分类号 TP391.1 密级 公开

U D C 681.37

博士学位论文

中文文本信息抽取模型与方法研究

于 江 德

导师姓名(职称) 樊孝忠（教授） 答辩委员会主席 林宗楷

申请学科门类 工 学 论文答辩日期 2007-9-10

申请学位专业 计算机应用技术

2007 年 7 月

中文文本信息抽取模型与方法研究

北京理工大学

DISSERTATION

Models and Methods for Chinese Text Information Extraction

Yu Jiangde

Supervisor:	Prof. Fan Xiaozhong
Speciality:	Computer Application
Academic Degree Applied for:	Doctor of Engineering
Degree Conferred by:	Beijing Institute of Technology

July, 2007

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签名：　　日期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的，复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签　　名：　　日期：

导师签名：　　日期：

摘 要

随着科学技术的迅速发展和信息传播手段的不断更新，每天都会产生大量文本数据，而其中很多却不能被浏览阅读，致使大量信息得不到有效利用。所以如何从海量文本中自动抽取特定类型的信息，已经成为急需解决的问题。这样的需求促使文本信息抽取成为自然语言处理领域的研究热点。文本信息抽取是从文本中自动抽取用户感兴趣的信息，并将这些信息转换为结构化数据的过程。

本文依据所处理的文本对象的不同将文本信息抽取分为两类：一类是半结构化文本信息抽取，所处理的文本句法结构不完整，具有明显的版面结构和一些特定的标识信息，通常从这类文本中抽取连续的信息域。例如从科研论文中抽取头部信息和引文信息。另一类是自由文本信息抽取，所处理的文本由自然语言形式的语句组成，具有完整的句法结构，主要研究从这类文本中抽取特定类型的事件信息，称为文本事件信息抽取。例如从新闻报道中抽取“职务变动”类事件信息。本文围绕这两类文本信息抽取中的语言模型和机器学习方法，面向中文文本进行了深入的研究和探索，取得了如下创新成果：

（1）针对隐马尔可夫模型不能充分利用对抽取有用的上下文特征，提出了一种基于条件随机场的半结构化文本信息抽取方法。该方法能充分利用半结构化文本中的版面结构等特征，并针对中文文本的特点，先利用分隔符、特定标识符对文本进行分块，然后利用条件随机场模型进行信息抽取。对中文科研论文的头部信息和引文信息抽取的实验结果表明，该方法抽取性能要明显优于基于隐马尔可夫模型的方法。

（2）提出了一种从未标注的中文文本中基于自扩展策略自动获取事件抽取模式的算法，该算法从少数几个种子抽取模式开始，通过一个增量迭代的过程发现新的抽取模式，在每一轮迭代中采用类似于 TF/IDF 的评估方法对产生的候选模式进行排序，选择最优的模式并入当前模式集。应用该方法从人民日报语料中自动获取“职务变动”类事件的抽取模式，实验结果表明，该方法产生的抽取模式在中文文本事件抽取中具有较好的抽取性能，综合指标 F 值达到 66.3%。

（3）事件探测和分类是事件信息抽取中的首要任务，对事件抽取的后继任务至关重要。传统的事件分类仅仅依据事件表述语句中的触发词，而忽略了触发词上下文中与事件密切相关的大量特征信息，致使分类效果不佳。为此提出了一种基于最大熵

模型的事件分类方法，该方法能够综合事件表述语句中的触发词信息及各类上下文特征对事件进行分类。应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行的分类实验表明，该方法的分类效果明显优于传统的分类方法。

（4）提出了一种基于触发词的论元结构，利用条件随机场模型来识别事件要素及其角色的方法。该方法利用事件表述语句中触发词的论元和要抽取的事件要素之间的对应关系，以浅层句法分析为基础，把短语或命名实体作为识别的基本单元，选择基于句法成分的、基于谓词的、句法成分-谓词关系、语义四类特征作为模型特征集，将条件随机场模型用于事件要素及其角色的识别。应用该方法对“职务变动”和“会见”两类事件的事件要素和角色进行识别，在各自的测试集上分别获得了 77.3%和74.2%的综合指标 F 值。

（5）针对句式简单的事件表述语句，提出了一种基于隐马尔可夫模型的中文文本事件抽取方法，该方法首先通过触发词探测从文本中发现特定类型事件的候选语句，然后利用隐马尔可夫模型从中抽取每个候选事件的事件要素。为每一类事件要素构建一个独立的隐马尔可夫模型用于该类事件要素的抽取，构建模型时采用随机优化的方法优化模型结构。实验结果表明，该方法能较好地实现这类事件的抽取。

最后，设计并实现了一个面向 Web 的文本信息抽取原型系统。该系统由图形用户界面、数据获取、文本粗加工和信息抽取四个模块组成，用于对不同类型的文本信息抽取模型与方法进行实验。

关键词：自然语言处理；文本信息抽取；事件信息抽取；语言模型；机器学习；半结构化文本；自由文本

Abstract

With the rapid development of science and technology and improvement for transmitting information, large amount of text data appears in the whole world every day, however much of it cannot be read which leads to a lot of information being ignored. The study of automatically extracting special information from text is becoming a crucial problem. Owing to the strong requirements, extracting information from text has been one of the hottest research domains in Natural Language Processing (NLP). Text information extraction can automatically get user-needed information from the text, present the information to users in a structured form.

In this dissertation, the text information extraction is divided into two categories according to different texts. The first one is semi-structured text information extraction. The semi-structured text has fragmentary syntactic structure, obvious layer structure, and some special labels information etc. features, continuous information fields are extracted from this kind of text. For example, the header information and citation are extracted from research papers. The other category is free text information extraction. Texts to be dealt with consist of some natural language sentences, which have complete syntactic structure. The research mainly focuses on extracting event information from the free text, which is known as text event information extraction. For example, the information about *management succession* is extracted from a series of news. In this dissertation, a series of studies and experiments have been done on language models and machine learning methods for information extraction from Chinese text, and the main achievements of this dissertation can be described as follows:

(1) As the context-sensitive features cannot be fully used for information extraction by using Hidden Markov Model (HMM), a method based on Conditional Random Fields (CRFs) is proposed to extract information from semi-structured text. The method can make full use of these features such as layer structure etc. of semi-structured text, and consider the characteristics of the Chinese text. The method firstly makes use of the format information of list separators and special-labels to segment the text, and then combines

CRFs for special-fields extraction. Experiments were done that extracting information from header and citation information of Chinese research papers, the results show that the proposed method could get better performance than that based on HMM.

(2) An algorithm based on bootstrapping strategy is presented to acquire extraction patterns automatically from un-annotated Chinese texts. Starting with a small set of extraction patterns as seeds, the algorithm applies an incremental iterative procedure to find new extraction patterns. During the process of the each iteration, the system evaluates the quality of the candidate patterns based on TF/IDF scoring, selects the top-ranked patterns and adds them to the set of current patterns. Experiments are performed on automatic acquisition of extraction patterns of *management succession* from the People Daily corpus. Experimental results show that extraction patterns generated with the algorithm have good result for Chinese event information extraction, and the F measure achieves 66.3%.

(3) Event detection and classification are principal tasks in event information extraction, and they are crucial for these subsequent tasks of the event extraction. A traditional method for event classification is based on the trigger in the event mention sentence which ignores a lot of special information of trigger in event classification, and it leads to lower classification performance. In consideration of the previous shortcomings, an approach based on maximum entropy model is proposed for event classification. This approach can classify the events by merging the features about trigger and it's context in event mention sentence. Experiments are performed on *management succession*, *meeting*, *terror attack*, *judicial adjudicate* and *natural disaster* etc. five types of events in the People Daily corpus, and the results show that the method in this paper is much better performance than the traditional approach apparently.

(4) A method based on argument structure of the trigger is proposed and it applies CRFs for identifying event argument and its semantic role. The method uses the corresponding relation between trigger's arguments and event arguments in the event mention sentence, takes shallow syntactic parsing as the foundation, phrases or named entities as the labeled units. Four categories features, including features based on sentence constituents, features based on predicate, predicate-constituent features, and semantic features are selected as a set of features. The CRFs model is applied to identify event

argument and its semantic role. This method is used to identify event argument and its roles on two test sets of *management succession* and *meeting*, and the F measure is 77.3% and 74.2% respectively.

(5) For the simple event mention sentences, a method based on Hidden Markov Models (HMMs) is proposed for extracting the event information from Chinese texts. Firstly, a candidate sentence, which contains a description of specific events via trigger detecting is found. After that, extract event argument form these candidate sentences by using HMMs. For every kind of event argument, a separate HMM is constructed to extract event arguments. During the process of constructing model, stochastic optimization is applied to optimize the structure of the model automatically. The experimental results show that the method has comparatively good performance for simple event extraction from Chinese texts.

At last, a prototype system of information extraction has been designed and fulfilled. The system consists of four modules: Graphics User Interface (GUI), acquisition of text data, text preprocessing and information extraction etc. The experiments about the different language models and methods for text information extraction can be done in the system.

Key words: natural language processing; text information extraction; event information extraction; language model; machine learning; semi-structured text; free text

目 录

图索引

表索引

第1章 绪 论

1.1 研究背景及意义

当今信息社会，大量的有用信息存在于自然语言形式的文本之中。但面对海量文本，通过浏览阅读获取其中的有用信息，几乎是不可能的。为了应对信息爆炸带来的严峻挑战，迫切需要一些自动化的工具帮助人们从海量文本数据中快速、准确地找到真正需要的信息。如果能够借助计算机对这些文本进行处理，从中抽取出那些对用户有用的信息，并转化为一种结构化的形式供用户分析和利用，无疑会大大加快处理速度。信息抽取（Information Extraction，IE）研究正是在这种背景下产生的。

文本信息抽取是以半结构化或无结构的自然语言文本为处理对象的文本挖掘技术，能够从文本中自动地抽取用户感兴趣的事实或事件信息，并将这些信息转换为结构化的数据并存储的过程 [1][2][3]。这些信息可以是对一个事物的参数具体描述或它的一系列属性，也可以是一些特定类型事件的事件要素（event argument）信息 [4]。例如，从产品评测文档中抽取产品的详细参数；从科研论文中抽取该论文的标题、作者、隶属单位、摘要等头部信息；从新闻报道中抽取出恐怖事件的详细信息：时间、地点、作案者、受害者、袭击目标、袭击方式等；从新闻文本中抽取职务变动事件的详细情况：人员、组织、职位、时间等。通常，抽取出来的信息以结构化的形式表示，可以直接存入数据库中，供用户查询或进一步分析利用。

面对日益增加的文本数据，开展中文文本信息抽取的研究有着重要的理论研究价值和现实意义，有着广阔的应用前景，对推动我国的信息化进程和维护国家信息内容安全意义重大。下面从三个方面进一步阐述。

