**介绍**

**1.1：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，基本上涵盖了整个生物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

**1.2：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。

**1.3：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。

1.3.1：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括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，随着语料库分布的不同，会使得最终的结果产生震荡。

1.3.2：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规则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存分析，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”句法依存分析的结果。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存关系，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存关系，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存关系路径。

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存关系路径来提取关系。如图3所示，对于实体sigmaB和yvyD，它们之间的依存关系路径为“The sigmaB-dependent promoter---nsubj---drives---dobj---expression---of---yvyD”（图中的黄色箭头部分）。可以看出这个路径中包含关键词drives，同时包含依存关系nsubj，因为nsubj这个依存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之间存在依存关系。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，句法依存分析等等。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存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。

1.3.3：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，第二个实体后面的三个特征集合。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，语法依存树等语法信息，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，在蛋白质和蛋白质关系抽取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存分析，对于句法依存的结果，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，而每个依存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，并且结合依存分析（dependency）和深度依存（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。

**1.4：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，这给生物领域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很好的效果。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。

**1.5：生物实体关系抽取面临的一些挑战**

尽管生物实体关系抽取工作取得了一定的成就，但是考虑到生物领域文本的多样性和复杂性，生物关系抽取仍然面临着以下的问题和挑战。

1. 标记数据的代价过于昂贵。因为生物实体关系反映的是客观真理，所以标记数据不仅需要从字面层次去理解，还需要有丰富的生物领域内的知识来判断关系是否真的合理，这些原因导致生物实体关系的数据的数量太稀少。这使得不管是基于规则还是统计机器学习算法都无法充分概况或者学习数据的内部特征；
2. 基于规则和机器学习的算法大多依存于句法分析这些外部工具，但是针对于生物领域的句法分析器，一般来说效果很难得到保证，这种错误的传播会影响这些系统的性能；
3. 大多数基于机器学习的系统对于特征的选择太过于直觉，很少有工作去分析哪种特征适用于什么样的情况，这导致特征工程就像一个黑盒子，无法直观的去解释和理解算法本身；
4. 单词的语义（semantic）信息大多被忽略，这使得在判断是否存在关系的时候（二分类），这种语义信息的丢失可能影响并不明显，但是在判断实体之间是什么关系类型的时候（多分类），语义信息会显得比较重要。

**1.6：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生物实体关系抽取的框架。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，我们定义了一些简单的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，我们将构建好的基于规则的关系抽取系统应用到跟“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，并通过一系列网络分析的算法来分析实体之间的关系。

在第三章节，我们定义了一个基于神经网络的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输入的句子进行编码；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，并提出了一些可以改进的部分和未来可以研究的一些方向。

**基于规则和机器学习的生物实体关系抽取系统**

2.1：本章引论

从引言部分可以看出，句法树分析和句法依存树分析在基于规则和机器学习的生物实体关系抽取系统中占有重要地位。在本章节中，在语法分析基础上，并结合现有的一些工作，本文将从规则和机器学习两个角度来构建两个不同的生物实体关系抽取系统。章节2.2介绍了如何对一个输入句子进行句法树分析以及依存树分析，同时详细地说明了如何利用语法分析的结果来生成需要的特征。章节2.3在提取到的特征的基础上，1）定义了少量的规则来进行关系抽取（基于规则的关系抽取系统）；2）提出了一个新的基于编辑距离的核函数和一个混合核函数来进行关系抽取（基于机器学习的关系抽取系统）。章节2.4将两个关系抽取系统应用到了标准的数据集，并同现有的基于规则或机器学习的关系抽取系统做了一些比较，并详细分析了实验结果。章节2.5展示了如何应用基于规则的关系抽取系统去抽取原始文本中实体之间的关系，同时本章节提出了从图的角度来分析抽取到的实体关系对，并通过一系列的图的分析的指标来判断一个实体在一个关系作用图中的地位。章节2.6是对构建的系统做一个简单的总结。

2.2：语法分析以及特征构建

语法分析作为整个系统最重要的环节，它能够分析句子的单词之间的结构关系，从而帮助我们了解整个句子结构，这对于我们从语法的层次去理解句子至关重要，所以说语法分析作为本系统核心环节，它的使用和性能直接影响着后面的处理环节。



图6：利用Stanford parser对句子“The gerE gene are transcribed by sigmaK gene”句法分析的结果。被红色方框圈住的部分是抽取的名词短语块。

在本文中，语法分析分成两个方面：句法分析和依存分析。注意到句法分析重点分析的是句子的语法结构关系，比如哪些单词可以组合成一个短语，哪些词语是动词的主语或者宾语，比如从图6可以看出，句法分析是一种树结构关系。而依存句法理论认为，句法结构本质上是词和词之间的关系，一个依存关系连接两个词，分别是核心词（head）和修饰词（modifier），核心词被修饰词修饰。如果将一对依存关系中的两个单词当作节点，而依存关系作为边，那么依存分析更像是一个图，而不是句法树这种层次结构关系。

前面提到依存分析是从分析单词之间的语法关系，考虑到生物文献中的句子一般较长，为了减少不必要的短语层面的语法关系，本文也从短语层面去分析一个句子，这样就相当于减少了需要分析的句子的长度，一方面可以加快语法分析器的处理速度，另一方面也可以降低依存分析的错误率。一般来说，需要使用特定的工具来进行短语的抽取。但是由于在基于规则的生物实体关系抽取系统中，动词的作用比较明显，所以在实际的应用系统中，应该尽量保留动词的原始特性，因此本文只考虑抽取名词短语，而动词短语不作为抽取的范围。本文主要是从句法分析的结果基础上进行名词短语的抽取，如图6所示，被红色方框圈住的部分就是抽取到的名词短语。从图6可以看出，只要抽取NP（必须限定该节点的深度为2）节点下的所有叶子节点，组合在一起，就可以作为一个名词短语使用。这样就可以得到一个新的句子输入：[“The gerE gene”, “are”, ” transcribed”, “by”, “sigmaK gene”]，我们可以在这个输入上进行依存分析。

短语层面的依存分析主要应用到基于规则的系统中，因为这样可以去除短语层面冗余的依存关系，从而简化规则的制定。但考虑到单词层面依存分析对于基于机器学习的算法更有用，因此，本文从基于短语和基于单词两个层面来分析一个输入句子。图7展示了两个版本的依存分析结果。在下个章节中，这些依存分析会作为特征进行关系抽取。



