硕 士 学 位 论 文

基于组合学习和自训练的
生物医学事件抽取研究

The Research of Biomedical Event Extraction Based on Combinational Learning and Self-training

作 者 姓 名：李浩瑞

学科、 专业：计算机应用技术

学　　号：21109223

指 导 教 师：王健 副教授

完 成 日 期：2014 年 4 月 30 日

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　要

生物医学文献数量的急剧增加，使得生物医学从业者在海量生物医学文献中快速地获取大量的感兴趣的信息变得困难。因此，快速有效地从海量无结构化的文本中抽取出便于管理、查询的结构化信息成为生物医学信息抽取的热点的研究课题。生物医学事件抽取属于生物医学信息抽取的范畴，旨在从分子层面在无结构化的文本信息中抽取出有蛋白质有关的结构化生物事件信息。

在生物医学事件抽取的研究中，机器学习的方法得到了广泛的应用。本文在研究过程中主要利用了机器学习的方法对生物医学事件进行抽取，涉及到组合学习，模型自训练以及核方法的机器学习方法。在事件的处理流程上采用了常用的文本预处理、事件触发词检测、事件元素识别以及整体后处理的步骤。本文在生物医学事件触发词检测的阶段采用了基于不同决策规则的学习器进行组合学习、使用模型自训练的方法在触发词检测阶段引入了未标注语料实现半监督学习。在触发词的检测过程中，采用了建立触发词字典来判断文档中词是否为候选触发词，对所选候选触发词进行特征提取进行分类任务，确定候选词是否为触发词并指定相应的触发词类型。在事件元素检测的阶段，构造触发词和蛋白质关系对，借鉴蛋白质交互关系抽取的方法对触发词蛋白质对之间的关系进行检测。根据事件的定义类型将事件分为简单事件和复杂事件分别进行元素的检测。在简单事件中直接鉴定触发词蛋白质的关系，在复杂事件中采用了先鉴定是否存在关系再鉴定存在哪一种关系的方法。最终采用核函数的方法对触发词蛋白质对进行关系检测，来确定事件的元素。

本文是在 BioNLP’09 和 BioNLP’11 共享任务提供的公开的语料集上进行训练和测试的，同时实验过程中采用的未标注语料来源于 PubMed 中的摘要文档。利用本文的方法在采用的语料集上进行模型建立和验证，结果表明本文采用的方法对事件抽取系统的性能有所改善，取得了不错的抽取效果。

关键词：生物医学；事件抽取；组合学习；自训练；核方法

The Research of Biomedical Event Extraction Based on Combinational learning And Self-training

Abstract

With dramatic increasing in amount of biomedical literature, it becomes difficult for biomedical practitioners to efficiently access to the information which they are interested in for such a mass of biomedical literature. Therefore, it turns into a arrestive direction in biomedical information extraiction field that efficiently extracted managable and structured information from massive and unstructured text. Biomedical Event Extraction belongs to the scope of biomedical information extraction, and its objective is to extracte structured biological events information related to protein in unstructured text information on the molecular level.

Machine learning methods have been widely utilized in biomedical event extraction research. In this paper, we also make use of machine learning approachs in biomedical event extraction, involving combination of learns, self-training model and kernel method. We separate event extraction into four common steps, including text pre-processing, trigger detection, argument detection and post-processing. In the phase of trigger detection, we combine different classifiers based on their decision theory and employe self-training approach to implement semi-supervised method and make use of unlabeled data. A trigger dictionary, which can decide whether a word is a candidate trigger, is established from the training corpus. And then features are extracted for ecah candidate triggers. After classification we determines whether a candidate trigger is a real trigger, and also assign an event type to each real trigger. In the argument detection stage, we construct trigger-protein pairs and draw on approach used in protein-protein interaction extraction to achieve the goal of argument detection. We divide argument dectction into simple detction and complex detection based on the definition of event types. In simple detection, we directly decide the whethe an argument belongs to a trigger. However, in comlex detection, we primarily judge whether a trigger-argument pairs exist a relationship. And if it exists, we use the same method which use in first stage to identify the exact type between these two entities in this pair. We use kernel method in this paper to recognise the relationship of trigger-argument pairs.

We carry out our experiments on public corpus of BioNLP'09 and BioNLP'11 shared task. While unlabeled data corpus is selected from PubMed abstract document. Finally, result

shows that our method can improve the performance of biomedical event extraction system and achieve a relative good extraction reslut.

Keywords: Biomedical; Event Extraction; Combination of learners; Self-training; Kernel method

目　　录

1 绪论

1.1 研究背景

随着互联网和信息技术的应用和发展，当今世界的信息数量呈现指数级增长。与此同时，生物医学文献数量也急剧增加，截止 2013 年年底，当前国际上最权威的生物医学文献数据库 MEDLINE 已经有超过两千三百万的生物医学文献。仅在 2013 年就有超过七十万的生物医学文献被添加到 MEDLINE 数据库中。因此，快速有效的从数量如此庞大的生物医学文献中获得感兴趣的信息，成为了一个热点的研究课题。因而生物医学信息抽取在上述环境下应运而生。

生物医学信息抽取的主要目的是从非结构化的生物医学文本中抽取出结构化的信息，从而能够方便信息的管理、分析和查询。生物医学信息抽取相继出现了命名实体识别、实体关系抽取以及生物医学事件抽取等相关子领域。并且随着命名实体识别系统性能达到能够支持实际应用的标准，研究的重点开始向关系抽取和事件抽取转移。与关系抽取只是抽取出一对有关系的实体对不同，事件抽取是要抽取出事件的完整信息，包括事件的类型和参与事件不同实体的作用。生物医学事件抽取所完成的任务就是从非结构化的生物医学文本中抽取出细粒度的信息。

生物医学事件抽取是在 BioNLP'09 共享任务[1]中首次提出。鉴于 MUC[2]，TREC[3]，ACE[4]以及 BioCreative[5]等公开评测任务促进了各自相关领域技术。东京大学组织了此次共享任务，旨在号召全领域共同努力来推动生物医学事件抽取技术的发展，并使得生物医学事件抽取能够支持更详细和更加结构化数据库的发展。本次共享任务的语料来源于 GENIA 语料，并从 GENIA 本体中选定了九种事件类型。从元素参与复杂性角度九类事件可大体分为三类。其中基因表达类型（Gene_expression）等五种类型属于简单事件，调控类型（Regulation）等三类事件属于复杂事件，还有一类由于参与元素的特殊性属于绑定类型（Binding）。BioNLP'09 共享任务取得了较好的效果，极大的提高了生物医学事件抽取系统的性能，在本次任务中获得最好性能的系统取得了 51.95%的 F 值。受到此次共享任务的影响，生物医学事件抽取成为了生物医学信息抽取领域的热点问题，在此次任务之后有更多的生物医学事件抽取系统提出。两年之后举办的 BioNLP'11[6]共享任务中，在 BioNLP'09 的语料基础上保留了全部的摘要并添加了部分全文。这既可以测试当前系统性能较之前一次任务是否有所提高，同时又可以通过全文检测系统的通用性。同时此次任务也在面向整个生物医学领域进行事件抽取进行了尝试。最终本次公开任务性能最好的系统取得了 57.46%的 F 值，较上次共享任务有提高。BioNLP'13 共享

任务[7]在 2013 年举办。本次任务的主要目标是将生物医学事件抽取更好的应用的知识库建设方面。此次任务较前两次任务在数据集和事件类型上有了较大的变化。在数据集方面摒弃了摘要，全部采用了全文的语料，且标注中加入了新的信息。事件类型方面也进行了调整，将九种事件类型调整成为十三种。由于事件类型的增加和语料类型的变化，本次任务中性能最好的系统取得了 50.97%的 F 值。

随着生物医学文献的增加，人工管理和构建数据库变得费事费力。例如当前基因本体数据库（GO）和蛋白质关系数据库的人工管理效率较低。同时，系统生物学中的通路管理[8]也成为一个研究的热点。许多的通路模型整合了成百上千篇文章的知识，这些通路的管理是需要持续不断的人力[9]。如果在这些知识库的构建和管理的方面能够使用生物医学事件抽取的技术作为一种手段，将在很大程度上提高通路管理的研究效率[10]。然而目前的生物医学事件抽取系统的总体性能距离能够应用于实际应用的的标准还相差很多。因此提高生物医学事件抽取系统的性能成为一个很有必要的研究方向。

1.2 研究现状

通用领域的事件抽取的研究早于生物医学事件抽取的研究，目前通用领域的方法主要分为基于机器学习的方法和基于模板规则的方法。赵妍妍[12]等人结合了二元分类和最大熵多元分类的中文事件抽取方法是一种基于机器学习的抽取方法。肖升[13]等所做的基于动词论元结构的中文事件抽取的研究是基于模板规则的方法。Alan Ritter[14]等人使用了基于非监督方法的潜在变量模型并结合了基于规则的方法实现了在 Twitter 上进行事件抽取。

生物医学事件抽取的研究从 2009 年主办的 BioNLP’09 共享任务开始成为研究的热点。此次共享任务吸引了很多实验室的参与，取得了较好的效果，并为之后生物医学事件抽取的研究奠定了基础。本次任务中的图尔库系统[11]将事件抽取分为了触发词检测和元素检测，取得了此次评测中最好的性能。随着领域内继续研究和以及 BioNLP’11 和 BioNLP’13 的推动，生物医学事件抽取得到了蓬勃的发展。目前生物医学事件抽取的系统主要分为基于机器学习的系统和基于规则的系统。

其中基于规则的 ConcordU 系统[1]在 BioNLP’09 共享任务中获得了第三名，是性能最好的基于规则的系统。Bui[15-17]等人采用的基于规则的方法在 BioNLP’11 和 BioNLP’13 的共享任务中均取得了不错的效果。这些方法都有着较高的准确率，召回率比较低。

基于机器学习的系统数量也比较多。大部分的基于机器学习的系统是将事件抽取分为几个步骤来处理的。有一部分基于机器学习的系统是将事件抽取看成一个整体的过

程，通过寻找全局最优解来达到抽取事件的目的。其中图尔库系统[11]是把顺序处理生物医学事件的代表系统。该系统的主要流程是首先将给定的语料进行预处理，然后从标注语料中学习模型进行触发词检测，然后进行事件元素的检测，最后根据不同事件类型定义的元素个数和类型进行后处理。MineEvent[18]系统采用同样流程并在图尔库系统的基础上增加了丰富的特征，并增加了指代消解的处理过程，将语料中出现的指代词与实体进行了处理，提高了系统的性能。之后图尔库系统形成了较为成熟的版本 TEES[19]和 TEES2.1[20]。TEES2.1 参与了 BioNLP'13 的共享任务，取得了第二名的成绩。值得一提的是 BioNLP'13 共享任务的第一名系统是基于 TEES2.1 的 EVEX[21]系统。EVEX 系统在 TEES2.1 的基础上利用了事件从排序，在 BioNLP'13 的共享任务中获得了比 TEES2.1 较好的效果。在 BioNLP'11 中获得第二名的 Umass[22]系统，将事件量化成触发词、元素以及绑定限制之间的关系，通过对偶分解的方式求得全局最优时触发词和元素之间边的状态，从而来获得整个事件。Michel[23]等人利用马尔科夫逻辑网，将事件的相关部分表示成一阶的逻辑形式，然后将特征带入其中，利用相关的性质求得全局最优解，在最优解的状态下获得整个事件。

1.3 本文的工作

在生物医学事件抽取的方法中，机器学习的方法是比较普遍的做法。传统的系统只选用了单一的分类器。由于不同的分类器在实验的过程中考虑的侧重点不同，单个分类器对特征的利用可能不够全面。同时生物医学事件抽取面临语料标注复杂，因此标注语料的规模也限制了系统学习的范围。本文采用组合学习器的方式来充分的利用不同学习器的特点，同时利用自训练的方法引入未标注语料对已有的标注语料进行补充，达到改善事件抽取系统性能的效果。

本文的对生物医学事件的处理过程采用了图尔库系统处理方式。在触发词识别阶段主要是利用了丰富的特征和不同学习器的组合构建触发词检测模型，同时采用了自训练的方法来引入未标注语料的信息来进行触发词检测的模型构建。在生物医学事件元素检测的阶段利用核函数的方法来进行模型构建。核函数可以将低维线性不可分的问题转化为高维线性可分的问题，是解决非线性分类的一个有效的方法，并且设计灵活可以根据不同的需求选择不同的核函数。本文首先利用丰富的特征构建基于特征的核函数，并与图核模型进行融合，从不同角度挖掘更为全面的信息构建事件元素的识别模型。

1.4 本文的结构

论文分为四个部分，主要论述了组合学习器、核函数和学习器自训练在生物医学事件抽取的应用以及取得的效果，主要内容包括了生物医学事件抽取以及相关技术的介绍、算法设计及实验和系统的性能评估，具体的章节安排如下：

第一章，绪论部分，主要详细的介绍生物医学事件抽取的研究背景、研究现状、及本文在生物医学事件抽取上所做的工作。

第二章，主要详细的阐述了与生物医学事件抽取相关的背景知识、BioNLP共享任务、研究过程中常用的工具和方法以及实验语料和相关的性能评价指标。

第三章，详细介绍基于组合学习器的生物医学事件触发词识别方法。详细的阐明了生物医学事件触发词检测遇到的问题及组合学习器的实验及系统性能。

第四章，主要介绍了学习器自训练的情况下结合未标注语料的生物医学事件触发词检测。以及详细介绍基于核方法的生物医学事件元素识别。并给出了系统的总体实验结果和性能比较。

结论部分，对本文相关工作进行总结，对生物医学事件抽取方向进行展望。

2 生物医学事件抽取相关技术

2.1 信息抽取技术与文本挖掘相关知识

生物医学事件抽取隶属生物医学信息抽取的范畴，而信息抽取也又是是文本挖掘领域中比较重要的关键技术。因此生物医学事件抽取和信息抽取、文本挖掘有着密切的联系。同时在文本挖掘中也使用了很多机器学习的技术。在此简单的介绍信息抽取和文本挖掘的相关知识。