（1）从满足用户信息需求的角度看，文本信息抽取是其他信息获取手段的一种有益补充。面对互联网时代庞杂无序的海量信息，用户要在信息海洋里查找信息，就象大海捞针一样。搜索引擎服务在一定程度上满足了用户查找信息的需要，但 Internet 的飞速发展和搜索引擎日趋庞大，进一步凸显出海量信息和人们获取所需信息能力之间的矛盾——搜索结果仍然充满大量的无用信息，获取全面、准确、有效的信息仍然需要花费大量的时间和精力。同样，其他信息获取技术，如信息检索（Information Retrieval，IR）、文本分类、文本过滤、文本聚类等仅仅可以从一个大的文档集合中

找出用户需要的相关文档，而信息抽取技术却可以从相关文档中抽取出粒度更小、精度更高的信息，满足用户更深层次和更具体的信息需求。

（2）从技术实现的角度看，文本信息抽取可以为其他信息获取技术提供支持。信息抽取作为一种将无结构化信息转换为结构化信息的一种手段，为进一步的信息处理打下了基础。文本信息抽取对搜索引擎、问答系统（Question Answering, QA）、文本数据挖掘、网络信息安全、自动文摘、数字化图书馆、生物信息学、商业智能系统等许多应用领域的实现起功能上的支持作用[5][6][7][8][9]，或者有助于提高它们的性能。

（3）从维护国家信息安全的角度看，面向中文文本的信息抽取研究有重大社会价值。在当今信息爆炸式增长的态势下，全社会的信息化进程对语言信息处理技术提出了强烈的需求，这是容易理解的。因为信息的表述、传播、转换都要依赖语言文字作为最主要的载体。当前任何一个国家、民族、社团，要跟上文明进步的节奏，就必须努力使自己所使用的语言文字的计算机处理技术与全社会的信息化同步前进。在我国，中文信息处理已经进行了大量卓有成效的研究并取得了许多成果[①]，同时也面临着一系列挑战[10][11][12]。作为中文信息处理研究的一部分，中文文本信息抽取的研究对促进中文信息处理和维护国家信息安全意义重大。

文本信息抽取虽然是一种用于处理各种类型文本文档的有效方法，然而建立一个信息抽取系统却是非常费时费力的。早期出现的信息抽取系统往往依赖于人们手工建立的抽取规则或模式[13]，然而，通过手工创建的规则很难保证具有整体的系统性和逻辑性，并且这些规则一般具有高度的领域相关性和较差的可移植性[14][15]。因此，迫切需要探索一些更加有效的方法来实现文本信息抽取。而信息抽取中语言模型和机器学习方法的研究对提升信息抽取系统的可移植性、普适性和抽取性能意义重大。

语言模型（language model）是自然语言处理（Natural Language Processing, NLP）的基础和核心。要实现计算机对自然语言的处理，就必须采用数学的或逻辑的方法对自然语言进行精确的描述和刻画，以便用计算机对其进行自动处理。这种对语言进行描述和刻画的数学公式或形式系统称为语言模型。文本信息抽取作为自然语言处理的一个应用领域，语言模型在其中的作用不言而喻，所以对中文文本信息抽取中语言模型的研究非常必要。

机器学习（Machine Learning）是研究计算机怎样模拟和实现人类的学习行为，建立能够通过学习自动提高自身水平的计算机算法的理论方法的学科。它是人工智能

[①] 见中文信息学会成立二十五周年学术会议《中文信息处理重大成果汇报展》资料汇编。

的核心，也是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。为了提高信息抽取系统的性能和效率，国内外也进行了大量信息抽取中机器学习方法的研究[16][17][18]。本课题面向中文文本信息抽取的机器学习方法及其应用的研究对深入广泛地应用信息抽取技术必将有深远的影响。

综上所述，中文文本信息抽取是一个具有广泛应用前景的研究方向，而它的研究又离不开语言模型和机器学习方法的研究，语言模型理论是指导更好地开展这一研究的基础，机器学习方法对提高信息抽取系统的可移植性、普适性作用重大。并且语言模型和机器学习又是相辅相成的，互相支持的。本课题就是围绕文本信息抽取这一研究热点，深入研究能够实现自动信息抽取的语言模型和机器学习方法，并在半结构化文本和自由文本数据集上验证这些模型和方法的性能。

1.2 文本信息抽取基础

1.2.1 信息抽取和信息检索的区别

与信息抽取密切相关的一项研究是信息检索，但信息抽取与信息检索存在许多差异[3][19][20]，主要表现在如下三个方面：

① 功能不同。信息检索系统主要是从大量的文档集合中找到与用户需求相关的文档列表；而信息抽取系统则旨在从文本中直接获取用户感兴趣的事实或事件信息。

② 处理技术不同。信息检索系统通常利用统计及关键词匹配等技术，把文本看成词的集合，不需要对文本进行深入分析理解；而信息抽取往往要借助自然语言处理技术，通过对文本中的句子以及篇章进行一定的分析处理后才能完成。

③ 适用领域不同。由于采用的技术不同，信息检索系统通常是领域无关的，而信息抽取则是领域相关的，只能抽取系统预先设定好的特定类型的事实事件信息。

另一方面，信息检索与信息抽取又是互补的。为了处理海量文本，信息抽取系统通常以信息检索系统的输出作为输入，可以看作对信息检索获得的信息进一步加工的过程[21]；而信息抽取技术又可以用来提高信息检索系统的性能。二者的结合能够更好地服务于用户的信息处理需求。另外信息抽取技术并不试图全面理解整篇文档，只是对文档中包含特定类型信息的部分进行分析，即一般只需要浅层的句法分析和语义分析。所以从某种意义上说，信息抽取技术又是完全文本理解的基础[22]。

1.2.2 信息抽取的分类

可以将信息抽取的抽取对象分为结构化信息、半结构化信息和无结构信息[5][23]。相应地，按照待抽取对象的不同，信息抽取可以分为三类：结构化信息抽取、半结构化信息抽取、无结构信息抽取。

结构化信息是指具有良好的结构布局，通常指存放在数据库中的数据信息，或者根据事先规定的格式生成的文本数据。其中待抽取的目标信息易于通过一个固定的模式进行抽取，抽取的准确度较高[23]。

半结构化信息是指在一定程度上具有某种结构特征的信息。该类信息的结构化程度比结构化信息弱，但比无结构化信息好。其中的待抽取目标可通过基于分隔符或某种特定标签来定位。网页由于缺少规范的语法结构，并且包含有大量的标签信息，所以一些研究者将其归入半结构化信息[24]。而有些自然语言形式的文本数据，由于具有一定的结构特征和一些特定的标签信息，本文将这类文本数据称为半结构化文本，也归入半结构化信息之列。

无结构信息指新闻报道、纪实文学、研究报告等自由文本，通常由符合某种语言表达规范的自然语言语句组成。从这类文本数据中抽取特定类型的信息具有很大难度，涉及到自然语言处理领域的诸多问题。

综上所述，细化之后的信息抽取分类情况如图 1.1 所示。本文研究的文本信息抽取包括半结构化文本信息抽取和自由文本信息抽取两类（粗体部分），并且是面向中文文本的，在后面的章节里，如无特殊说明所处理的文本均指中文文本。

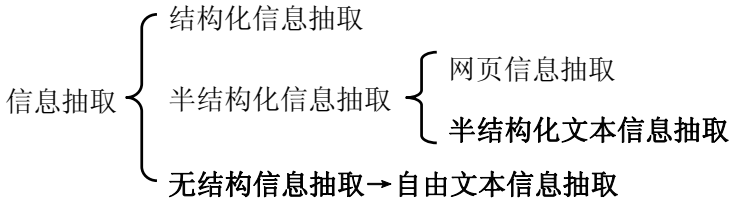

图 1.1 信息抽取分类

1.2.3 文本信息抽取的目标

文本信息抽取的目标是从自然语言形式的文本中抽取特定类型的文本信息片断，并将这些信息转换为结构化的数据存储以便进一步利用。本文所研究的文本信息抽取

分为两类，一类是半结构化文本信息抽取，通常从这类文本中抽取一些连续的信息域，这些连续的信息域可通过信息之间的分隔符或特定标识符等结构特征来辅助进行抽取。例如，从科研论文中抽取论文的头部信息，论文的头部信息包括论文的标题、作者、隶属单位、摘要、关键词、中图分类号等十几个信息域。如图 1.2 所示。

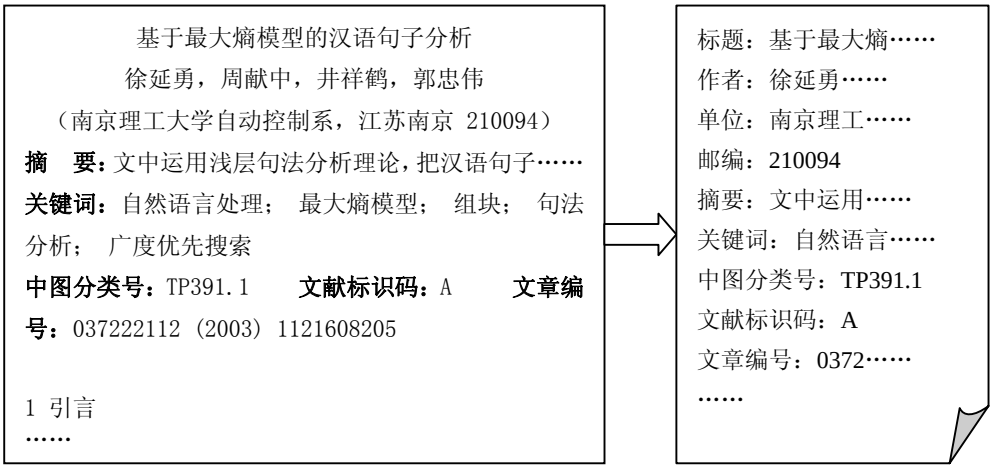

图 1.2 半结构化文本信息抽取示例

另一类是自由文本信息抽取，主要研究从这类文本中探测特定类型的事件并抽取相关的事件信息，这就是所谓的文本事件信息抽取（简称事件抽取，event extraction），是本课题研究的重点部分。例如从职务变动相关的新闻报道中抽取“职务变动”类的事件信息。如图 1.3 所示。