图7：利用Stanford parser对句子“The gerE gene are transcribed by sigmaK gene”进行单词层面（左图）和短语层面（右图）的依存分析结果。图中用红色标记的是生物实体，两个生物实体之间的依存路径用黄色箭头表示。

2.3：基于规则和机器学习的关系抽取系统

从公式1可以看出，关系抽取系统的关键在于特征的抽取以及映射函数的构建。本文主要依赖句子依存分析的结果，抽取实体之间的最短依存路径作为特征用来进行生物实体关系抽取。已经有大量的研究表明[B]，实体之间最短依存路径可以提供充分的信息来定义一对实体之间的关系类型，这也是我们使用最短依存路径的动机所在。另外考虑到基于规则和机器学习的关系抽取系统的差异性，在基于规则的抽取系统中，本文采用短语层面的依存分析结果，而在基于机器学习的关系抽取系统中，则采用单词层面的依存分析结果。

2.3.1：基于规则的关系抽取系统

现有的研究表明[]，绝大部分的生物实体之间的关系是通过动词相互作用的(从表1可以看出)，还有很少的一部分是通过介词之间来体现关系的（章节1.3.2可以看出）。根据这两个特点，在依存分析的基础上，可以制定两个相应的规则来进行关系抽取。

策略1：如果实体之间存在相互作用动词，并且第一个实体跟该动词之间是主语和谓语的关系的时候，则认为这对实体之间存在关系（主要抽取geneA activates geneB类型）。

以图7右图为例子，我们可以抽取实体gerE和sigmaK之间的最短依存路径为：

The gerE gene---------nsubjpass---------transcribed---------agent---------sigmaK gene

通过观察这条最短依存路径可以看出，该路径中包含相互作用动词“transcribed”，同时包含依存关系“nsubjpass”，而nsubjpass就表明The gerE gene和transcribed之间是被动的关系。这时候则认为gerE和sigmaK存在关系。

策略2：如果两个实体包含在一个名词短语块中，抽取该名词短语块的依存关系，如果这个名词短语块与另一个单词之间通过介词作用，而这个单词又包含相互作用类型的词语，则认为这对实体存在关系（主要抽取The activation between geneA and geneB类型）

以图4左图为例子，可以抽取路径：

The stimulatory Effects---------of--------- TNF and IL-1

TNF and IL-1和The stimulatory Effects通过介词of连接，而短语块The stimulatory Effects又包含相互作用类型的词语The stimulatory，此时认为实体TNF和IL-1存在关系。

从上面的策略可以看出，相互作用词语在策略中占有重要地位，常见的相互作用词包括：bind，depend，rely，activate等等，这些词语从侧面反映了实体之间存在的生物方面的作用关系，所以这些词语是判断实体是否存在关系的重要依据。本文主要参照了Katrin Fundel等人构建的相互作用动词表，在此基础上，我们利用WordNet，采用一个类似“滚雪球”一样的算法，然后，通过人工选择的方法，拓展了这个相互作用的词表。算法描述如下：

1. 将Katrin Fundel等人构建的相互作用动词作为“种子词”加入到堆栈中和一个哈希表中
2. 判断堆栈是否为空，如果为空，则转向4。否则，从堆栈中取出一个词语，利用WordNet，找出该词语的同义词，然后将这些同义词全部加入到了堆栈和哈希表中。
3. 判断哈希表长度是否大于2000，如果不是，重复2，否则，转向4。
4. 手工去除不合理的词语。

2.3.2：基于机器学习的关系抽取系统

前面提到，基于机器学习的关系抽取系统主要是基于单词层面的句法分析结果。以图8为例，该例子中包含KaiC，SasA，KaiA，KaiB四个生物实体，因此，总共存在条最短依存路径，这些路径如下所示：

1. KaiC - nsubj - interacts - prep with - SasA
2. KaiC - nsubj - interacts - prep with - SasA - conj and - KaiA
3. KaiC - nsubj - interacts - prep with - SasA - conj and - KaiB
4. SasA - conj and - KaiA
5. SasA - conj and - KaiB
6. KaiA - conj and - SasA - conj and - KaiB

同时，为了减少数据的稀疏性，可以将路径中的生物实体名称用特殊符号进行替换，在本文中，用*PROTEIN\_1*代表第一个实体，*PROTEIN\_2*代表第二个实体，而位于两个实体之间的则标记为*PROTEIN\_0*，这样上面的6条最短依存路径写成下面的形式。

1. *PROTEIN\_1*- nsubj - interacts - prep with - *PROTEIN\_2*
2. *PROTEIN\_1*- nsubj - interacts - prep with - *PROTEIN\_0* - conj and - *PROTEIN\_2*
3. *PROTEIN\_1*- nsubj - interacts - prep with - *PROTEIN\_0* - conj and - *PROTEIN\_2*
4. *PROTEIN\_1* - conj and - *PROTEIN\_2*
5. *PROTEIN\_1*- conj and - *PROTEIN\_2*
6. *PROTEIN\_1* - conj and - *PROTEIN\_0* - conj and - *PROTEIN\_2*

可以看出前面三个样本表明两个实体之间存在关系，而后面的三个样本则表明两个实体并不存在关系。在本文中，存在关系的样本被标价为正样本，不存在关系的样本则被标记为负样本。



图8：利用Stanford parser对句子“The results demonstrated that KaiC interacts rhythmically with KaiA, KaiB, and SasA”进行单词层面分析的结果。其中KaiC，SasA，KaiA，KaiB都为生物实体。

前面提到，基于机器学习的系统大多利用核函数来衡量两个输入样本的相似性，在本文中，也就是需要找到核函数来衡量两条输入的最短依存路径之间的相似性。我们使用余弦核函数以及编辑距离核函数来衡量输入样本相似性。

假设和是两条输入路径的one-hot表达形式，其中，，则代表着词表的数量（注意到这里面的依存关系nsubj，conj等都作为一个单词使用）。则代表着该路径包含在词表位置上的单词，所以是一串0,1类型的向量。可以定义余弦核函数如下：

余弦核函数的值是一个归一化的值，它的区间为，可以看出如果输入和完全相同时候，此时余弦核函数的值为1，如果两个输入完全不相同，此时余弦核函数的取值为0.

