单位代码：10359　　密　级：公 开
学　　号：2014110527　　分类号：TP391

合肥工業大學

Hefei University of Technology

硕士学位论文

MASTER'S DISSERTATION

论文题目：基于机器学习和神经网络的关系抽取研究

学位类别：学术硕士

专业名称：信号与信息处理

作者姓名：华磊

导师姓名：任福继 全昌勤

完成时间：2017 年 4 月

合　肥　工　业　大　学

学术硕士学位论文

基于机器学习和神经网络的关系抽取研究

作者姓名：华磊

指导教师：任福继　教授

全昌勤　教授

学科专业：信号与信息处理

研究方向：自然语言处理

2017 年 4 月

A Dissertation Submitted for the Degree of Master

Research of relation extraction based on machine learning and neural network

By

Hua Lei

Hefei University of Technology

Hefei, Anhui, P.R.China

April, 2017

致　谢

时间总是能很轻易地改变一个人，一转眼，在工大已经待了七年，把我从一个毛头小伙变成了一个胡子邋遢的大叔。然而，这就是青春，是时间的力量，让人敬畏。

犹记得刚读研的时候，迷茫的有点无所适从，那时候混迹在学校的各个角落，但生活的重心不会总是迷茫，很高兴在这段时间里遇到了一些志同道合的朋友，是他们改变我生活的节奏，让我重新审视自己。不管是足球，学习，篮球，游戏还是其他的一些活动，他们总是会带给我不一样的感受，带给我不一样的激情。我们会聚在一起讨论一些久远的东西，作为一群对代码有执着追求的人，我们总是很执着地认为技术可以改变世界。于是总是忘不了熬夜写代码的日子，辛苦却快乐着。大概就是这些很琐碎的东西填补了我的生活，让我觉得生活不再单调。感谢我的那些朋友们，在研究生平淡如水的日子里，他们给了我太多的惊喜和快乐。

研二和研三时候开始写论文，于是有了不一样的生活。科研之路，似乎漫长无尽头。大概只有跟老师的每次交谈中才能依稀知道点自己要干嘛。尽管外面吵吵闹闹，但还是努力让自己静下心来，让自己忙起来，整个研二，是自己最专注的一年，也是这样的经历让我明白，时间固然让人敬畏，但也能激发一个人，去散发属于自己的光芒。

只是转眼，研究生生活就要结束。我真的很感谢那些出现在我生命中的人，亲人，老师，室友，朋友，学长学姐，尽管生活会经历浮浮沉沉，尽管日子还是这样的单调和匆忙，尽管生命会出现这样或者那样的重逢和离别，但你们，才是我不断前行的动力。

作者：华磊

2017 年 2 月 28 日

摘　要

自动生物实体关系抽取技术对于构建生物领域知识库，知识图谱，提高搜索引擎检索效率具有重要意义。通过总结现有工作，本文从三个方面进行自动实体关系抽取。

基于规则的关系抽取。在句子句法分析和依存分析的基础上，通过观察动词和介词在生物实体关系中的重要作用，本文首先构建了一个相互作用词词表，然后制定了两条通用的策略来判断一对实体之间是否存在关系。同其他基于规则的关系抽取系统相比，本文提出的基于规则的系统在 LLL-challenge 数据集上取得了第二好的效果。

基于核函数的关系抽取。在句子依存分析基础上，可以发现，实体对在依存分析中的最短路径一般包含足够的信息去判断这对实体是否存在关系。基于这样的观察，本文提出了一种编辑距离和余弦核函数相结合的混合核函数，提出的混合核函数能充分抓住数据的结构信息以及单词层面的信息。实验结果表明，同其他基于核函数的系统相比，本文提出的方法在 BioInfer 标准数据集上取得了最优的结果。

基于神经网络的关系抽取。为了避免特征工程，缓解生物关系抽取领域数据稀少问题，在卷积神经网络(CNN)和循环神经网络(RNN)两种结构基础上，本文提出了 multi-CNN 以及单通道 RNN 两个模型。实验结果表明，提出的模型在 DDIExtarction，Aimed 以及 BioInfer 数据集上都取得了最优结果。

关键词：关系抽取；网络分析；神经网络

ABSTRACT

Automatic biological entity relation extraction technology is of great significance to construct the knowledge database, knowledge graph and further improve the efficiency of search engine retrieval in the field of biology. By summarizing the existing work, this paper introduce the entity relation extraction technology from three aspects as follow：

Rule-based relation extraction. On the basis of sentence syntax and dependency analysis, by observing the important role of verbs and prepositions in biological entities relation extraction, this paper first constructs an interactive word list, and then develops two general strategies to judge whether there is a relationship between a pair of entities. Compared with other rule-based relational extraction systems, the method proposed in this paper achieves the second best performance on the LLL-challenge task.

Kernel-based relation extraction. Based on sentence dependency analysis, we can observe that the shortest dependency path between two entities generally contains enough information to judge whether there is a relationship between a pair of entity, on the basis of this observation, in this paper, we propose a kernel function that combines the edit distance kernel and cosine kernel, the proposed hybrid kernel function can fully grasp the data structure and word level information. The experimental results show that compared with other kernel function-based systems, the proposed method achieves the best results on the BioInfer standard data set.

Neural network-based relation extraction. To avoid data sparseness and feature engineering problems in traditional relation extraction systems, by integrating convolutional neural network (CNN) and recurrent neural network (RNN), in this paper, we propose two models for biological relation extraction: multi-CNN and single-channel RNN. The experimental results show that the proposed models achieve the best results on DDIExtraction, Aimed and BioInfer data sets.

KEYWORDS: relation extraction; network analysis; neural networks

目录

第一章　绪论.....8
1.1：课题研究背景以及意义.....8
1.2.　生物实体的关系抽取.....9
1.3.　生物实体关系抽取工作的研究现状.....10
1.3.1. 基于实体共现的方法.....10
1.3.2. 基于语法规则匹配的方式.....11
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.....15
1.4.　基于神经网络的生物实体关系抽取技术.....18
1.5.　生物实体关系抽取面临的一些挑战.....18
1.6.　本文的研究重点以及后续章节安排.....19
第二章　基于规则和机器学习的生物实体关系抽取系统.....20
2.1.　本章引论.....20
2.2.　语法分析以及特征构建.....20
2.3.　基于规则和基于机器学习的关系抽取系统.....22
2.3.1. 基于规则的关系抽取系统.....22
2.3.2. 基于机器学习的关系抽取系统.....23
2.4.　实验结果.....26
2.4.1. 基于规则的生物实体关系抽取系统的实验结果.....26
2.4.2. 基于机器学习的关系抽取系统的实验结果.....27
2.5.　大规模生物实体关系网络的构建.....29
2.5.1. 生物文献摘要的获取以及预处理.....29
2.5.2. 生物实体关系网络的构建和分析.....29
2.6.　本章小结.....34
第三章　基于神经网络的生物实体关系抽取技术研究.....35
3.1.　本章引论.....35
3.2.　词向量.....35
3.2.1. 稠密表示的基本概念.....35
3.2.2. 词向量训练.....37
3.3.　神经网络模型.....38
3.3.1. 卷积神经网络（CNN）.....39
3.3.2. 循环神经网络（RNN）.....40

3.4. 基于神经网络的生物实体关系抽取框架.....43
3.5. 基于 CNN 的生物实体关系抽取系统.....43
3.5.1. 标准数据集.....43
3.5.2. 数据的预处理.....45
3.5.3. 实验设置.....46
3.5.4. 模型训练.....47
3.5.5. DDI 实验结果.....47
3.5.6. DDI 实验对比.....48
3.5.7. PPI 实验结果.....50
3.5.8. PPI 实验结果对比.....50
3.5.9. 基于 CNN 的关系抽取系统错误分析和总结.....51
3.6. 基于 RNN 的关系抽取系统.....52
3.6.1. RNN 在 DDI 任务上的实验结果.....52
3.6.2. RNN 模型分析和总结.....53
3.7. 本章小结.....53
第四章 总结和展望.....55
参考文献.....57
攻读硕士学位期间的学术活动及成果情况.....62

插图清单

图 1.1 基于关键词“breast cancer”的 PubMed 的检索结果 8
图 1.2 利用“论元结构”以及规则来进行生物实体关系抽取 12
图 1.3 句子依赖分析结果.....13
图 1.4 实体对与介词连接的句法分析结果.....14
图 1.5 卷积树核计算.....16
图 2.1 利用 Stanford parser 对句子句法分析的结果 21
图 2.2 利用 Stanford parser 对句子进行单词层面（左图）和短语层面（右图）的依存分析结果.....22
图 2.3 利用 Stanford parser 对句子进行单词层面分析的结果 24
图 2.4 度中心性.....30
图 2.5 度中心性分布.....30
图 2.6 中间中心性.....32
图 2.7 中间中心性分布.....32
图 3.1 词向量和 one-hot 表示 36
图 3.2 词向量聚类结果.....38
图 3.3 一个简单的 CNN 进行卷积池化的示例 39
图 3. 4 RNN 的基本结构 41
图 3.5 基于神经网络的生物实体关系抽取.....43

表格清单

表 1.1 生物实体之间典型的关系.....9
表 1.2 根据实体的位置构建的 unigram，bigram，tri-gram 特征.....15
表 2.1 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.....26
表 2.2 标准数据集数据统计.....27
表 2.3 单个余弦核函数以及编辑距离核函数在 Aimed 和 BioInfer 上表现.28
表 2.4 本文提出的 hybrid 核函数跟其他基于核函数的方法比较.....28
表 2.5 度中心性排名前 10 基因.....31
表 2.6 中间中心性排名前 10 的基因.....32
表 2.7 加权排名最靠前的 15 个基因.....33
表 3. 1 DDI 中 5 种关系的描述以及例子.....44
表 3. 2 DDI 数据集的数据统计.....44
表 3.3 不同版本的词向量的统计特性.....46
表 3.4 基于随机初始化向量，单通道向量以及多通道向量的 CNN 模型的结果.....47
表 3.5 四个系统所采用的特征.....48
表 3.6 本文提出的 multi-CNN 跟其他四个系统的比较.....48
表 3.7 经过数据预处理和没有经过预处理的 multi-CNN 的实验结果对比.49
表 3.8 交叉训练测试结果（F 值）。.....49
表 3. 9 Baseline，单通道以及 multi-CNN 在 Aimed 和 BioInfer 两个数据集上的实验结果.....50
表 3. 10 multi-CNN 跟其他 PPI 系统的性能对比（F 值）.....51
表 3. 11 baseline 模型，单通道 RNN 模型，以及 multi-CNN 实验结果.....52

第一章　绪论

1.1：课题研究背景以及意义

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。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。

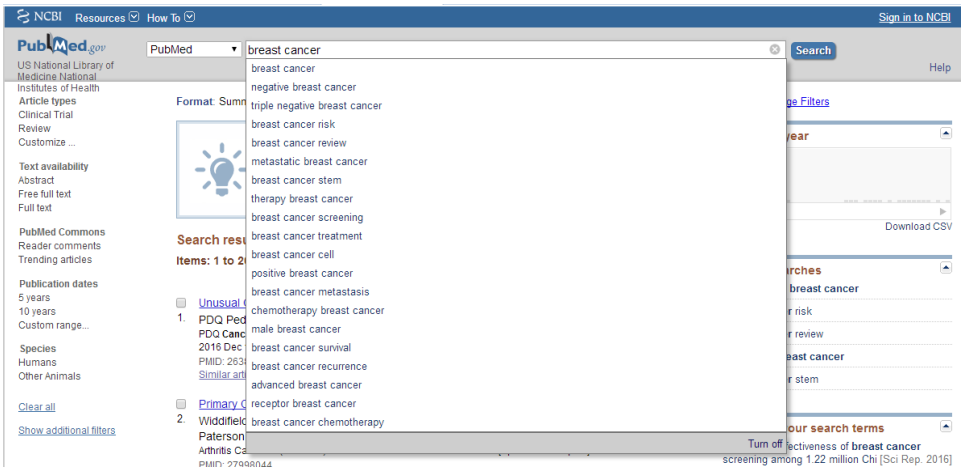

图 1.1 基于关键词“breast cancer”的 PubMed 的检索结果

Fig 1.1 Search results based on key words “breast cancer” on PubMed

21 世纪以来，随着生物技术的迅猛发展，相关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。因此，如何利用自然语言处理技术以及数据挖掘算法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 等几个主流生物文献的数据库。此外，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”（乳腺癌）为例，图 1.1 展示了 PubMed 的检索结果。从图中可以看出，在搜索框键入关键词“breast cancer”以后，搜索引擎能够联想出很多关于乳腺癌的主题，比如乳腺癌的治愈，乳腺癌的风险等等。但是想要获取更深层次的主题，比如同乳腺癌相关的基因，如果没有丰富的生物医学相关知识，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乳腺癌相关的基因实体，那么搜索引擎就可以从这些实体中选择合适的内容，为用户展示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。

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名称，比如基因（gene），蛋白质（protein），疾病（disease）或者药物（drug）等。如果想要进行生物实体关系的抽取，第一步需要识别出这些生物实体。已经有大量工作[1][2]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。定义中另一个概念则是关系，对于生物实体关系抽取系统而言，我们在本文中把关系定义为生物实体之间的相互作用，而这种相互作用可以体现在多个方面，从表 1.1 可以看出一些典型的实体关系。