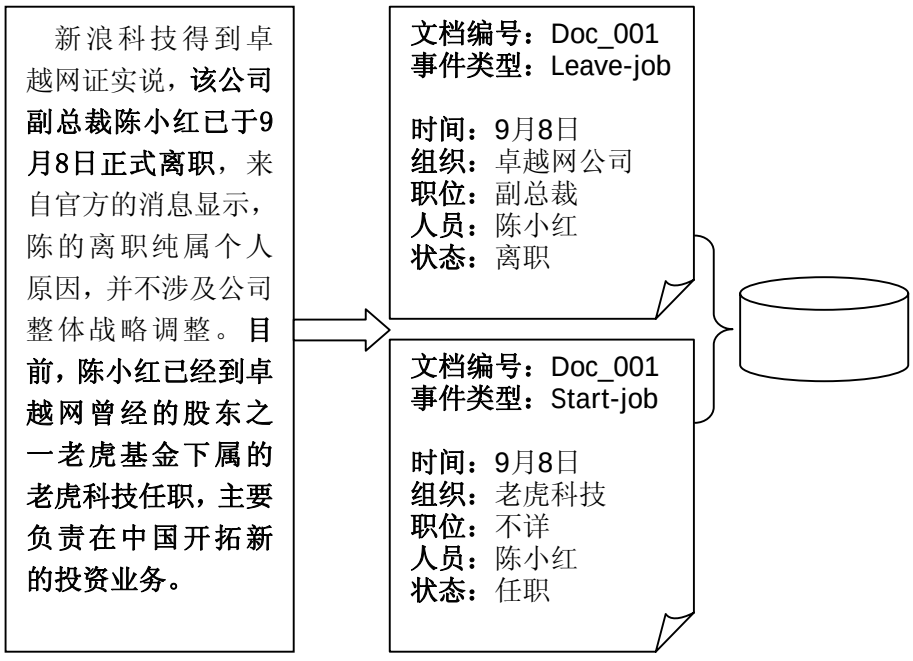

图 1.3 文本事件信息抽取示例

1.2.4 文本信息抽取的基本概念

为了进一步对文本信息抽取加深认识，本节介绍该领域一些基本概念：命名实体、实体关系、场景模板、主题领域、抽取模式、事件表述语句、事件要素等。并通过一些示例来阐述这些概念的含义。

命名实体（Named Entity）是文本中基本的信息元素，是正确理解文本的基础。狭义地讲，命名实体是指现实世界中的具体的或抽象的物体，包括各种专有名词、时间词、数量词和名词词组等，如人、组织、公司、地名、职位名等。命名实体识别（Named Entity Recognition，NER）就是从文本中识别并分类命名实体的过程。例如，图 1.3 左侧的文本经过命名实体识别之后被标注如下：

> 新浪科技/ORG 得到卓越网/ORG 证实说，该公司/PRO 副总裁/POS 陈小红/PER 已于 9 月 8 日/DATE 正式离职，来自官方的消息显示，陈/NOM 的离职纯属个人原因，并不涉及公司整体战略调整。目前，陈小红/PER 已经到卓越网/ORG 曾经的股东之一老虎基金/ORG 下属的老虎科技/ORG 任职，主要负责在中国/LOC 开拓新的投资业务。

实体关系（Entity Relation）是指信息抽取过程中所抽取出的各种命名实体之间、实体及其属性之间的各种关系。如组织和地点之间的位于关系（location-of）、组织和雇员之间的雇佣关系（employee-of）、公司和产品之间的生产关系（product-of）、“上下级”之间的从属关系（affiliation-of）等。例如，图 1.3 左侧文本中的命名实体之间就存在雇佣关系、从属关系等。实体关系抽取（Entity Relation Extraction）就是从文本中识别出实体之间或实体及其属性之间的关系的过程。

模板（Template）是信息抽取系统抽取出的事实或事件信息输出时所采用的结构化形式，由一系列槽（Slot）组成。待抽取的特定事实（关系）或事件称为一个场景（Scenario），例如职务变动事件、恐怖袭击事件、自然灾害事件等都是特定的场景。而主题领域（Subject Domain）的概念要广一些，指被处理的一类文本。通常一个领域可以包含多个场景，如 IT 领域可以包含有职务变动场景、推出新产品场景、公司合并场景等。简单地说，场景模板（Scenario Template，ST）任务是指从文本中抽取出感兴趣的信息块（事实或事件），并填入模板的过程。由此可见事件信息抽取是场景模板任务的主要部分。

抽取模式（Extraction Pattern）是信息抽取过程中使用的匹配规则，是句法形式到

语义含义的映射。抽取模式可以传递特定领域中待抽取的事实事件信息，即有相应语义含义的句法形式。例如，图 1.3 的示例中可用的抽取模式有：①某公司总裁×××离职；②×××到某公司任职。

事件表述语句（event mention sentence）是指对一件客观发生的事件的参与者、发生时间、发生地点、原因、过程和结果等所进行描述的语句。本文将事件表述语句中具体的参与者、发生时间、发生地点、原因、过程和结果等信息称为事件要素（event argument），有些文献也可以称为事件论元、事件变元等。事件信息抽取的过程就是从自然语言形式的文本中找到一个具体事件的表述语句并识别出该事件的事件要素的过程。下面是三个“职务变动”类事件的表述语句。

（1）12 月 26 日，在中共安阳市第九届委员会第一次全会上，靳绥东当选安阳市委书记。

（2）1 月 1 日，球王贝利被巴西新总统任命为体育部长。

（3）白宫首席大厨称因无法满足第一夫人要求被解雇。

1.3 信息抽取研究的发展历程

信息抽取的研究最早开始于 20 世纪 60 年代中期，这被看作是信息抽取技术研究的初始阶段，它以两个长期的、研究性的自然语言处理项目为代表 [20][25][26]。一个是美国纽约大学开展的Linguistic String项目 [25]，该项目开始于 60 年代中期并一直延续到 80 年代。它的主要研究内容是建立一个大规模的英语计算语法，与之相关的应用是从医疗领域的X光报告和医院出院记录中抽取结构化信息。

另一个相关的长期项目是由耶鲁大学Roger Schank及其同事在 20 世纪 70 年代开展的有关故事理解的研究。由他的学生Gerald De Jong设计实现的FRUMP系统 [26]是根据故事脚本理论建立的一个信息抽取系统。该系统从新闻报道中抽取信息，内容涉及地震、工人罢工等很多领域或场景。FRUMP系统把有线新闻网络作为数据源，使用一些新闻故事的简单脚本来对有线新闻网络进行监测。

除了这两个典型的系统外，在 1981 年，Cowie研制了一套系统 [20]，主要从和动植物相关的正规结构描述中抽取一些简单信息填入一个具有固定记录格式的数据库中。而ATRANS是一个商品化产品系统 [20][27]，主要用于处理国家银行中金融转帐信息，采用类似于FRUMP系统的概念句子分析技术。

从 20 世纪 80 年代末 90 年代初，信息抽取研究进入了快速发展阶段。在信息化浪

潮的推动下，西方发达国家都十分重视信息抽取技术的研究和应用，把它列为与信息检索、自然语言理解、文档分类和摘要、语音识别等并重的语言工程项目。美国政府有专门的文本处理研究计划（例如 Tipster 计划），内容包括了信息抽取、文档检索、文献摘要等，以期提高政府部门的信息处理速度和质量。美国许多大学和公司的研究机构也都开展了有计划的、长期的、系统的信息抽取及其应用研究工作，并且有专门的机构组织各种评测活动对当前的研究进展进行评估，例如著名的消息理解系列会议（Message Understanding Conference，MUC）以及后来的自动内容抽取（Automatic Content Extraction，ACE）评测会议等。这些评测极大地推动信息抽取研究向前发展。

1.3.1 MUC 系列评测会议

MUC 是美国政府支持的一个专门致力于真实新闻文本理解的评测会议。从 1987 年到 1998 年，共举行了七届，它由美国国防高级研究计划委员会（DARPA，the Defense Advanced Research Projects Agency）资助。MUC 评测的目的是为信息抽取研究提供公共测试平台。从历次 MUC 会议，可以清楚地看到信息抽取技术发展的历程。

1987 年 5 月举行的首届 MUC 会议基本上是探索性的，没有明确的任务定义，也没有制定评测标准，总共有 6 个系统参加评测，所处理的文本是海军军事情报，每个系统的输出格式都不一样。

MUC-2 于 1989 年 5 月举行，共有 8 个系统参加，处理的文本类型与 MUC-1 一样。MUC-2 开始有了明确的任务定义，规定了模板以及槽的填充规则，抽取任务被明确为一个场景模板填充的过程。

MUC-3 于 1991 年 5 月举行，共有 15 个系统参加，抽取任务是从新闻报告中抽取拉丁美洲恐怖事件的信息，定义的抽取模板由 18 个槽组成。从 MUC-3 开始引入正式的评测标准，如召回率（Recall）和准确率（Precision）等。

MUC-4 于 1992 年 6 月举行，共有 17 个系统参加，任务与 MUC-3 一样，仍然是从新闻报告中抽取恐怖事件信息。但抽取模板变得更复杂了，总共由 24 个槽组成。

MUC-5 于 1993 年 8 月举行，共有 17 个系统参加。此次会议设计了两个目标场景：金融领域中的公司合资情况、微电子技术领域中四种芯片制造处理技术的进展情况。MUC-5 的一个重要创新是引入了嵌套的模板结构，信息抽取模板不再是扁平结构（flat structure）的单个模板，而是借鉴面向对象和框架知识表示的思想，由多个子

模板组成。

MUC-6 于 1995 年 9 月举行，训练时的目标场景是劳动争议的协商情况，测试时的目标场景是公司管理人员的职务变动情况，共有 16 家单位参加了这次评测。MUC-6 的评测更为细致，强调系统的可移植性以及对文本的深层理解能力。除了原有的场景模板填充任务外，又引入三个新的评测任务：命名实体识别、共指（Coreference）关系确定、模板元素（Template Element，TE）填充等。

最后一届MUC会议——MUC-7 于 1998 年 4 月举行。训练时的目标场景是飞机失事事件，测试时的目标场景是航天器（火箭/导弹）发射事件。除MUC-6 已有的四项评测任务外，MUC-7 又增加了一项新任务——模板关系任务。共有 18 家单位参加了 MUC-7 评测。其中台湾大学的资讯工程研究所的自然语言处理实验室在MUC-7 上提交了一个TE系统参加了测试，测试了中文命名实体（人名、地名、时间等名词性短语）的识别，取得了与英文命名实体识别系统相近的性能 [28]。

MUC系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用。MUC定义的信息抽取任务以及确立的评价体系已经成为信息抽取研究事实上的标准。到MUC-7 时信息抽取研究的内容按照不同层次分为五大部分 [29]：