从直观的角度来看，余弦核函数并没有将输入路径的顺序考虑进去，它只是考虑两个输入样本共同的单词部分。因此在此基础上，为了获取样本的结构信息，本文采用编辑距离核函数来衡量样本之间的相似性。

编辑距离(edit distance)又称Levenshtein距离，它衡量的是一个字符串通过插入，删除，替换，从而转化成一个新的字符串所需要的最小的操作步数。最初的编辑距离是为了衡量两个字符串之间转化需要的步数，本文进行了拓展，将字符串拓展到了字符串列表，也就是计算从一个字符串链表转化成另一个字符串列表的最少步数。

比如对于下面所示的和两条输入最短依存路径，第一条样本可以通过两次插入操作转化为第二条样本，因此可以定义输入和的编辑距离，考虑到这个数值不是一个归一化数值，我们将这个数值除以输入和中最大的长度。

*PROTEIN\_1*- nsubj - interacts - prep with - 插入(*PROTEIN\_0)* - 插入(conj and) -*PROTEIN\_2*

*PROTEIN\_1*- nsubj - interacts - prep with - *PROTEIN\_0* - conj and - *PROTEIN\_2*

综合考虑，本文定义编辑距离核函数如下，其中是一个衰减因子，它能控制编辑距离对于分类器最终的贡献，在实际应用中，这个参数是可以调整的，一般来说这个参数在1-3之间。

另一方面，为了充分利用余弦核函数以及编辑距离核函数，最后，本文提出了一个混合的核函数。该核函数的定义如下：

注意到这里面的也是一个超参数，它能控制余弦核函数对于最终分类器的贡献，通过实验发现，编辑距离的核函数效果相对于余弦核函数来说会稍微好点，因此，在这种情况下，可以适当的降低的数值，从而突出编辑距离核函数的作用，具体的参数设置以及实验结果会在下一个章节详细说明。

2.4：实验结果

本章节重点分析了基于规则和机器学习关系抽取系统在标准数据库上的实验结果，并同现有的关系抽取系统做了比对。在实验对比部分，本文都是采用准确率，召回率以及F-值三个指标作为衡量标准。

2.4.1：基于规则的生物实体关系抽取系统的实验结果

本文选择LLL-challenge 2005作为标准数据集来测试提出的基于规则的关系抽取系统，因为LLL-challenge 2005任务的主题就是通过学习规则来抽取蛋白质或者基因之间的作用关系，此外LLL-challenge规模相对较小，便于手动分析系统的性能高。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 准确率 | 召回率 | F-值 |
| Autopat | 0.58 | 0.63 | 0.60 |
| Rule-based | 0.56 | 0.30 | 0.39 |
| RelEx | **0.82** | **0.72** | **0.77** |
| **Our system** | 0.77 | 0.70 | 0.73 |

表3：本文提出的方法在LLL-challenge任务上同现有基于规则系统的性能比较。

表3列举了一些基于规则的现有生物实体关系抽取系统。从表3中可以看出RelEx系统在LLL-challenge数据集上取得了最好的效果，我们的系统取得了第二好的效果。本文提出的系统跟RelEx最大的区别在于我们尝试用最简单的规则（两条）来实现关系抽取，而RelEx在制定规则的基础上，通过对数据观察，定义了一系列的额外的过滤规则，这些规则一方面可以用来抽取规则没有匹配到的生物关系作用对，另一方面可以过滤错误的抽取到的实体关系对，这些措施能够确保RelEx获得很好的准确率和召回率。

另外，我们通过错误分析发现，导致系统出错的原因包含以下几个方面。

1：大量的错误是由依存分析器错误分析造成。比如依存分析器无法生成实体跟动词之间的依存关系（比如主谓关系），这导致两个实体之间最短依存路径上就不会包括动词，这样，实体之间的关系就没法检测出来。

2：由实体的指代产生错误。在句子中，经常出现which，that这类指代词，这些词一般指代一个生物实体，由于系统只抽取两个实体之间的最短依存路径，所以此时指示代词和实体之间的关系会被忽略。

3：有些关系并不是通过动词或者介词相互作用。

尽管从语法角度来设计规则已经可以大大减少规则的数量，但是，考虑到语言的多样性，有限的规则还是很难归纳所有的情况，本文在设计基于规则的关系抽取系统时，重点关注的还是关系表达最普遍的情况，对于其他的比较少见的表达方式，并没有将这些表达考虑在内。  
 2.4.2：基于机器学习的关系抽取系统的实验结果

考虑到LLL-challenge数据集规模较小，在基于机器学习的关系抽取实验中，为了更好的体现机器学习算法的归纳数据能力，本文选择数据集规模较大的Aimed和BioInfer数据集。Aimed数据集由Bunescu等人标记，大约包含200篇生物文献的摘要，总共1900左右的句子。它是蛋白质-蛋白质关系抽取领域标准的数据集。而BioInfer数据集是由Turku实验室开发标注，它的规模相对较大。在本文中，一对存在关系的蛋白质被标记成正样本，否则，为负样本。Aimed和BioInfer的统计数据如表4所示：

|  |  |  |
| --- | --- | --- |
| 数据集 | 正样本 | 负样本 |
| BioInfer | 2512 | 7010 |
| Aimed | 995 | 4812 |

表4：Aimed和BioInfer数据集统计

|  |  |  |
| --- | --- | --- |
| 数据集 | 余弦核 | 编辑距离核 |
| Aimed | 55.3 | 59.8 |
| BioInfer | 62.3 | 70.0 |

表5：单个余弦核函数以及编辑距离核函数在Aimed和BioInfer的表现（以F-值衡量）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Aimed | | | BioInfer | | |
| 核函数 | 准确率 | 召回率 | F-值 | 准确率 | 召回率 | F-值 |
| Shallow kernel | 60.9 | 57.2 | 59.0 | - | - | - |
| Graph kernel | 52.9 | 61.8 | 56.4 | - | - | - |
| edit-distance | 58.4 | 61.2 | 59.6 | - | - | - |
| Hybrid kernel | 55.0 | **68.8** | 60.8 | 65.7 | **71.1** | 68.1 |
| Multiple kernel | - | - | **64.2** | - | - | 67.6 |
| Walk weight kernel | 61.4 | 53.3 | 56.6 | 61.8 | 54.2 | 57.6 |
| 提出的混合核函数 | **85.3** | 49.5 | 62.5 | **86.0** | 60.0 | 70.3 |