表 1.1 生物实体之间典型的关系

Tab 1.1 The typical relations between two entities

实体类别	关键词(中文)	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，实体关系抽取问题本质上可以转化成分类问题[3]。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）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$s = w_1, w_2, \dots, e_1 \dots w_j \dots e_2 \dots w_n$，其中$e_1$和$e_2$是命名实体，可以定义映射函数$f$如公式(1.1)所示（这里面为了问题描述方便，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）：

$$f(T(s)) = \begin{cases} +1 & \text{如果}\mathrm{e}_1\text{和}\mathrm{e}_1\text{存在关系} \\ -1 & \text{其他} \end{cases} \tag{1.1}$$

$T(s)$可以看作是从一个包含实体对的句子中提取到的特征，而映射函数f则决定了两个实体之间是否存在关系，因此可以认为它是一个分类器。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f。如果存在大量的标记数据，那我们就可以训练f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f是一系列的规则或者模板，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，f的不同选择会使得系统有所差异，具体的f选择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[5][6][7]的方法（Co-occurrence based），基于规则[8][9][10]的方法（Rule-based）以及基于机器学习[11][12][13][14]的方法（Machine Learning based）。本章节重点介绍这三方面的现有工作。

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文档层面上去考虑一对实体对是否存在关系， Bunescu 等人[5]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）等估计方法都可以用来检测两个实体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 可以通过式（1.2）来计算。

$$\mathrm{PMI}(p_1,p_2)=\log\frac{P(p_1,p_2)}{P(p_1)P(p_2)}=log\frac{n_{12}}{n_1n_2} \quad (1.2)$$

其中$P(p_1,p_2)$代表着两个生物实体p_1,p_2同时出现的概率，它可以利用p_1,p_2同时出现在一起的次数n_{12}来近似估计。同理$P(p_1)$和$P(p_2)$作为两个实体各自出现的概率，同样也可以利用p_1和p_2各自出现的次数n_1和n_2来近似估计。Bunescu[5]指出 PMI 这种方法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，但是利用各种估计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。更深层次，比如句子层面的语法或者句法信息并没有被包括进去，这使得这类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，但是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[16]，生物领域内的 GENIA[17]等等。这些语法分析工具为分析文本提供了便利，这个章节描述的基于语法规则的方法大多基于这些语法分析工具的结果进行。

最早利用句法分析和规则来进行关系抽取的方法是由 Yakushiji 等人[15]提出的。他们首先利用大规模通用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（因为大多数实体之间的作用关系都是通过动词体现出来的，所以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 1.2 展示该系统的一个流程[15]。

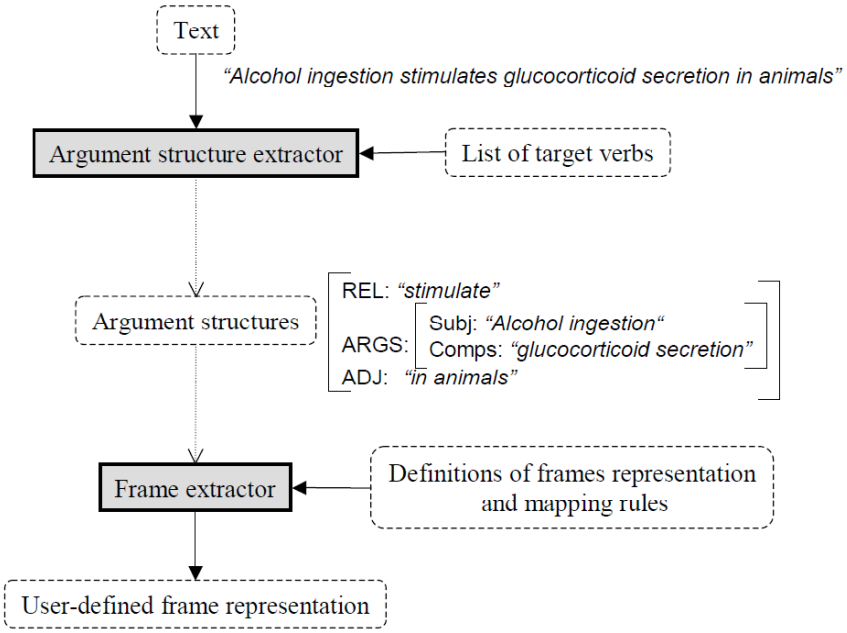

图 1.2 利用"论元结构"以及规则来进行生物实体关系抽取

Fig 1.2 Relation extraction with "argument structure" and rules

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：

1. P1 activates P2 （简单的主动表达）
2. P2 is activated by P1 （一般的被动表达）
3. P1 activating P2 （动名词形式的表达）
4. 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高级别的（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[8]提出了一种利用句法依存分析（dependency parser）来进行关系抽取的方法。他们首先利用 Stanford parser 对句子进行依存分析，通过生成的句法依存树（如图 1.3 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 0.82 的准确率。

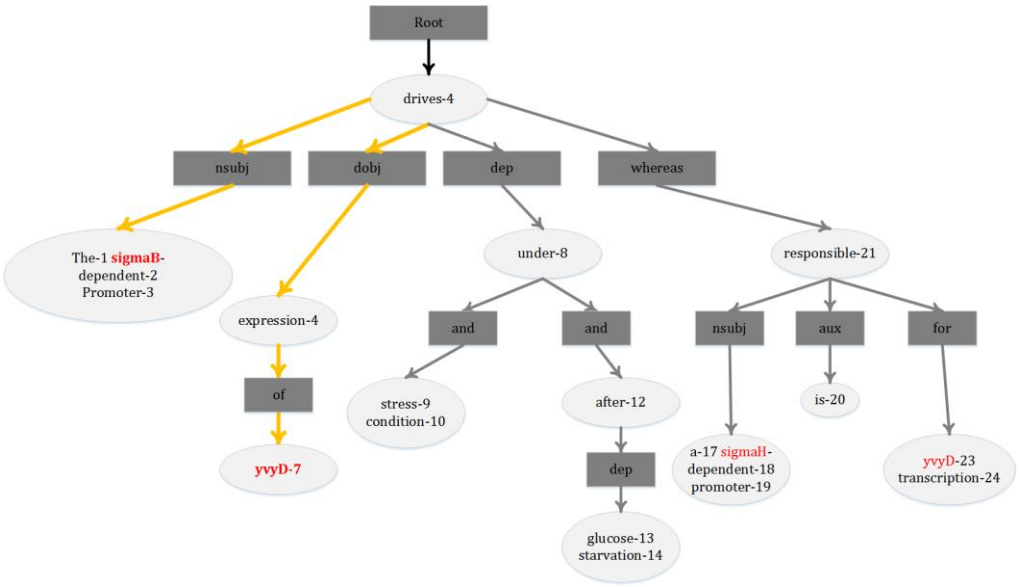

图 1.3 对句子“The sigmaB-dependent promoter drives expression of yvyD under stress conditions and after glucose starvation whereas a sigmaH-dependent promoter is responsible for yvyD transcription”句法依存分析的结果。图中所有的生物实体都用红色进行了标记，椭圆中每个单词后面的数字代表着一个单词在句子中的位置，深灰的矩形方框代表着依存关系，每个椭圆形则代表着一个短语块，句法分析是在短语块的基础上进行的，其中黄色箭头部分代表着一对实体之间的依存关系路径

Fig 1.3 The dependency parser results for sentence “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”. All entities are labeled with red color, the number after word in each ellipse represents the word position in a sentence， while the deep gray rectangle represents the dependency relation between a pair of words, the dependency analysis is based on noun-phrase chunk, and the yellow arrows are the shortest dependency path between two entities.

Fundel 等人[8]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：

1. P1-relation-P2 （比如：P1 activates P2）；
2. Relation-of-P1-by-P2 （比如：activation of P1 by P2）；
3. Relation-between-P1-and-P2 （比如：Interaction between P1 and P2）。

根据这些关系类型，在句法依存分析的基础上，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 1.3 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 1.3 中的黄色箭头部分）。可以看出这条路径中包含关键作用动词 drives（生物中代表驱动的意思），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，*sigmaB*，drivers，以及 *yvyD* 三者之间是主谓宾的关系，由此可以判定 *sigmaB* 和 *yvyD* 之间存在相互作用关系。对于第二个和第三个通过介词作用的关系类型，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。第二、三类型考虑的是介词，因此主要抽取实体和介词之间的依存路径。如果检测到实体之间通过介词“of，by，to，for，in”连接到关键词，则认为这对实体存在关系。具体的例子可以参考图 1.4。

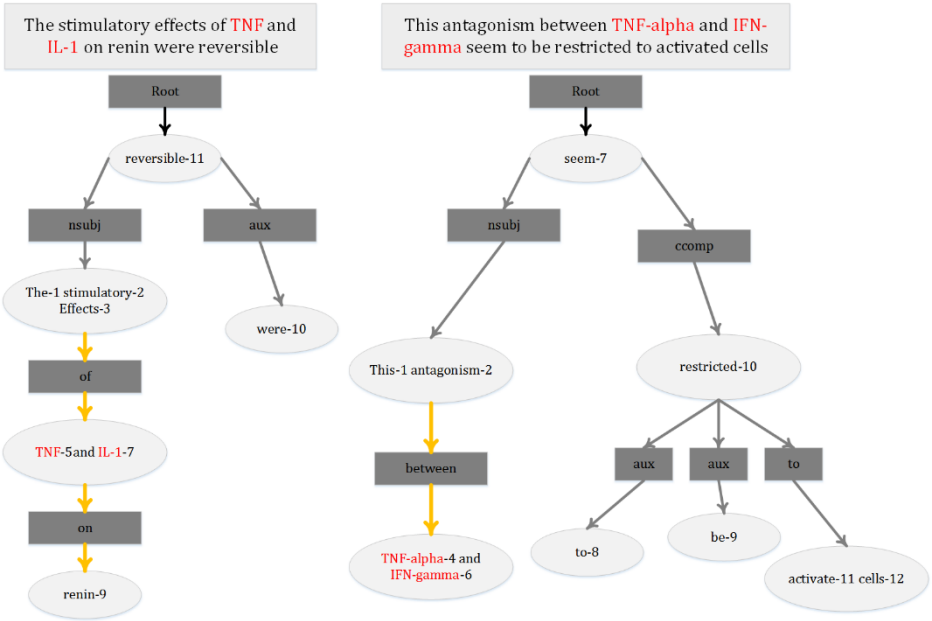

图 1.4 实体对与介词连接的句法分析结果。黄色部分代表着一对实体跟介词连接的路径。可以看出在左边图中实体 TNF 和 IL-1 通过介词 of 连接到关键词 stimulatory (刺激)，因此判定这两者之间存在关系；在右边图中，实体 TNF-alpha 和 IFN-gamma 通过介词 between 连接到关键词 antagonism（拮抗）

Fig 1.4 The dependency parser results when two entities are connected to preposition. The yellow parts represent the path when two entities are linked with prepositions. In the left of figure, entities TNF and IL-1 are linked with key word “stimulatory” through preposition “of”, while for right part of figure, entities TNF-alpha and IFN-gamma linked to key word “antagonism” through preposition “between”

跟基于“论元结构”的方法相比，Fundel 等人[8]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，语法分析，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[8]的方法中，规则已经被限制到只有 3 条，跟句子层面的模板相比，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 在标准测试数据集上已经取得超过 0.90 的准确率），但是考虑到生物领域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传播，会影响整个生物实体关系抽取系统的性能，这也是目前基于语法规则的方法面临的一个挑战。

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

大多数基于统计机器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$T(s)$以及选择合适的分类器f。对于分类器f而言，支持向量机（SVM）[31]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优势，因此 SVM 也在生物实体关系抽取系统中被广泛应用，并且体现出良好性能。在基于 SVM 分类器的基础上，特征$T(s)$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，因此，核函数的选择对于 SVM 分类问题也至关重要。基于以上原因，本章节重点介绍了如何利用抽取到的特征$T(s)$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$K(x,y)$来衡量输入样本x和y之间的相似性，其中K可以认为是某一种核函数。下面重点描述了在关系抽取任务中一些常用的核函数。

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。Sætre 等人[18]提出可以根据两个生物实体的位置关系，将 n-gram 特征分成三个方面：即位于第一个实体前面、两个实体之间和第二个实体后面的特征集合。例如对于句子 Activation of P2 by P1 has been confirmed，其中 P2 和 P1 是两个生物实体，则可以构造如表 1.2 所示的 unigram（一元语法）、bigram（二元语法）、tri-gram（三元语法）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(比如两个 one-hot 向量的点乘)实现自动关系抽取。

表 1.2 根据实体的位置构建的 unigram，bigram，tri-gram 特征

Tab 1.2 The unigram, bigram and tri-gram features based on entities position

特征类型	unigram 特征	bigram 特征	tri-gram 特征

前向	Activation, of	Activation of	-
中间	by	-	-
后向	has, been, confirmed	has been, been confirmed	has been confirmed

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，会对系统的性能产生一定影响。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。