- 命名实体识别（Named Entity Recognition, NER），抽取文本中出现的命名实体并进行分类。
- 模板元素填充（Template Element Construction, TE），找出所有特定类型的命名实体及其属性。
- 指代消解（Coreference Resolution），找出文本中指向同一个实体的词或词组。
- 模板关系填充（Template Relation Construction, TR），抽取文本中两个实体间或实体及其属性之间的特定关系。
- 场景模板任务（Scenario Template, ST），是最高级别的信息抽取任务，从文本中发现特定类型的事件并填充进模板中。

通常将上面的五大部分任务归并为 3 个不同的任务（由简到难）：命名实体识别、实体关系抽取、场景模板任务。在MUC系列评测会议中英文信息抽取的各项指标（最好水平）大体上如下 [28]：命名实体识别 90%，属性识别 80% (TE任务)；事实识别(Facts) 70% (TR任务)；事件识别(Events)60% (ST任务)。这些指标也自然地反映了英文中自然语言处理在各个层次上的难度。

在最后一届MUC（MUC-7）上表现最好的是SRA公司的系统 [28][30]，其所有的三

项IE指标都是最高的。其评测结果如表 1.1 所示：

表 1.1 MUC-7 中 SRA 公司系统的评测结果

	Recall	Precision	F-Score
TE	86%	87%	86.76
TR	67%	86%	75.63
ST	**42%**	**65%**	**50.79**

1.3.2 ACE 系列评测会议

随着MUC会议的停办，由美国国家标准技术局（NIST）组织的ACE评测会议成为推动信息抽取研究的主要动力 [3][4]。ACE评测的目标是推动自动内容抽取技术的发展，从而实现文本形式的自然语言的自动处理。ACE评测 1999 年 7 月开始酝酿，2000 年 12 月正式启动，迄今已经举办过七次评测（2000 年 5 月、2002 年 2 月、2002 年 9 月、2003 年 10 月、2004 年 8 月、2005 年 11 月、2007 年 1 月）。ACE 评测提供的语料不仅有英文语料，还包括中文和阿拉伯文语料。目前，除强烈的应用需求外，ACE 评测会议是推动信息抽取研究进一步发展的主要动力。

经过前面几届ACE评测会议，ACE将评测任务进行了合并和重新划分。ACE2007（ACE07）评测计划 [31]中有五个主要评测任务：实体探测与识别（EDR，Entity Detection and Recognition）、特定类型数值探测与识别（VAL，Value Detection and Recognition）、时间表达式探测与识别（TERN，Time Detection and Recognition）、关系探测与识别（RDR，Relation Detection and Recognition）和事件探测与识别（VDR，Event Detection and Recognition）。另外还有三个提及级（mention-level）的任务：实体提及探测（Entity Mention Detection）、关系提及探测（Relation Mention Detection）和事件提及探测（Event Mention Detection）。其中，对每类任务都进行了详细的定义和划分，详细信息可参考ACE07 评测计划 [31]。

1.3.3 信息抽取系统的评测

在对信息抽取系统抽取性能进行评估时，通常采用 3 个评测指标 [3][4]：准确率（Precision，P）、召回率（Recall，R）、综合指标F值（F）。准确率表示在抽取的全部信息条数中，正确的所占的比值。召回率是指在所有应该抽取出的信息中(包括得到的和不应该忽略的)，正确抽取出的信息条数所占的比值。准确率描述系统抽取的信

息中，正确的、有用的占多少。召回率表示应该得到的信息中，系统抽取出了多少。计算公式如下：

$$P=\frac{\text{准确抽取的信息条数}}{\text{抽取出的所有信息条数}}, \quad (1.1)$$

$$R=\frac{\text{准确抽取的信息条数}}{\text{所有正确的信息条数}}, \quad (1.2)$$

两者取值在 0 和 1 之间，通常存在反比的关系，即P增大会导致R减小，反之亦然。一般来说，对于一个信息抽取系统，片面追求一个指标的提高而忽视另一个指标是无意义的，应该同时追求较大的P和R。所以，实际评估一个系统时，应同时考虑P和R，但同时要比较两个数值，很难做到一目了然。许多人提出综合两个值进行评价的办法，综合指标F值就是其中一种。计算公式[3][4]如下：

$$F=\frac{(\beta^2+1)PR}{(\beta^2P+R)}, \quad (1.3)$$

其中 β 决定对 P 侧重还是对 R 侧重，通常设定为 1、2 或 1/2。若 β ＝1，则将 P 和 R 视为同等重要，本文中如没有特殊声明 β 取值均为 1；若 β ＝2，则 R 的重要程度是 P 的 2 倍；若 β ＝1/2，则 P 的重要程度是 R 的 2 倍。

1.4 文本信息抽取的研究现状

下面首先对中文文本信息抽取研究现状简要概述并分析中文的一些特点，接着，由于从半结构化文本和自由文本中抽取信息存在较大的差异，所使用的具体抽取方法及技术也不相同，所以分别对半结构化文本信息抽取和自由文本信息抽取的研究现状进行介绍。

1.4.1 中文信息抽取研究概要

目前，由于强烈的应用需求，文本信息抽取已成为自然语言处理和文本数据挖掘领域的研究热点。国内外投入了大量的人力物力财力从事相关技术的研究。国外在信息抽取研究上由于开始较早，所以研究的深度和广度要大的多，并有许多成型的系统产生。国内的信息抽取研究起步较晚，并且大部分研究工作集中在命名实体识别[32][33][34][35][36][37][38][39][40][41]，近些年来，较深层次的信息抽取研究工作和成果也大量涌现，如实体关系研究[5][37][42][43][44][45][46]、指代消解研究[47][48][49][50][51]、事件信息抽取

研究[52][53][54]、信息抽取与知识库[55]、信息抽取中的机器学习方法研究[16]等。并且设计并实现完整的中文信息抽取系统也处于探索阶段[5][56][57][58]。

现有的自然语言处理理论和技术大多都是以英语为研究对象发展起来的，而汉语无论在语音、文字上，还是在词汇、语法、语义及其语用等各个层面上都与英语存在着较大的差异。这使得无法完全套用英语文本信息抽取中已成熟的一些理论和技术，在面向中文文本信息抽取的语言模型和机器学习方法研究中应该注意中文文本或汉语的如下特点[11][59]：

（1）中文文本中的词、短语之间界限不清晰。汉语是以字为基本单位，词之间没有明显的标记。所以中文信息处理的基础课题和特有的问题就是中文分词，分词本身也有一定的错误率[60]，并且可以导致部分语义信息的损失[61]，这无疑降低了后续处理的实际效果。

（2）词类缺乏形态标志和形态变化。这主要表现在两方面：一方面在汉语中，不能从词形上分辨出哪个是名词、动词、形容词等；另一方面词语在句子中担任不同成分和时态不同时，在形式上也完全一样，没有形态变化。

（3）在中文句子中，词类与句法成分是一对多的对应。而在印欧语系里，词类与句法成分基本上是一对一的对应关系。

（4）语义灵活，一方面语法的灵活主要来源于语义的灵活；另一方面同一结构可以表达不同的意思，同一意思可以用不同结构表达。

（5）语序固定，语序成为汉语表示语法意义的重要手段。汉语的基本语序是[59]：主语在谓语之前，宾语在动词之后，修饰语在中心语之前，补语在动词或形容词之后。语序变动，结构关系和意义随之改变。

（6）只要语境允许，句法成分，包括重要的虚词，都可以省略。

1.4.2 半结构化文本信息抽取的研究现状

半结构化的文本数据一般不具有完整的句法结构，但具有明显的版面布局结构和特定的标签信息。并且这些数据一般布局紧凑，通常是一个连续的序列。针对半结构化文本的这些特点，可以将半结构化文本信息抽取看成序列数据标注，并使用统计语言模型来实现。另外，随着基于机器学习的分类算法的逐渐完善，也有一些学者开始将这类问题转化为分类问题来求解[62]。

统计语言模型也称概率模型或经验模型，主要是利用概率论、数理统计和信息论方法，从大规模语料库中获取蕴含在其中的知识，并依据这些知识构造用于信息抽取的语言模型。这类方法的最早的形式是Hobbs[63]提出的基于有限状态自动机的信息抽取方法。该方法使用有限状态自动机对文本进行标注，从而识别不同类型的信息域。随后，Hsu等人[24]利用单层或多层的有限状态传感器判断目标信息的边界，并将该方法用于网页信息抽取。

基于有限状态自动机的方法不需要对文本进行复杂的自然语言处理，抽取效率高并且易于开发。然而，这类方法需要对有限状态自动机的结构进行手工调节，对不同类型的文本数据的适应能力差。因此，学者们转而利用另一种有限状态自动机——隐马尔可夫模型（Hidden Markov Model，HMM）[64]来实现半结构化文本信息抽取，该模型可以通过机器学习方法自动学习模型结构和参数。

国内外对基于隐马尔可夫模型的半结构化文本信息抽取进行了大量的研究和实验。例如，利用HMM模型对科研论文头部信息和引文信息进行抽取[65][66][67][68]；基于HMM模型的学术报告信息抽取[69][70][71]；利用HMM模型从电子版的个人简历中抽取信息[72]等等。

在基于HMM的信息抽取方法中，文本切分的粒度、HMM模型的结构和数据稀疏问题的解决是影响信息抽取准确性的主要因素。大部分基于HMM的信息抽取方法采用词语作为文本特征，并利用词语在文本中出现的词频来计算概率值。HMM的模型结构是体现文本上下文关系的重要部分。一些学者通过对数据的分析手工确定HMM的模型结构[73]。但是，手工确定的结构不仅难以扩充到大规模的数据集，而且人对数据的先验知识往往与真实的数据不符合。因此，很多学者研究了HMM结构的自动学习方法[67][70]。与自然语言文本中较大的词语特征维数相比，HMM使用的训练数据常常过于稀疏，从而导致学习模型参数时缺乏足够的训练数据。为了解决这一问题，人们通常使用平滑方法解决在训练语料中没有出现的词语的输出概率。

这些研究都把半结构化文本信息抽取看成是一个序列标注问题，并采用 HMM 模型来实现。尽管 HMM 被广泛应用于序列数据标注中，然而作为一种生成模型(Generative Model)，它并不是最合适的模型。HMM 通过定义标记序列和观察序列的联合概率 P(S,O)来搜索最佳的标记序列，定义这样的联合概率意味着所有可能的观察序列都应该被枚举出来，然而如果观察元素间具有长距离依赖性，这个任务将是很困难的。因此，为了保证推导的正确性，HMM 进行了两个独立性假设：一是假设每个状态仅仅依赖于它

的前一时刻状态；二是假设时刻 t 的观察值仅仅依赖于时刻 t 的状态。这两个假设简化了推导，但同时也降低了标注性能。事实上，真实的序列数据不仅存在长距离依赖性，而且观察序列中各种有利于提升标记序列精确度的上下文特征也未能被整合到模型中去。因此，研究更适合于半结构化文本信息抽取的语言模型非常必要。另外已有研究的训练语料和测试语料基本上都是英文语料，针对中文半结构化文本的特点进行适当的调整也相当必要。