表6：本文提出的hybrid核函数跟其他基于核函数的方法比较

从表5可以看出，编辑距离核函数的性能要远好于余弦核函数，在Aimed数据集，采用编辑距离核函数可以提出4.5左右的F-值，而在BioInfer数据集上，采用编辑距离核函数，F-值可以进一步提高7.7。基于这样的观察，我们可以适当的减少公式AA中的数值，从而突出编辑距离核函数的作用，在本实验中设置为0.001，而编辑距离中的超参数则设置为0.5。

从表6可以看出，跟单个核函数相比，提出的混合核函数进一步提高了系统的性能。尤其是在Aimed数据集上，跟单个编辑距离核函数相比，F-值提高了2.7，而在BioInfer数据集上，性能也有稍微的提高。跟其他基于核函数的方法相比，本文提出的系统在Aimed数据集上取得了第二好的效果，仅次于Miwa等人提出的Multiple核函数。而在BioInfer数据集上，我们的系统取得了最优的结果，并进一步将F-值提高了2.2左右。这说明本文提出的混合核函数是有效的。此外，通过跟其他核函数方法对比，可以看出本文提出的混合核函数取得最高的准确率（Aimed数据集上85.3，BioInfer数据集上86.0），这样的特点适用于某些对准确率有较高要求的系统。

总结来说1）单个编辑距离核函数的性能要远远优于余弦核函数；2）多个核函数的整合有利于提高系统的系能。

2.5：大规模生物实体关系网络的构建

在本章节中，我们将2.3.1中构建的基于规则的关系抽取系统应用到了实际的生物文献中，通过抽取实体关系，构建关系作用图，从而找出了同乳腺癌（breast cancer）相关基因。

2.5.1：生物文献摘要的获取以及预处理

本文以breast cancer gene作为关键词，利用网络爬虫的方式从pubmed数据库上抓取了约47000条左右的生物文献摘要，利用Stanford parser，我们对这些摘要做了标准化的文本处理，包括：分句，分词，词性标注，句法分析。另一方面，为了识别文本中出现的生物实体名词，利用爬虫的方式，我们抓取了OMIM数据库上的基因和蛋白质的名称，构建了生物实体词典，然后通过词典比对的方式来识别生物实体。

2.5.2：生物实体关系网络的构建和分析

将2.3.1中的系统应用到处理好的文本，可以抽取一系列的生物实体关系。如果将每个实体都当作一个节点，每对实体之间的关系当作边，可以构建一个生物实体的相互作用图。本文将从度中心性（degree centrality）和中间中心性（betweenness centrality）两个角度来分析构建好的。

度中间性主要是衡量节点的中心性趋势。度中心性越高则意味着节点的中心性越强，节点的地位相对越高。图8是利用Cytoscape以度中间性为指标绘出的基因网络图。

图8中总共包括845个节点，每个圆形的节点都是一个基因实体，存在相互作用关系的节点之间通过直线连接。形状的大小代表着节点度中间性数值的大小，另一方面，不同的度中心性数值通过不同的颜色来表示。度中间性数值最小是绿色，最大情况是红色。颜色可以认为是从浅色到深色的变化。其中图8右上角部分是为了清晰展示分析效果，对于图片的局部进行了放大处理。



图8：度中心性



图9： 度中心性分布

图9是对度中心性分布的一个概括，可以看出大部分的节点分布在10以下，事实上，对于最后的分析而言，如果一个基因实体节点的度中心性小于或者等于2，那么可以认为这个节点是处于整个基因网络的边缘部分，对于网络的中心性没有太大的贡献。本文重点关注的是节点中度中间性超过10或者说在整个度中间性中排名靠前的节点，这些节点对于分析很有意义。选取度中心性排名前10的基因，表7展示了这些基因

|  |  |  |
| --- | --- | --- |
| 基因名称 | 度中心性 | 是否于breast cancer相关 |
| BRCA1 | 149 | YES |
| HER2 | 143 | YES |
| MB | 133 | NO |
| EGFR | 73 | YES |
| BRCA2 | 62 | YES |
| ERBB2 | 62 | YES |
| D1 | 56 | YES |
| MYC | 53 | YES |
| PR | 46 | YES |
| VEGF | 43 | YES |

表7： 度中心性排名前10基因

对于上述通过度中间性找到的前10个基因。其中BRCA1和BRCA2的全称是breast cancer 1 基因和breast cancer 2 基因，它们是公认的会对乳腺癌产生影响的基因。HER2中文名称为人类表皮生产因子接收器，英文全称为Human epidermal growth factor receptor-2，目前已经证实了该基因在乳腺癌中为致癌基因，而且已经有实验室研制出了合适的抗体来对抗HER2基因。MB全称为 MDA-MB-231细胞。Zhou[11]  等人已经说明了该细胞与乳腺癌之间的关系，但是这里面MDA-MB-231是作为一个细胞存在，所以我们认为这是系统的一个误判。EGFR中文名称是表皮生长因子受体，英文全称为：epidermal growth factor receptor，该基因已经证实跟乳腺癌有关联。ERBB2中文名称为人类表皮生长因子受体2，Omenn Gs[13]已经说明该基因会在乳腺癌的细胞水平上表现出来。D1其实应该是cyclin D1,中文名称为细胞周期素D，该基因同样与乳腺癌相关。剩下的MYC中文名称为原癌基因，VEGF中文名称为血管生长因子，它们均同乳腺癌的产生或者表达存在关系。

与度中间性不同的是，中间中心性衡量的是一个节点对于资源控制能力。可以将中间中心性看作一个交通路口的人流量，如果说一个交通路口的人流量大，那么则认为该交通路口在整个城市中处于一个重要的地位。同理，反映到一个基因网络图中，当我们认为一个节点处于一个关键位置时候，那么其他两个节点的路径要经过该节点。

图11是以中间中心性为指标，同样每个节点代表一个基因实体，节点形状的大小代表着中间中心性的大小，节点的颜色深浅同样代表着节点中间中心性的大小。利用Cytoscape可以对数据进行可视化，如图10所示，其中右上角是对部分图片的区域放大效果。



图10：中间中心性（中间中心性）



图11： 中间中心性分布

对于中间中心性，同理，需要找到最合适的能够控制整个网络资源的节点，所以中间中心性数值越大越好，这里面可以选择图11中的中间中心性在区间(0.2,0.3)的节点，忽略其他的节点，找出排名最靠前的十个基因，这些基因通过表格8展示。