2.1.1 文本挖掘

随着文本信息数量的增加，大量的有用信息隐含在海量的文本中。从数目庞大的文本中挖掘出有用信息成为了一门热门研究领域。文本挖掘的目的通过自然语言处理技术和相关分析方法将文本转换成结构化数据，然后用于之后的查询和管理。因此文本挖掘也称为文本数据挖掘。文本挖掘典型的方法包括文本的分类、聚类，实体识别，观点分析，文档摘要和实体关系抽取等。通常涉及输入文本的处理，产生结构化数据，系统性能的评价和解释等。同时由于文本挖掘的过程中涉及到了有关信息检索、数理统计学、模式识别与机器学习、计算语言学以及数据库等相关的理论和技术，使得文本挖掘成为涵盖多学科的交叉研究领域。随着各行业数据的不断增多，文本挖掘得到了广泛的应用到研究，商业以及政府等不同的领域，例如科学发现、生物医学、企业商业智能、出版行业、社会舆论监督甚至是在国家安全以及情报等方面。图 2.1 展示了文本挖掘的基本流程。

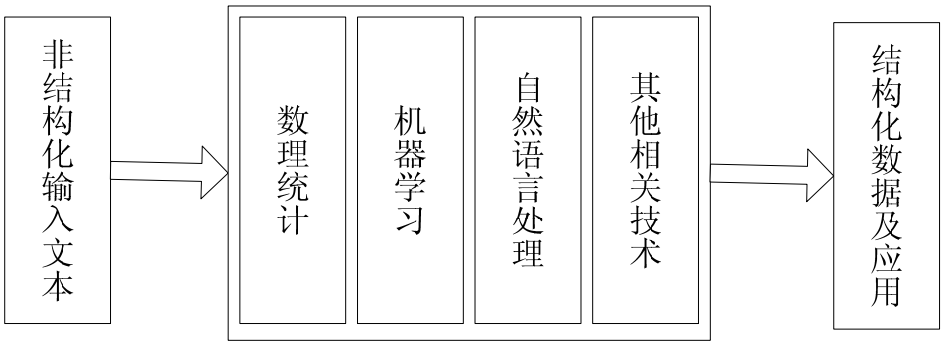

图 2.1 文本挖掘过程

Fig. 2.1 Pipeline of Text Mining

文本挖掘包含了很多学科的知识，如图 2.1 中中间部分所示，很多的文本挖掘方法的理论基础是基于统计学原理实现的；自然语言处理方法是数据挖掘中必不可少的技术，是对非结构化文本中相关信息的识别必要手段；机器学习技术是文本挖掘中常用的技术，使用机器学习的方式从处理过的文本中学习相关的知识和规律、优化相关参数、建立相应的模型达到预测的目的。通过这些技术的应用，实现包括实体识别，信息抽取等相关的文本挖掘过程。

2.1.2 信息抽取

信息抽取（Information extraction，IE）是从可机读的非结构化或者半结构化的文档中自动的抽取抽取结构化的信息的一个过程。这些信息包括实体的识别、实体交互关系识别以及事件识别等。信息抽取所做的就是使用自然语言处理技术处理人类语言文档。在信息抽取的过程中并不是全面理解整篇文档，只是对文档中包含感兴趣的相关信息部分进行分析、提取。面对海量的文档，使用信息抽取获取信息是行之有效的一种手段。抽取的信息能够使得用户快速的获取感兴趣的信息并便于使用统一的格式进行存储，并用于之后的查询、管理和分析。

以信息抽取的形式从海量文本中获取信息，要比依靠关键词进行搜索更为可靠。信息抽取与信息检索（Information retrieval，IR）两者有着密切的关系，两者都是获取有用信息的重要手段。但是两者之间也存在很到的差异。信息检索的主要功能是根据用户提供的关键词，查找并返回与关键词相关性高的文档集合。对于与关键词相关的内容还需要通过用户自己分析相关文档得到。信息抽取的主要功能是从文档中直接返回用户感兴趣的信息。相比信息检索，信息抽取用到更多的自然语言处理的技术来对文档进行分析和理解。通常来说，信息抽取更加的领域相关的信息，并且系统只能抽取预先定义好类型的事件，信息检索系统领域相关性较低，所返回的是与关键词相关的文档集合，并没有领域限制。在某种程度上信息检索和信息抽取存在功能互补，面对海量文本可以通过检索获得相关的文档集合，将这个相关的文档集合作为信息抽取的对象。同时信息检索的性能也可以通过信息抽取进行相应的改善。

信息提取可以追溯到二十世纪 70 年代末，自然语言处理（NLP）兴起的早期[24]。信息抽取的技术有了很大的进步，有两个主要因素对信息抽取的发展有重要影响：一个是在线和离线文本数量的激增使得对信息抽取的需求越来越迫切，另一个原因是 MUC[2]（Message Understanding Conference）会议和 ACE[4]（Automatic Content Extraction）会议对信息抽取领域的长期关注和持续推动。美国政府对会议进行了大力支持，希望研究的信息抽取系统能够在世界的报纸新闻中准确地发现与恐怖主义相关的事件。在 MUC

和 ACE 会议的推动下，信息抽取的范畴从最初的命名实体识别如人名、地名组织机构名，逐步向较为复杂的实体之间的关系抽取如人名和组织机构之间的关系，再到更为复杂的事件抽取如会议的时间、地点、内容等方向发展。同时在这个过程中引入了代词、名词共指以及相同事件合并的概念，使得抽取的信息更加的全面。

随着互联网的出现，信息呈现几何级增长。面对如此海量的数据，挑战不仅仅来自如何存储、传输信息，如何高效的管理信息也是面临的一个巨大的挑战。随着自然语言处理以及数据库的发展，使用信息抽取技术以结构化的形式存储和管理从文本中抽取的信息成为现实。同时通过在海量文本数据上建立一定的规则，使得人们根据自己的兴趣从海量的文本中快速获取相关信息成为可能。如表 2.1 所示。

表 2.1 抽取的结构化信息

Tab. 2.1 Extracted Structed information

原始文本文档	抽取的结构化信息
搜狐体育： 北京时间 11 月 9 日 20 点整，2013 赛季亚冠联赛决赛第二回合的比赛中，中国广州恒大队坐镇主场广州天河体育场迎战韩国首尔 FC 队。第 57 分钟，恒大率先由埃尔克森反击破门。4 分钟后，首尔队德扬扳平了比分。最终双方踢成了 1-1 平，总比分战成 3-3，恒大依靠客场进球的优势夺得了亚冠联赛冠军。这是自从亚冠联赛改制以来，中国球队首次登顶亚洲之巅。	类型：体育 实体：广州恒大队 实体：首尔 FC 队 实体：天河体育场 实体：亚冠联赛冠军 …… 时间：北京时间 11 月 9 日 20 点整 事件：2013 赛季亚冠联赛决赛 ……

表 2.1 中是一个有关信息抽取的举例，左侧是一篇搜狐体育关于 2013 赛季亚冠联赛决赛第二回合报道的摘要。这样的报道在生活中随处可见。对于关心体育的人们主要关心的是比赛的时间、参加的队伍、比赛结果等相关信息。通过信息抽取，可以得到表 2.1 右侧的结构化信息，可以一目了然的发现比较感兴趣的信息，比如抽取出了文本中涉及到的实体以及文本中涉及到的事件 0。通过这种方式可以从海量的文本中抽取出结构化数据，方便之后的在这些结构化数据上进行处理，从而实现海量文本文档的管理和分析。

2.2 生物医学事件抽取

生物医学事件抽取就是一个在医学研究文章中自动检测分子交互关系描述的过程[19]。具体是指从非结构化的生物医学文献中自动的抽取特定类型的结构化事件信息，包

括事件的类型，事件的触发词以及参与事件的相关实体。从事件的定义可以看出，生物医学事件的参与实体是多种的，各个实体在事件中的作用也是不同的。甚至一个事件可能成为另一个事件的参与者，这样就形成了复杂的嵌套或者网状的形式。相比于生物医学命名实体识别和生物实体间的关系抽取，生物医学事件抽取不仅要识别出实体和实体之间具有关系，并且还要识别出具有哪种关系和参与这些关系的实体有哪些。从这个方面来说，生物医学事件抽取相比于简单的二元关系抽取如蛋白质-蛋白质交互关系更能反映原始的生物数据和生物过程, 所以从文本中自动的识别生物事件变的非常有意义。因此生物医学事件抽取是一种更为复杂和细粒度的关系抽取。

每一个事件都会有一个触发词，触发词可以是一个的单词也可以是一个词组。触发词描述了一个生物医学事件的事件类型。同时一个触发词又有可能属于不同事件，也就是说两个类型不同的事件可能共享同一个触发词。每个事件都至少有一个参与实体，而更为复杂的事件会有更多的参与实体。而一个事件的参与实体甚至可能是另一个事件。例如生物医学文献中的一个句子“DeltaFKH did not have a discernable affect on Tax expression.”其中蕴含了几个生物医学事件。可以通过生物医学事件抽取得到其中的事件。在这个句子当中“DeltaFKH”和“Tax”是给定标注好的蛋白质。表 2.2 展示了该句子进过生物医学事件抽取得到的结果，也就是通过结构化的形式展现句子中蕴含的事件信息。

表 2.2　事件抽取展示

Tab. 2.2　Exhibition of event extraction

	事件 1	事件 2
事件类型	Gene_expression	Regulation
触发词	expression	affect
主题	Tax	E1（事件 1）
目标		DeltaFKH

从表 2.2 的展示中可以清楚的看到，在这个句子中还有两个生物事件，事件 1 和事件 2。事件 1 的事件类型是“Gene_expression”，触发词是句子最后的一个单词“expression”，参与的实体只有一个标注的蛋白质“Tax”，并且是以 Theme（主题）的形式参与事件。事件 2 的事件类型是“Regulation”，触发词是 affect，参与的实体有

两个，一个是事件 1 是以 Theme（主题）的形式参与到实践中。另一个参与实体是标注的蛋白质“DeltaFKH”。从抽取的结果中可以清楚的看到，事件 2 是一个嵌套事件，它的参与实体是包含另一个事件的。可以看到以结构化形式呈现出的信息比原始文本文件所要表达的信息更清晰，同时结构化的信息也便于存储以及方便查询。

事件抽取以其有表现力的结构化呈现而流行，广泛地应用于系统生物学，涉及到从对通路的产生和标注提供支持到数据库自动产生母体数据和丰富数据库数据等领域。特别是随着生物医学文献的增加，实体的自动标注以及蛋白质、基因本体等数据库构建也对生物医学事件抽取提出了更多的需求。因此，生物医学事件抽取逐渐吸引了研究者的关注。

正是在这种背景下，BioNLP’09 共享任务的组织者借鉴了之前通用领域的一些评测任务成功推动相关领域技术发展和进步的经验，组织了第一次专门致力于细粒度信息抽取的生物医学事件抽取的共享任务。同样生物领域一些评测任务成功的推动相关领域技术发展的成功案例，也为此次共享任务提供了相关技术上支持。例如致力于生物医学检索的 TREC Genomics track[25]，致力于生物医学命名实体识别的 JNLPBA[26]以及致力于生物分子间关系抽取的 LLL[27]和致力于蛋白质-蛋白质关系数据库的构建的 BioCreative[28]。正式由于这些评测任务推动了生物医学事件抽取相关技术，特别是命名实体识别系统性能的不断提高，为生物医学事件抽取向着应用方向发展提供了技术上的保证。

基于生物医学事件抽取研究的现状，BioNLP’09 的组织者在语料范围、语料标注和相关事件类型的定义方面提供了较为明确的限定。语料选自结构相对规范的生物医学文献摘要，给定了准确的蛋白质位置标注，事件类型也选了特定的几种。随着 BioNLP 共享任务的不断发展，语料范围从摘要发展到全文和事件类型的种类也有了增加。同时，共享任务的连续举办很大程度上提高了领域内的研究生物医学事件抽取的积极性。成功的成为了一个促进生物医学事件抽取技术进步以及相关技术应用的平台。

2.3 句法分析

词通过一定的组织形式构成句子，要想较好的理解一个句子，从词的层面上找到同一个句子中不同词语之间的句法关联是一种行之有效的方法。句法分析的主要目的就是通过一定的规则寻找同一个句子的词与词之间的句法关联。并且按照一定的层次结构将这些关联关系表示成图样式的句法结构。随着领域需求的不断发展，不同领域对句法分析的不同需求促使多种类型句法分析器的出现。这些句法分析器输出的句法分析结构的

图层也相同。因此根据领域的特点和句法分析器的特点选择适合领域的句法分析器，才能更好的使用句法分析来理解领域中出现的句子的句子结构。

GDep[29-31]句法分析器，是东京大学 Tsujii 实验室开发的基于概率统计的 LR 分析法实现依存句法分析的解析器，它是以生物医学语料 GENIA 为训练数据，专门用于生物医学文本句法分析的依存分析器。该句法分析将一个句子的解析结构构建依存关系树。依存关系树是用来表示一个句子中词与词之间的语法关系。在依存分析树中每一个节点代表一个词，每一条边代表了两个词之间的关系，并且这种关系是由引导词指向从属词的。图 2.2 展示的是一个句子经过 Gdep 解析得到的依存分析树状表示的结果。这个句子的文本形式是“These data suggest a crucial role for IRF-4 in the function of immune cells.”图中边所表示的关系是由父节点指向子节点。