1.4.3 自由文本信息抽取的研究现状

自由文本信息抽取的研究非常多，在各个层面上都有一些的研究，但是相对而言事件信息抽取层面上的研究较少。已有的文本事件抽取方法概括起来有两类：基于抽取模式的事件信息抽取和基于触发词探测的事件信息抽取。除此之外，北京大学的孙斌提出了基于“状态跃迁链”的事件抽取方法，并开发了一个信息提取原型系统InfoX[56]。

1.4.3.1 基于抽取模式的信息抽取方法

传统的信息抽取系统几乎都是基于模式匹配的，其核心问题是抽取模式的获取[14][74][75][76][77]。为了进行抽取模式的学习，人们先后在不同的信息抽取系统中采用过各种抽取模式获取方法，按照这些系统所需要的用户辅助工作的不同和对用户工作量大小和技能要求高低的不同，可将这些系统分为五类：

（1）手工构建抽取模式的信息抽取系统

这些系统主要出现在MUC评测会议的早期，它们采用手工的方式定制抽取模式或模式结构，系统移植性较差。这类系统有PLUM[78]、FASTUS[79]、GE NLTOOLSET[80]、PROTEUS[81]等。

（2）基于人工语料标注的抽取模式学习系统

这些抽取模式学习系统进行模式学习的一般做法是：设计一种信息抽取模式表示方式；人工标注训练语料；使用一种机器学习方法从中学出相应的信息抽取模式。这类系统有AutoSlog[82]、PALKA[83]、CRYSTAL [84]、LIEP[85]、HASTEN[86]等，以AutoSlog和PALKA最为典型。

（3）基于人工语料分类的抽取模式学习系统

这类模式学习系统首先由人工对训练语料进行领域相关与否的分类，然后根据人工提供的这种粗浅的分类，从这些训练语料学习出领域相关的抽取模式。典型的系统为AutoSlog-TS[87]，该系统是AutoSlog的后继产品。

（4）基于 WordNet/HowNet 和语料标注的抽取模式学习系统

这类系统的共同之处是在学习抽取模式时要用领域无关的概念层次知识库WordNet或HowNet作为支持。该类信息抽取模式学习系统的典型代表是TIMES[88]和GenPAM[5]，GenPAM是中科院计算所的姜吉发提出的一种基于领域无关概念知识库的事件抽取模式学习方法。姜吉发 [5]将GenPAM用于从MUC-7 提供的英语飞行事故训练语料中进行抽取模式获取研究，并在这些模式指导下从MUC-7 提供的测试语料中进行坠机事件的抽取实验，平均抽取结果F值得到了 63％。

（5）基于种子模式的自扩展抽取模式获取系统

这种信息抽取模式学习系统的代表是ExDisco[76][89]、Snowball[90][91][92]等，这两个系统获取抽取模式的机理都是基于种子模式的自扩展（bootatrapping）方法。其中ExDisco用于从英文文本中获取事件抽取模式，然后基于这些模式进行事件信息抽取。而Snowball用于从英文文本中获取实体关系的抽取模式，利用这些关系模式可以从新的文本文档中发现新的实体关系。 基于ExDisco的“职务变动”场景模板任务的抽取性能如表 1.2 所示 [76]：

表 1.2 ExDisco 在“职务变动”场景下的抽取性能

抽取模式库	训练结果			测试结果		
	P	R	F	P	R	F
Seed	0.830	0.380	0.526	0740	0.270	0.396
ExDisco	0.800	0.620	0.699	0.720	0.520	0.602
Union	0.790	0.690	0.735	0.730	0.570	0.636
Manual-MUC	0.710	0.540	0.619	0.700	0.470	0.564
Manual-NOW	**0.790**	**0.690**	**0.739**	**0.750**	**0.560**	**0.640**

另外山西大学的郑家恒等 [93]采用聚类的方法自动生成针对中文文本的信息抽取模式，通过计算模式实例间的相似度来划分实例类别，以农作物信息文本为实验语料进行了抽取模式的获取研究。

虽然抽取模式的自动获取研究已经有很多，但仍然有许多问题需要进一步研究，其中尤其以自动获取模式的机器学习方法和候选模式的评估最为重要。而针对中文的事件抽取模式获取研究较少，所以中文文本事件抽取模式获取的研究就尤为重要。

1.4.3.2 基于触发词探测的事件信息抽取方法

基于触发词探测的事件信息抽取思想是综合ACE评测 [31]对事件探测与识别的要求提出来的。ACE评测会议对事件探测与识别任务（VDR）进行了详细规定，要求从文本文档中探测和识别特定类型的事件，并将事件信息抽取分为事件属性的识别、事件要素及其角色的标注、事件提及的识别等三个方面。综合ACE评测对事件抽取的要求将事件抽取分为两步，第一步是特定事件的探测和事件的分类，主要探测特定事件的表述语句并确定事件的类别或子类别；第二步是从事件表述语句中识别出事件的要素及其语义角色并填充到预定义的事件模板中。

事件探测和事件分类是事件抽取的基础，事件探测旨在发现特定类型事件的表述语句，这些语句是进一步抽取的数据源。而事件分类是用于确定该事件表述语句所叙述的事件类别，确定的事件类别正确与否对事件模板的选择以及究竟要抽取哪些事件要素来填充模板至关重要。传统的事件探测和事件分类主要依据触发词（trigger）来确定 [54]，触发词是能够很好地概述出该事件的中心意义的词。例如，职务变动事件中的“任命”、“辞去”等词语。基于触发词的事件探测和分类是将含有特定触发词的语句作为候选事件语句并依据触发词对事件进行分类。例如，文本中的语句“方正集团董事长魏新辞去方正科技董事长职务”包含触发词“辞去”，就认为该语句是个离职类事件的表述语句。这样仅仅依据触发词来进行事件分类效果不佳，需要寻找更适合的方法来确定事件表述语句的类别。

从事件表述语句中识别事件要素并判定其语义角色是文本事件信息抽取的难点，其中涉及到自然语言处理中许多核心问题，特别是动词的论元结构 [94][95][96]和语义角色标注 [97][98][99]。论元结构理论在国内外语言学界已经有大量研究，但是将这一理论应用到自然语言处理领域的研究较少，应用到文本信息抽取的研究就更少。语义角色标注是近几年自然语言处理领域中的研究热点之一，国内外已经进行了一些卓有成效的研究和实验 [97][98][100][101][102][103][104]。目前人们大多采用统计学习的方法解决语义角色标注问题。基本的思想是把句子中连续的句子成分作为标注的基本单元，然后根据一定的语言学知识列出该单元的各种特征，并与该单元的语义角色类型组成学习的实例，最后使用某种统计学习方法对这些实例进行自动的学习，以便对新的实例进行预测。

使用统计学习的方法进行语义角色标注，离不开标注好语义角色的语料资源，英

语中较为知名的有FrameNet[105]和PropBank[106]两种。对于汉语的语义角色标注资源也有一些，Chinese PropBank是UPenn基于Chinese Penn TreeBank标注的汉语浅层语义标注资源[②]；台湾科学研究院的中文句法结构树库（Sinica TreeBank）[107]，利用74种语义角色，在句法关系链上添加了语义标签；山西大学的刘开瑛等人[③]正在构建的汉语框架语义知识库（Chinese FrameNet，简称CFN）是一个以框架语义学为理论基础、以真实语料为事实依据的语义词典，其资源用语义Web标记语言描述。

综上所述，分析事件表述语句中动词（大部分是触发词）的论元结构，并探讨事件表述语句中的语义角色标注问题对准确抽取特定类型事件的事件要素并判定其语义角色至关重要。

1.5 论文的研究内容和组织结构

1.5.1 研究内容

本文围绕文本信息抽取这一研究热点，针对汉语的特殊性，对中文文本信息抽取中的语言模型和机器学习方法进行了深入的研究和探索。重点研究两类文本信息抽取，一类是半结构化文本信息抽取；另一类是自由文本中特定类型的事件信息抽取。具体研究内容如下：

（1）为提高半结构化文本信息抽取的性能，研究能够综合待抽取信息的版面结构等上下文特征，并充分利用半结构化文本中的分隔符和特定标识符的语言模型来实现半结构化文本信息抽取。

（2）为了增加信息抽取系统的可移植性，研究中文自由文本中事件抽取模式的自动获取方法和候选模式的评估方法，并将这些抽取模式用于中文文本事件信息抽取。很多模式获取技术是基于有指导学习的，所以往往需要利用大量有标注的数据。显然大量标注数据所需的代价往往是昂贵的，怎样从少量的标注数据甚至不标注数据自动学习获取事件抽取模式是需要研究的问题。

（3）对大量事件表述语句研究发现：仅仅依据触发词就判定一个语句是某类候选事件语句很容易出错，而触发词的上下文中包含了对事件类别确定有重大价值的各类特征。为确定候选事件语句具体表述的事件类别，研究利用最大熵原理，建立统计

[②] 详见 http://www.cis.upenn.edu/~Chinese/
[③] 见刘开瑛在中文信息学会二十五周年学术研讨会上的讲稿“汉语框架语义知识库构建工程介绍”。

语言模型，选择合适的特征用于事件表述语句类别确定。

（4）为实现事件表述语句中的事件要素及其角色的识别，研究和分析事件表述语句中触发词的论元结构，并结合事件要素上下文特征利用统计语言模型进行事件要素及其语义角色的识别。

（5）针对一些简单的事件表述语句，研究采用隐马尔可夫模型进行事件信息的抽取。为每一类事件要素构建一个独立的 HMM 模型用于这类要素的抽取，用机器学习的方法从训练语料中学习模型的结构。