表8： 中间中心性排名前10的基因

|  |  |  |
| --- | --- | --- |
| 基因名称 | 中间中心性 | 是否于breast cancer 相关 |
| HOX | 1.0000 | YES |
| SYK | 1.0000 | YES |
| MB | 0.2370 | NO |
| BRCA1 | 0.2218 | YES |
| HER2 | 0.2015 | YES |
| EGFR | 0.0729 | YES |
| MYC | 0.0683 | YES |
| PR | 0.0678 | YES |
| ERBB2 | 0.0576 | YES |
| D1 | 0.0544 | YES |

中间中心性排名前10的基因同度中间性结果相比多出了一个HOX和SYK基因，事实上，除了这两个基因，其他基因同度中间性中的结果很相似，只是位置排名上稍微有点区别。对于HOX基因,中文同源异形盒，SYK中文为增值细胞核抗原，参考一些英文文献，同样可以得出这两个基因跟乳腺癌息息相关。对于其他的基因，已经在度中间性分析部分已经进行了叙述。都证明了这些基因于乳腺癌之间确实是存在关系的。从另一方面来说，选择度中间性和中间中心性作为度量的标准是合理的，因为，通过这两个度量，都可以得到相似的结果。基于以上的观察，为了充分利用度中心性以及中间中心性两种度量标准，本文将两种度量方式加权处理，并找出了加权数值排名最靠前的15个基因，如表9所示：

表9：加权排名最靠前的15个基因

|  |  |  |
| --- | --- | --- |
| 基因名称 | 加权数值 | 是否与breast cancer相关 |
| BRCA1 | 0.6109 | YES |
| HER2 | 0.5806 | YES |
| MB | 0.5648 | NO |
| HOX | 0.5067 | YES |
| SYK | 0.5067 | YES |
| EGFR | 0.2814 | YES |
| ERBB2 | 0.2368 | YES |
| BRCA2 | 0.2298 | YES |
| D1 | 0.2151 | YES |
| MYC | 0.2120 | YES |
| PR | 0.1882 | YES |
| VEGF | 0.1665 | YES |
| EGF | 0.1617 | YES |
| PRO | 0.1497 | NO |
| SRC | 0.1410 | YES |

15个结果中存在两个错误，其中MB已经进行过说明，是由于MB属于细胞水平，不属于基因层次。而后面的PRO则是由于系统在进行命名实体识别时候出现的错误，一方面以PRO命名的基因跟乳腺癌之间不存在任何的联系，更多时候，在本系统的语料库中，PRO不是一个基因实体。除了这两个基因，大部分的基因都是度中间性和中间中心性的组合，这进一步说明了加权分析的合理性。

总结来说，在本章节，通过网络分析的方式，一方面，可以找出与乳腺癌相关的基因，另一方面，通过计算每个基因节点在网络中的地位，可以衡量一个基因实体的重要性，这也从侧面反映出当前对于乳腺癌研究的一些热点基因。最主要的，这些实体关系的抽取可以为搜索引擎提供技术支持，可以进一步用于构建知识库，从而方便用户检索，因此存在一定的实际意义。

2.6：本章小结

本章介绍了基于规则和基于机器学习的两个实体关系抽取系统。基于规则的系统主要采用语法分析加规则的方式，而基于机器学习的系统主要利用最短依存路径来构造核函数。从实验结果来看，基于规则的系统在LLL-challenge数据集上取得了较为理性的效果，而提出的混合核函数模型在BioInfer数据集上则取得了很好的效果，从而说明本文提出的混合核函数是有效的。另一方面，通过构建大规模的关系作用图，利用图的分析方法，本文找出了与疾病相关的一些基因，这一分析方法对于现实中的搜索引擎具有一定的借鉴意义。

**基于神经网络的生物实体关系抽取技术研究**

3.1：本章引论

第二章节详细描述了基于规则和机器学习的关系抽取算法。从中可以看出，一方面，这两种方法都需要手动构建特征，而且不同的特征对于结果的影响很大，所以算法的移植性较差；另一方面，在基于机器学习算法的关系抽取系统中，大部分系统只能利用标记好的数据进行有监督的学习，但是考虑到生物领域标记数据成本太高，标记数据规模一般不大，这会影响机器学习算法的性能。基于以上提到的两点局限性，在本章节，我们提出了一个利用神经网络来进行生物实体关系抽取的框架。提出的框架具有以下的特点：

1. 可以利用各种神经网络结构进行自动特征提取工作，避免手动特征提取；
2. 可以使用大量的无标签数据，利用神经网络来训练语言模型，从而充分抓住单词之间的语义信息；
3. 具有良好的性能，并在多个生物实体关系抽取的相关数据集上取得非常好的效果；
4. 具有很好的移植性，可以实现跨领域。

本章节的剩余部分安排如下，在3.2章节会重点介绍词向量的概念，在3.3章节，我们介绍了两种神经网络结构卷积神经网络（CNN）和循环神经网络（RNN），在3.4章节中，结合3.2中的词向量以及3.3节中的神经网络，我们提出了一个用于生物关系抽取的框架，3.5章节是我们在标准数据集上做的一些实验结果，最后在3.6章节，我们对基于神经网络的生物实体关系抽取技术做了一个简单的总结。

3.2：词向量（word embedding）

3.2.1：词向量的基本概念

在传统的机器学习算法中，当面对自然语言处理问题时候，一般都是将输入特征，比如词性，词根或者其他的语义信息用1维向量来表示，这种表达方式称之为one-hot表示。但是大多数基于神经网络的模型则是将每个特征映射成一个低维稠密向量（low-dimensional dense vector），这样的向量称为嵌入（embedding），其中将每个单词映射而成的单词向量称为词向量（word embedding）。图12用例子的形式展示两种特征表达方式的不同。

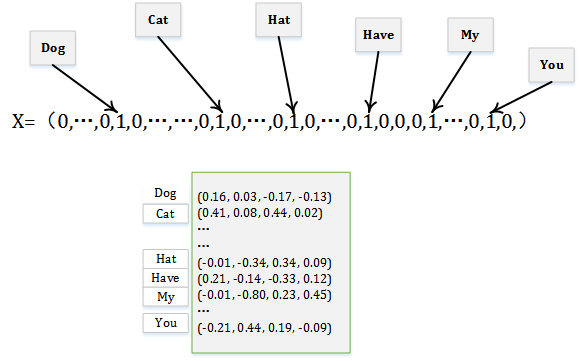


图12：图的上部分是one-hot表达方式，可以看出每个单词都是用1维向量表示，也就是用0,1代表这个特征是否存在，所以称这样的表达方式为one-hot表示；图的下半部分则是embedding的表达方式，每个单词都是用一个4维的稠密向量来表示。