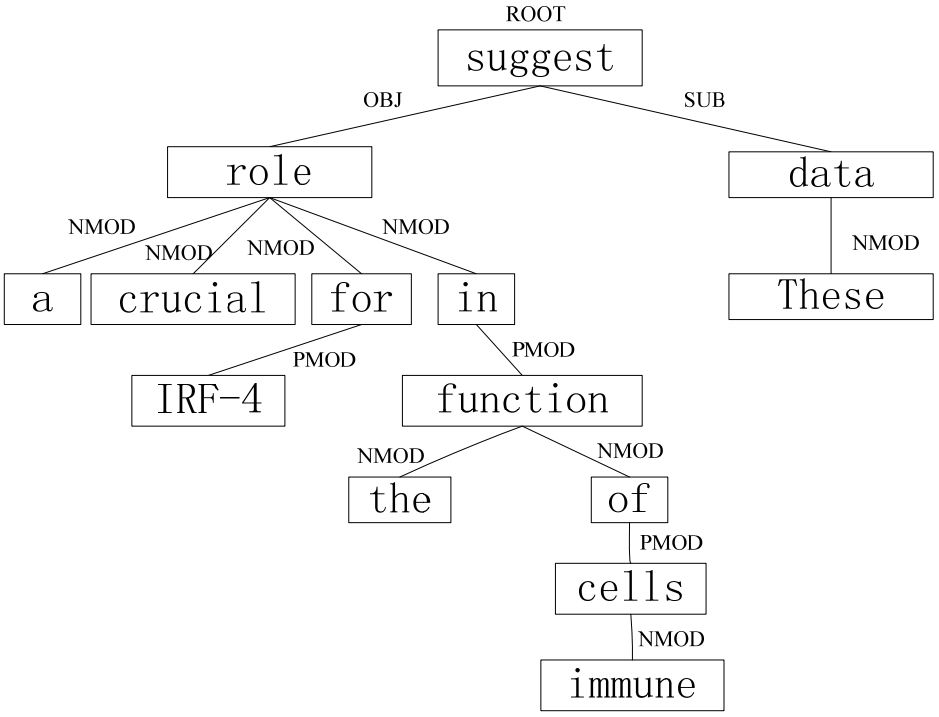

图 2.2　依存分析树

Fig. 2.2　Dependency tree

Gdep 得到的每一个句法分析树都有一个根节点，根节点在一个句子的依存结构中处于中心地位。如图 2.2 中所示，“suggest”就是依存分析树的根节点也就是“ROOT”。“data”是“suggest”的子节点，并且依存关系为“SUB”即表示两者之间为从属的关

系。“role”和“suggest”的依存关系为“OBJ”，表示两者之间的为目标指向关系。“crucial”和“role”之间的依存关系为“NMOD”表示两者之间为名词关系，两者可以构成名词性短语。同理“in”和“function”之间是“PMOD”的依存关系，表明两者之间是介词关系。通过对句子进行句法解析，构成依存分析树，就可以从词的层面找到各个词之间的依存关联分析，从而能够更好的理解句子的信息。

2.4 相关机器学习方法

2.3.1 支持向量机

支持向量机是一种基于线性判别的分类模型，它使用 Vapnik 原则，即在解决实际问题之前总会把解决一个较为简单的问题作为第一步[32]。相比于感知机，间隔最大化是支持向量机与感知机最大的区别。同时核方法的引入，使得支持向量机能够很好的应用到非线性分类上。支持向量机的学习过程是将间隔最大化转换成求解凸二次规划最优化。在最优化的过程中最终要求解的是最小化一个经过正则化之后的合页损失函数。

支持向量机可以分为线性支持向量机和非线性支持向量机。线性支持向量机在训练数据线性可分情况下通过硬间隔最大化得到线性可分支持向量机，在训练数据近似线性可分的情况下，通过软间隔最大化得到线性支持向量机。通过引入核方法，Boser[33]等人提出了非线性支持 向量机。

2.3.1.1 最优分离超平面

支持向量机的基础是线性可分支持向量机。它的目的是学习一个能够将训练集里的正例和负例正确分开的超平面。超平面到任意一边离超平面最近点的距成为间隔。支持向量机的目的是找到能够使得间隔最大化的最优间隔超平面，同时又使得分类器的泛化误差最小。最终利用学习到的超平面对未标注的数据进行预测。

假设有训练样本(x_t,y_t)，x_t 是一个 n 维特征空间中的一个向量，y_t 是类别标签-1 代表负例(图 2.3 中的白色实例)，+1 代表正例（图 2.3 中的黑色实例）。图 2.3 展示了一组二维空间上线性可分的数据，感知机是通过误分类实例最少的策略来寻找超平面的，在这组数据上可以有很多超平面将两组数据进行正确的分离，误分类均为 0。在这些超平面中有着唯一的一个超平面，它是的两边的实例距离这个超平面的距离都达到最大值，如图 2.3 所示，$w*x+w_0=0$ 是其中一个能将训练样本正确分离的超平面，它的特点是最大化了超平面 $w*x+w_0=1$ 和 $w*x+w_0=-1$ 之间的间隔（margin）。由图 2.3 可知，要求解的超平面是最大了两组实例的间隔，因此将最大间隔的问题转化成（公式 2.1），转换成最小化问题：

$$\min \frac{1}{2} \| w \|^2$$

$$\text{s.t.} \quad y_i(w^* x_i + w_0) \geq 1 \quad \forall t \qquad \text{(公式 2.1)}$$

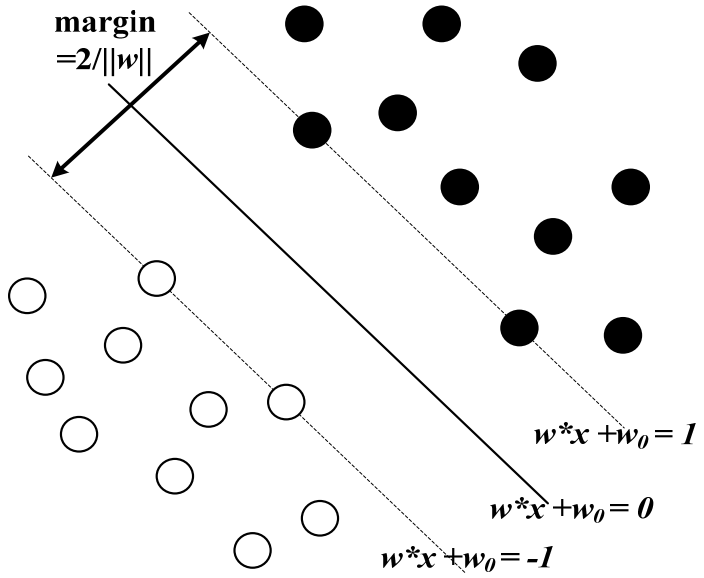

图 2.3　线性可分情况下支持向量机的最优分离线

Fig. 2.3　Optimal separating line of lineraly separable case for SVM

公式 2.1 是一个二次凸规划问题，由最优化理论可知，该优化问题必然存在全局最优解且唯一。对于训练集中的实例，处于正负例边界的点是最难正确分割的点。如果对于那些处于正负例边界的实例，有一个超平面都能够以充分大的置信度来将他们分离开来，这样的超平面对位置数据也将有很好的分割效果[34]。这也是对于最大间隔分类超平面很好的直观阐述。而这些距离最优分离超平面的实例就是所说的支持向量。同时反过来，在最右分离超平面的决策中只有支持向量起到了作用，其他的实例并不起作用。由于在一个训练样本中支持向量的个数一般很少，所以支持向量机是有很少的满足要求的训练样本来确定的，这在一定程度上也减轻了运算量。

2.3.1.2 线性支持向量机

通常情况下，训练样本线性可分的情况是非常少见的，存在一些特殊的点使得训练样本点不能简单的线性可分。这些点不满足硬间隔的约束条件。这种情况下通过对每个样本点(x_i,y_i)引入一个大于等于 0 的松弛变量 ξ_i 来解决这个问题。也就是让间隔函数加

上引入的松弛变量满足硬间隔的条件，同时对于松弛变量引入一个代价因子 C。这样最优化问题变成了（公式 2.2）

$$\min(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i)$$

$$\text{s.t.} \quad y_i(w^* x_i + w_0) \geq 1 - \xi_i \qquad i = 1,2,\ldots,N$$

$$\xi_i \geq 0 \qquad i = 1,2,\ldots,N \qquad \text{(公式 2.2)}$$

从公式 2.2 可以看出通过引入松弛变量和惩罚因子，从而在错误分类给出一定惩罚的基础之上，在尽可能小的惩罚代价条件下寻找最大间隔。在这种条件下寻找到的最大间隔称之为软间隔最大。由于实际应用中大部分的训练数据不是线性可分的，因此软件个最大化的支持向量机的应用场合更为普遍[35]。

2.3.1.2 核方法

非线性分类也是研究过程中经常会遇到的分类问题，通过分线性模型能对非线性分类问题进行很好的求解。图 2.4 展示了一个二维空间中的数据集，有白色三角和黑色圆点组成。左边的图可以清楚的看到，不能够找到一条直线将两类实例准确无误的分开，然而抛物线可以轻而易举的将两种实例分开。但是如果构造另一个空间，让左边的空间映射到右边的空间，可以很简单的将横轴从一次映射到二次，这样我们就可以很方便的使用线性可分的方法来求解最优化问题。

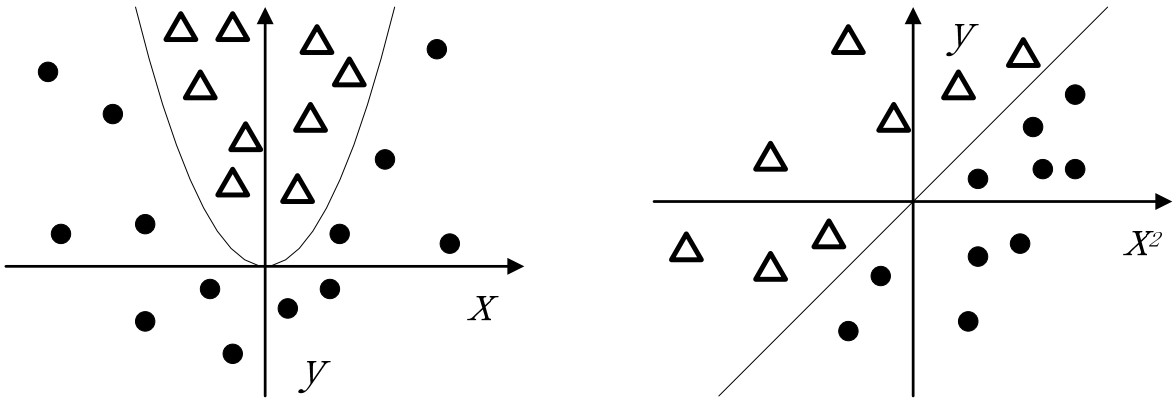

图 2.4　核方法举例

Fig. 2.4　Example of kernel method

图 2.4 给出的启示是，求解非线性可分的问题可以通过以下方法实现。首先将原空间通过一个变换映射到另一个新的空间。使用线性可分的求解方法在新空间内求解问题。这就是核方法的技巧。

将核函数的方法引入到支持向量机中，就可以使用支持向量机来求解非线性可分的问题。在公式 2.2 求解的过程中通过引入拉格朗日因子 α，最终将公式 2.2 表示成只涉及到实例与实例内积的形式。将涉及到内积的部分使用核函数代替，那么超平面可以表示为：

$$f(x) = sign(\sum \alpha^t y^t K(x^t, x) + w_0) \quad \text{(公式 2.3)}$$

公式 2.3 中的 $K(x^t,x)$被称为核函数。核函数可以不显示的定义原有空间与映射空间具体的映射过程，可以直接经过计算核函数而得到结果。很大程度上减小了计算量。这样通过定义相应的核函数，将核方法引入到支持向量机中，是的支持向量机可以用来求解非线性可分的问题。支持向量机有很多常用的核函数，例如线性核、多项式核以及以高斯函数为基础的径向基核等。对于新的预测数据，经过核函数的计算然后根据 $f(x)$给出的符号给待预测点 x 分配相应的类别标签。

2.3.2 随机森林

随机森林（Random Forests）是一种使用了一组未修剪的决策树的分类算法。它是决策树预测的组合，其中每一棵森林中的每一棵决策树的预测值都依赖于一个独立同分布的抽样随即向量。每一棵分类树都是使用了数据的引导样例，并且在每一个数据分割中变量的候选集是整体变量的一个随机子集[36]。随机森林使用两种方法来构建树：一种是装袋法，它是一种对于组合不稳定学习器比较有效的方法[37-38]；另一种是随机变量选取法。

假设给定一组分类器 $C_1(x)$, $C_2(x)$,…, $C_k(x)$和从随机向量的分布中随机抽取的训练集 Y，X，定义间距函数为：

$$mg(X,Y) = av_k I(C_k(X) = Y) - \max_{j \neq Y} av_k I(C_k = j) \quad \text{(公式 2.4)}$$

此处 $I(x)$是指标函数。所谓间距，是用来衡量给一个样本 X 投票 Y 时，投它正确类票数平均数超过投它是其他类票数平均数的程度。间距越大，学习器在分类时得到的结果就越可信。随机森林的泛化误差可以表示成在训练样本 X，Y 空间中的概率形式，如公式 2.5 的所示：

$$PE^* = P_{XY}(mg(X,Y) < 0) \quad \text{(公式 2.5)}$$

在随机森林中，第 k 个分类器可以表示成另一种形式，即 $C_k(x) = C(X, \Theta_k)$。在森林中树的数目较多的情况下，随机森林遵循强大数定理并遵循如下的结构：随着数的数量增加，可以肯定的是对于所有的 Θ 序列，PE^*收敛于 H[39]。其中 H 可表示为下式：

$$P_{X,Y}(P_\Theta(C(X,\Theta)=Y)-\max_{j\neq Y}P_\Theta(C(X,\Theta=j)<0)) \quad \text{（公式 2.6）}$$

此处 X，Y 标明了概率是在 X，Y 空间上的。泛化误差的上限可以表示成两个参数的形式，这两个参数分别表示了每一个单独分类器的准确性和各分类器之间的依赖性。通过公式 2.6 可以证明，当更多的树木加入到森林中时随机森林是不会过拟合的，泛化误差会有一个极限值。