1.5.2 论文组织

基于以上的研究内容，本论文由以下八章构成，论文的组织及各部分之间的关系如图 1.4 所示，具体安排如下：

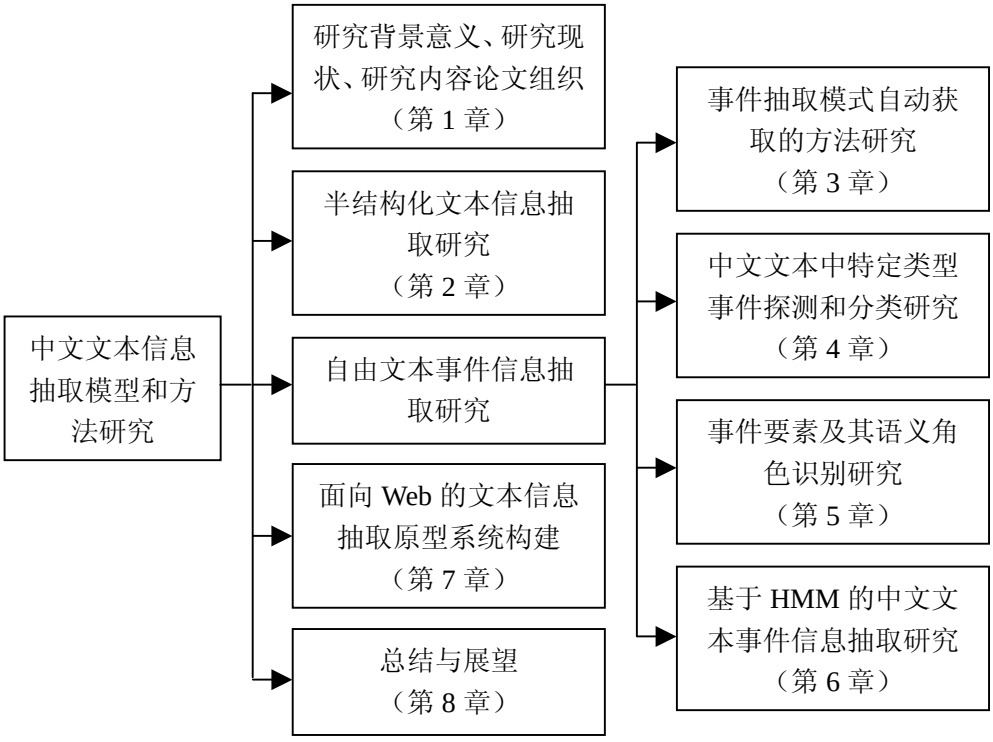

图 1.4 本文组织结构

第 1 章：分析了本课题研究的背景和意义，介绍了文本信息抽取的基础知识及其发展历程，简要概述了当前的研究现状，最后给出了主要研究内容和本文的组织结构。

第 2 章：分析了半结构化文本数据的特点，以从中文科研论文中抽取头部信息和引文信息为例，针对隐马尔可夫模型不能充分利用对抽取有用的上下文特征，提出了

一种基于条件随机场的半结构化文本信息抽取方法，并针对中文文本的特点，在进行信息抽取时先利用分隔符、特定标识符对文本进行分块，在分块的基础上利用条件随机场模型进行信息抽取。

第 3 章：分析了已有的抽取模式获取方法和中文文本中抽取模式获取的难点，提出了一种从未标注的中文文本中基于自扩展策略自动获取事件抽取模式的算法，该算法从少数几个种子抽取模式出发，通过一个迭代的过程发现新模式，每一轮迭代从三个层次对抽取模式进行扩展，然后采用类似于 TF/IDF 的评估方法对产生的候选模式进行评估，选择得分最高的几个模式并入到当前模式集。最后进行了抽取模式自动获取的实验，并验证这些抽取模式。

第 4 章：事件探测和分类是事件信息抽取中的首要任务，对事件抽取的后继任务至关重要。传统的事件分类仅仅依据事件表述语句中的触发词，而忽略了触发词上下文中与事件密切相关的大量特征信息，致使分类效果不佳。为此提出了一种基于最大熵模型的事件分类方法，该方法能够综合事件表述语句中的触发词信息及各类上下文特征对事件进行分类。应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行了分类实验。

第 5 章：事件要素及其语义角色的识别是事件信息抽取的关键。详细介绍了论元结构理论及其在信息抽取中的具体应用，分析了事件表述语句中触发词所支配的论元和事件要素之间的对应关系。将条件随机场用于信息抽取中事件表述语句的事件要素及其语义角色识别，对“职务变动”和“会见”两类事件的表述语句进行了分类实验。

第 6 章：提出了一种基于隐马尔可夫模型的中文文本事件抽取方法，该方法首先通过触发词探测从文本中发现特定类型事件的候选语句，然后利用隐马尔可夫模型从这些语句中抽取每个候选事件的事件要素，为每一类事件要素构建一个独立的隐马尔可夫模型用于该类事件要素的抽取。最后进行了“职务变动”和“自然灾害”两类事件的抽取实验。

第 7 章：设计并实现了一个面向 Web 的文本信息抽取原型系统。该系统由图形用户界面、数据获取、文本粗加工和信息抽取四个模块组成，用于对不同类型的文本信息抽取模型与方法进行验证。

第 8 章：总结了本文取得的研究成果，并给出了进一步的研究思路。

第2章 基于条件随机场的半结构化文本信息抽取

本章首先以科研论文的头部信息和引文信息为例分析了半结构化文本的特征，并简要概述了从这类文本数据中抽取信息的相关研究。然后以从中文科研论文中抽取头部信息和引文信息为例，针对隐马尔可夫模型不能充分利用对抽取有用的上下文特征，提出了一种基于条件随机场模型的半结构化文本信息抽取方法，该方法的关键在于模型参数估计和特征选择。实验中采用 L-BFGS 算法学习模型参数，并选择局部、版面、词典、状态转移四类特征作为模型候选特征集，采用增量特征选择算法选择特征。在抽取信息时先利用分隔符、特定标识符等格式信息对文本进行分块，在分块基础上利用条件随机场进行特定信息域的抽取。实验表明，该方法抽取性能要明显优于基于隐马尔可夫模型的方法，且加入不同的特征集对抽取性能有不同提升作用。

2.1 半结构化文本特征分析

半结构化文本数据一般不具有完整的句法结构，但具有明显的版面布局结构和特定的标签信息。常见的这类文本有科研论文的头部信息和引文信息、学术报告公告、个人简历、招聘信息、产品参数信息等，其中科研论文的头部信息和引文信息是最重要的一类半结构化文本数据。下面以中文科研论文头部信息和引文信息为例，对半结构化文本的特征进行分析。

论文的头部信息（paper header）是指从论文开始到论文的正文引言部分的一段文本，一般包括论文标题、作者、作者隶属单位、单位地址、邮编、E-mail、摘要、关键词、中图分类号、文献标识码、文章编号等十几个信息域。图 2.1 是一篇中文计算机类科研论文[④]的头部信息。从图中可以看出，头部信息具有以下几个特征：

（1）一篇论文的头部信息可以按照内容划分为若干类信息，每类信息称为一个信息域。如“标题”信息域、“作者”信息域、“邮编”信息域等。这些信息域一般按一定顺序分布在不同位置，在位置上彼此独立，或者交叉在一起。例如，标题、作者、摘要等许多信息域位置彼此独立；而作者单位、单位地址、邮编则交叉在一起。每个信息域有一个或多个信息项组成，有的信息域可以有多个信息项，如作者、单位、

[④] 该论文见《电子学报》2003 年 11 期，其电子版可从中国期刊全文数据库获得。

关键词等；而有的信息域只能有一个信息项，如中图分类号、文献标识码等。

（2）大部分信息域所对应的文本不具有完整的句法结构，即不是由一个或多个完整的自然语言形式的语句组成。例如，“标题”、“地址”、“作者”等。

（3）头部信息作为一个整体具有一定的版面布局结构，这些布局包括信息域之间的顺序、所处版面位置、字体大小粗细、信息项之间的分隔符等。

（4）有些信息域前面有特定的标识信息。例如，“摘要：”、“关键词：”、“中图分类号：”等分别出现在相应的信息域前面。

（5）不同信息域的内容有各自的特征。例如，“邮编”一定是 6 位数字组成；“单位”信息域中多含有“大学”、“学院”、“系”这样的字词。

基于最大熵模型的汉语句子分析

徐延勇，周献中，井祥鹤，郭忠伟

（南京理工大学自动控制系，江苏南京 210094）

摘　要：文中运用浅层句法分析理论，把汉语句子……

关键词：自然语言处理；　最大熵模型；　组块；　句法分析；　广度优先搜索

中图分类号：TP391.1　　**文献标识码：**A　　**文章编号：**037222112

1 引言

……

图 2.1 科研论文头部信息示例

科研论文的参考文献一般以引文的格式放在正文后面，称为论文的引文信息。它是对参考文献的作者、标题、发表期刊、期次、页码等信息的引文格式表述。图 2.2 是几个参考文献的引文信息示例。从图中可以看出，引文信息具有以下几个特征：

[1] 刘云中，林亚平，陈治平. 基于隐马尔可夫模型的文本信息抽取[J].系统仿真学报，2004,16(3):507-510.

[2] 张玲,黄铁军,高文.基于隐马尔可夫模型的引文信息提取[J].计算机工程，2003,29(20):33-34.

[3] 林亚平，刘云中，周顺先，等.基于最大熵的隐马尔可夫模型文本信息抽取[J].电子学报，2005,33(2):236-240.

图 2.2 科研论文引文信息示例

（1）和论文的头部信息类似，一条引文信息可以按照内容划分为若干类信息。如“作者”、“标题”、 “发表期刊”、“发表年卷”、“期次”、“页码”等。

（2）引文信息所对应的文本的句法结构更弱，即不是由完整的自然语言形式的语句组成，甚至可以说几乎不出现完整的语句。

（3）具有较严格的版面格式，即引文格式。不同的信息域、信息项之间有较严格的编排顺序、分隔符等格式信息。

（4）不同类别的信息在内容上有各自的特征。例如，“年代”有特定的表示格式；“期次”和“页码”都有数字组成等。

综合科研论文头部信息和引文信息的特征分析，可以总结出半结构化文本具有以下几个特征：

（1）半结构化文本一般可以按内容分为不同的信息域，不同信息域之间按一定的顺序排列，排列紧凑，通常待抽取的特定信息出现在连续的信息域中。

（2）大部分信息域不具有完整的句法结构，即不是有完整的自然语言形式语句组成。所以处理这类文本一般不涉及深层次的自然语言处理技术。

（3）具有较丰富的版面结构特征，这些版面结构特征反映了不同信息域、信息项的上下文顺序、类别归属等。

（4）有些信息域之前有特定的标识符，这些标识符有助于识别该信息域。

（5）每类信息在内容上有各自的特征，这些特征对标注信息的类别有重要作用。

科研论文的头部信息和引文信息是最常见的一类半结构化文本数据。随着科学研究的快速发展和信息更新速度的加快，全世界发表的科研论文也越来越多，这使得人们对科研论文的检索、统计、分析技术要求越来越高。而科研论文的头部信息和引文信息对基于域（例如，标题、作者、关键词）的论文检索、论文信息统计和引用分析是不可缺少的。通过获取科研论文的头部信息和引文信息，可以有效地组织和管理这些论文，便于提高用户检索论文的效率，而且还能够做大量统计工作。比如：（1）论文主题分析及相关论文统计；（2）对期刊、科研单位、某篇论文或某个作者进行引用分析；（3）发现研究热点和研究趋势等。所以说，中文科研论文的头部信息和引文信息也是一种最重要的半结构化文本数据，研究从科研论文中自动抽取头部信息和引文信息有着重要的现实意义。本章下面部分以从中文科研论文中抽取头部信息和引文信息为例来研究半结构化文本信息抽取中的语言模型和机器学习方法。