Choi 等人[19]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子树的相似性，从而判断两个输入样本的相似性。图 1.5 展示了如何计算两棵句法树之间相似性。

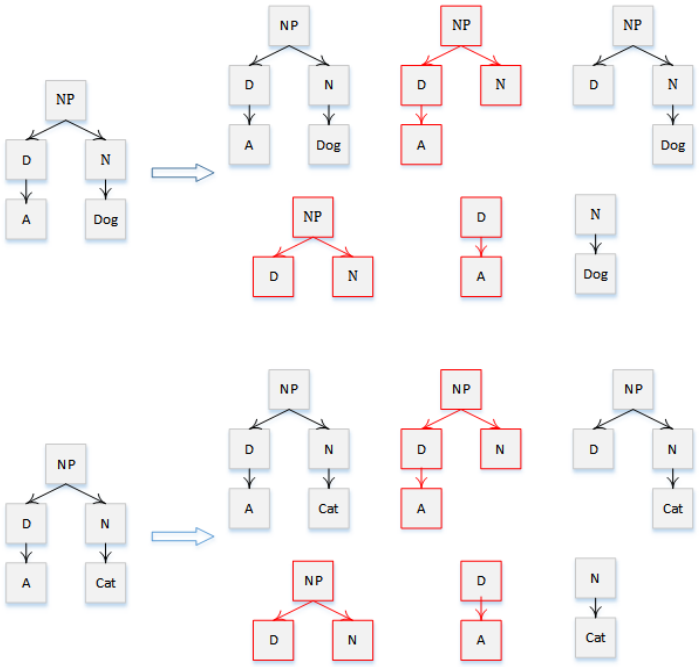

图 1.5 箭头右边代表着一个句法树的所有子树结构，图的上部分是短语“A Dog”的所有子树结构，图的下部分是“A Cat”的所有子树结构，两者之间相同的子树结构都通过红色进行标记，因此可以计算两个树的卷积核$K(x,y)=3$

Fig 1.5 The right parts of the arrows represent the sub-tree structures of a complete parser tree. The upper parts are the sub-trees for phrase “A dog”, while the bottom of the figure represents the sub-tree structures for phrase “A cat”. The common sub-trees are labeled with red color, and thus the kernel value $K(x,y)$ for two input samples are 3

对于每对蛋白质，Choi 抽取包含这对蛋白质的子树，同时使用一些规则对

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[20]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$A \in \mathbb{R}^{|V|\times|V|}$来表示这个有向图，其中$V$代表有向图节点，另外用$L$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$A(i,j)$代表顶点i和j之间的权重，而$L(k,i)=1$则代表着节点i拥有类别k。他们首先计算出所有的顶点相连接的情况，如公式（1.3）所示。

$$(\mathrm{I}-\mathrm{A})^{-1} = I + A + A^2 + \cdots = \sum_{k=0}^{\infty} A^k \tag{1.3}$$

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公式（1.4）表示。

$$W = (I-A)^{-1} - I \tag{1.4}$$

根据上面最终的邻接矩阵以及L，可以定义基于图的核函数为公式（1.5）所示，其中$G^x = LWL^T$，x和y则是两个输入的样本。

$$K(x,y) = \sum_{i=1}^{|L|}\sum_{j=1}^{|L|} G_{i,j}^x\, G_{i,j}^y \tag{1.5}$$

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，可以计算出两个输入样本之间的相似性。同样的，Airola 等人[20]也利用 SVM 作为基本分类器，实现实体关系抽取。

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 核函数可以获得单词层面的信息，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[21][22]充分利用 bag-of-word (bow)核、Subset tree (ST)核以及 graph 核，并且结合依存（dependency）分析和深度依存（deep dependency）分析，提出了以下的混合核函数（公式（1.6））。

$$K(x,y)=\sum_{i\in\{bow,ST,graph\}}\ \sum_{j\in\{dependency,\ deep\}}\frac{K_{ij}(x,y)}{\sqrt{K_{ij}(x,x)K_{ij}(y,y)}} \quad (1.6)$$

可以看出 Miwa 等人提出的核函数是一种归一化的核，因为这些核函数存在一个归一化因子$K_{ij}(x,x)K_{ij}(y,y)$。Li 等人[23]则提出了一种加权的混合核函数，他们定义混合函数公式（1.7）所示：

$$K(x,y)=K_a(x,y)+\alpha K_b(x,y) \quad (1.7)$$

其中K_a和K_b代表着不同的核函数，注意到这里面的α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，不管是在词层面还是语法层面。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，这已经在很多工作中得到了证实[21][22] [23]，目前最先进的基于机器学习的生物实体关系抽取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[26]和语音识别[27]领域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[28]、问答系统[29]、句子建模[30]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）[24]或者循环神经网络（RNN）[25]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[23]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，也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括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赖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系统的性能；
3. 大多数基于机器学习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. 单词的语义信息（semantic information）大多被忽略，虽然判断一对实体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后续章节安排

本文主要从三个方面来实现生物实体关系的自动抽取：1）结合句子依存分析和特定规则来实现关系抽取；2）提出一个新的混合核函数来进行关系抽取；3）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，在句法分析的基础上，我们定义了两条通用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 LibSvm 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同“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方法来分析实体跟实体以及实体跟疾病之间的关系。

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的概念以及如何训练词向量；其次我们介绍了在词向量的基础上，如何利用 CNN 或者 RNN 模型对输入进行编码。在此基础上，我们提出了两个模型：multi-CNN 和 RNN；最后我们详细地分析了提出的模型在标准数据集上的实验结果。

在第四章节，我们对已做的工作做了一个简单的概括，并提出了一些未来可以改进的方向。

第二章　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，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s进行句法树分析以及依存分析；章节 2.3 详细说明了如何利用语法分析的结果来生成最短依存路径特征：$T(s)$，并在提取到的特征$T(s)$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，同现有的基于规则或机器学习的关系抽取系统做了比较，并详细分析了实验对比结果；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同乳腺癌相关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 是对本章内容的简单总结。

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。

在本文中，语法分析分成两个方面：句法分析和依存分析[1]。注意到句法分析重点关注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 2.1 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，一条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。

前面提到依存分析是分析单词之间的关系，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

[1] https://en.wikipedia.org/wiki/Dependency_grammar

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。因此，在实际的应用系统中，应该尽量保留动词的原始特性。本文只考虑抽取名词短语块，而动词短语不作为抽取的范围。本文主要是在句法分析的基础上进行名词短语的抽取，如图 2.1 所示，被红色方框圈住的部分就是抽取到的名词短语块。从图 2.1 可以看出，只要抽取 NP（必须限定该节点的深度为 2，其中 NP 是 noun phrase 的缩写，代表名词短语的意思）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列表：[“The gerE gene”, “are”, “transcribed”, “by”, “sigmaK gene”]，我们可以在这个输入上进行句法依存分析。

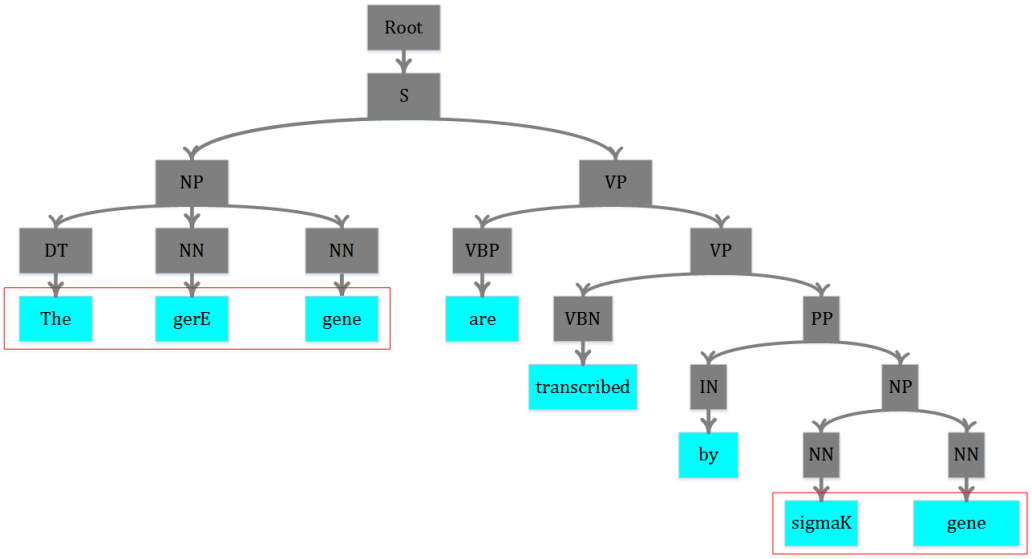

图 2.1 利用 Stanford parser 对句子“The gerE gene are transcribed by sigmaK gene”句法分析的结果。被红色方框圈住的部分是抽取的名词短语块。

Fig 2.1 Parser tree results for sentence “The gerE gene are transcribed by sigmaK gene” generated by Stanford parser. The noun-phrase chunks are circled by red rectangle.

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算法更有用，因此，本文从基于名词短语块和基于单词两个层面来分析一个输入句子。图 2.2 展示了两个版本的依存分析结果。在下个章节中，这些依存分析结果会用来构建关系抽取的特征。

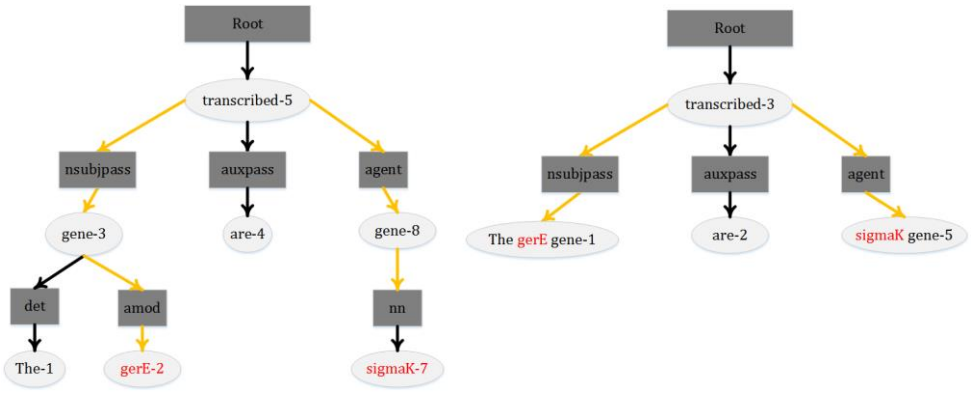

图 2.2 利用 Stanford parser 对句子“The gerE gene are transcribed by sigmaK gene”进行单词层面（左图）和短语层面（右图）的依存分析结果。椭圆形代表一个输入单词，矩形代表单词之间的依存关系。图中用红色标记的是生物实体，两个生物实体之间的依存路径用黄色箭头表示。

Fig 2.2 The dependency parser results generated by Stanford parser for sentence “The gerE are transcribed by sigmaK gene” on both word and noun-phrase levels. Each ellipse in figure represents an input word, while the rectangle is the dependency relationship between two words. In this figure, the entities are labeled with red color, the shortest dependency paths between two entities are all represented by yellow arrows.

2.3. 基于规则和基于机器学习的关系抽取系统

从公式（1.1）可以看出，关系抽取系统的关键在于特征$T(S)$的抽取以及映射函数f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。已经有大量研究表明[8][10][12][32]，实体之间最短依存路径可以提供充分的信息来判断一对实体之间的关系类型，这也是我们使用最短依存路径作为主要特征的依据。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

现有研究表明[8][15]，绝大部分生物实体之间的关系是通过动词相互作用的（从表 1.1 可以看出），还有很少一部分是通过介词来体现关系（章节 1.3.2 可以看出）。根据这两个特点，在依存分析和最短依存路径特征的基础上，本文制定了两个相应的策略来进行关系抽取。

策略 1：如果实体的最短依存路径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 2.2 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”（生物中的专业术语，表示转录的意思），同时包含依存关系“nsubjpass”，而 nsubjpass 就表明短语块 The *gerE* gene 和动词 transcribed 之间是被动的关系。所以这时候 *gerE*，transcribed，*sigmaK* 这三者之间是主谓宾的关系，此时则认为 *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 1.4 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 两个名词短语块是通过介词 of 连接的，而名词短语块 The stimulatory Effects 又包含相互作用类型的词语 stimulatory（在生物学上表示刺激的意思），此时可以认为实体 *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。这些词语从侧面反映了实体之间存在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 Fundel 等人[8]构建的相互作用动词表，在此基础上，利用 WordNet[2]，采用一个类似“滚雪球”的算法。然后通过人工选择，拓展了这个词表。算法步骤描述如下：

1. 将 Fundel 等人[8]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依存分析结

[2]在 WordNet 中，同义词会被放在一个集合中，具体的介绍可以参考 https://wordnet.princeton.edu/.