通过上面的例子可以看出，两种表示最大的区别在于从稀疏的特征表示变成了低维稠密的特征表示。使用低维稠密向量来表达特征的一个最直观的好处在于降低计算量，比如常用的英文单词词表的大小接近100000，对于基于神经网络的方法而言，在稀疏表示中，这个节点数目意味着大矩阵的计算（这个矩阵的某一维数是100000），而且很高维的特征空间上，神经网络的性能不是特别好，相比而言，稠密表示由于是低维（一般50-300维）表示，因此具有计算优势。

稠密表示的最大优点在于其归纳能力：也就是如果我们相信某些特征之间存在联系，那么这些联系可以在稠密表示中表现出来。打个简单比方，假设存在单词“dog”和“cat”，可以看出在one-hot的表示形式中，是看不出dog和cat之间的关系，但是在稠密表示中，考虑到dog和cat两者之间语义上的相似性，可以通过算法（具体的算法会在后面的章节描述）使得训练好的dog和cat两个词向量之间的余弦距离足够大，从而体现语义上的相似性，值得注意的是，这种单词之间的语义性在生物关系抽取任务中具有重要地位，打个比方，如果系统需要判断一对生物实体之间的关系是“依赖”（rely on）还是“绑定”（bind），这需要系统知道单词bind和rely之间的关系，而词向量就可以抓住这些单词之间的关系。类似的对于单词的词性“VBD”（动词的过去分词形式）和“VB”（一般动词形式），如果认为这两个词性有相似的方面，那么在训练词性的向量时候，可以使得这两个向量在向量空间上是相近的。总结来说，稀疏表示丢弃了特征之间的联系，而稠密表示则可以通过特定的算法使得特征之间保持联系。另外，从风险均摊理论来说，one-hot表示相当于把所有鸡蛋放在一篮子里面，而稠密表达则是将风险均摊在每一维数上，因此，直观上来看，稠密表示的鲁棒性相对较好。一般来说，如果特征离散，特征维数较少，可以采用one-hot的表达形式，但是如果认为特征之间存在联系，而且特征维数较高，则可以考虑采用稠密表达。当然这只是一个直观的选择，但是具体到实际情况，这两种表达的选择仍然是一个开放问题。

3.2.2：词向量训练

前面提到，词向量作为稠密表示的一种，可以抓住词语之间的语义信息。所以问题的关键就在于如何训练这些词向量。这个章节将详细叙述如何获得词向量。

最简单获取词向量的方式是采用随机初始化。当标记数据足够的情况下，可以随机初始化词向量，然后将这个词向量作为网络中的一个参数，当训练模型的时候，可以微调（fine-tuning）这个参数，使得这个参数最终能够去拟合训练数据集。这时候得到的词向量可以认为是同任务相关的（task specific），也就是适用于当前任务的。

从前面章节的叙述可以看出，生物领域标记数据还是太过于稀少，因此采用随机初始化的方式会使得在训练模型的时候绝大部分的词向量得不到或者很少更新，这使得在小规模任务上，基于词向量的系统的性能会受到影响。一个比较折中的方法则是采用无监督的预训练（pre-trained）方式来获取词向量，然后利用训练数据进行微调。

无监督训练词向量算法核心的观点是：使得语义相似的单词具有相似的词向量。但是单词之间的相似性在大部分时候是很难定义和衡量的。当前无监督算法的假设都是基于Harris等人提出的观点：words are similar if they appear in similar contexts（具有相同上下文的单词具有相同的语义信息）。举个简单例子，对于句子“the cat sat on the mat”和句子“the dog sat one mat”，根据Harris等人的观点，那么单词“cat”和“dog”应该有类似的语义，因为这两个单词具有相同的上下文“the”和“ sat on the mat”。现有的无监督学习算法也都是通过上下文信息来预测当前单词，或者利用当前单词来预测上下文。图13展示了训练好的词向量聚类结果。



图13：词向量聚类结果，可以看出，语义相近的单词都被聚在了一起。比如右上角被红色方框圈起来的单词induce，response和inhabit，就具有相似的语义。

可以看出，无监督训练的最大好处在于可以充分利用大量的无标记的数据，这使得训练好的词向量可以提供有监督的训练数据中没有的信息。在理想情况下，训练好的词向量可以抓住单词之间的语义信息，而这语义信息可以认为是为训练数据提供了一个先验，这个额外的先验信息可能对于特定任务是有用的。

常见的无监督的词向量训练算法有：word2vec，Glove，Senna。这些算法都是基于神经网络算法，并采用随机梯度下降算法（SGD）优化参数。具体的词向量训练算法的描述已经超出了本文叙述的范围，在这里，本文只给出词向量优化的目标。

给定一个单词和它的上下文，其中每个单词都通过一个维数的随机初始化的稠密向量表示。在大多数无监督学习算法模型中，都是去最大化条件概率，也就是给定一个上下文，去预测单词（在整个词表空间上）的概率分布。这里面上下文的选择是多样性的，但是最多的还是选择目标单词（待预测单词）的左右单词作为上下文。

3.3：神经网络模型

在本章节，结合3.1介绍的词向量，我们重点描述了两种主流的神经网络结构：卷积神经网络（CNN）以及循环神经网络（RNN）。从结构上来说，CNN能够很好的抓住输入样本局部信息，而RNN则能捕捉远距离的依赖信息，因此，两者之间各有特点，具体的结构特点会在下面章节详细描述，同时为了描述的方便，本文定义一个输入句子，其中为对应的每个输入单词的词向量，则为输入样本的单词个数，所以如果将所有的词向量连接在一起，就可以获得一个输入样本的矩阵表示。

3.3.1：卷积神经网络（CNN）

有时候我们希望对一个有序的集合（如由单词组成的句子序列）做出预测，最常见的情况，比如预测句子的情感倾向（积极，消极，中性），可以发现，大部分时候，一个句子中只有少量几个单词提供有用的信息，其他的单词基本不提供信息。例如对于句子“我很高兴”，词语“高兴”就已经提供了足够的信息来表明这个句子表达积极的情感。所以问题的关键在于如何选择这些信息量大的词语，本文主要利用CNN来自动提取这些有用的信息。