同时作为一个有效的分类算法，随机森林还具有很多优点。比如，随机森林的算法具有较高的精度、能够处理较高维度的变量、能有效的对较大的数据量进行处理、在分布不平衡的数据集中平衡相应的误差等。随机森林有很强的理论基础，数学上严密的证明使得随机森林在算法可信度上有所保证。同时，随机森林在保证不会过拟合的情况下有着较快的运行速度并可以根据实验的需要控制森林中数目的个数来达到性能和复杂度上的平衡。对于特征维数较高、数据量较大同时语料分布不均匀的分类任务，随机森林是很好的选择。

2.5 评价指标和语料

2.5.1 评价指标

本文采用准确率（P），召回率（R）和 F 值（F-score）三个常用的性能评价指标。其中准确率是系统查准能力的体现，即体现当系统查找到一个实例，该实例与原数据一致的置信度。召回率体现的是系统的查全性能，即体现了系统能够查找到与原数据相一致的实例的能力。在分类任务中，系统对测试数据进行预测，得到预测的正例样本的集合为 A。预测数据真正的正例集合为 B。通过集合 A 和集合 B 的交集得到系统预测为正例同时在原数据集中为真正的正例的的集合计为 TP。分别使用|A|，|B|，|TP|来表示对应集合所包含的实例的个数。则可以通过公式 2.7 和公式 2.8 可以计算出系统的准确率和召回率。

$$P=\frac{|TP|}{|A|} \quad \text{（公式 2.7）}$$

$$R=\frac{|TP|}{|B|} \quad \text{（公式 2.8）}$$

通常情况下，一个系统中的准确率和召回率在一定程度上是相互影响的。在相同的参数条件下，当准确率提高的时对应的召回率相应的降低。反之召回率提高的准确率相应减小。虽然两者不是呈现严格的反比例关系，但这种影响的确存在。这样就是的只使用准确率和只使用召回率来评价一个系统的性能显得不够合理。为了能够综合准确率和召回率来衡量一个系统的性能，引入了准确率和召回率的调和平均数来评价系统性能，这个调和平均数就是 F 值。如公式 2.9 所示，为 F 值的定义。

$$F-score=\frac{2\times P\times R}{P+R} \qquad \text{（公式 2.8）}$$

F 值的引入使得能够使用单一的数字来评价系统的整体性能。采用调和平均数而非算数平均数能够较好的反应系统性能较弱的一方面，从而更为全面的评价系统的性能。例如，有一个测试集合包含 100 个正例，某系统的预测结果中只有一个正例，并且这一个正例是真正的正例。根据准确率召回率的计算公式可得，该系统的准确率为 100%,而对应的召回率只有 1%。根据公式 2.8 可知该系统的 F 值不足 2%，然而如果采用算数平均值得到的结果大于 50%。由此可见，采用调和平均数来定义 F 值能较好的考虑性能较弱的一方面，从而更全面的考虑系统的整体性能。

2.5.2 语料

本文的采用的语料包括了 BioNLP’09 和 BioNLP’11 的 Genia Event 的公开语料集。同时在实验中用到了未标注语料，未标注语料选自 PubMed 当中的摘要。由于规模的而考虑选择了能够获取到的最近两年的摘要信息作为未标注语料。

BioNLP’11 的语料包含 BioNLP’09 的语料，它是在 BioNLP’09 语料的基础上经过添加部分的全文类型的语料得到的。这两个数据集有着相同的标注形式，事件类型。为了保持与对比实验选用数据集，本文同时选用了这两个数据集。

本文中所使用的 BioNLP 的两个语料都包含了九种事件类型，其中包含了五种件的事件类型，一种绑定事件类型和三种复杂的调节事件类型。每个事件类型都从不同的角度来描述一种类型的生物事件。同时根据事件所包含的元素，又可以将事件按照如下的方式进行分类。五种简单的事件类型的事件中包含一个触发词和一个主题元素且主题元素是蛋白质。绑定事件类型也是包含了一个触发词，在绑定事件中主题虽然也只是蛋白质但是个数可能不止一个。对于复杂的调节事件，在包含一个触发词的同时还包含了主题和目标两个元素。而之所以称之为复杂事件是由于在三种调节事件中事件的主题和目标元素不仅可以是蛋白质而且可以是一个事件，从而在复杂的事件中形成了事件嵌套的形式。

BioNLP’11 的语料是来自于 GE 语料的摘要和全文。由于主办方并没有将最终的测试集的答案进行公布，所以实验中使用的测试集是官方公布用于调参的发展集。相比于测试集发展集有了最终的答案，使得可以清楚的统计实验中的各个指标。表 2.3 是有关 BioNLP’11 训练集和发展集中语料的大体情况，其中本文将发展集当做测试集。语料中加入的全文有分成了标题、方法、结论等几个不同的部分。表 2.3 中所展示的在训练集和测试集中摘要的部分也是 BioNLP’09 语料的统计情况。

表 2.3　BioNLP’11 语料统计数据

Tab. 2.3　Statistic of corpus of BioNLP’11

	训练数据		测试数据	
	摘要部分	全文部分	摘要部分	全文部分
文献总数（篇）	800	5	150	5
事件总数（个）	8615	1695	1795	1455
蛋白质总量（个）	9300	2325	2080	2610
单词总量（个）	176,146	29583	33827	30305

3 组合学习器的生物医学事件触发词检测

事件抽取通过识别文本触发词和参与的实体来发现实体和实体之间的关系。事件触发词的类型决定了整个事件的类型。作为整个事件抽取流程中的基础步骤，事件触发词检测的性能对整个事件抽取过程的性能有着至关重要的影响。研究的过程中生物医学事件触发词识别受到了词的歧义性问题的困扰[40]。在触发词检测过程当中，语义歧义使得触发词检测有一定的难度。如下面的（1）、（2）和（3）例句中，单词“expression”在（1）和（3）中是触发词，而在（2）中不是触发词。而是触发词的情况下，该单词在（1）和（3）标识的事件类型也是不同的类型。因此，很难判定诸如“expression”这类单词是否是触发词或者在是触发词的情况下它们标识的触发词的类型。

（1）It activates Prot18 gene expression in T lymphocytes.

（2）…… , the expression was enhanced at 30 min.

（3）the expression of c-fos mRNA was suppressed at 30 min

在 BioNLP’11 共享任务中，FAUST 系统[41]的原理是融合了 Umass 的系统和斯坦福事件预测器结果，取得了较好的结果。本文利用组合学习器的方法，使用从原始句子和句子依存分析树中产生的特征来进行触发词检测。组合总是做出类似决策的学习器是毫无意义的[42]。将决策原则不同的分类器进行组合，分类器在决策时可以进行互补。本文采用了两个基础的分类器：一个是支持向量机，它是基于线性判别的决策理论；另一个是随机森林，它是基于决策树的决策理论。这两个分类器在决策原理上是不相同的。

在实验的过程中，除了使用一些常用的文本特征，如词特征，词袋特征等，还从依存分析树中发掘了相关的语义特征。把这些特征应用到两个判别原则完全不同的学习器中，即支持向量机（SVM）和随机森林（Random Forest）。最终，根据每个学习器单独预测性能的好坏指派权值，对两个分类器输出的结果进行线性加权组合得到最终的输出结果。图 3.1 展示了本文在基于组合学习器的方法进行生物医学事件抽取的流程图。

从图 3.1 中可以看出本文实验过程中的主要过程。首先，对原始的输入语料进行分句子、标注实体重定位、蛋白质及词组触发词的替换和去停用词等相关预处理工作。从处理的句子中提取上下文特征，同时对预处理过的句子进行句法分析并从句法分析的结果中提取语义特征。将两部分特征形成的特征集合分别用来训练支持向量机和随机森林。分别使用 SVM 和 RF 对测试数据进行预测，并最终通过结果线性加权融合来得到最终的预测结果。

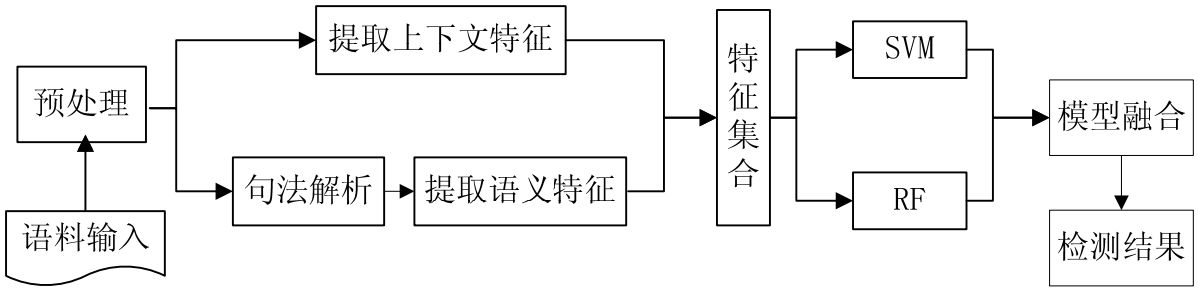

图 3.1　基于组合学习器的生物医学事件触发词检测流程

Fig. 3.1　Pipleline of biomedical event trigger detection based on Combination of Learners

3.1　语料预处理

BioNLP'09 的语料中对一个文档的标注信息分为两部分。对于一个以“.txt”结尾的文档会有两个标注文档，一个是以“.a1”结尾的蛋白质标注文件，在这个文件中准确的标注了蛋白质相对于文档开始的位置，另一个是以“.a2”结尾的文件，文件中详细的标注了事件的触发词的相对于文章开始位置的偏移量和类型以及详细的事件类型和事件元素的标注信息。

由于句法解析对于单个句子解析的结果效果较好，同时跨句子的事件在训练集中的比例不到 1%，因此在整个触发词检测和事件抽取的过程中处理的单位是一个句子。所以预处理的第一步是将给定的语料进行分句处理。同时由于标注文件中提到的标注实体所标注的位置是相对于整个文档开头的偏移量，因此在分句的过程中需要重新调整实体的标注位置，位置需要是相对于每一新的句子的开始位置的便宜。标注位置的调整涉及到蛋白质以及触发词。 并且要根据原本位置的和句子长度确定每个蛋白质、触发词和事件属于的句子。

句法分析的结果是将每个词解析成一行，然而在标注的蛋白质中会存在多个词构成的蛋白质。如果将这种蛋白质分成单个词进行解析，那么就失去了蛋白质本身所包含的信息。因此在句法解析之前将蛋白质进行替换，并建立相应的蛋白质字典。替换之后的蛋白质将以“Prot+序号”的形式出现在句子中，并根据替换前后的长度差调整句子中蛋白质和触发词的标注位置。

经过处理好的句子使用 Gdep 进行解析，解析的结果得到相应的解析信息，每一行代表了句子中的一个词，每一个字段代表了不同的信息。不同的句子之间的解析结果用空行加以区分。图 3.2 是 Gdep 解析的一个实例，是句子“These data suggest a crucial role for IRF-4 in the function of immune cells.”的解析结果展示。

```
1  These    These    B-NP   DT  O          2   NMOD
2  data     datum    I-NP   NNS O          3   SUB
3  suggest  suggest  B-VP   VBP O          0   ROOT
4  a        a        B-NP   DT  O          6   NMOD
5  crucial  crucial  I-NP   JJ  O          6   NMOD
6  role     role     I-NP   NN  O          3   OBJ
7  for      for      B-PP   IN  O          6   NMOD
8  Prot3    Prot3    B-NP   NN  B-protein  7   PMOD
9  in       in       B-PP   IN  O          6   NMOD
10 the      the      B-NP   DT  O          11  NMOD
11 function function I-NP   NN  O          9   PMOD
12 of       of       B-PP   IN  O          11  NMOD
13 immune   immune   B-NP   JJ  B-cell_type 14  NMOD
14 cells    cell     I-NP   NNS I-cell_type 12  PMOD
15 .        .        O      .   O          3   P
```

图 3.2　解析结果举例

Fig. 3.2　Example of Parsing Result

解析结果中每一行从左到右涉及到不同的信息，如第一个字段是单词在句子中的标号。还有包含单词在句子中的形式、单词的原型形式、单词的词性标注、单词再依存分析树中的父节点以及与父节点之间的依存关系。根据这些信息就可以构建图 2.2 所示的依存分析树。

在触发词识别的过程中采用基于词典匹配的方式选择候选触发词。触发词字典就是从训练集中出现的触发词得到的。将训练语料出现的触发词加入到字典中，在对训练语料进行预测是，选定在字典中出现过的词作为候选词，通过提取特征表示候选触发词，然后对候选触发词进行识别。

3.2　特征提取

3.2.1　上下文特征

上下文特征是指从候选触发词所在文本语句的上下文中提取到的特征。包括候选词本身，词性，词周围的特殊信息以及词袋信息等特征。本文在触发词检测中用的的上下文特征包括如下几种：

（1）词特征

词特征主要包含候选词本身以及经过 Gdep 解析后产生的词干和这个词在句子中的词性。由于与候选词最接近的几个位置上对候选词的影响比较大，所以引入候选触发词附近特定位置上的词，同词袋特征相比，该类特征增加了位置信息。

（2）词袋特征

词袋特征是指候选词周围的词，包括了候选词前边和后边的个数为 N 以内的词。考虑到特征的维数和特征的表现能力，本文将 N 设定为 6。

（3）距离特征

距离特征主要指的是衡量候选词和最近的蛋白质之间的距离。因为触发词是和蛋白质紧密相关的，因此一个距离蛋白质近的候选词比一个距离蛋白质远的候选词更有可能是触发词。本文定义的距离指的是在原始语句中候选词到最近蛋白质所包含的单词的个数。当与候选词相邻的词是蛋白质是增加特征“nextToPro”，当候选词与蛋白质在一定距离是增加特征“nearToPro”经过研究，在 BioNLP’09 的训练集中大部分的触发词是靠近蛋白质的。图 3.3 中表示的是在 BioNLP’09 的训练集中触发词和其距离最近的蛋白质的分布图，例如有超过 1200 个触发词与蛋白质相邻，距离定义为 1，超过 1600 个触发词与蛋白质距离是 2。