果。以图 2.3 为例，该例子中包含 *KaiC*，*SasA*，*KaiA*，*KaiB* 四个生物实体，因此，总共存在$C_4^2 = 6$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，可以将路径中出现的生物实体用特殊符号进行替换[12]。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全部标记为 *PROTEIN_0*，这样上面的 6 条最短依存路径可以写成下面的形式。

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之间不存在关系。在本文中，存在相互作用关系的样本被标记为正样本（+1），不存在关系的样本则被标记为负样本（-1）。

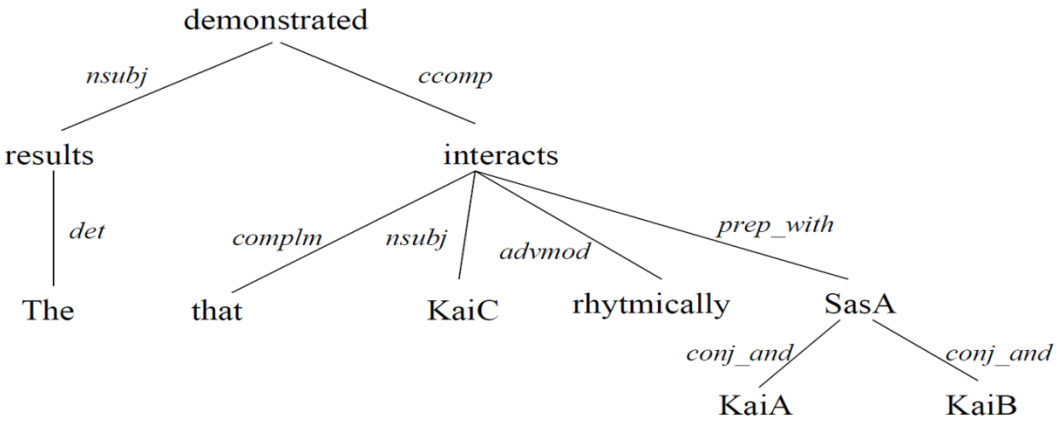

图 2.3 利用 Stanford parser 对句子“The results demonstrated that KaiC interacts rhythmically with KaiA, KaiB, and SasA”进行单词层面分析的结果。其中 KaiC，SasA，KaiA，KaiB 都为

生物实体，依存关系都用斜体表示。

Fig 2.3 Dependency parser results for sentence “The results demonstrated that KaiC interacts rhythmically with KaiA, KaiB, and SasA” by Stanford parser. KaiC, SasA, KaiA and KaiB are gene entities, all dependency relations are italicized.

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x和y是两条输入最短依存路径的 one-hot 表达形式，其中$x \in \mathbb{R}^n, y \in \mathbb{R}^n$，$n$则代表着词表的数量（注意到这里面的依存关系 nsubj、conj_and 等都作为一个单词使用）。$x_i = 1$则代表着该路径包含在词表i位置上的单词，所以x是一串 0,1 类型的向量。可以定义余弦核函数K_{cos}如下（公式 2.1）：

$$K_{cos}(x, y) = \frac{x \cdot y}{||x||\,||y||} \tag{2.1}$$

余弦核函数的值是一个归一化的值，它的区间为[0,1]。当输入x和y完全相同时候，此时余弦核函数的值为 1，如果两个样本输入完全不相同，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，为了抓住输入样本的结构信息，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，替换这三种操作，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最小步数，本文进行了拓展，将字符串拓展到了单词列表，也就是计算从一个单词列表转化成另一个单词列表需要的最少步数。

比如对于下面所示的x和y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x和y的编辑距离$edit(x, y) = 2$，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x和y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公式（2.2）所示，其中λ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，一般来说在 0.1-1 之间。

$$K_{edit}(x,y) = e^{-\lambda edit(x,y)} \quad (2.2)$$

另一方面，为了充分利用余弦核函数以及编辑距离核函数，本文提出了一个混合的核函数K_{hybrid}。该核函数的定义如公式（2.3）所示：

$$K_{hybrid}(x,y) = K_{edit}(x,y) + \alpha K_{cos}(x,y) \quad (2.3)$$

注意到这里面的α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，单个编辑距离核函数效果相对于余弦核函数来说会有很大的提高，因此，在这种情况下，可以适当地降低α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基于机器学习关系抽取系统在标准数据集上的实验结果，并同现有的关系抽取系统做了对比。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。另外，在基于机器学习的关系抽取系统实现方面，本文主要通过修改 LibSvm[3]的核函数（替换成本文提出的混合核函数）来实现。具体的代码实现方案可以在这里[4]发现。

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[5]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 2005 数据规模相对较小，便于手动分析系统的性能。

表 2.1 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。

Tab 2.1 Comparisons with other rule based systems on LLL-challenge task

方法	准确率	召回率	F-值
Autopat[10]	0.58	0.63	0.60
Rule-based[10]	0.56	0.30	0.39
RelEx[8]	**0.82**	**0.72**	**0.77**

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[4] https://github.com/coddinglxf/RelationExtractionMatrerial
[5] http://genome.jouy.inra.fr/texte/LLLchallenge/

Our system	0.77	0.70	0.73

表 2.1 列举了一些基于规则的现有生物实体关系抽取系统。从表 2.1 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，此外，在制定规则的时候，我们重点关注通过动词相互作用的关系。而 RelEx 在制定三条规则的基础上，通过对数据观察，定义了一系列额外的过滤规则。这些额外的规则一方面可以用来抽取一般规则没有匹配到的生物关系作用对，另一方面可以过滤抽取到的错误实体关系对，这些措施能够确保 RelEx 系统获得很好的准确率和召回率。

另外，我们通过错误分析发现，导致本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（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. 由实体的指代问题产生错误。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，重点关注的还是最常见的关系表达方式，对于其他比较少见的关系表达方式，我们并没有考虑在内。

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相对较大的 Aimed 和 BioInfer 数据集[6]。Aimed 数据集由 Bunescu 等人[33]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[7]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 2.2 所示：

表 2.2 Aimed 和 BioInfer 数据集统计

Tab 2.2 Statistics for Aimed and BioInfer datasets

[6] mars.cs.utu.fi/PPICorpora/
[7] http://bionlp.utu.fi/clinicalcorpus.html

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

从表 2.3 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高 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 2.3 中的α数值，从而突出编辑距离核函数的作用。在本实验中α设置为 0.001，而编辑距离核函数中的超参数λ则设置为 0.5，这些参数的设定都是通过验证集交叉测试得出的。

表 2.3 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）

Tab 2.3 Performances with single kernel function on Aimed and BioInfer datasets（measured with F-scores）

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

从表 2.4 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[22]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了最高的准确率（Aimed 数据集上 85.3，BioInfer 数据集上 86.0，这两个数值都远高于其他的系统），这样的特点适用于某些对准确率有较高要求的系统。

表 2.4 本文提出的 hybrid 核函数跟其他基于核函数的方法比较

Tab 2.4 Comparisons with other kernel based methods

	Aimed			BioInfer		
核函数	准确率	召回率	F-值	准确率	召回率	F-值
Shallow kernel[34]	60.9	57.2	59.0	-	-	-
Graph kernel[20]	52.9	61.8	56.4	-	-	-
Edit-distance[12]	58.4	61.2	59.6	-	-	-
Hybrid kernel[21]	55.0	**68.8**	60.8	65.7	**71.1**	68.1

Multiple kernel[22]	-	-	**64.2**	-	-	67.6
Walk weight kernel[35]	61.4	53.3	56.6	61.8	54.2	57.6
本文提出的混合核函数	**85.3**	49.5	62.5	**86.0**	60.0	70.3

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性能。

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 上抓取了约 47000 条的生物文献摘要，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，同样利用爬虫的方式，我们抓取了 OMIM[8]数据库上的基因和蛋白质的名称，构建了生物实体词典，通过词典比对的方式来识别原始文本中的生物实体。

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G。本文将从图分析中的度中心性（degree centrality）和中间中心性（betweenness centrality）两个角度来分析构建好的G。

度中心性主要衡量的是某个节点中心性趋势。如果有很多其他的节点连接到一个节点，那么这个节点所对应的度中心性越高，这也意味该节点在网络中的中心性越强，节点的地位相对越高。图 2.4 是利用 Cytoscape[9]以度中间性为指标绘出的网络图。

图 2.4 中总共包括 845 个节点，基因实体通过圆圈表示，每对存在关系的实体都通过实线连接。每个节点度中心性的大小可以通过圆圈的大小和颜色来衡量。圆圈尺寸越大，颜色越深，对应的度中心性数值越大。图 2.4 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。

[8] https://www.omim.org/
[9] http://www.cytoscape.org/

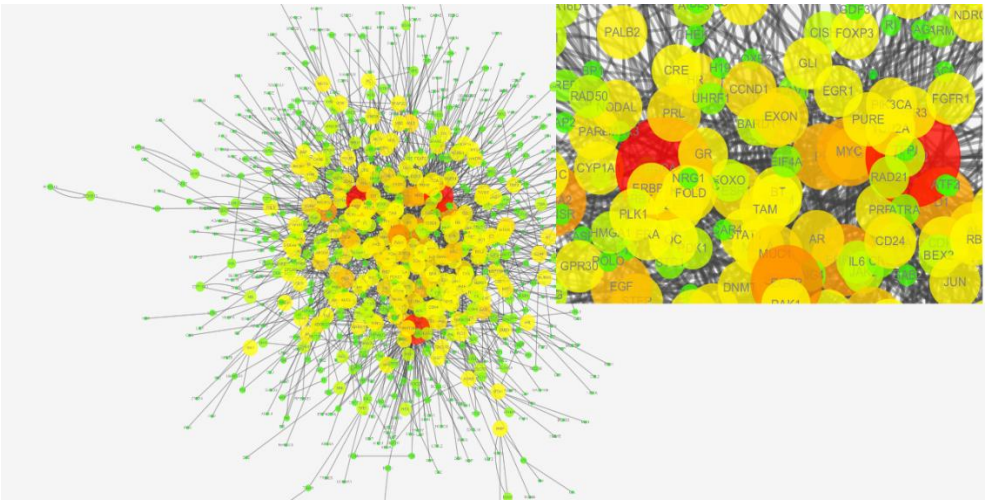

图 2.4 度中心性

Fig 2.4 Degree centrality

图 2.5 是对度中心性分布的一个概括，可以看出大部分节点的度中心性分布在 10 以下，事实上，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，那么这个节点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，选取度中心性排名前 10 的基因，表 2.5 展示了这些基因。

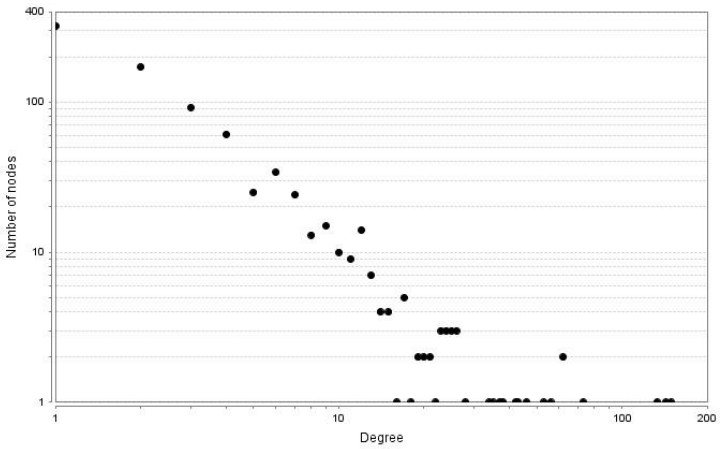

图 2.5 度中心性分布

Fig 2.5 The distribution of degree centrality

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。已有研究表明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，现有研究已经表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。

表 2.5 度中心性排名前 10 基因

Tab 2.5 Top 10 gene measured by degree centrality

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

与度中心性不同的是，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，反映到一个基因相互作用网络图中，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 2.6 是以中间中心性为指标，同样每个节点代表一个基因实体，节点形状的大小以及节点的颜色深浅都代表着中间中心性的大小。利用 Cytoscape 可以对数据进行可视化，如图 2.6 所示，其中右上角是对部分图片的区域放大效果。