2.2 相关研究

针对半结构化文本的特征，可以将半结构化文本信息抽取看成序列数据标注问题，并使用统计语言模型来实现这些信息的抽取。其中，由于隐马尔可夫模型（Hidden

Markov Model，HMM）[64]具有坚实的统计学基础，可以成功地处理未知数据，并且易于开发，因此，国内外基于HMM的半结构化文本信息抽取的研究非常多。例如，利用HMM模型对科研论文头部信息和引文信息进行抽取 [65][66][67][68]；基于HMM模型的学术报告信息抽取 [69][70][71]；利用HMM模型从电子版的个人简历中抽取信息 [72]等等。

国内外对基于HMM模型的科研论文头部信息和引文信息抽取有大量的研究和实验，并对其中的文本切分粒度、HMM模型的结构、模型参数学习和数据稀疏等问题进行了深入研究。Seymore等人 [67]用一个HMM模型对计算机科研论文头部信息的所有域进行抽取，取得了 92.9%的抽取精度。并深入探讨了HMM模型结构的机器学习问题，从最大化最细化的HMM模型开始使用横向合并（Neighbor-merging）和纵向合并（V-merging）进行状态合并。横向合并将紧相邻的几个同类状态合并为一个状态，合并后的状态通过自转移来模拟一个信息域中多个信息项之间的转移。纵向合并将那些从同一个状态转移而来或要转移到同一个状态去的且具有相同类标签的状态合并为一个状态。经过这些合并后，最终就得到优化的HMM模型结构，如图 2.3 所示（该图引自文献 [67]）。林亚平等 [69]用基于最大熵的HMM从科研论文头部信息和学术报告信息中抽取信息。刘云中等 [66]用HMM模型并利用文本排版格式、分隔符等信息，对科研论文头部信息进行抽取。张玲等 [65]采用基于符号特征提取的HMM模型结构学习方法，进行引文信息抽取。Yin P等 [68]采用Bigram HMM从论文的参考文献中抽取元数据。Han等 [62]将该问题看作分类问题，采用支持向量机（Support Vector Machines, SVM）从论文的头部信息中抽取元数据。

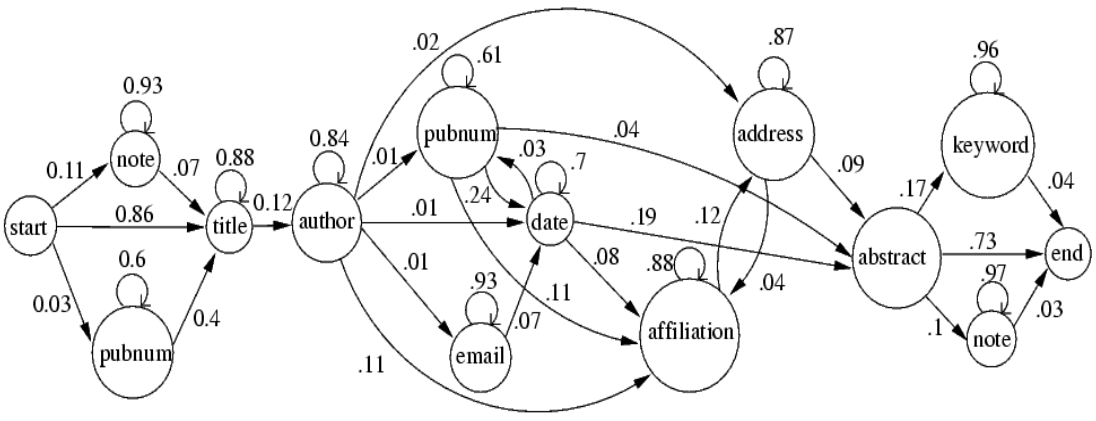

图 2.3 科研论文头部信息抽取中优化的 HMM 模型结构示例

这些研究都把该问题看成序列标注或分类问题，并采用序列标注模型或分类器来实现，尤其以 HMM 模型居多。作为一种生成模型（Generative Model），尽管 HMM 被

广泛用于序列标注，但它并不是最合适的模型。HMM 通过定义标记序列和观察序列的联合概率 P(S,O)来搜索最佳的标记序列，定义这样的联合概率意味着所有可能的观察序列都应该被枚举出来，这个任务是很困难的，当观察元素间具有长距离依赖性时尤其如此。所以，为了保证推导的正确和简便，HMM 给出两个独立性假设：一是假设每个状态仅仅依赖于它的前一时刻状态；二是假设时刻 t 的观察值仅仅依赖于时刻 t 的状态。这两个假设简化了推导，但同时也降低了标注性能。事实上，真实的序列数据不仅存在长距离依赖性，而且观察序列中各种有利于提升标注序列精确度的版面结构特征、信息域内容特征等上下文特征也未能被整合到 HMM 模型中去。而且无论国外还是国内，这些研究大部分都使用英语语料进行研究和实验，那么对中文科研论文头部信息和引文信息抽取效果需要研究验证。

基于以上的分析，针对HMM的这一缺点和中文文本的一些特性，本章提出了一种基于条件随机场（Conditional Random Fields，CRFs）的中文科研论文头部信息和引文信息抽取方法。CRFs是Lafferty等人[108]于 2001 年提出的一种用于序列数据标注的条件概率模型，是一种判定性模型（Discriminative Model）。CRFs通过定义标记序列和观察序列的条件概率P(S|O)来预测最可能的标记序列。CRFs不仅能够将丰富的上下文特征整合到模型中，而且还克服了其他非产生性模型（例如最大熵马尔可夫模型）的标注偏差问题（label bias problem）[108]。近几年来，CRFs已经被成功地应用到许多自然语言处理领域，例如，词语切分及词性标注[108][109]、浅层句法分析[110][111]、组块分析和短语识别[112][113]、命名实体识别[114][115][116][117]、信息抽取[118][119]等。应用CRFs从科研论文抽取头部信息和引文信息的关键是模型参数估计和特征选择与归纳。本章采用L-BFGS（Limited-memory BFGS）算法[120]对模型参数进行估计，并选择四类特征作为模型的候选特征集，然后再采用增量特征选择算法来选择有效特征。从中文科研论文抽取信息时先利用信息项之间的分隔符、信息域之前的特定标识符等格式信息对文本进行分块，在分块基础上用CRFs模型进行特定信息域的抽取。实验结果表明，该方法抽取性能明显优于HMM模型。

2.3 条件随机场

条件随机场是一种以给定的输入结点值为条件来预测输出结点值概率的无向图模型。用于模拟序列数据标注的 CRFs 是一个简单的链图或线图（如图 2.4 所示），

它是一种最简单也最重要的 CRFs 模型，称为线链 CRFs（linear-chain CRFs）。在模型的图形结构中随机变量之间通过指示依赖关系的无向边所连接。线链 CRFs 假设在各个输出结点之间存在一阶马尔可夫独立性，其输出结点被边连接成一条线性链，这种 CRFs 可以被理解为条件训练的有限状态机（FSMs）。

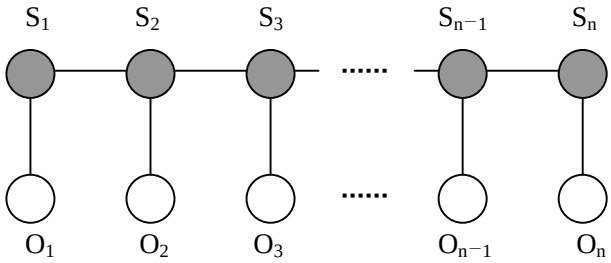

图 2.4 线链 CRFs 的图形结构

设 O = $\{o_1,o_2,\ldots\ldots,o_T\}$表示被观察的输入数据序列，例如论文头部信息的词或信息项序列。S = $\{s_1,s_2,\ldots\ldots,s_T\}$表示被预测的状态序列，每一个状态均与一个标记（例如标题、作者）相关联。这样，在一个输入序列给定的情况下，参数为 $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_K\}$的线链 CRFs，其状态序列的条件概率为：

$$P_\Delta(S\mid O)=\frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right), \qquad (2.1)$$

其中，Z_O 是归一化因子，它确保所有可能的状态序列的条件概率和为 1，即它是所有可能的状态序列的“得分”的和：

$$Z_O=\sum_{S}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right)。 \qquad (2.2)$$

$f_k(s_{t-1},s_t,o,t)$是一个任意的特征函数，通常是一个二值表征函数。λ_k是一个需要从训练数据中学习的参数，是相应的特征函数 $f_k(s_{t-1},s_t,o,t)$的权重，取值范围可以是-∞到+∞。特征函数 $f_k(s_{t-1},s_t,o,t)$能够整合任何特征，包括状态转移 $s_{t-1}\to s_t$特征，以及观察序列 O 在时刻 t 的所有特征。例如：当 s_{t-1}是标记“标题”，s_t是标记“作者”，并且 o_t中第一个字在姓氏词典中时，则相应的特征函数取值为 1。而且如果这种情况出现，相应 λ_k将有正的较大的权重。

给定一个由公式（2.1）定义的条件随机场模型，在已知输入数据序列 O 的情况下，最可能的标记序列可以由下式求出：

$$S^*=\arg\max_{S} P_\Delta(S\mid O)， \qquad (2.3)$$

最可能的标记序列可以由上式通过类似于 HMM 中的 Viterbi 算法动态规划求出。

要建立 CRFs 模型还有两个关键的问题：参数估计和特征选择。参数估计是从训练数据集学习每一个特征的权重参数，即求解向量 $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_K\}$ 的过程。而特征选择是根据待处理的问题筛选出对 CRFs 模型有表征意义的特征。

2.3.1 参数估计

参数估计就是从训练数据集学习权重参数向量 Λ 的过程，这个过程一般通过最大对数似然估计实现。设训练数据集表示为：$D=\{O^{(i)},S^{(i)}\}_{i=1}^{N}$，其中，每个 $O^{(i)}=\{o_1^{(i)},o_2^{(i)},...,o_T^{(i)}\}$ 是一个输入数据序列；$S^{(i)}=\{s_1^{(i)},s_2^{(i)},...,s_T^{(i)}\}$ 是相应的输出数据序列。在训练数据集 D 下条件对数似然为：

$$L_\Delta=\sum_{i=1}^{N}\log P(S^{(i)}\mid O^{(i)}), \qquad (2.4)$$

将 CRFs 模型中的条件概率公式（2.1）代入（2.4）式，可得：

$$L_\Delta=\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\log Z_{O^{(i)}}\right) \qquad (2.5)$$