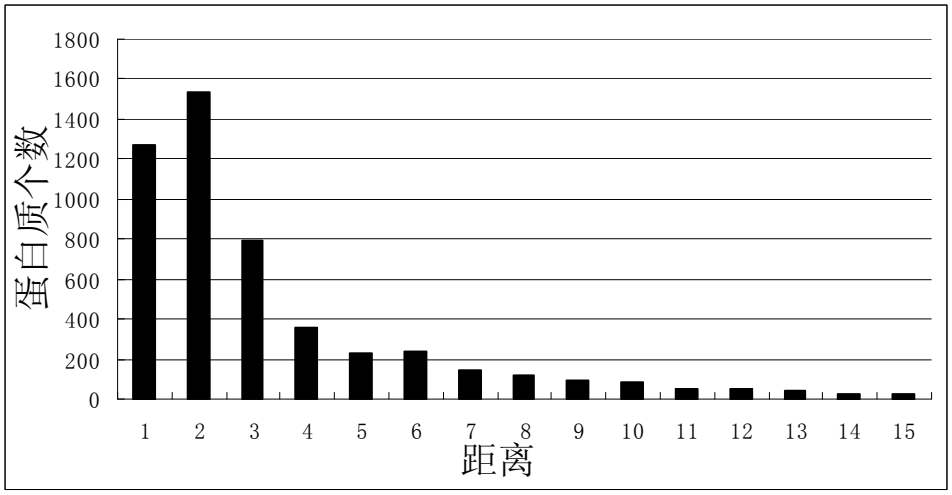

图 3.3 蛋白质个数与距离最近的触发词的关系图

Fig.3 The number of the proteins with the different distance to the trigger

3.2.2 语义特征

语义特征指的是不能从文本上下文中直接发掘的信息。语义特征蕴藏着一些特殊的信息，这些信息是在上下文特征无法发现的，同时这些特征可以不受修饰词长短的影响

准确的找到中心词[43]。所以实验中采用了 Gdep 都处理之后的句子进行句法解析，来发掘相应的语义特征。本文在触发词检测的实验中用的语义特征包括以下几个方面。

（1）依存分析特征

依存分析特征主要来自于 GDep 解析器的解析结果，包括了候选词与依存分析父节点的依存信息和候选词在依存分析树中的路径信息以及候选词在依存分析树中的父节点和子节点的息。

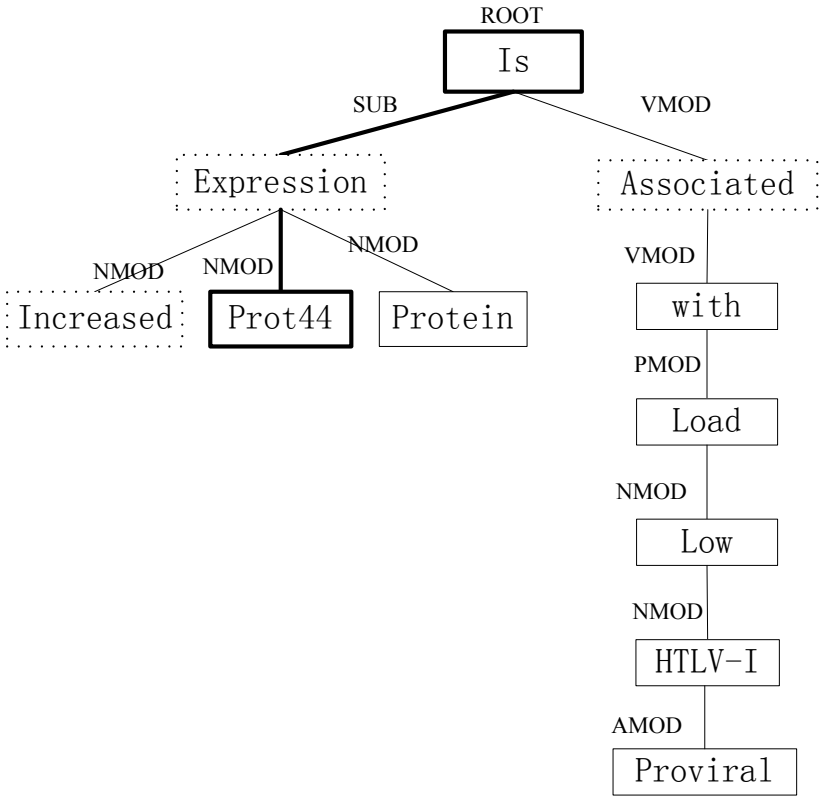

图 3.4　依存分析树中的语义特征

Fig. 3.4　Semantic features in Dependency tree

图 3.4 所示的是“Increased Foxp3 Protein Expression Is Associated with Low HTLV-I Proviral Load”的 Gdep 解析的结果。其中“Foxp3”在原始语料中被标注为蛋白质，经过预处理之后替换为图 3.4 中的“Prot44”。根据触发词字典，该句子中有三个单词属于候选触发词，它们分别是图 3.4 中用虚线框标注出来的“Increased”、“Expression”以及“Associated”三个单词。在提取相关的依存分析是都会考虑到候选词本身的依存

信息以及每个词的父节点出现的什么词，父节点的依存信息，因为在句子的解析结构中相邻的父子节点会有更紧密的语义关系。

（2）依存路径特征

相同的候选词在一个句子里是触发词而在另一个句子里不是触发词。经过研究，在如下的两个句子中，expression 在第一个句子中是触发词而在第二个句子中不是触发词。

(a)Prot24 can directly inhibit STAT-dependent early response gene expression induced by both IFNalpha and Prot25 in monocytes by suppressing the tyrosine phosphorylation of Prot23.

(b)IL-10 preincubation resulted in the inhibition of gene expression for sev-eral IFN-induced genes……

使用了依存分析器之后，在依存分析树中构建从蛋白质到根节点的路径。第一个句子中的 expression 在该路径上并且是触发词。而第二个句子中的相同单词没有在这个路径上并且不是触发词。

实验中通过构建蛋白质到根节点的路径来加入了路径的相关特征。这里用到了两者路径的信息。首先构建蛋白质到根节点之间的路径。然后判断一个候选触发词是否在该路径中，如果在该路径中添加在相应在路径中的信息，如果不在路径中判断其父节点在不是根节点的情况下是否在该路径中，如果父节点在路径上，说明跟蛋白质的关系也较为紧密，添加相应的特征来表示这种关系。从图 3.4 中可以看到，该例中只包含一个蛋白质，由“Prot44”到根节点“Is”构建了一条路径，在图中用粗线表示。图中的“Increased”、“Expression”满足在路径上或者父节点再不是根节点的情况下在路径上。“Associated”不满足上述条件。查阅真实的标注语料会发现经过“Increased”、“Expression”是两个经过标注了的事件触发词，而“Associated”不是触发词。经过对训练集的统计，发现有超过 70%的触发词有这个特性，因此实验中本文采用了这种路径特征来鉴别一个候选词是否是触发词。

3.5 实验过程及结果分析

3.5.1 实验过程

本文使用了当前性能先进的和多分类速度最快的分类器 LibSVM[44]和在控制过拟合上有很大优势并且泛化误差有上限的随机森林作为组合学习的基础学习器。多分类支持向量机有一个正则化参数，这个参数决定了在模型的复杂度和训练误差之间的权衡。多分类支持向量机在训练样本的数量和每一个训练样本非零特征的平均数量上都是线性增长的，这个性质使它成为更适合本文目的的学习方法。在支持向量机模型的构建过

程中，本文实验采用了径向基（RBF）核函数并且把 shrinking 和概率参数设为 1，调整其他支持向量机的参数来进行支持向量机的模型训练。随机森林的最主要参数有两个，对于随机森林通常会随着森林中树数量的增加随机森林的精确率随之提高，但随机森林模型的复杂度也随之提高。另一个参数是用来在分类过程中随机选择属性的个数的数量。实验将随机森林的随机数种子设为默认值。出于模型的精度和计算的复杂性考虑，经过初步的实验和相关的经验，本文选择了 150 棵树和 150 个随机特征的随机森林来进行随机森林模型的构建。

实验使用相同的特征集合对两个学习器分别进行模型训练，并用训练好的模型对 BioNLP'09 的发展集进行预测。之所以使用发展集是因为 BioNLP'09 测试集的答案组织者是没有给出的。在经过两个模型的预测分别得到对应的输出结果后，根据模型的准确率将两个模型的输出中同一个实例的结果进行线性加权组合。两个模型输出的结果的每一行都是一个实例属于每一类的概率，因此具体的组合方法就是对每一个候选实例通过把两个模型的输出各个类别的概率对应进行加权相加，重新计算该候选实例属于各个类别的概率。最终将候选实例分配给重新计算后概率最大的一类，并将该候选触发词标记为该类对应的事件类型。

3.5.2 实验结果分析

本文对组合模型和每个单独的模型在 BioNLP'09 发展集上的输出结果进行了分析。本文系统和语义消歧系统（以下简称 WSD）的性能比较如表 3.1 所示，其中描述了本文与 WSD 方法在每一个事件类型的性能，以及系统总体性能上的对比。

从表 2 中可以看出，本文方法跟 WSD 方法在在总体的准确率上几乎相同，但是本文方法获得了比 WSD 方法高很多的总体召回率，从而在一定程度上提高了系统的总体 F 值。

通过表 3.1 和图 3.5 可以发现，regulation, positive regulation, negative regulation 这三类事件相比于其他类型的事件是更难检测的。这三个类型的 F 值都在 55%以下，而其他类型的 F 值在 60%以上。导致这种情况的主要原因是这三个类型是复杂的事件类型，它们包含了网状的关系和更多的事件元素，因此更难检测。本文系统性能最好的事件类型 WSD 方法相同是 **Protein_catabolism** 类型。值得注意的是，本文的系统在 regulation, Positive regulation, Negative regulation 这三类复杂事件的检测上相比于 WSD 有较好的性能。正是由于这三类复杂事件较高的性能，本文的系统才能在整体的性能上超过 WSD 系统。

表 3.1 本文方法与 WSD 方法的性能比较

Tab.3.1 The comparison of the performance of our method and WSD's

类型	本文方法			WSD 方法		
	准确率 (%)	召回率 (%)	F 值 (%)	准确率 (%)	召回率 (%)	F 值 (%)
Protein_catabolism	100	78.9	88.2	100	84.6	91.7
Gene_expression	70.9	66.5	68.1	75.9	77.4	76.7
Phosphorylation	73.3	82.5	77.6	82.8	70.6	76.2
Localization	69.2	67.5	68.4	72.7	61.5	66.7
Binding	73.5	55.6	63.3	78.7	52.9	63.3
Transcription	65.6	58.8	62.0	64.0	61.5	62.7
Regulation	46.8	37.7	41.8	51.2	25.6	34.1
Positive_regulation	56.9	51.7	54.2	64.9	42.2	51.2
Negative_regulation	49.6	43.8	46.5	50.0	23.3	31.8
OverAll	70.0	63.9	66.8	70.2	52.6	60.1

表 3 所呈现的是每一个单独的学习器和组合后的学习器的最好的性能。与 WSD 方法相比较，本文的支持向量机使用了训练集的全部实例来构建模型，以及比较多的特征，并对参数和核函数进行了调整。在随机森林模型中，使用 30 多组实验来调整树的数目和随机属性选取个数这两个参数。最终，综合考虑性能和时间消耗，实验选取了 150 棵树和 150 个随机属性的随机森林模型，它的性能如表 3 所示。在组合了两个学习器之后，实验得到了比单独使用任何一个学习器性能都好的结果。本文获得了 66.8%的 F 值，比单独使用支持向量机的方法高出 1%,比 WSD 方法高出了 6.7%。

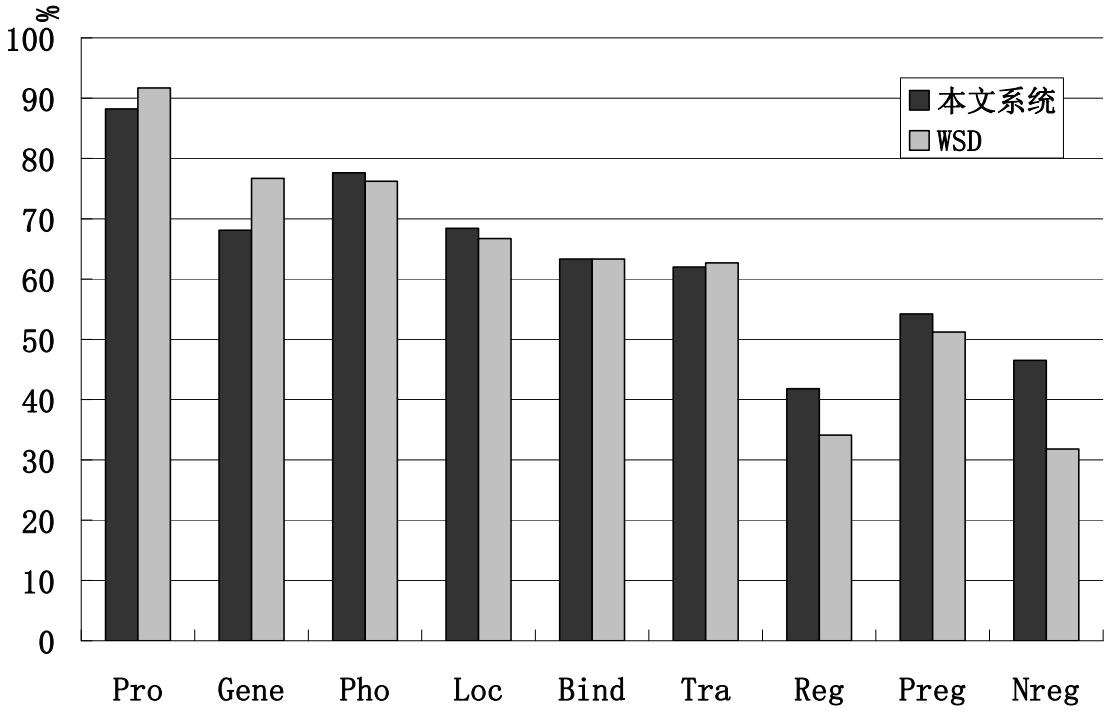

图 4 本文系统和 WSD 系统在各个事件类型上的性能比较

Fig.4 The comparison of each class of the two systems

表 3.2 单个学习器与组合学习器的性能比较

Tab.3.2 The performance of separate learners and combination

	准确(%)	召回(%)	F 值(%)
SVM	72.2	60.5	65.8
RF	72.6	35.6	47.7
Combination	70.0	63.9	66.8