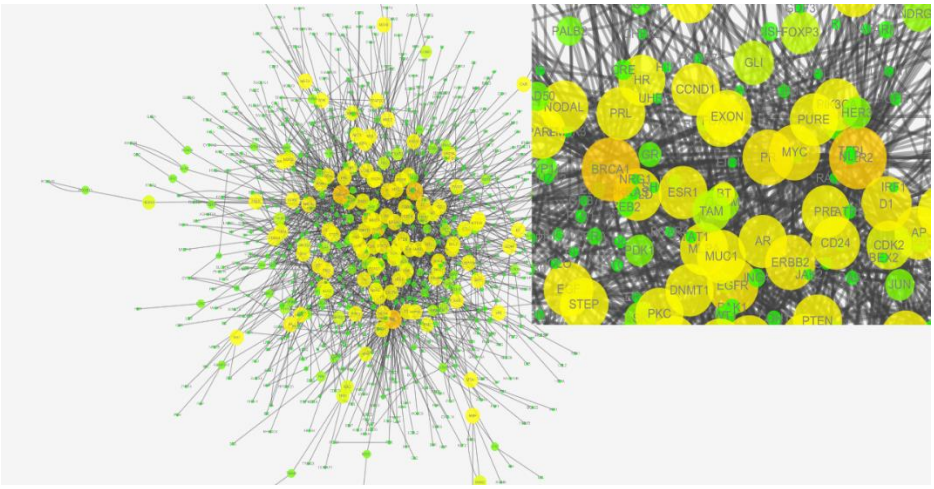

图 2.6 中间中心性

Fig 2.6 Betweenness centrality

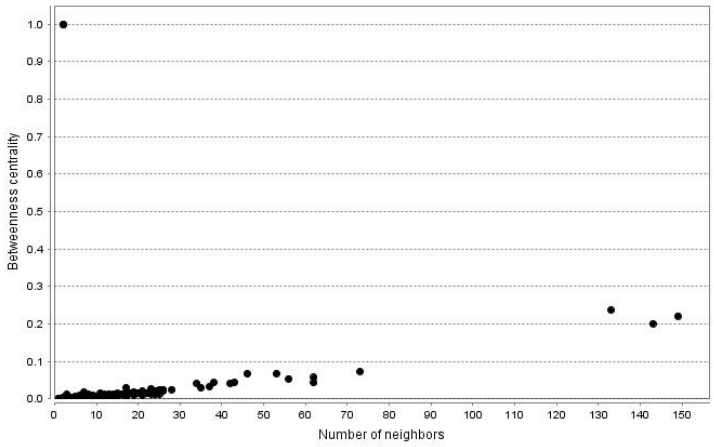

图 2.7 中间中心性分布

Fig 2.7 The distribution of betweenness centrality

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，找出排名最靠前的十个基因，这些基因通过表格 2.6 展示。

表 2.6 中间中心性排名前 10 的基因

Tab 2.6 Top 10 gene measured by betweenness centrality

基因名称	中间中心性	是否于 breast cancer 相关
HOX	0.8320	YES
SYK	0.6212	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，中文为同源异形盒，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进行了叙述，都证明了这些基因同乳腺癌之间确实是存在关系的。从另一方面来说，选择度中心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 2.7 所示：

表 2.7 加权排名最靠前的 15 个基因

Tab 2.7 Top 15 genes measured by weighted centrality

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，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，进一步用于构建知识库或者知识图谱，方便用户检索，提高检索效率和准确性。此外，这种网络分析方式可以适用于其他类型疾病，所以这种分析方式存在一定的普适性，具有一定的实际意义。

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和规则相结合的方式，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想的效果，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，本文找出了与乳腺癌相关的一些基因，这一分析方法对于现实中生物文献方面的搜索引擎具有一定的借鉴意义。

第三章　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$T(s)$，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，这在一定程度上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完成自动特征提取工作，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 节中的词向量以及 3.3 节中介绍的神经网络，我们提出了一个用于生物关系抽取的框架；3.5、3.6 章节是我们在标准数据集上做的一些实验结果和分析；最后在 3.7 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总结。

3.2. 词向量

3.2.1. 稠密表示的基本概念

在传统的机器学习算法中，当面对自然语言处理问题时，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 0，1 这种布尔类型来表示，这种表达方式称为 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 3.1 用例子的形式展示两种特征表达方式的不同。

通过图 3.1 中的例子可以看出，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），而且在高维的特征空间上，神经网络的性能不是很理想，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。

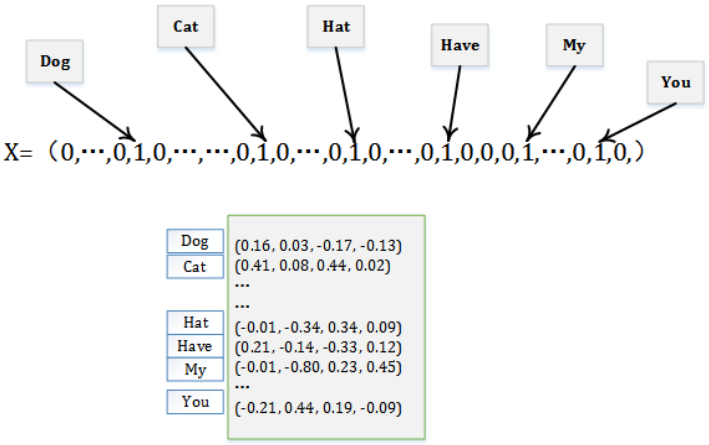

图 3.1 图的上部分是 one-hot 表达方式，可以看出每个单词都是用 1 维向量表示，也就是用 0,1 代表这个特征是否存在，所以称这样的表达方式为 one-hot 表示；图的下半部分则是 embedding 的表达方式，每个单词都是用一个 4 维的稠密向量来表示。

Fig 3.1 The upper part of the figure is the “one-hot” representation, each word is represented by a bit. The bottom of the figure is the “embedding” representation, each word is represented by a 4-dimension dense vector.

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， dog 和 cat 之间是不存在任何关系的，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（图 3.2 展示了训练好的词向量聚类结果，可以看出，语义相近的单词会被聚类在一起），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）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，那么在训练词性向量时，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，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[37]。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[36]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。

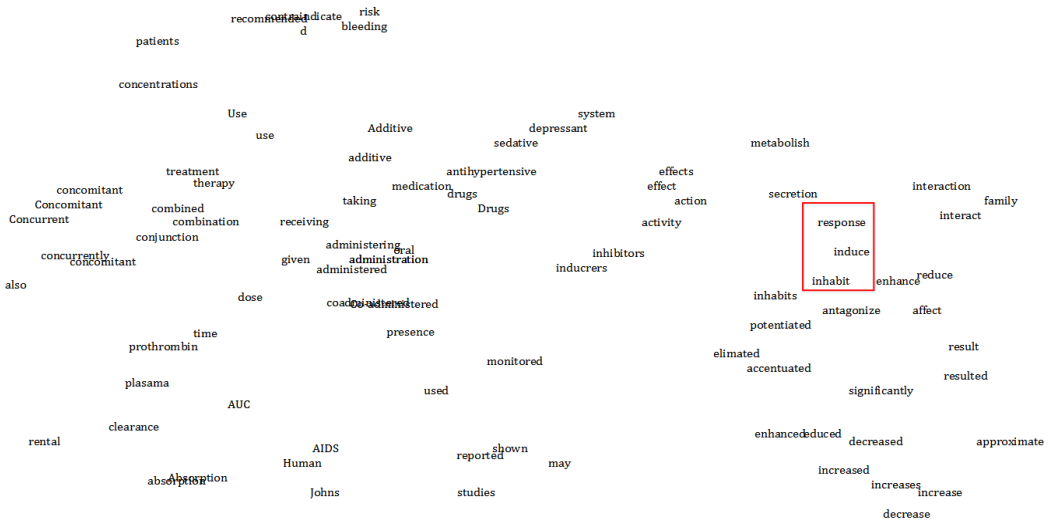

图 3.2 词向量聚类结果，可以看出，语义相近的单词都被聚在了一起。比如右上角被红色方框圈起来的单词 induce，response 和 inhabit，就具有相似的语义。

Fig 3.2 The cluster results of word embedding. Words with similar semantic are clustered together. The words “induce”, “response” and “inhabit” which are circled by red rectangle have the similar meanings.

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对于特定任务可能是有用的。

常见的无监督词向量训练算法有：Word2vec[38][39]，Glove[40]，Senna[41]。这些算法都基于神经网络算法，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w和它的上下文c，其中每个单词都通过一个d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$p(w|c)$，也就是给定一个上下文c，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，结合章节 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 能够很好地抓住输入样本的局部信息，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，同时为了描述方便，本文定义一个输入句子$X = \{x_1, x_2 \dots\dots, x_N\}$，其中$x_i \epsilon \mathbb{R}^d$为对应的每个输入单词的词向量，$N$则为输入样本的单词个数。

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（比如由单词组成的句子序列）做出预测，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或者提供很少信息。例如对于句子“我今天很高兴”，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[42][43]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（k-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k-单词的窗口数据转化成一个m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（mean pooling）和最大池化（max pooling），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，也就是相当于选择一句话中的关键词，这个跟本章开头的想法是吻合的。图 3.3 对如何进行卷积和池化操作做了一个简单的示例。

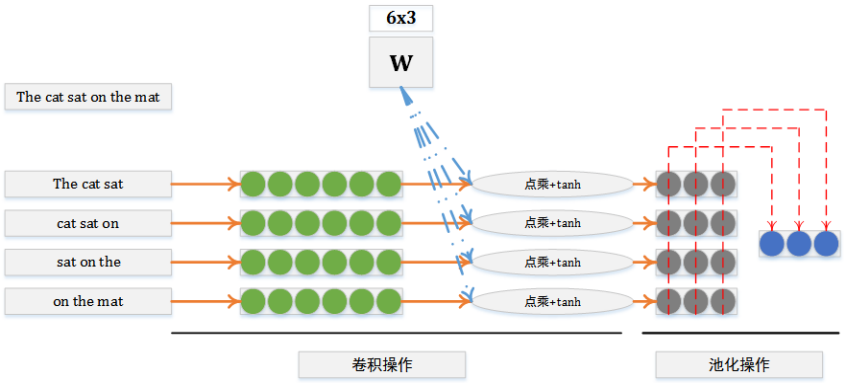

图 3.3 一个简单的 CNN 进行卷积池化的示例。在本例子中，输入句子为“the cat sat on the mat”，其中单词窗口k选择为 3，因此存在四个窗口输入，如图最左边所示。假设每个单词都用一个 2 维的向量表示，因此 3 个单词的窗口都可以用一个 6 维的向量表示（如图中的绿色部分），卷积操作相当于对每个单词窗口应用卷积核W，这里面将m选择为 3，因此，通过卷积操作，一个 6 维的向量会转化成一个 3 维的向量，在本例中，4 个窗口数据会转化成

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。

Fig 3.3 An example for a simple CNN. In this example, the input sentence is "the cat sat on the mat", the word window size k is 3, and thus input sentence will generate 4 word windows (as shown in the left part of the figure). If each word is represented with a 2-dimesion vector, a 3-word windows can be represented with a 6-dimsion dense vector (the green part in figure), the convolution operation is equal to apply filter kernel W to each word windows, in this example, we choose m as 3, and thus 4 word windows will be transformed into 4 3-dimension (gray part in figure) vectors. The last pooling takes the maximum value among the columns of the gray parts, and this will generate a 3-dimension vectors (blue part in figure), such vector is the extracted features by CNN.

下面将从数学公式的角度来阐述 CNN 中卷积和池化的过程。对于输入样本$X = \{x_1, x_2 \dots\dots, x_N\}$（在本文中，就是包含一对实体的句子），如果单词窗口长度为k，那么对于长度为N的输入样本则存在$N-k+1$单词窗口，对于第i个单词窗口输入$w_i \in \mathbb{R}^{k \cdot d}$，它是由窗口中$k$个单词的词向量拼接而成。如果定义卷积核为$W \in \mathbb{R}^{k \cdot d \times m}$，则卷积结果$p_i \in \mathbb{R}^m$可以定义如公式（3.1）所示：

$$p_i = f(w_i W + b) \qquad (3.1)$$

其中f是一个非线性函数，像常见的 sigmoid，tanh 函数等，而b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p_i，所以$N-k+1$单词窗口则对应个$N-k+1$个卷积输出，最终的卷积输出$\mathrm{P} \in \mathbb{R}^{(N-k+1) \times m}$是一个矩阵。

从图 3.3 的例子可以看出，最大池化操作是对卷积矩阵P每列取最大值的结果。因此，池化输出$\mathrm{C} \in \mathbb{R}^m$可以定义如公式 3.2 所示：

$$C = max(P[:, i]) \qquad i = 1,2,3 \dots\dots m \qquad 3.2$$

其中$\mathrm{P}[:, i]$是矩阵P的第i列。

由最大池化生成的向量C可以认为是对一个句子全局的表示（sentence level 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更加拟合当前任务，并且这种特征提取是自动进行的。

3.3.2. 循环神经网络（RNN）

从章节 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，所以如果输入的两个单词不包含在一个单词窗口中，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 可以循环地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 2.4 展示了 RNN 的基本结构。