CNN已经在图像处理领域展现了非常优异的性能，包括图像分类，目标分割，目标识别等。应用到自然语言处理领域，CNN核心的思想则是对输入句子的每个单词窗口（-单词的窗口）应用非线性函数，这个非线性函数一般称为卷积核（在图像处理中被称为滤波器），这个操作则被称之为卷积（convolution）操作。这样通过应用滤波器，可以将一个-单词的窗口数据转化成一个维的向量，在标准的CNN结构中，卷积操作后一般接上池化（pooling）操作，最常见的池化操作包括平均池化和最大池化，在自然语言处理领域，最大池化被广泛应用，因为通过选择由卷积操作生成的特征中的最大值，相当于获得了信息量最大的特征，这个跟本章开头的想法是吻合的。图14对如何进行卷积和池化操作做了一个简单的示例。



图14：一个简单的CNN进行卷积池化的示例。在本例子中，输入句子为“the cat sat on the mat”，其中单词窗口选择为3，因此存在四个窗口输入，如图最左边所示。假设每个单词都用一个2维的向量表示，因此3个单词的窗口都可以用一个6维的向量表示（如图中的绿色部分），卷积操作相当于对每个单词窗口应用卷积核，这里面将选择为3，因此，通过卷积操作，一个6维的向量会转化成一个3维的向量，在本例中，4个窗口数据会转化成4个3维的向量，也就是图中的灰色部分。最后的池化操作则是选择灰色部分列最大数值，最终会生成一个3维的向量（图中的蓝色部分），这个向量就是CNN提取出来的特征。

下面将从公式的角度来阐述CNN中卷积和池化的过程。对于输入样本，如果窗口为，那么对于长度为的输入样本则存在单词窗口，对于第个单词窗口输入，它是由个窗口中的词向量拼接而成。如果定义卷积核为，则卷积结果可以定义如下：

其中是一个非线性函数，像常见的sigmoid，tanh函数等，而则是一个偏置。

从上面的叙述可以看出，每个输入的单词窗口都会生成一个卷积结果，所以单词窗口则对应个个卷积输出，最终的卷积输出是一个矩阵。

从图14的例子可以看出，最大池化操作是对卷积矩阵的每列取最大值的结果，因此，池化输出可以定义如下

其中是矩阵的第列。

由最大池化生成的向量可以认为是对一个句子全局的表示（sentence representation），针对不同的任务，这个表示可以有不同的用法，比如在分类任务中，这个特征表示可以作为某个分类器的特征，在聚类任务中，这个表示向量则可以作为聚类算法的特征输入。在基于神经网络的模型中，根据反向传播算法，由于误差的存在，CNN中的参数会随着任务的不同而进行微调，这会使得抽取的特征C更加的拟合当前的任务，而且可以看出，这种特征提取是自动进行的。

3.3.2：循环神经网络（RNN）

从3.3.1的描述可以看出，CNN主要是对一个单词窗口做非线性变换，这使得CNN模型在一定程度只能看到当前单词窗口内的信息，而窗口外的信息无法看到，所以如果输入的样本中两个单词不包含在一个窗口中，那么这两个单词之间的关系CNN是很难学习到的，因此可以认为CNN学习到的特征是局部的。

本章节介绍的循环神经网络（RNN）结构则可以较好地解决单词之间的长距离依赖问题，由于RNN可以循环的处理输入序列的每个单词，因此RNN对于处理序列数据具有天然的优势，它能将任意长的句子编码成一个固定大小的特征向量。图15展示了RNN的基本结构。



图15：RNN的基本结构

在本文中，我们使用来表示一个输入的向量列表，RNN将一个有序的向量列表和一个初始化的隐含状态向量（是RNN的隐含节点数）作为输入，对于每个输入，RNN都会对应的产生一个输出向量，一般是由通过变化生成，这个可以进一步的作为特征，去完成特定的任务。比如对于词性标注任务，后可以接一个Softmax分类器，用来判断当前输入的词性标签。进一步通过观察图15中的RNN结构可以看出，对于每一个输入，需要利用上一个时刻的隐含状态，然后重新生成新的隐含状态，所以是跟相关的，类似的是跟状态相关的，通过这样链式推断下去，可以看出是跟前面的所有个状态都相关的。所以RNN不需要任何的马尔科夫假设，在每个输入上的输出都是跟当前输入的历史有关的，这使得RNN可以充分利用历史信息。

为了从数学的角度来描述RNN，本文定义一个循环函数和一个映射函数。其中循环函数以隐含状态和输入作为输入，然后生成新的隐含状态，而映射函数则用来将映射到输出向量，具体的公式可以描述如下：

可以看出对于循环函数的不同定义会导致RNN的结构有所不同，而RNN也经历了从最简单的RNN（simple-RNN）到LSTM（Long-Short Term Memory）的改变。

Simple-RNN的结构相对简单，循环函数只是一个对输入和进行非线性变换的函数。可以定义simple-RNN的计算方式如下：

其中和都是simple-RNN的参数，它们分别用来对输入数据和前一个隐含状态做线性变换。可以看出，虽然simple-RNN的结构相对简单，但是simple-RNN在序列标注，语言模型等自然语言处理任务上取得了非常好的效果。

尽管simple-RNN性能比较优异，但是由于梯度弥散（vanishing）问题的存在，导致很难训练一个好的simple-RNN模型，正是由于这样的原因，在simple-RNN的基础上，出现了许多RNN的变种，其中LSTM作为变种的一种，被认为可以有效缓解梯度弥散问题。

为了缓解梯度弥散问题，LSTM提出了一个“memory cell”的结构，在这个memory cell中，梯度信息可以有效的保存下来。而memory cell中的信息的访问，都需要通过“门”结构来控制的。LSTM中包含三种门结构：输入门（input gate），输出门（output gate），以及遗忘门（forgot gate）。其中输入门可以控制新的输入中有多少可以加入到memory cell中，输出门则控制可以输出多少的信息，而遗忘门则控制着旧的memory cell加入到新的memory cell中的信息。标准的LSTM的定义如下：

表示向量之间的点乘，代表memory cell，为sigmoid函数，所以，，，的数值都在*[0,1]*之间，因此可以认为是控制的是输入，遗忘，输出信息的比例。可以认为是输入到LSTM中的信息，那么则代表着输入到memory cell中的信息，则代表旧的memory cell可以保留的信息，所以新的 是和这两部分信息之和。同理由于控制着输出信息比例，所以则代表着最终输出的信息部分。总结来说，LSTM中门的结构本质上可以认为是对信息流的一种控制，而memory cell结构通过存储梯度信息，则有效地避免了梯度弥散问题，这使得LSTM在很多自然语言处理任务包括：序列标注，语言模型，依存分析等任务上取得了非常优异的成果。