表 3.3 中给出了比较详细的错误分析。实验的目的是找到每一个触发词并给它们标注一个事件类型，因此只对发展集上 624 个错误实例进行分析。首先，有 113 个实例是有多于一个单词的词组触发词构成的，这种情况正本文是做了简化处理的，所以在检测的过程中根本不能发现词组触发词。另外，有 18 个触发词是包含“-”的，这些单词在实验中也是被忽略的。有 198 个触发词是在发展集中出现过，而没有在训练集中出现过，这些词就不会出现在触发词字典中，从而无法检测到它们。还有 295 个触发词被错误的分类，包括将类型标注错误和将触发词标注成为非触发词。

表 3.3 错误分析
Tab.3.3 Error analysis

错误类型	数量
多个词的触发词	113
包含“-”的触发词	18
发展集出现训练集未出现的触发词	198
误分类的触发词	295

最后，从表 3.4 中可以看到，虽然随机森林在发展集中找到了比较少的触发词，但随机森林仍然将支持向量机认为是触发词的 25 个词正确的排除，使得在找到正确触发词个数相同的情况下，提高了组合学习器系统的召回率。

表 3.4 各模型找到的触发词的个数
Tab.3.4 The tokens which find by models

模型	数量
SVM	1243
RF	680
组合模型	1218

3.5.3 小结

选用不同决策原理的学习器进行加权组合，能够在分类决策时进行决策互补。本文的实验结果表明组合学习器方法对于提高生物医学事件抽取的性能有一定的作用。同时从句法分析中挖掘的语义特征也具有一定的辨识性。在接下来的阶段，可以从理论眼里上寻找有理论依据的决策互补的学习器进行组合学习，并可以根据一定的扩展性将组合学习器可以用到抽取整个生物医学事件的工作中来。

4 自训练和核方法的生物医学事件抽取

生物医学事件抽取的过程的过程就是检测出事件触发词以及与触发词共同参与事件的相应元素。在这个过程中，首先要找到在一个句子中那个词是触发词、并且要确定它是哪一个类型的触发词，与此同时还要找出这个触发词与哪些实体发生了关系也就是寻找哪些实体参与了由触发词引导的事件，与触发词构成一个完整的事件。

生物医学事件抽取的所抽取的结构化信息比较丰富，这导致了抽取的过程比较复杂，同时语料的标注工作也十分繁杂。基于监督学习的生物医学事件抽方法面临一个问题，训练模型所用的标注语料只能来自于 BioNLP 共享任务提供的公开语料集中的训练集，从而导致了标注数据集有限，使得模型不能充分的得到学习和训练。引入未标语料，从未标注语料中获取相应的信息来支持生物医学事件抽取性能的改善成为一个可行的选择。同时在事件元素的检测中相当于寻求触发词与蛋白质以及触发词与相应事件的关系抽取，在五种简单事件和绑定事件中触发词只与一个蛋白质实体发生作用关系形成事件，这就相当于从给定的标注语料中寻找触发词和蛋白质对的交互关系。对于复杂事件，可能含有一个主题类型元素，也可能同时含有主题和目标两个元素。同时元素类型可能是蛋白质实体，也可能是事件。因此对于复杂类型事件来说，寻求的是触发词与蛋白质，触发词与触发词的多类型交互关系。因此，在这个过程中可以寻求在其他领域取得较好效果的关系抽取方法来支持事件元素的检测。

本文在生物医学事件抽取的两个阶段分别做了如下处理：在触发词检测的阶段引入了模型自训练方法，通过模型自训练的方法引入未标注语料对触发词检测进行补充；在事件元素检测的过程中根据不同事件类型定义的特点，分别对触发词蛋白质对以及触发词事件对进行相应的关系抽取，引入核函数方法进行事件元素的关系抽取，从而达到检测事件元素的目的。

图 4.1 展示了本文基于自训练和核方法的生物医学事件抽取的主要过程，主要包括了基于自训练的触发词检测过程、基于核方法的事件元素检测过程以及基于规则的后处理过程。触发词检测的过程中根据预先定义的特征空间，在未标注语料抽取与标准语料相同的特征，用于触发词检测模型的自训练过程。事件元素检测是在触发词检测的基础之上实现的。将检测出的触发词根据事件类型的不同，与蛋白质和事件构成相应的关系对进行相应的关系检测。最终根据事件类型的定义来进行相应的后处理，剔除不符合定义的事件。

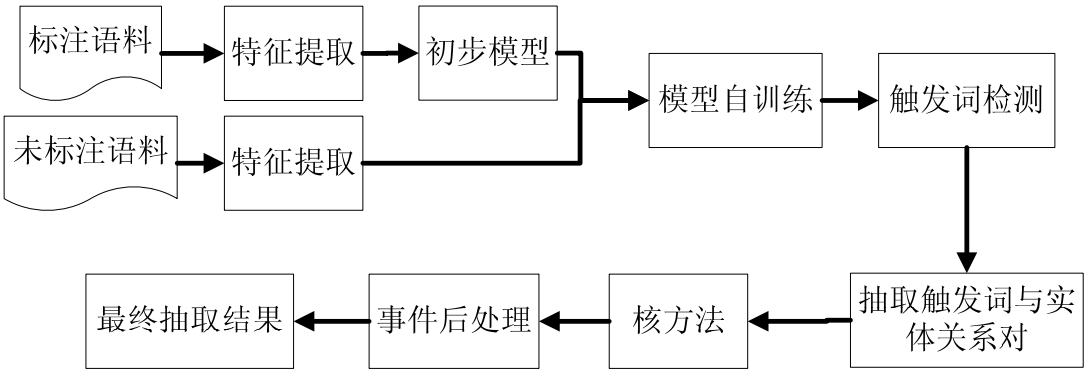

图 4.1　基于自训练和核方法的事件抽取系统流程

Fig. 4.1　Pipleline of event extraction system based on bootstrapping and kernel

4.1　基于自训练的触发词检测过程

事件触发词检测是生物医学事件抽取的一部分，也是比较关键的一部分。在这个过程中要识别句子中的哪些词是事件的触发词。同时要给识别出的触发词标记一个预先设定的类别标签。在利用监督学习方法生成的触发词检测系统，在监督学习训练模型的过程中由于训练语料的局限性，只能利用 BioNLP 共享任务给定的标注语料来进行模型的训练和学习。标注语料的有限性使得在模型训练和学习的过程中面临着模型在训练过程中得不到充分学习的情况。因此，有必要通过一定的手段从未标注语料中选取一定的样本加入到训练样本中，从而扩大训练数据语料来更好的训练模型。在这个过程中本文首先使用了丰富的特征，并使用模型自训练的方法将未标注语料按照一定的标准添加到训练集中来扩充训练集。

4.1.1　半监督方法和自训练学习

随着生物医学文献的增加，文本信息的获取将越来越方便。未标注语料的获取已经变的非常的方便和廉价。相比于未标注语料，标注语料在标注过程中需要耗费大量的人力和财力。因此如果能将未标注数据运用到监督学习上，使得在利用较少标注语料的情况下对实现对监督学习方法进行改善和提升。这种结合标注语料和未标注语料来进行机器学习的思想称之为半监督学习方法。

常用的半监督方法有很多，早在 1968 年就研究者通过极大似然的方式利用标注语料和未标注语料来训练分类器[45]。随着研究的发展，越来越多的半监督方法得到了提出，其中包含了贡献较大的 Co-Training 算法[46]，该算法使用了两个相互独立的学习器，在两个不同的充分冗余的视图上通过引入未标注语料对分类问题进行求解。同时在

Co-Training 的基础上又研究者提出了 Tri-Training[47]的方法。该方法使用了三个分类器在结合标注语料和未标注语料的情况下对未标注语料进行进行标注，当两个分类器对一个未标注实例的标注结果相同，就将该实例标记为置信度较高的实例添加的第三个学习器的训练集中。

自训练方法是一种基于半监督的机器学习方法。自训练方法通过给定的少量的训练样本，选取一个基础机器学习方法，在给定的样本上进行模型训练，得到基础模型。通过基础模型对未标注语料进行预测，之后按照一定的置信度阈值选取预测的未标注语料的样本加入到训练集中，然后使用添加未标注语料的训练集来对模型进行重新训练，如此迭代下去，直到达到必要的结束条件停止，完成训练语料的添加和模型的训练。通常这种方法也被称为 Bootstrapping 的方法。同时自训练学习方法在人工智能领域广泛的应用[48]。

4.1.2 未标注语料

本文的未标注语料选自 PubMed 的摘要文本中的两年的文本信息。选取 PubMed 为未标注语料主要的原因是由于 PubMed 是与生物医学高度相关的语料，同时标注数据中的摘要部分同样是来着 PubMed 部分，因此在数据分布式方面选取未标注语料与标注数据具有一定相关性。由于 PubMed 的文献了过大，因此本文只选了共享任务语料发布之前的两年的语料作为候选未标注语料集合。也就是包含了 PubMed 中 2007 和 2008 两年的所有的 PubMed 的摘要。

由于生物医学事件抽取的测试集中是提供了相关蛋白质的标注信息的，这样在触发词的检测中才能更好的利用触发词与蛋白质之间具有一定的关系来发掘相应的特征，因此本文对未标注语料进行了相应的处理。首先是对 Pubmed 格式的语料进行初步的处理，包括文本的提取，提取文本的分句相关工作。并根据给定的标注语料建立标注语料的触发词字典和蛋白质字典。使用建立的触发词字典和蛋白质字典进行语句的过滤，把整个句子中不包含触发词字典和蛋白质字典中词的句子剔除，保留至少一个候选触发词和包含候选蛋白质的句子。同时对分句后的文本进行 Gdep 的解析工作，以备从解析的依存分析树中挖掘特征。根据建立的蛋白质字典和 Banner[49]命名实体识别系统的标注对过滤后的文本进行蛋白质标注，然后再根据蛋白质字典和标注的结果对文本进行过滤，只留下句子中出现了蛋白质字典中的单词并被标注成为蛋白质的句子。由于可能出现多个词构成的蛋白质，因此要根据蛋白质字典对句子中的标注蛋白质进行相应的替换，并更改相应的标注位置。

经过相应的预处理过后，根据解析的 Gdep 文件和替换后的文本信息，进行特征的提取。对未标注语料的特征提取过程中，用了与标注语料相同的特征和特征空间。在提取特征的过程中，对未标注语料也是主要涉及了以下特征特征。首先是从文本上下文中提取的上下文特征，包括了词本身，词干，相应的词性以及特殊位置的词。同时还包含了候选的触发词与标注出来的蛋白质之间的位置距离特征。另外还包含了在 Gdep 解析生成的依存分析书中挖掘的语义信息，涉及到了相关的依存分析特征，比如一个单词在依存分析树中雨父节点的依存关系，以及依存分析中父节点信息等，同时也使用了从依存分析树中构建的蛋白质到根节点的依存路径，分析候选触发词是否在该依存路径当中来产生相应的特征。最终将提取的未标注语料和相应的实例的特征表示均匀的分成了包含句子个数相等的 20 等份，并分别为各个文件生成各自的特征文件以备自训练过程中的使用。

4.1.2 自训练方法算法及实验步骤

本文采用的自训练方法是首先根据给定的标注语料对模型进行初步的训练，然后使用产生的模型对第一份未标注语料进行预测，在预测的结果中选的一定符合标准的未标注实例添加到当前的训练语料中，形成新的训练语料集合，并采用新的训练语料对模型就行训练，如此反复迭代达到相应的停止条件停止。图 4.2 展示了基于自训练的触发词检测的流程。

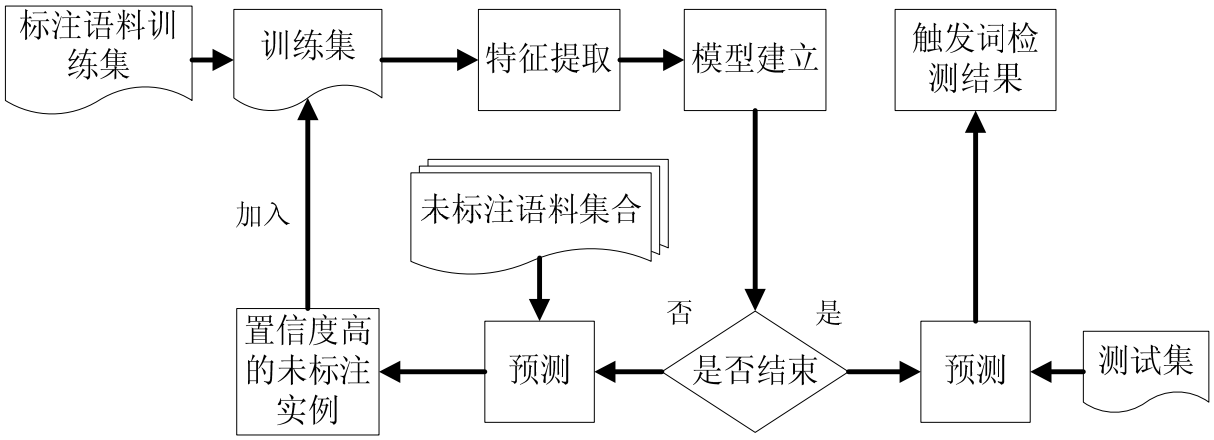

图 4.2 触发词检测自训练流程

Fig. 4.2 Pipleline of Bootstrapping in trigger detection

本文中采用的是 Bootstrapping 的方法对模型进行自训练的。主要原因是 Bootstrapping 方法循环迭代的停止条件较为简单，同时又能够解决因为添加重复的样本