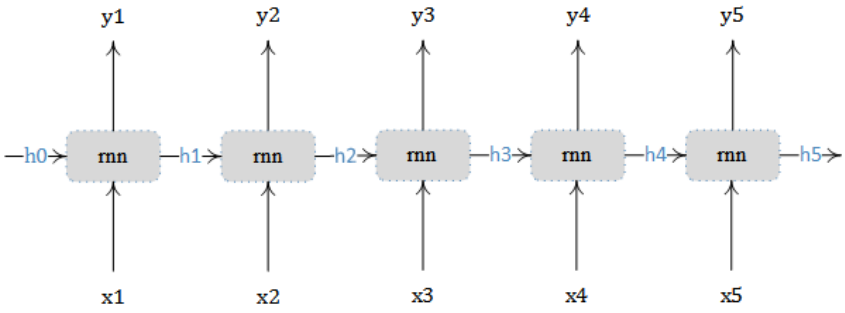

图 3. 4 RNN 的基本结构

Fig 3.4 Basic structure of RNN

在本文中，我们使用$X_{i:j}$来表示一个输入的向量列表$[x_i, \ldots\ldots x_j]$，RNN 接收一个有序的向量列表$X_{1:N}$和一个初始化的隐含状态向量$h_0 \in \mathbb{R}^n$（n是 RNN 的隐含节点数，h_0一般都初始化为 0 向量）作为输入，对于每个输入x_i，RNN 都会对应地产生一个输出向量y_i和一个隐含状态h_i，y_i一般是由h_i通过非线性变化生成，这个y_i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，y_i后可以接一个 Softmax 分类器，用来判断当前输入单词的词性标签。进一步通过观察图 3.4 中的 RNN 结构可以看出，对于每一个输入x_i，需要利用上一个时刻的隐含状态h_{i-1}，然后重新生成新的隐含状态h_i，所以h_i是跟h_{i-1}相关的，类似的h_{i-1}是跟h_{i-2}状态相关的，通过这样链式推断下去，可以看出h_i是跟前面的所有$i-1$个状态都相关的。所以 RNN 不需要任何的马尔科夫假设，在每个输入上的隐含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R和一个映射函数O。其中循环函数R以隐含状态h_i和输入x_{i+1}作为输入，然后生成新的隐含状态h_{i+1}，而映射函数O则用来将h_{i+1}映射到输出向量y_{i+1}，具体的描述可以参考公式（3.3）：

$$\begin{aligned} h_{i+1} &= R(h_i, x_{i+1}) \\ y_{i+1} &= O(h_{i+1}) \end{aligned} \tag{3.3}$$

可以看出对于循环函数R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）[44][45]到LSTM（Long-Short Term Memory）[46]的改变过程。

Simple-RNN的结构相对简单，循环函数R只是一个对输入h_i和x_{i+1}进行非线性变换的函数。可以定义simple-RNN的计算方式如公式（3.4）：

$$h_{i+1} = R(h_i, x_{i+1}) = f(x_{i+1}W^x + h_iW^s + b) \tag{3.4}$$

其中W^x和W^s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“记忆单元”(memory cell)的结构，在这个记忆单元中，梯度信息可以有效地保存下来。而记忆单元中的信息的访问，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记忆单元中，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记忆单元加入到新的记忆单元中的信息，标准的LSTM的定义如公式（3.5）所示：

$$\begin{gathered} c_{i+1} = c_i \odot f + g \odot i \\ i = \sigma\left(x_{i+1}W^{xi} + h_iW^{hi}\right) \\ f = \sigma(x_{i+1}W^{xf} + h_iW^{hf}) \\ o = \sigma(x_{i+1}W^{xo} + h_iW^{ho}) \\ g = tanh(x_{i+1}W^{xg} + h_iW^{hg}) \\ h_{i+1} = tanh(c_{i+1}) \odot o \end{gathered} \tag{3.5}$$

$\odot$表示向量之间的点乘，c代表记忆单元，σ为sigmoid函数，所以i，f，o，的数值都在[0,1]之间，因此可以认为控制的是输入，遗忘，输出信息的比例。g可以认为是输入到LSTM中的信息，那么$g \odot i$则代表着输入到记忆单元中的信息，$c_i \odot f$则代表旧的记忆单元可以保留的信息，所以新的c_{i+1}是$g \odot i$和$c_i \odot f$这两部分信息相加结果。同理由于o控制着输出信息比例，所以$tanh(c_{i+1}) \odot o$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，而记忆单元结构通过存储梯度信息，则可以有效缓解梯度弥散问题，这使得LSTM在很多自然语言处理任务包括：序列标注[49]、语言模型[48]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

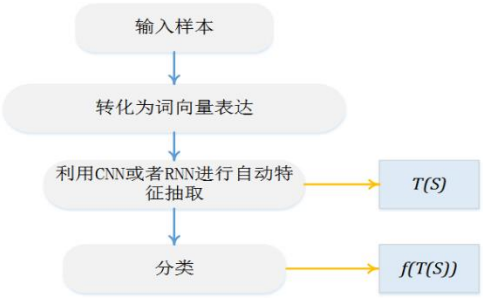

图 3.5 基于神经网络的生物实体关系抽取

Fig 3.5 Biomedical relation extraction via neural networks

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 3.5 所示。

从图中可以看出，这个框架跟我们最初在公式（1.1）中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，提出的框架在词向量基础上，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。关于本章节代码实现部分可以参考这里[10]。

3.5.1. 标准数据集

为了测试算法性能，本文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。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[50]。DDI 这个任务主要抽取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之间关系分成 5 类：advice，effect，mechanism，int 和 other。在表格 3.1 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，

[10] https://github.com/coddinglxf/DDI

DDI 任务是一个细颗粒的多分类任务，因此相对难度加大。

表 3. 1 DDI 中 5 种关系的描述以及例子，其中药物实体都用斜体表示。

Tab 3.1 The five relations and corresponding examples in DDI task, the drug entities are all italic.

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	-

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 3.2 的上半部分对这个数据集做了详细统计。通过统计可以发现，在训练集中，负正样本的比率约为 5.88：1，考虑到数据的不均衡分布会影响到算法性能，而我们重点关注的是正样本情况，因此本文通过定义一些简单的规则，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。

表 3. 2 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，负样本则表明一对药物之间不存在关系。

Tab 3.2 The statistics for DDI corpus, the upper figure is the original statistics of corpus, while the bottom of the figure shows the results after pre-processing. The positive samples indicate there is a relationship between two entities, while the negative samples indicate no relation between two drug entities.

	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
Int	178	10	188	94	2	96
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

3.5.2. 数据的预处理

标准的数据预处理包括分句和分词，为了进一步减少数据的稀疏性，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。但是这种处理方式可能会产生一些噪声数据，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，这些噪声数据会对系统产生以下影响：1）导致数据的分布不均衡，影响神经网络算法的性能；2）进一步增加训练时间。基于以上原因，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. 可以看出这些斜体标记的实体都处于并列结构，因此这些实体对不会存在关系。

表格 3.1 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准确率。

3.5.3. 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。在 3.3.1 章节，我们详细介绍了如何利用 CNN 进行特征抽取，可以看出，在 CNN 中，每个单词首先都需要转化成词向量的表达。考虑到单一词向量表达的局限性，本文提出了一种多通道的 CNN（multi-CNN）。多通道的启发主要来自于数字图像处理中 RGB 三通道彩色图像处理，回归到本文中，这里面多通道的含义则表示采用多个版本的词向量。因此，每个单词会转化成多种词向量形式的表达。由于本文主要的应用范围是生物领域的关系抽取，因此，我们抓取了 PMC，PubMed，MedLine，Wikipedia 这些语料，利用工具 word2vec，训练了 5 个版本的词向量，每个版本的词向量都是 200 维。这些词向量的统计特性如表 3.3 所示：

表 3.3 5 个版本的词向量的统计特性

Tab 3.3 The statistic for 5 word embeddings

类别	包含的词表大小	训练的语料
1	2515686	PMC
2	2351706	PubMed
3	4087446	PMC 和 PubMed
4	5443656	Wikipedia 和 PubMed
5	650187	MedLine

引入多个版本的词向量会带来以下几个优势：1）PMC，PubMed 和 MedLine 基本上涵盖了绝大部分生物领域方面的文献，因此训练出来的词向量对于生物方面单词的语法特性具有很好的归纳性；2）一些常用的单词会出现在 5 个版本的词向量中，这些单词将会提供更加丰富的语义信息；3）采用多版本的词向量，单词之间的信息可以实现共享，而且可以扩大单词的覆盖范围，确保标准数据集中大部分单词能够包含在训练好的词向量词表中。

在系统实现过程中，为了采用多通道，需要对公式（3.1）做一个简单地修改，如公式（3.6）所示。这个公式表明，每个版本的词向量都会将每个单词窗口转化成w_i^j，而对应的都有特定的卷积核W^j与之对应，最终的结果则是 5 个版本的卷积结果的求和。

$$p_i = f(\sum_{j=1}^{5} w_i^j W^j + b) \tag{3.6}$$

3.5.4. 模型训练

从公式 3.2 抽取到的特征C最终会被送到一个 Softmax 分类器（对应于引言部分提到的映射函数f），Softmax 分类器可以输出每个类别的概率。在模型训练时候，我们采用交叉熵作为损失函数，模型中的参数更新都是采用随机梯度下降（SGD）算法。

3.5.5. DDI 实验结果

在 DDI 实验中，本文首先比较了 Baseline，单通道 CNN 以及 multi-CNN 三个模型的实验结果。其中 Baseline 模型采用了随机初始化的单通道词向量，单通道模型则是只利用表格 3.3 中提到的由 Wikipedia 和 PubMed 语料训练而成的词向量，而 multi-CNN 模型则是利用表格 3.3 中提到的 5 个版本的词向量。此外，在每次实验中，对于每种类别，本文都计算了准确率（P），召回率（R）以及最终的 F-score。表格 3.4 展示了这些结果。

表 3.4 基于随机初始化向量，单通道向量以及多通道向量的 CNN 模型的结果。

Tab 3.4 The experimental results based on randomly initialized embedding, one-channel embedding and multi-channel embedding.

	Baseline			单通道 CNN			Multi-CNN		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Advice	**89.39**	53.88	67.24	80.77	67.12	73.32	82.99	73.52	**77.97**
Effect	56.32	57.42	56.87	60.46	73.67	66.41	**67.03**	69.47	**68.23**
Mechanism	78.33	53.36	63.47	64.72	70.81	67.63	**85.00**	62.75	**72.20**
Int	**93.55**	30.21	45.67	82.05	33.33	47.41	75.51	38.54	**51.03**
Micro-F	70.00	52.68	60.12	66.50	67.31	66.90	**75.99**	65.25	**70.21**

Baseline 模型由于只利用随机初始化的单通道词向量，因此单词之间的语义信息不会被模型考虑进去。从表格 3.4 可以看出，跟 Baseline 模型相比，单通道 CNN 模型效果更好，可以将 F 值从 60.12 提高到 66.90，这表明在 DDI 任务中，语义信息相对比较重要。此外，跟单通道 CNN 模型相比，Multi-CNN 模型取得了更好的效果，可以提高约 3.31 左右的 F 值。并且，在单个类别上，Multi-CNN 模型都取得了最优的 F 值，这组对比试验则表明，本文提出的 Multi-CNN 模型是有效的，Multi-CNN 更加丰富的语义信息确保了模型取得了最优的结果。

另外一个值得注意的地方在于，本文提出的三个模型在 Int 这个类别上的准确率都相对较差，最高的 F 值只有 51.03，这个跟其他系统[51][52][53]报告的结果是类似的。最大的原因可能还是由于 Int 类别的训练数据太少造成的，从表格 3.2

可以看出，Int 类别在训练数据集中只有 188 个样本，而在测试数据集中仅有 96 个样本。

经过上面的对比实验，我们可以做出以下的一些结论：1）在 DDI 任务中，语义信息相对比较重要。2）更加丰富的语义信息（多通道词向量）可以提高模型的性能。3）数据规模的大小会影响模型的性能。

3.5.6. DDI 实验对比

在这个章节，我们比较了 multi-CNN 与 DDIExtraction 任务中排名前三的模型（FBK-irst[51]，WBI[52]以及 UTurku[53]），以及最近 Kim 等人[4]提出的线性 SVM 模型。这四个模型都是利用 SVM 作为基准分类器。其中 FBK-irst 和 Kim 模型都是先检测实体对之间是否存在关系（二分类），然后判断实体之间存在什么样的关系（在判断有关系的基础上进行多分类）。跟 FBK-irst 模型中 one-against-all 策略不同的是，Kim 等人在 SVM 中利用 one-against-one 的策略，他们称这样的策略可以有效地解决数据分布不均衡问题。WBI 和 UTurku 系统则完全忽略了策略问题，他们直接利用 SVM 多分类器。具体的特征选择在表格 3.5 中做了详细的描述。

表 3.5 四个系统所采用的特征

Tab 3.5 The features used in four systems

方法	特征
Kim[4]	单个单词，依存关系图，单词对，语法树，名词短语块
FBK-irst[51]	线性特征，树核，浅层语义信息
WBI[52]	由其他 DDI 系统提取出的特征组合而成
UTurku[53]	外部资源，单词特征，依存关系图，线性特征

从表格 3.5 可以看出，列举的这些系统还是严重依赖于特征工程。一些常用的特征，比如单词特征、句法树特征、依存分析图等都被广泛的应用到这些系统中。本文提出的 multi-CNN 可以有效地避免手动提取特征过程。另外，从表格 3.6 的实验对比可以看出，multi-CNN 在 Advice，Effect 以及 Mechanism 三个类别上取得了最好的 F 值，并且进一步将整体 F 值提高了 3.2。

表 3.6 本文提出的 multi-CNN 跟其他四个系统的比较，所有的数据都是 F 值。其中 Detection 则代表着检测任务，它主要检测一对实体是否存在关系，因此是二分类任务

Tab 3.6 Comparison with other systems measured by F scores. The Detection indicates detection task, the main purpose of this task is to decide whether there is a relation between two entities, and thus this is a binary classification task

	Advice	Effect	Mechanism	Int	Detection	总体 F 值
Kim[4]	72.5	66.2	69.3	48.3	77.5	67.0
FBK-irst[51]	69.2	62.8	67.9	**54.7**	**80.0**	65.1
WBI[52]	63.2	61.0	61.8	51.0	75.9	60.9
UTurku[53]	63.0	60.0	58.2	50.7	69.6	59.4
multi-CNN	**78.0**	**68.2**	**72.2**	51.0	79.0	**70.2**