3.4：基于神经网络的生物实体关系抽取框架

C:\Users\Administrator\AppData\Local\Microsoft\Windows\INetCache\Content.Word\神经网络框架.tif

图16：基于神经网络的生物实体关系抽取

在前面几个章节，我们详细介绍了词向量（word embedding），卷积神经网络（CNN），循环神经网络（RNN）的基本概念。在此基础上，本文定义了用于生物实体关系抽取的框架。如图16所示：

从图中可以看出，这个框架跟我们最初在公式1对于生物实体关系抽取的定义是吻合的。跟传统的基于机器学习算法的关系抽取系统相比，区别在于这里面在词向量基础上，利用CNN或者RNN神经网络进行了自动特征抽取，从而避免了手动提取特征的过程。在下面的章节，我们详细介绍了如何利用这个框架进行生物实体关系抽取。

3.5：基于CNN的生物实体关系抽取系统

在本章节中，我们首先介绍了使用的数据集以及数据集的预处理方法；然后我们介绍了如何利用CNN进行实体关系抽取，最后，我们介绍了实验结果。

3.5.1：数据集

|  |  |  |
| --- | --- | --- |
| 类别 | 描述 | 示例 |
| Advice | 可以同时使用两种药物的推荐或者建议 | Interaction may be expected, and *UROXATRAL* should not be used in combination with other *alpha-blockers* |
| Effect | 药物之间相互作用影响的描述 | *Methionine* may protect against the ototoxic effects of *gentamicin* |
| Mechanism | 药物之间存在药理作用 | *Grepafloxacin*, like other quinolones, may inhibit the metabolism of *caffeine* and *theobromine* |
| Int | 药物之间关系的声明 | The interaction of *omeprazole* and *ketoconazole* has been established |
| Other | 不存在关系 | - |

表10：DDI中5中关系的描述以及例子，其中药物实体都用斜体表示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 训练 | | | 测试 | | |
| DrugBank | Medline | 总计 | DrugBank | Medline | 总计 |
| 摘要数目 | 572 | 142 | 714 | 158 | 33 | 191 |
| 正样本 | 3788 | 232 | 4020 | 884 | 95 | 979 |
| 负样本 | 22118 | 1547 | 23665 | 4367 | 345 | 4712 |
| Advice | 818 | 8 | 828 | 214 | 7 | 221 |
| Effect | 1535 | 152 | 1687 | 298 | 62 | 360 |
| Mechanism | 1257 | 62 | 1319 | 278 | 24 | 302 |
| 经过规则过滤和预处理 | | | | | | |
| 正样本 | 3767 | 231 | 3998 | 884 | 92 | 976 |
| 负样本 | 14445 | 1179 | 15624 | 2819 | 243 | 3062 |
| Advice | 815 | 7 | 822 | 214 | 7 | 221 |
| Effect | 1517 | 172 | 1669 | 298 | 62 | 360 |
| Mechanism | 1257 | 62 | 1319 | 278 | 21 | 299 |
| Int | 178 | 10 | 188 | 94 | 2 | 96 |

表11：DDI数据集的数据统计，表格的上部分原始数据统计情况，表格的下半部分是经过预处理后的统计情况。其中正样本代表药物对之间存在关系，负样本则表明药物对之间不存在关系。

本章节主要使用两个任务上的数据集：Protein-protein interaction（PPI）关系抽取以及Drug-drug interaction（DDI）关系抽取数据集。其中PPI任务包括两个数据集Aimed和BioInfer，这两个数据集已经在2.4.2章节做了详细的介绍。这里面主要介绍DDI这个任务上的标准数据集：DDIExtraction 2013。DDI这个任务主要抽取的是药物之间的相互作用关系，该任务要做的是将一对药物实体分成5类：advice，effect，mechanism，int和other。在表格10中有对这5个类别详细的描述。可以看出，跟PPI只判断实体之间是否存在关系相比，DDI任务是一个细颗粒的多分类任务，因此相对难度加大。

DDIExtraction 2013这个数据集分成两个部分：DrugBank和MedLine，其中训练集包含714篇摘要，而测试集包含191篇摘要。表格11的上半部分对这个数据集做了详细的统计。通过统计可以发现，负正样本的比率约为5.88：1，考虑数据的不均衡分布可能会影响到算法性能，本文利用一些简单的规则，可以过滤掉一部分负样本，在下个章节，我们仔细阐述了数据的预处理方法。

3.5.2：数据的预处理

标准的数据预处理包括分句和分词，进一步的为了减少数据的稀疏性，本文同样利用2.3.2中的预处理方式，把一对实体名称用特殊符号代替。但是这种处理方式是可能产生一些噪声数据的，比如说如果抽取的两个实体实际上是同一个实体，那么这个实体对是不可能存在关系的，这样的数据可以称为噪声数据。噪声数据会对系统产生以下的影响：1）会导致数据的分布不均衡，影响系统性能；2）会增加训练时间。基于以上的原因，本文定义了两条规则来过滤掉噪声数据。

规则1：如果一个实体对表示的是同一个实体，那这个实体对为噪声数据。比如对于：**Anesthetics**, general: exaggeration of the hypotension induced by general **anesthetics**.实体对“**Anesthetics**”和“**anesthetics**”不会存在关系，应该这两者表示同一个实体。

规则2：如果几个实体处于并列结构，那么这些实体对为噪声数据。比如对于：**Potentiation**, **barbiturates**, **antihistamines**, **narcotics**, **hypotensive** agents or phenothiazines should be used with caution.这些实体都是处于并列结构，因此这些实体对不会存在关系。

表格11展示了应用规则到DDIExtraction 2013数据集的结果，可以看出，定义的两条规则是非常有效的。在训练数据集上，负样本数目从23665变成了15624，减少了约34.0%，但是正样本数目只有22个被错误的过滤掉。在测试数据集上，负样本数目减少了约35.0%的比率，并且只有3个正样本被错误过滤。从后面的实验部分可以看出，这种过滤方式不仅可以缩短训练时间，还可以有效地提高系统性能。

3.5.2：实验设置

在本章节，我们详细叙述了如何利用CNN进行生物实体关系抽取。