导致过拟合的现象。本文选用的基础分类器是具有多分类功能的支持向量机 LibSVM, 图 4.3 展示的是算法的详细描述。

输入：标注语料训练集TR，未标注语料U，置信度条件C，迭代次数N=20。
步骤：1.将未标注语料U按照句子个数等分为20份$\{u_1, u_2, \cdots, u_N\}$。
2.使用数据集TR训练模型得到模型M_0。
3.使用M_0预测u_1。
4.取符合置信度条件C的未标注实例及其预测结果的集合u_{1C}。
5.令$TR = TR \cup u_{1C}$。并在TR上重新训练得到模型M_1。
6.重复3、4、5，知道循环次到达N。
7.使用训练得到的所有模型预测测试集中的实例。
输出：测试集TE的预测结果

图 4.3 自训练算法描述
Fig. 4.3 Details of self-training algorithm

4.1.3 实验结果及分析

多分类支持向量机对于每个实例的预测结果是给定该实例属于某个类别的概率。通常采用的标记方法是对于一个实例将概率值最大所对应的类别标签分配给该实例。因此在未标注实例根据处理时，所选的置信度就是分类器输出的概率大小。对于一个实例属于每一个类别的概率之和为 1。通常来说一个实例属于某个类别的概率大于了 0.5，那么它在该实例的预测结果中肯定是属于概率最大的类别。同样可以理解的就是，如果一个实例属于某个类别的概率越大，那么该实例属于这个类别的置信度就越大。如果一个实例的预测结果中有两个类别的结果的概率值是非常接近的，那说明相对于此处的训练集来说该实例处在分类的边界部分。该类实例如果处于训练集当中会对支持向量机的分类超平面影响比较大。

本文在未标注语料添加的置信度上进行了不同的选择。添加不同置信度的实例集合对自学习的训练有一定的影响。本文在实验过程中尝试满足不同类型置信度标注的未标注实例的集合，其中包括以下几个置信度标准。（a）选取未标注语料中置信度高于某个一阈值的所有正例进行添加；（b）选取最大概率和次大概率差别最小的一定数量正例进行添加；（c）选取最大概率和次大概率差别最小的一定数量正负例按照一定的比

例进行添加；（d）选取一定比例置信度高的正负例和一定数量最大概率和次大概率差别最小的正负例进行添加。图 4.4 展示了是不同置信度阈值的实验结果。

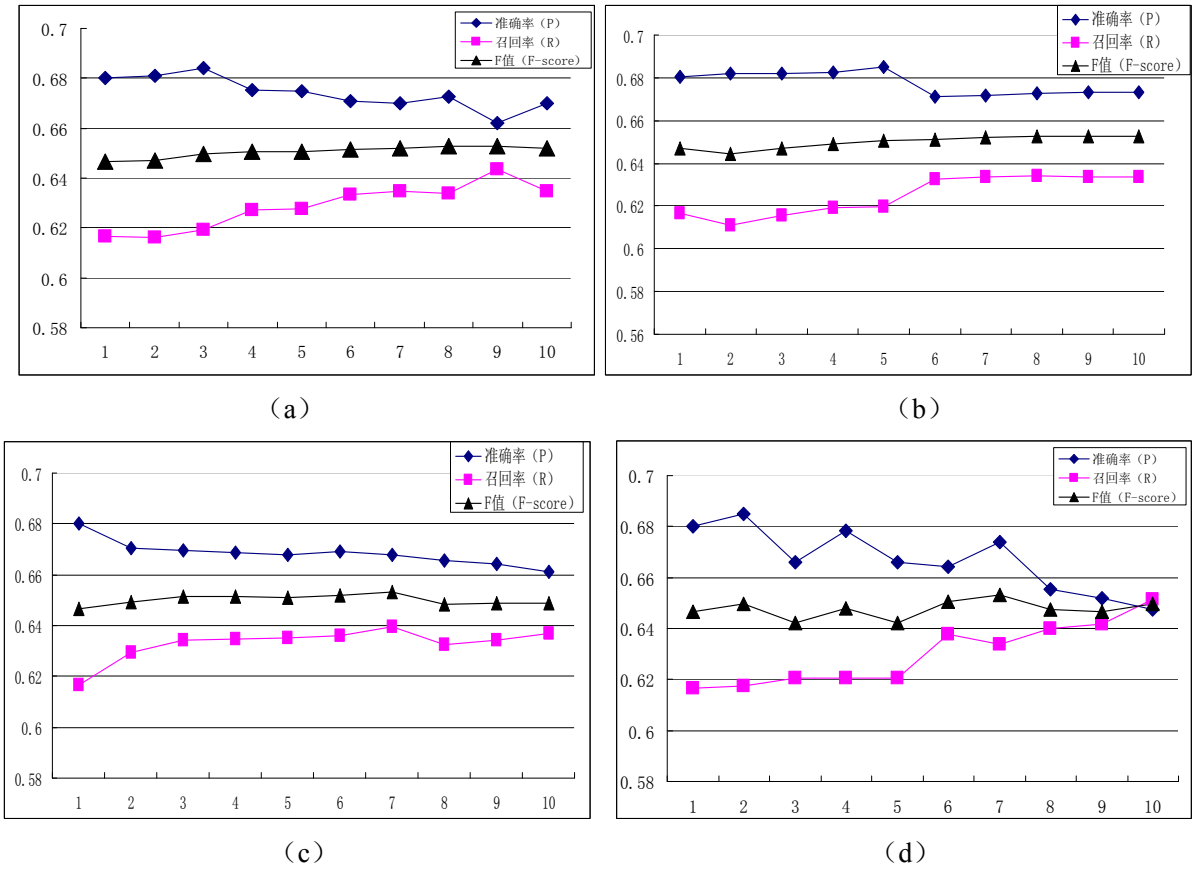

图 4.4　不同置信度阈值的实验结果

Fig. 4.4　Results of self-training algorithm with different threshold values

由于迭代次数较多，因此只选取前 10 次的迭代结果进行展示。从图 4.3 可以看出（a）选取未标注语料中置信度高于某个一阈值的所有正例进行添加时在 F 值有小幅提升的情况小，随着语料规模的增大，召回率有逐渐上升的趋势。此处为了避免添加过的的实例选取了较高的置信度阈值，选取预测类别概率大于 0.95 的实例进行添加。（b）选取最大概率和次大概率差别最小的一定数量正例进行添加时各种评价指标的变化情况。此处根据原来训练集的规模选取了预测类别最大概率和最小概率差值小于 0.1 的；（c）选取最大概率和次大概率差别最小的一定数量正负例按照一定的比例进行添加时系统性能指标变化情况。（d）选取一定比例置信度高的正负例和一定数量最大概率和次大概率差别最小的正负例进行添加时在召回率方面会有较大的提示。

4.2 基于核方法的事件元素检测

4.2.1 图核

图核是用来计算两个图之间相似度的一种方法。它是通过对两个图当中的共同节点的关系进行比较实现相关计算，本文实验中采用的是全路径依存图核[50]。图核包含三种类型的数据：有向子图、分析结构子图、线性顺序子图。其中包含有与句子有关的依存句法结构信息的是分析结构子图。在这个图中包含有两种类型的节点，一种是包含有最短路径信息和词信息的词顶点，另一种是包含的是相关的最短路径信息和词信息的链节点。

图核是在计算机中是通过矩阵的形式在存储图的结构并计算两个图之间的相似程度。在这个过程中包含了两个矩阵，其中一个是包含了词、词性以及词之间的依存关系等信息的标签矩阵 T，另一个是节点与节点之间的边矩阵 E。标签矩阵是一个 M*T 的矩阵，在这里 M 表示的是节点的个数，T 表示的是标签的数目。该矩阵是一个稀疏的矩阵，它表示了标签与节点之间的关联属性。其中 T_{ij} 表示了第 j 个标签是否出现在了第 i 个节点中，出现为 1，不出现为 0。变矩阵 E 是一个表示顶点之间是否存在边的一个 M*M 的一个稀疏矩阵，其中 E_{ij} 表示的是 i 节点和 j 节点时候有边存在。

因此一个图的矩阵 G 可以定义成公式 4.1 的形式：

$$G = T^T \sum_{n=1}^{\infty} E^n T = T^T ((I-E)^{-1} - E)T \qquad \text{（公式 4.1）}$$

其中第 i 个标签和第 j 个标签之间存在的所有的连通路径权重总和用 G_{ij} 来表示。而给定两个图 G，G^*，则图核可以通过公式 4.2 的形式来定义出来。

$$K(G, G^*) = \sum_{i=1}^{|T|} \sum_{j=1}^{|T|} G_{ij} G^*_{ij} \qquad \text{（公式 4.2）}$$

4.2.2 实验方法

在生物医学事件抽取的过程中，BioNLP'11 的语料中将事件类型分成了 9 类，其中有五类属于简单的事件类型，事件的元素只包含一个类型为主题（Theme）的元素，并且该元素仅限于标注的蛋白质实体。而绑定事件是介于简单事件和复杂事件之间的一种事件类型。该类型也只包含了由标注蛋白质构成的主题（Theme）类型的元素，但个数可以是一个也可以是多个。三种复杂事件表示的事件更为复杂，首先从元素类型方面出现了主题（Theme）和目标（Cause）类型之分，其次是在元素的个数上是至少包含主题

元素可以不包含目标元素，还有就是在元素的实体类型上也可能是标注的蛋白质实体，也可能是另一个事件。图 4.5 展示了 BioNLP'11 语料中涉及到的事件类型以及各个类型元素的数目和类型。

表 4.1　BioNLP'11 语料的事件类型和元素类型

Tab. 4.1　Event types and argument types in BioNLP'11 Corpus

事件类型	元素类型
Gene expression	主题（蛋白质）
Transcription	主题（蛋白质）
Protein catabolism	主题（蛋白质）
Phosphorylation	主题（蛋白质）
Localization	主题（蛋白质）
Binding	主题（蛋白质　一个或多个）
Regulation	主题（蛋白质/事件），目标（蛋白质/事件）
Positive regulation	主题（蛋白质/事件），目标（蛋白质/事件）
Negative regulation	主题（蛋白质/事件），目标（蛋白质/事件）

本文的实验中将简单事件和绑定事件归于一类进行处理，由于简单事件和绑定事件都不涉及到事件的嵌套，事件的元素只是主题类型的蛋白质。因此，在训练集中构建简单事件的触发词列表，从训练集里边寻找简单事件的候选触发词与标注的蛋白质实体之间的关系，如果候选词与标注实体是语料中标注的事件，则将该触发词蛋白质对标注为正例，如果没有找到相应的关系则将该触发词蛋白质对标注为负例。采用提出的实体对进行图核模型的训练。对于测试集，从已经检测出的触发词中寻找被标记为简单事件的词，并与标注的蛋白质实体组成触发词蛋白质对，通过在训练集上训练好的图核模型进行元素的预测。在最终的结果中，如果元素被标记成为正例，则说明触发词与蛋白质之间存在关系，并且由于是简单事件只能是主题（Theme）的关系。从而与触发词构成一个简单事件。

处理复杂事件采用的也是图核的方式。由于使用的图核是借鉴了蛋白质交互关系这种二元关系信息抽取的方法，因此在处理复杂事件是进行了特别的处理。在该过程中首先从训练集中构建复杂事件的触发词列表。根据此列表和简单事件集合以及标注的蛋白质集合，构建触发词与蛋白质，触发词与简答事件以及触发词与触发词之间的关系对。对这三种标注关系对在标注的训练语料中查找，如果存在这种标记关系则将该触发词实体对标记为正例。在这里不区分触发词与实体之间是主题（Theme）还是目标（Cause）

关系。同时对测试集中已经检测出来的复杂事件的触发词进行相同方式的处理。在训练集上训练模型进行识别。

接下来在训练集当中，对标注的复杂事件进行分析，将触发词与实体之间存在主题（Theme）关系的实体对标记为正例，同理将存在目标（Cause）关系的实体对标记为负例。同样对训练集检测出的与复杂事件触发词存在关系的实体进行图核检测。最终通过预测结果的输出来判断触发词与各个实体之间存在的具体关系。最终将简单事件和复杂事件进行合并，形成最终的候选抽取事件的集合。

4.2.3 事件后处理

在经过触发词，事件元素检测的步骤之后，会产生最终的候选抽取事件。这些事件中包含了触发词与蛋白质，触发词与事件等的实体对。在后处理的过程中，主要根据表4.1 给出的事件元素类型中元素的数目和相应类别进行处理。主要包括了以下几种事件的过滤。

第一种是无元素事件。这种类型的事件出现的原因是因为在触发词检测阶段某候选词被判断为触发词并标定类别，但在元素检测的阶段并没有检测到与该触发词存在交互关系的蛋白质或者事件，使得该触发词只存在了一个事件类别而不存在元素。因此，对于这种类型的事件直接剔除。

第二种是多主题（Theme）元素事件。由表 4.1 可知，每一个事件有且只有一个主题类型元素。该类事件的处理比较复杂。首先对于简单事件，选择在元素检测过程中置信度较大的一对触发词蛋白实例进行最终事件的形成。对于绑定事件由于事件的性质决定该事件类型可以存在多个主题元素，因此在处理该类事件是主要是判断元素的个数，如元素个数较少则直接认为是一个事件，当元素个数较多是进行置信度选择构成单元素事件。

第三种后处理事件是缺失元素事件。这类事件主要针对的是复杂事件缺失了事件中的主题（Theme）元素的情况。由表 4.1 可知，在一个事件类型中必然会存在一个主题元素，对于目标（Cause）类型的元素的存在性不是必要的。对于这种只存在目标类型元素的事件，在后处理的过程中直接将该事件剔除。

4.2.4 实验结果及分析

本文基于核方法的事件元素检测是在 BioNLP’11 的 GE 任务的语料上进行的。表 4.2 展示的是本文基于自训练和核方法的生物医学事件抽取在每个类上的性能以及最终的总体性能。

表 4.2　基于自训练和核方法的生物医学事件抽取性能

Tab. 4.2　Performance of Event Extraction based on self-training and kernel method