此外，在 detection 任务中，multi-CNN 同样取得了第二好的 F 值，仅次于 FBK-irst[51]的 80.0。从直观上来说，detection 由于是一个二分类任务，所以它主要关注的是如何区分正样本和负样本。对于大多数基于机器学习的传统方法而言，最好的方法就是利用 depend、bind 等这样的关键词来区分一对关系，因为这些关键词不大可能包含在一个负样本中，此外由于在 detection 任务中，系统并不需要区别 depend 和 bind 之间的语义区别，因此语义信息在 detection 任务中所起到的作用是有限的。但是在细颗粒的分类任务中，语义信息则变得重要，因此 multi-CNN 取得相对理想的总体 F 值。

另一方面，为了验证语料预处理是否有作用，本文对比了经过数据预处理和没有预处理的实验结果。从表格 3.7 可以看出，本文提出的语料预处理是有效的，通过过滤噪声数据，可以有效地提高 2.41 左右的 F 值。

表 3.7 经过数据预处理和没有经过预处理的 multi-CNN 的实验结果对比

Tab 3.7 Comparisons between preprocessing and un-preprocessing on multi-CNN model

方法	总体 F 值
Multi-CNN（经过语料预处理）	70.21
Multi-CNN（未经过语料预处理）	67.80

前面提到 DDIExtraction 的语料库是由 DrugBank 和 MedLine 两部分组成，相对于 DrugBank（训练集包含 572 篇摘要）而言，MedLine（只包含约 142 篇摘要）数据规模要小很多，为了对比不同的数据规模对于 multi-CNN 最终效果的影响。本文在 DrugBank 和 MedLine 的语料基础上做了一个交叉测试的实验。从表格 3.8 可以看出，multi-CNN 在 DrugBank 数据上取得了 70.8 的 F 值（相对于 Kim 系统的 69.8，FBK-irst 系统的 67.6），而在 MedLine 数据集上，multi-CNN 只取得了 28.0 的 F 值（相对于 Kim 系统的 38.2，FBK-irst 系统的 39.8）。Segura-Bedmar 等人[9]指出，取得这样结果的一个重要原因在于数据量的减少，导致关键词语的丢失，从而影响了系统的性能。这进一步说明，数据规模的大小对于最终的实验结果影响较大。

表 3.8 交叉训练测试结果（F 值）。其中第一行代表测试数据集，而第一列则代表训练数据集

Tab 3.8 Cross validation results on separated DrugBank and MedLine corpus. The first column indicates the train datasets while the first row is the test datasets.

	DrugBank	MedLine
DrugBank	70.8	52.6
MedLine	10.0	28.0

3.5.7. PPI 实验结果

本章节介绍了 CNN 在 PPI 任务上的实验结果[11]。由于 PPI 是一个二分类任务，因此没有在每个类别上的 F 值统计。表格 3.9 展示了 baseline，单通道 CNN，multi-CNN 三种模型在 Aimed 和 BioInfer 数据集上的实验结果。从表格 3.9 可以看出，单通道 CNN 模型要比 baseline 模型效果好，在 Aimed 和 BioInfer 模型上，分别提高了 1.31 和 1.60 左右的 F 值，而通过引入多通道词向量，multi-CNN 取得了最好的结果，进一步地提升了最终的 F 值。

表 3. 9 Baseline、单通道以及 multi-CNN 在 Aimed 和 BioInfer 两个数据集上的实验结果

Tab 3.9 Baseline, one-channel and multi-CNN experimental results on Aimed and BioInfer datasets.

	Baseline			单通道 CNN			Multi-CNN		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Aimed	71.62	61.25	64.27	72.28	60.82	65.58	**76.41**	69.00	**72.45**
BioInfer	78.13	73.00	75.47	76.06	79.43	77.07	**81.30**	78.10	**79.62**

3.5.8. PPI 实验结果对比

表格 3.10 详细对比了 multi-CNN 跟其他 PPI 系统的性能。在引言部分，我们已经介绍，核函数方法被广泛地应用到关系抽取领域。Choi 等人[19]提出了卷积树的算法，通过对树的剪枝以及调整树的衰减因子，他们的算法在 Aimed 和 BioInfer 数据集上取得了很好的效果。Erkan 等人[12]则提出了基于编辑距离的核函数。此外，混合核函数[21][22]也被广泛应用到 PPI 领域，这些方法都取得了相对不错的实验结果。值得注意的是词向量信息也被 Li 等人[23]整合到了系统中，所不同的是，他们通过聚类的方法，给每个单词分配了一个类别。这种方式可以有效改善传统 one-hot 模型中单词之间语义信息丢失的情况，聚类的方法可以大致保证单词相近的单词拥有相同的类别，从本质上来说，这类聚类算法还是为了整合语义信息。而且他们的实验结果也表明（跟他们的 baseline 模型相比，词向

[11] 考虑到 Aimed 和 BioInfer 的数据规模相对与 DDI 数据集来说较小，因此，我们并没有使用 3.5.2 提出的预处理方式

量的融入可以在 Aimed 和 BioInfer 数据集上分别提高 7.1 和 4.9 左右的 F 值），通过融入词向量，可以取得更好的结果。

但是 Li 等人[23]提出的词向量聚类特征，从本质上来说还是一种硬性的分配，也就是每个单词都被分配一个类别标签，所以提取到的特征仍然是离散的。得益于词向量以及 CNN，本文提出的 multi-CNN 可以在连续的空间上去训练一个模型，学习到的模型可以认为是一个自动平滑的模型，这可以有效地避免硬性分配的过程。此外，从表格 3.9 的实验对比结果来看，本文提出的 baseline 模型已经取得了相当不错的 F 值。通过引入单通道词向量，系统已经可以在 BioInfer 数据集上取得了最优的结果。最后通过整合多通道词向量，multi-CNN 在 Aimed 和 BioInfer 数据集都取得了最优的结果。

表 3.10 multi-CNN 跟其他 PPI 系统的性能对比（F 值）

Tab 3.10 Comparisons with other systems on PPI task（F-score）

方法	Aimed	BioInfer
Choi 等人[19]	67.0	72.6
Yang 等人[54]	64.4	65.9
Li 等人[23]	69.7	74.0
Erkan 等人[12]	59.6	-
Miwa 等人[21]	60.8	68.1
Miwa 等人[22]	64.2	67.6
本文的 Multi-CNN	**72.4**	**79.6**

3.5.9. 基于 CNN 的关系抽取系统错误分析和总结

鉴于生物领域关系抽取的复杂性，从原始文本中抽取实体之间的关系仍然是一个比较大的挑战。本章节详细分析了 multi-CNN 模型容易出错的地方。总结来说，错误可以分成两个大的方面。

1. 输入句子的长度比较长（超过 60 个单词），而两个实体在句子中的距离又很接近。
2. 输入句子的长度比较长（超过 70 个单词），而两个实体在句子中的距离又很远。

如果两个实体在句子中的距离很接近，这两个实体很容易被一个单词窗口包括进去，从概率的角度来说，在 CNN 最大池化阶段，包含实体的这个单词窗口是有可能会被系统作为噪声丢弃的；相反，如果两个实体之间的距离太远，这两个实体又很难被固定单词窗口捕获信息。作为一个可行的解决方案，可以考虑引

入实体之间的最短依存关系路径，这相当于进一步压缩句子长度，从而减弱长句子所带来的影响。在未来工作中，我们会尝试整合依存分析信息。

3.6. 基于 RNN 的关系抽取系统

本章节首先介绍了如何利用 RNN 进行关系抽取，然后分析了 RNN 模型在 DDI 数据集上的实验结果，以及和 multi-CNN 的对比结果，最后总结了 RNN 的优缺点以及可行的改进方案。

在 multi-CNN 模型中，池化的结果C会被作为最终的特征传到一个 Softmax 分类器，从而实现分类。在 RNN 模型中，给定一个包含实体对的句子$X = \{x_1, x_2 \dots\dots, x_N\}$，由图 2.4 可以看出，最终可以生成一个隐含状态h_N，这个h_N相当于编码了整个序列的信息，因此可以作为一个句子级别的全局特征。但是考虑到 RNN 只能逐一的，一个接一个的处理单词序列，这使得在处理当前单词的时候，是无法看到后面单词信息的。一个可行的解决方案是将输入的句子X反序，将反序的X输入到 RNN 模型中，则可以得到另外的一个隐含状态$h_N^{'}$，这个隐含状态则可以包含未来信息。因此，在模型实现时候，可以利用$h = [h_N, h_N^{'}]$作为最终的特征来编码一个句子（在相关工作上，h被称为双向 RNN 特征，这种模型称为 Bi-directional-RNN）。跟 multi-CNN 模型类似，特征h最终会被传到一个 Softmax 分类器，从而进行分类。

3.6.1. RNN 在 DDI 任务上的实验结果

对于 RNN 而言，我们的输入是一个单词向量序列$X = \{x_1, x_2 \dots\dots, x_N\}$，在 baseline 模型中，我们采用随机初始化的词向量，在单通道 RNN 模型中，我们只利用表格 3.3 中提到的 Wikipedia 和 PubMed 训练而成的词向量。表格 3.11 展示了这两个模型的实验结果。

表 3.11 baseline 模型，单通道 RNN 模型，以及 multi-CNN 实验结果

Tab 3.11 The experimental results of proposed baseline, one-channel RNN and the multi-CNN models.

	Baseline			单通道 RNN			Multi-CNN		
	P	R	F	P	R	F	P	R	F
Advice	76.44	60.45	67.51	71.74	75.00	73.33	82.99	73.52	**77.97**
Effect	59.59	57.75	58.66	69.80	69.01	**69.41**	67.03	69.47	68.23
Mechanism	69.81	51.21	59.08	78.92	55.71	65.31	85.00	62.75	**72.20**
Int	71.43	36.46	48.28	80.77	43.75	**56.76**	75.51	38.54	51.03

Micro-F	66.88	54.27	59.92	73.23	63.85	68.22	75.99	65.25	**70.21**

可以看出，单通道 RNN 模型的性能全面超过 baseline 模型，这进一步说明语义信息在 DDI 任务中是非常重要的。另外通过对比表格 3.11 和 3.6 可以看出，单通道 RNN 模型的 F 值已经超过了基于 SVM 的传统核函数（Kim 等人的 67.0）方法。跟 multi-CNN 相比较，单通道 RNN 的总体性能有所下降，但是在单个的类别 Effect 和 Int 上仍然取得了最优的 F 值。此外，结合表格 3.11 和 3.4 可以看出，单通道 RNN 的效果已经超过单通道 CNN 模型，而 multi-CNN 性能的提升主要得益于多通道的词向量，这进一步说明，多通道的丰富的语义信息在 DDI 任务中占有重要地位。

3.6.2. RNN 模型分析和总结

从上面的实验结果可以看出，简单的单通道的 RNN 模型已经可以在 DDI 数据集上取得比较理想的结果。这主要得益于 RNN 对于处理序列数据具有天然的优势，它可以处理任意长度的句子，因此在 3.5.9 章节中描述的 multi-CNN 出现的问题在 RNN 中不会那么明显。但是本文使用的单通道的 RNN 还是存在下面一些问题。

1. 虽然前面描述的改进的 RNN（比如 LSTM）可以缓解梯度弥散和长距离依赖问题。但是如果输入句子太长，这类问题并不能杜绝。
2. 本文只选择最终的隐含状态作为特征进行分类。这需要保证最终的隐含状态能够编码整个句子的信息，但事实上，这种编码可以认为是一个信息压缩过程，在这个过程是存在信息丢失的。

第一个问题从目前来看，还是一个开放的问题，一些改进的 RNN 变种结构，比如 GRU（gate recurrent unit）[47]，或者改进的记忆网络（memory network）[54]，在特定任务下，可能会缓解这类问题。第二个问题其实可以引入 attention（注意力）机制[56]，与其选择最终的隐含状态，也可以选择所有的隐含状态的加权结果，这在一定程度上可以缓解由于信息压缩而造成的信息丢失问题。

此外，在改进方案上，从 multi-CNN 的实验结果来看，我们已经知道多通道的词向量可以提升效果，基于这方面的考虑，一个可行的改进方案是同时训练多个单通道的 RNN，每个单通道 RNN 都由一种词向量来初始化，然后利用这些单通道 RNN 抽取到的特征进行分类。这样做的好处是可以充分利用多个版本的词向量，但是同时也会带来运算量的成倍增加。

3.7. 本章小结

本章节重点介绍了基于 CNN 和 RNN 的生物实体关系抽取系统。通过实验

结果可以看出，提出的 multi-CNN 以及单通道的 RNN 模型都在标准数据集上取得了最优的效果。同时，本章也对 multi-CNN 以及单通道 RNN 模型的优缺点做了简单地分析，并提出了一些改进的方案。总结来说，基于 CNN 和 RNN 的生物实体关系抽取系统，可以结合无监督学习，充分利用外部资源（词向量），避免手动特征抽取，从而实现生物实体关系自动抽取。这类处理问题的框架在很多生物领域的自然语言任务中，具有很好的借鉴意义。

