工作相对较少，这也跟前面提到的生物文献特点有关：句子比较长、专有名词多。这给关系抽取工作带来一定的难度。目前比较成功的方法是Li等人利用词向量（word embedding）作为额外特征，并将这个特征融合到SVM中，他们提出的方案在多个标准的蛋白质关系抽取的数据集上取得了优秀的效果。但是注意到这里词向量只是作为额外特征加入，没有使用任何基于神经网络的算法。

**生物实体关系抽取面临的一些挑战**

**本文的研究重点以及章节**

本文主要从两个方面来实现生物实体关系的自动抽取：1）结合句法规则和核函数来进行关系抽取；2）提出了一个利用神经网络模型来进行关系抽取的框架。本文的剩余章节安排如下。

在第二章节，利用句法分析的结果，在此基础上，我们定义了一些规则来进行生物实体的关系抽取，其次我们提出了一个新的混合核函数，通过修改SVM的核函数来实现自动关系抽取。在实验分析部分，我们与现有的关系抽取系统做了详细的比较。此外，我们利用构建好的基于规则的关系抽取系统，并将此系统应用到跟“breast cancer”相关的英文文献下，利用抽取到的关系，构建实体相互作用网络，从而分析得出与“breast cancer”相关的基因。

在第三章节，我们定义了一个用于生物领域实体关系抽取的框架。首先，我们详细介绍了词向量方面的概念以及如何训练词向量；其次我们介绍了如何利用神经网络模型对句子进行编码；最后我们详细地分析了我们的实验结果。

在第四章节，我们对已做的工作做了一个简单的概括。