事件类型	准确率（P）	召回率（R）	F 值（F-score）
Protein_catabolism	51.85	54.49	53.14
Gene_expression	79.63	69.95	74.47
Phosphorylation	85.87	76.07	80.67
Localization	67.52	54.19	61.12
Binding	55.09	54.17	54.63
Transcription	58.21	51.13	54.43
Regulation	38.65	26.24	31.26
Positive_regulation	39.89	30.36	34.48
Negative_regulation	41.57	23.86	30.31
OverAll	**72.75**	**41.32**	**52.71**

从表 4.2 中，可以看出本文系统对于简单事件的预测性能要不复杂事件高很多，这是因为复杂事件中存在较多的嵌套事件导致了抽取的困难。

表 4.3 展示了本文系统与替他系统的性能比较。

表 4.3　本文系统与其他系统比较

Tab. 4.3　Comparison of other event extraction systems with ours

	总事件		
	R	P	F
FAUST[41]	49.41	64.75	**56.04**
Ours	41.32	**72.75**	**52.71**
MSR-NLP[51]	48.64	54.71	51.50
TM-SCS [52]	32.73	45.84	38.19

表 4.3 中展示了四种事件抽取系统的性能，其中包括了本文系统。FAUST 系统是 BioNLP'11 共享任务的第一名。第二名是与 FAUST 同源的系统，而 MSR-NLP 在 BioNLP'11 共享任务中获得了第四名。表 4.3 中 Uturku 经过开源工具还原所得，与本文的条件相同。从表 4.3 中可以看出我们的系统在整体 F 值方面与 FAUST 系统，但优于 MSR-NLP 和 Uturku 系统的总体性能。

从表 4.3 中可以看出，目前生物医学事件抽取的性能总体上还是比较低的。系统的 F 值均在 60%以下。这样的精度距离实际应用的要求还相差很远，因此生物医学事件抽取还有很大的研究潜力。

结　论

随着信息的电子化和互联网的普及，研究者可以很方便的获取大量生物医学文本信息。如何从获取的文本信息中快速地获取感兴趣的信息，成为一个亟待解决的问题。而生物医学信息抽取能将无序的非结构化文本中的有用信息，组织成为易于管理和查询的结构化信息。生物医学信息抽取也随着领域需求的改变和相关技术不断的成熟得到了进一步的发展。从开始的命名实体识别，再到交互关系的抽取，再到分子生物学层面的事件抽取，生物医学信息抽取不断发展，能够抽取的信息也越来越复杂。事件抽取旨在抽取分子层面上蛋白质之间发生的各种类型的相互作用关系。抽取的信息包含了事件的类型、参与的实体。与之前的信息抽取相比，事件抽取获得的信息更全面，更能反映原本的生物事件。

在面向生物医学事件抽取的研究中基于机器学习的方法是非常普遍的。而在机器学习的过程中单个学习器在分类任务中对特征的利用可能不够全面。同时监督学习在训练语料获取方面比较复杂而未标注数据的获取相对廉价。本文采用了基于组合学习器的方法，选取不同决策原理的分类器在分类任务中进行性能互补，同时引入了未标注语料，采用模型自训练的方法扩充训练语料。并将交互关系抽取中常用的核方法引入到事件元素的检测中来。

本文在研究的过程中采用了生物医学事件抽取常用的处理流程，将整个过程分为事件触发词检测和事件元素检测两个步骤。在事件抽取的过程中，触发词检测是整个事件抽取过程中的基础步骤，性能的好坏直接影响整个系统的性能，因此将事件触发词检测作为一个独立的研究问题。而触发检测过程中存在很严重的歧义问题，因此如何消除歧义准确的判定触发词成为研究的关键。本文首先在文本的上下文和语义方面构建了一个丰富的特征空间。利用从语料中提取的丰富特征，来训练决策原则截然不同的学习器。并在语料的测试集上进行验证，对同时被标记为同一类别的实例进行准确标注，对于类别不同的实例进行线性加权获取结果，来达到消除歧义检测触发词的目的。

在整个生物医学事件抽取的过程中，同样分为两个步骤，在触发词检测阶段因为未标注语料，通过模型自训练的方法，使用支持向量机作为基础分类器从未标注语料中获取满足一定置信度的实例加入到训练语料中来进行训练语料的扩充，达到触发词的检测工作。在元素检测的过程中，根据事件类型的定义将简单事件和复杂事件分开处理。构建触发词与蛋白质，触发词与事件之间构建交互关系对，通过交互关系抽取中常用的核函数的方法来进行交互关系的检测，确定触发词与蛋白质，触发词与事件之间是否存在

交互关系。对复杂事件中存在的交互关系，根据两种不同的关系类型，再次利用核函数的方法判定复杂事件中存在交互关系的实体的具体类型。实验结果表明，本文构建的系统取得了较好的 F 值。

综上所述，本文的主要工作主要包括如下几个方面：

（1）本文利用了组合学习器的方法，在丰富特征的基础上进行生物医学事件触发词检测的工作。

（2）本文引入未标注语料，使用模型自训练的方式从未标注语料中选择置信度较高的实例来扩充生物医学事件触发词检测的训练集。

（3）本文在事件元素检测的过程中将简单事件和复杂事件进行分开处理，并引入交互关系处理的思想，利用核函数的方法进行事件元素的检测。

对于组合学习器的方式进行生物医学事件抽取还有很多工作可以做，包括学习器的选择以及组合的方式。对未标注语料的利用方面还有很大的研究空间，包括添加未标注数据的置信度选取，以及未标注语料的选取等。同时可以引入更多在交互关系抽取中已经得到证的核方法来进行相关事件元素检测的工作。这些都将成为我们下一步的研究工作。

参 考 文 献

[1] Kim J D, Ohta T, Pyysalo S, et al. Overview of BioNLP'09 shared task on event extraction[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009: 1-9.

[2] Chinchor N. Overview of MUC-7/MET-2[C]. In Message Understanding Conference (MUC-7) Proceedings. 1998.

[3] Ellen Voorhees. Overview of TREC 2007. In The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings. 2007.

[4] Strassel S, Przybocki M A, Peterson K, et al. Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction[C]. LREC. 2008.

[5] Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, editors. Proceedings of the Second BioCreative Challenge Evaluation Workshop. CNIO Centro Nacional de Investigaciones Oncologicas. 2007.

[6] Kim J D, Pyysalo S, Ohta T, et al. Overview of BioNLP shared task 2011[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 1-6.

[7] Nédellec C, Bossy R, Kim J D, et al. Overview of BioNLP Shared Task 2013[J]. ACL 2013, 2013: 1.

[8] Ohta T, Pyysalo S, Ananiadou S, et al. Pathway Curation Support as an Information Extraction Task[J]. Proceedings of LBM 2011, 2011.

[9] Ohta T, Pyysalo S, Rak R, et al. Overview of the pathway curation (PC) task of bioNLP shared task 2013[J]. 2013.

[10] Wang X, McKendrick I, Barrett I, et al. Automatic extraction of angiogenesis bioprocess from text[J]. Bioinformatics, 2011, 27(19): 2730-2737.

[11] Björne J, Heimonen J, Ginter F, et al. Extracting complex biological events with rich graph-based feature sets[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009: 10-18.

[12] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3-8.

[13] 肖升, 何炎祥. 基于动词论元结构的中文事件抽取方法[J]. 计算机科学, 2012, 39(5): 161-164.

[14] Ritter A, Etzioni O, Clark S. Open domain event extraction from twitter[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1104-1112.

[15] Bui Q C, Sloot P. Extracting biological events from text using simple syntactic patterns[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 143-146.

[16] Bui Q C, Sloot P M A. A robust approach to extract biomedical events from literature[J]. Bioinformatics, 2012, 28(20): 2654-2661.

[17] Bui Q C, van Mulligen E M, Campos D, et al. A fast rule-based approach for biomedical event extraction[J]. ACL 2013, 2013: 104.

[18] Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution[J]. Bioinformatics, 2012, 28(13): 1759-1765.

[19] Björne J, Salakoski T. Generalizing biomedical event extraction[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 183-191.

[20] Björne J, Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task[J]. ACL 2013, 2013: 16.

[21] Hakala K, Van Landeghem S, Salakoski T, et al. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction[J]. ACL 2013, 2013: 26.

[22] Riedel S, McCallum A. Robust biomedical event extraction with dual decomposition and minimal domain adaptation[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 46-50.

[23] Riedel S, Chun H W, Takagi T, et al. A markov logic approach to bio-molecular event extraction[C]. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009: 41-49.

[24] Andersen P M, Hayes P J, Huettner A K, et al. Automatic extraction of facts from press releases to generate news stories[C]. Proceedings of the third conference on Applied natural language processing. Association for Computational Linguistics, 1992: 170-177.

[25] Hersh W R, Cohen A M, Roberts P M, et al. TREC 2006 Genomics Track Overview[C]. TREC. 2006.

[26] Kim J D, Ohta T, Tsuruoka Y, et al. Introduction to the bio-entity recognition task at JNLPBA[C]. Proceedings of the international joint workshop on natural

language processing in biomedicine and its applications. Association for Computational Linguistics, 2004: 70-75.

[27] Nédellec C. Learning language in logic-genic interaction extraction challenge[C]. Proceedings of the 4th Learning Language in Logic Workshop (LLL05). 2005, 7.

[28] Krallinger M, Morgan A, Smith L, et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge[J]. Genome Biol, 2008, 9(Suppl 2): S1.

[29] Sagae K, Jun'ichi Tsujii. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles[C]. EMNLP-CoNLL. 2007, 2007: 1044-1050.

[30] Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data[C]. Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005: 467-474.

[31] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii, Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics -10th Panhellenic Conference on Informatics, LNCS 3746, 2005: 382-392.

[32] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[33] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[C]. Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992: 144-152.

[34] Alonso-Atienza F, Rojo-Álvarez J L, Rosado-Muñoz A, et al. Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection[J]. Expert Systems with Applications, 2012, 39(2): 1956-1967.

[35] Pan S, Iplikci S, Warwick K, et al. Parkinson's Disease tremor classification - A comparison between Support Vector Machines and neural networks[J]. Expert Systems with Applications, 2012, 39(12): 10764-10771.

[36] Díaz-Uriarte R, De Andres S A. Gene selection and classification of microarray data using random forest[J]. BMC bioinformatics, 2006, 7(1): 3.

[37] Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning[M]. New York: Springer, 2009.

[38] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140.

[39] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[40] Martinez D, Baldwin T. Word sense disambiguation for event trigger word detection in biomedicine[J]. BMC bioinformatics, 2011, 12(Suppl 2): S4.

[41] Riedel S, McClosky D, Surdeanu M, et al. Model combination for event extraction in BioNLP 2011[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 51-55.

[42] Alpaydin E. Introduction to machine learning[M]. MIT press, 2004.

[43] Vlachos A. Two strong baselines for the BioNLP 2009 event extraction task[C]. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2010: 1-9.

[44] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[45] Hartley H O, Rao J N K.Classification and Estimation in Analysis of Variance problem[J].Review of International Statistical Institution,1968,36(3):141-147.

[46] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]. Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.

[47] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(11): 1529-1541.

[48] Li Y, Li H, Guan C, et al. A self-training semi-supervised support vector machine algorithm and its applications in brain computer interface[C]. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, 2007, 1: I-385-I-388.

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

[50] 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(Suppl 11): S2.

[51] Quirk C, Choudhury P, Gamon M, et al. Msr-nlp entry in bionlp shared task 2011[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 155-163.

[52] Kim J D, Wang Y, Takagi T, et al. Overview of genia event task in bionlp shared task 2011[C]. Proceedings of the BioNLP Shared Task 2011 Workshop. Association for Computational Linguistics, 2011: 7-15.

攻读硕士学位期间发表学术论文情况

1 李浩瑞，王健，林鸿飞，杨志豪，张益嘉. 基于混合模型的生物事件触发词检测.
主办单位：中国中文信息学会
刊物名称：中文信息学报
刊物类型：国内期刊
所属章节：论文第三章

致　　谢

时光荏苒，三年的硕士研究生生涯即将结束。回首往事，正是有了周围人的帮助才促使了我不断的成长，不断的进步。在我硕士毕业论文即将完成的时候，在此特别的感谢给予我帮助和支持的人们。

首先感谢的是我的导师王健副教授，在论文的撰写和审查方面王老师给了我很大的帮助，提出了很多意见。同时也正是王老师耐心的引导，才使得我能够一步一步的把握自己的研究方向。同时在这三年里王老师在学习和生活上对我关怀备至，给我提供了非常好的学习环境。同时王老师认真负责，精益求精的做人态度也给了我很大的影响，是我学习的楷模。

其次，感谢实验室的林鸿飞老师和杨志豪老师。在整个研究生学习期间，林老师渊博的知识和严谨的知学态度给予我很大的影响，同时林老师给实验室营造了很好的学术氛围，使得我在演讲和展示方面有了进步。同时在新生训练期间给予我的指导深深的影响我整个硕士阶段。杨老师是一个治学态度严谨，为人正直而又平易近人的一位老师。杨老师对我的教诲使的我在为人处事上有了进步。

感谢李彦鹏师兄、张益嘉老师，两位有着扎实的学术积累，为人谦逊，在学术上寄予了我很大的鼓励，是我学习的榜样。

感谢徐谦师兄、熊大平师兄，两位师兄是在平时学习和生活中给予我很大的帮助。从刚入学的新生训练，到平时学习研究中遇到的问题都得到了师兄的很多帮助。

感谢徐博师姐，以及我在读期间生物组的所有师兄师姐以及师弟师妹。为组里营造出一个很好的学习氛围和丰富多彩的学习生活。同时感谢实验室所有的同学，感谢大家营造出的愉快、乐观的学习氛围，是我能在这三年中愉快的度过我的硕士研究生生活。

最后要由衷的感谢我的家人，是你们给予我一路前行的动力。感谢你们一直以来的陪伴和鼓励，关心与支持。谢谢你们！

再次向所有给予我帮助和支持的人们表示由衷的感谢，祝你们健康平安！

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：____年____月____日

导 师 签 名：_____ 日期：____年____月____日