第四章　总结和展望

在互联网技术迅猛发展，数据量大规模增长的今天，自动生物实体关系抽取技术对于构建大规模生物领域的知识库，提高搜索引擎效率，理解生物实体之间的相互作用，具有重要的意义。

本文通过总结前人工作，在此基础上，提出了若干用于生物实体关系抽取的方法，并取得了如下的一些研究成果。

1：提出了一种利用规则进行生物实体关系抽取的方案，并且通过分析实体关系图的方式来衡量一个实体在关系图中的重要性。

通过分析生物领域实体关系之间的特点，考虑到实体之间关系大多以动词和介词相互作用，本文制定了两个通用的策略来进行关系抽取。通过应用对应的策略到句子的依存分析结果上，本文提出的系统在标准的 LLL-challenge 数据集上取得了第二好的效果。在此基础上，本文将提出的关系抽取系统应用到 PubMed 上跟乳腺癌相关的数据集上，构建出了与乳腺癌相关的基因关系图，并通过网络分析的方式，找出了同乳腺癌最相关的基因。文中提出关系抽取系统以及网络分析方法具有一定的通用性，可以应用到其他各种疾病。

2：提出了一种新的混合核函数来用于生物实体关系抽取。

通过观察生物实体之间的关系以及分析实体之间的最短依存路径，可以发现，最短依赖路径一般包含足够的信息来区别一对实体是否存在关系，因此，在最短依存路径基础上，为了体现路径的结构信息，本文提出利用编辑距离核函数来构建实体关系抽取系统。此外，为了充分利用单词信息，本文融合了编辑距离核函数以及余弦核函数，提出了一个新的混合核函数。实验结果表明：同其他核函数以及混合核函数相比，本文提出的方法在 BioInfer 数据集上取得了最优的结果。

3：提出了一个基于神经网络的关系抽取系统的框架

为了缓解生物关系抽取领域标记数据稀缺问题，本文通过结合词向量和神经网络模型。提出了一个基于神经网络的关系抽取系统的框架。在此框架下，本文使用 CNN 和 RNN 两种神经网络结构用于实体自动关系抽取，并在此基础上提出了两种模型：multi-CNN 以及单通道 RNN。实验结果表明，提出的模型在 DDIExtraction，Aimed 以及 BioInfer 等数据集上都取得了最优的结果，这充分说明本文提出的模型是有效的。

综上所述，本文提出的算法在一定程度上提高了关系抽取的准确率。但是考

虑到生物领域文本的复杂性，自动关系抽取技术仍然是一个具有挑战性的课题，针对于第三章提出 multi-CNN 以及单通道 RNN 模型，在以下几个方面存在改进的空间。

1. 除了单词层面的信息（词向量），其他类型的信息，比如单词的位置信息，词性信息等，也可以通过映射到向量的方式（稠密表达）来编码，并融合到模型；
2. 从第二章的叙述可以看出，语法信息在关系抽取中占有重要地位，在后期工作中，可以将语法信息融合到 multi-CNN 或者单通道 RNN 模型中；
3. 在单通道 RNN 模型中仅仅使用最终的隐含状态作为特征来判断实体之间的关系类型，作为一个可行的改进方案，可以引入 attention 机制。
4. 在单通道 RNN 模型中，RNN 的参数都是采用随机初始化，然后采用微调方式来训练参数的，考虑到 RNN 模型本身的参数较多，作为改进，可以训练一个语言模型来初始化 RNN，通过语言模型可以充分利用大量的无标记数据，这使得单通道 RNN 模型参数在训练开始时候就比较接近最优结果，这在一定程度上可以加快收敛速度，提高准确率。

参考文献

[1]. Settles B., Biomedical named entity recognition using conditional random fields and rich feature sets [C]. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics, 2004: 104-107.

[2]. Leaman R, Gonzalez G., BANNER: an executable survey of advances in biomedical named entity recognition [C]. Pacific symposium on biocomputing. 2008, 13: 652-663.

[3]. Bach N, Badaskar S., A review of relation extraction [J]. Literature review for Language and Statistics II, 2007.

[4]. Kim S, Liu H, Yeganova L, et al., Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach [J]. Journal of biomedical informatics, 2015, 55: 23-30.

[5]. R. Bunescu, R. Mooney, A. Ramani, E. Marcotte, Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline, in Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology (BioNLP'06), New York, NY, USA, 2006.

[6]. Blaschke C, Andrade M A, Ouzounis C A, et al., Automatic extraction of biological information from scientific text: protein-protein interactions [C]. 1999, 7: 60-67.

[7]. Blaschke C, Valencia A., The potential use of SUISEKI as a protein interaction discovery tool [J]. Genome Informatics, 2001, 12: 123-134.

[8]. Fundel K, Kuffner R, Zimmer R., RelEx—Relation extraction using dependency parse trees [J]. Bioinformatics, 2007, 23(3): 365-371.

[9]. Segura-Bedmar I, Martinez P, de Pablo-Sanchez C., A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents [J]. BMC bioinformatics, 2011, 12(2): S1.

[10]. Kao H Y, Tang Y T, Wang J F., Evolutional dependency parse trees for biological relation extraction [C]. Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on. IEEE, 2011: 167-174.

[11]. Cui B, Lin H, Yang Z., SVM-based protein-protein interaction extraction from medline abstracts [C]. Bio-Inspired Computing: Theories and Applications, 2007. BIC-TA 2007. Second International Conference on. IEEE, 2007: 182-185.

[12]. Erkan G, Özgur A, Radev D R., Semi-supervised classification for extracting protein interaction sentences using dependency parsing [C]. EMNLP-CoNLL. 2007, 7: 228-237.

[13]. Sun C, Lin L, Wang X, et al., Using maximum entropy model to extract protein-protein interaction information from biomedical literature [C]. International Conference on Intelligent Computing. Springer Berlin Heidelberg, 2007: 730-737.

[14]. Segura-Bedmar I, Martinez P, de Pablo-Sanchez C., Using a shallow linguistic kernel for drug-drug interaction extraction [J]., Journal of biomedical informatics, 2011, 44(5): 789-804.

[15]. Yakushiji A, Tateisi Y, Miyao Y, et al., Event extraction from biomedical papers using a full parser [C]. Pacific Symposium on Biocomputing. 2001, 6: 408-419.

[16]. De Marneffe M C, MacCartney B, Manning C D., Generating typed dependency parses from phrase structure parses [C]. Proceedings of LREC. 2006, 6(2006): 449-454.

[17]. Kulick S, Bies A, Liberman M, et al., Integrated annotation for biomedical information extraction [C]. Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL). 2004: 61-68.

[18]. Sætre R, Sagae K, Jun'ichi Tsujii., Syntactic features for protein-protein interaction extraction [J]. LBM (Short Papers), 2007, 319.

[19]. Choi S P, Myaeng S H., Simplicity is better: revisiting single kernel PPI extraction [C]. Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 206-214.

[20]. Airola A, Pyysalo S, Björne J, et al., All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning [J]. BMC bioinformatics, 2008, 9(11): S2.

[21]. Miwa M, Sætre R, Miyao Y, et al., Protein-protein interaction extraction by leveraging multiple kernels and parsers [J]. International journal of medical informatics, 2009, 78(12): e39-e46.

[22]. Miwa M, Sætre R, Miyao Y, et al., A rich feature vector for protein-protein interaction extraction from multiple corpora [C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 121-130.

[23]. Li L, Guo R, Jiang Z, et al., An approach to improve kernel-based protein-protein interaction extraction by learning from large-scale network data [J]. Methods, 2015, 83: 44-50.

[24]. Zeng D, Liu K, Lai S, et al., Relation Classification via Convolutional Deep Neural Network [C]. COLING. 2014: 2335-2344.

[25]. Miwa M, Bansal M., End-to-end relation extraction using lstms on sequences and tree structures [J]. arXiv preprint arXiv:1601.00770, 2016.

[26]. Hinton G E, Srivastava N, Krizhevsky A, et al., Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580, 2012.

[27]. Graves A, Mohamed A, Hinton G., Speech recognition with deep recurrent neural networks [C]. Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013: 6645-6649.

[28]. Yih W, He X, Meek C., Semantic Parsing for Single-Relation Question Answering [C]. ACL (2). 2014: 643-648.

[29]. Shen Y, He X, Gao J, et al., Learning semantic representations using convolutional neural networks for web search [C]. Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 373-374.

[30]. Kalchbrenner N, Grefenstette E, Blunsom P., A convolutional neural network for modelling sentences [J]. arXiv preprint arXiv:1404.2188, 2014.

[31]. Suykens J A K, Vandewalle J., Least squares support vector machine classifiers [J]. Neural processing letters, 1999, 9(3): 293-300.

[32]. Bunescu R C, Mooney R J., A shortest path dependency kernel for relation extraction [C]. Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005: 724-731.

[33]. Bunescu R, Ge R, Kate R J, et al., Comparative experiments on learning information extractors for proteins and their interactions [J]. Artificial intelligence in medicine, 2005, 33(2): 139-155.

[34]. Giuliano C, Lavelli A, Romano L., Exploiting shallow linguistic information for relation extraction from biomedical literature [C]. EACL. 2006, 18(2006): 401-408.

[35]. Kim S, Yoon J, Yang J, et al., Walk-weighted subsequence kernels for protein-protein interaction extraction [J]. BMC bioinformatics, 2010, 11(1): 107.

[36]. Harris Z S., Distributional structure [J]. Word, 1954, 10(2-3): 146-162.

[37]. Goldberg Y., A primer on neural network models for natural language processing [J]. Journal of Artificial Intelligence Research, 2016, 57: 345-420.

[38]. Mikolov T, Chen K, Corrado G, et al., Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

[39]. Mikolov T, Sutskever I, Chen K, et al., Distributed representations of words and phrases and

their compositionality [C]. Advances in neural information processing systems. 2013: 3111-3119.

[40]. Pennington J, Socher R, Manning C D., Glove: Global Vectors for Word Representation [C]. EMNLP. 2014, 14: 1532-1543.

[41]. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.

[42]. Krizhevsky A, Sutskever I, Hinton G E., Imagenet classification with deep convolutional neural networks [C]. Advances in neural information processing systems. 2012: 1097-1105.

[43]. Lawrence S, Giles C L, Tsoi A C, et al., Face recognition: A convolutional neural-network approach [J]. IEEE transactions on neural networks, 1997, 8(1): 98-113.

[44]. Elman, Jeffrey L., Finding Structure in Time. Cognitive Science. 1990, 14 (2): 179-211.

[45]. Jordan, Michael I., Serial Order: A Parallel Distributed Processing Approach, 1986.

[46]. Hochreiter S, Schmidhuber J., Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[47]. Chung J, Gulcehre C, Cho K H, et al., Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv:1412.3555, 2014.

[48]. Sundermeyer M, Schluter R, Ney H., LSTM Neural Networks for Language Modeling [C]. Interspeech. 2012: 194-197.

[49]. Graves A, Fernandez S, Gomez F, et al., Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]. Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 369-376.

[50]. Segura Bedmar I, Martinez P, Herrero Zazo M., Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013) [C]. Association for Computational Linguistics, 2013.

[51]. Chowdhury M F M, Lavelli A., FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information [J]. Atlanta, Georgia, USA, 2013, 351: 53.

[52]. Thomas P, Neves M, Rocktäschel T, et al., WBI-DDI: drug-drug interaction extraction using majority voting[C]. Second Joint Conference on Lexical and Computational Semantics (* SEM). 2013, 2: 628-635.

[53]. Björne J, Kaewphan S, Salakoski T., UTurku: drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge [C]. Second Joint Conference on Lexical and Computational Semantics (* SEM). 2013, 2: 651-659.

[54]. Yang Z, Tang N, Zhang X, et al., Multiple kernel learning in protein–protein interaction

extraction from biomedical literature [J]. Artificial intelligence in medicine, 2011, 51(3): 163-173.

[55]. Weston J, Chopra S, Bordes A., Memory networks [J]. arXiv preprint arXiv:1410.3916, 2014.

[56]. Bahdanau D, Cho K, Bengio Y., Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.

攻读硕士学位期间的学术活动及成果情况

1） 参加的学术交流与科研项目

（1）安徽省科技攻关项目(1206c0805039)：情感智能机器的基础理论与关键

2） 发表的学术论文（含专利和软件著作权）

1. Hua L, Quan C, Ren F. A hybrid kernel based method for relation extraction and gene-disease interaction network construction [C]. The 10th International Conference on Natural Language Processing and Knowledge Engineering, Sapporo, Japan. 2015.
2. Hua L, Quan C, Ren F. Gene-disease Relation Extraction and Gene Interaction Network Construction. The 7th International Conference on Bioinformatics and Computational Biology (BICoB), March 9-11, 2015.
3. Hua L, Quan C. A Shortest Dependency Path Based Convolutional Neural Network for Protein-Protein Relation Extraction [J]. BioMed Research International, 2016, 2016.
4. Quan C, Hua L, Sun X, et al. Multichannel Convolutional Neural Network for Biological Relation Extraction [J]. BioMed Research International, 2016, 2016.