为了避免估计大量参数时出现过拟合（over-fitting），对数似然经常需要将参数作先验分布的调整，采用高斯先验调整后，（2.5）式变为：

$$L_\Delta=\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\sum_{i=1}^{N}\log Z_{O^{(i)}}-\sum_k\frac{\lambda_k^2}{2\sigma^2}, \qquad (2.6)$$

其中最后一项是用于进行调整的特征参数的高斯先验值，σ^2 表示先验方差。优化(2.6)式需要一个迭代的过程，传统的办法是Della Pietra等人在GIS（Generalized Iterative Scaling Algorithm）算法 [121]的基础上提出的IIS（Improved Iterative Scaling Algorithm）算法 [122]。Lafferty等人 [108]在进行CRFs参数估计时采用了基于IIS的算法。IIS算法或基于IIS的算法的最大弱点是计算量大，求解速度慢。本章采用L-BFGS算法 [120]对CRFs的参数进行估计。L-BFGS是一种被实验验证了的训练速度明显快于IIS的算法 [123]，它是一种充分利用以前的梯度和修改值来近似曲率值的二阶方法，可以避免准确的Hessian矩阵的逆矩阵的计算。因而使用L-BFGS算法进行CRFs训练只要求提供对数似然函数的一阶导数，训练数据集的对数似然函数的一阶导数为：

$$\frac{\partial L_{\Lambda}}{\partial \lambda_k}=\sum_{i=1}^{N}\sum_{t=1}^{T} f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s,s'} f_k(s,s',o^{(i)},t)\,p(s,s'\mid o^{(i)})-\frac{\lambda_k}{\sigma^2},\quad(2.7)$$

其中，第一项为特征 f_k 在经验分布下的期望值；第二项为特征 f_k 在模型 Λ 下的期望值。对它们的计算，可采用动态规划高效实现。

2.3.2 特征构建与选择

建立条件随机场模型的另一个关键是如何针对特定的任务为模型构建并选择有效的特征集，用简单的特征集表示复杂的语言现象。一些实验 [117]已经显示了特征构建与选择对CRFs的性能有明显影响。针对从中文科研论文中抽取信息这一任务，在详细分析输入文本的上下文信息的基础上，结合 2.1 节对半结构化文本特征的分析，本章将用于中文科研论文信息抽取的候选特征分成四类：局部特征、版面特征、外部词典特征和状态转移特征。

2.3.2.1 局部特征

局部特征是指输入序列的局部（可以是一个词、一个信息项或一个信息域）所特有的拼写、文本内容、字符形式上的特征，这类候选特征是最基本的特征。例如，信息域“邮政编码”包含且仅包含六位数字；作者姓名包含 2-4 个汉字；中图分类号一般是一个大写字母等。本章中用到的部分局部特征如表 2.1 所示。

表 2.1 局部特征列表

特征函数	特征含义描述
ALLCHINESE	文本中的所有字符都是汉字
CONTAINSDIGITS	文本中包含一个或多个数字
ALLDIGITS	所有字符都是数字
SIXDIGITS	包含仅包含 6 个数字
CONTAINSDOTS	至少包含一个“.”
CONTAINSHYPHEN	包含连字符“-”
YEARFORMAT	年代格式，包含 4 个数字
CLCNUMBERFORMAT	包含字母
SINGLECHAR	仅仅包含一个字母
NAMELENGTH	包含 2－4 个汉字
TITLELENGTH	包含 6－24 个汉字
BIGLENGTH	超过 40 个汉字

2.3.2.2 版面特征

版面特征主要指要分析预测的信息项或信息域在输入序列中所处的版面位置、字体格式、分隔符等与版面布局相关的特征。例如，大多数标题处于一行的居中位置，并采用大字体、粗体等版面格式。本章中用到的部分版面特征如表 2.2 所示。

表 2.2 版面特征列表

特征函数	特征含义描述
LINE_START	一行的开始
LINE_IN	一行的中间
LINE_END	一行的末尾
BIG_SIZE	大字体
BOLD	粗体
ITALIC	斜体
END_ SEMICOLON	结尾是分号
END_ COMMA	结尾是逗号
BACK_ BLANK	后面是几个空格

2.3.2.3 外部词典特征

外部词典特征是指要分析的词语或信息项对象中的字、词是否出现在外部词典中。这些外部词典有：中国人名姓氏字典、中国人名常用字词典、中国地名词典、单位名称特征词词典、论文标题高频词词典等。例如，作者姓名域中的第一个字一般都出现在姓氏词典中；隶属单位域中有“大学”、“学院”、“研究所”等词条出现在相应词典中等。部分外部词典特征如表 2.3 所示。

表 2.3 外部词典特征列表

特征函数	特征含义描述
FAMILYNAME	在姓氏词典中
FIRSTNAME	在中国人名常用字词典中
AFFILIATION	包含“大学”、“学院”、“系”等词条
TITLETERM	包含“基于”、“研究”等词条
CLCNUMBER	在中图分类号中
ADDRESS	在中国地名词典中

2.3.2.4 状态转移特征

在条件随机场模型中，特征函数 $f_k(s_{t-1}, s_t, o, t)$ 不仅能够整合输入数据序列中出现

的特征，而且也能将状态转移作为特征整合到模型中。例如：当 s_{t-1} 是标记“标题”，s_t 是标记“作者”，并且 o_t 中第一个字在姓氏词典中时，此时特征函数 $f_k(s_{t-1}, s_t, o, t)$ 将取值“1”。这样，CRFs 模型就将状态转移 $s_{t-1} \to s_t$ 也整合进来。

2.3.2.5 特征的选择

利用上面构建的四类候选特征可以得到一个庞大的候选特征集，这个候选特征集通常包含许多特征，接下来要对训练语料中实际出现的特征进行特征选择，因为并不是所有的候选特征对模型都有较大的贡献，太多的特征会增加模型训练的时间。通过特征选择算法最终确定模型的有效特征集。McCallum[117]设计了一种用于条件随机场的特征自动选择和归纳算法，该算法是在初始原子特征集合的基础上，先通过对候选特征的初选和高增益特征间的连接构造复合特征，形成候选特征集。在此基础上再基于增益同时进行原子特征和复合特征的选择。但该算法计算量非常大，不易实现。本章借鉴最大熵模型（Maximum Entropy Model）中常用的增量特征选择方法 [124]（Incremental Feature Selection, IFS）的思想来选择特征。

增量特征选择算法的基本思想是：设 F 是一个候选特征集合，其中只有一部分特征是语言模型的有效特征，有效特征子集设为 S，特征选择就是从候选特征集 F 中选取有效特征子集 S 的过程。方法如下：开始设有效特征集 S 为空，然后不断地向 S 中增加候选特征，每次向 S 中增加的特征由训练数据决定。每增加一些特征需要对所有的候选特征调用 L-BFGS 算法，对 λ 重新计算，还要利用训练数据对新模型的对数似然进行计算，以判断该特征是否使模型的对数似然增量最大，实现也有一定困难，但能够获得有效的特征集。具体实现时采用简化的增量特征选择算法。

算法 2.1：简化的增量特征选择算法

输入：候选特征集合 F；训练数据集

（1）初始化。将模型的有效特征集合 S 设置为空，即令 S＝Φ；

（2）计算候选特征集 F 中每一个候选特征的增益；

（3）选出增益值大的一组候选特征，并将这些特征加入到有效特征集 S 中；

（4）使用 L-BFGS 算法重新调整选择特征集中的各特征参数值。

（5）对步骤（2）～（4）迭代，直至收敛。

输出：有效特征集合 S；整合了有效特征的 CRFs 模型

2.4 基于条件随机场的中文科研论文信息抽取

科研论文的头部信息可以看成是一个信息块序列，这些信息块分别属于标题、作者、隶属单位、邮编等不同的信息域。对不同期刊的大量科研论文头部信息进行综合分析后，本章对中文科研论文头部信息的抽取固定在标题（Title）、作者（Author）、隶属单位(Affiliation)、单位地址(Address)、邮政编码(Zip Code)、摘要(Abstract)、关键词（Keywords）、中图分类号（CLC Number）、文献标识码（Document Code）共 9 个信息域。利用 CRFs 从论文头部抽取信息时，每个状态和要抽取的一个信息域相关联。在抽取时我们充分利用逗号、分号、空格、括号、回车等分隔符将输入数据序列划分成为信息块，用这些可观察的信息块序列为条件来预测每个状态相关联的域。详细抽取过程描述如下：

（1）对论文头部进行预处理。主要是依据回车和逗号、分号等分隔符对头部信息进行信息块划分，然后在信息块基础上进行信息抽取。另外还要根据“摘要”、“中图分类号”、“关键词”等特定标签将相应信息域标记。

（2）从训练数据集 $D=\{O^{(i)},S^{(i)}\}_{i=1}^{N}$，采用 L-BFGS 算法学习出各特征函数的权重向量 Λ。

（3）在给定的 CRFs 模型 Λ 下，结合公式（2.1）和（2.3），用韦特比算法动态规划预测出最可能的状态序列，即信息块相关联的域，然后按关联情况抽取各个信息项或信息域。

上面详细分析了利用 CRFs 模型对中文科研论文头部信息抽取的过程。本章对中文科研论文引文信息的抽取固定在作者（Author）、标题（Title）、期刊或论文集（Journal）、发表年（Year）、卷（Volume）、页码（Page）共 6 个信息域。利用 CRFs 对引文信息的抽取机理与头部信息大同小异，这里不再赘述。

2.5 实验结果及其分析

为验证本章提出的基于条件随机场的头部信息和引文信息抽取方法，进行了多组实验。实验时首先将CRFs模型用于从不同语言（英语和汉语）数据集抽取头部信息和引文信息，比较该模型在两种不同语言中的抽取性能差异；接着分别用CRFs和HMM两种模型在中文数据集上进行了训练和评测，实验表明，CRFs的抽取性能要明显优于基

于HMM的性能；另外，在实验中还探讨了分块和不同特征集对CRFs模型抽取性能的影响。在对抽取性能进行评估时，采用了常用的 3 个评测指标（详见 1.3.3 节）：准确率（P）、召回率（R）、综合指标F值（F，其中 β 取值为 1）。

2.5.1 实验数据集

文献 [62][66][67][69]用于科研论文头部信息和引文信息抽取的语料是英文语料，都是基于美国卡内基梅隆大学（CMU）的两个数据集[⑤]进行实验。一个是计算机科研论文头部信息数据集，该数据集共 935 篇计算机科研论文的头部数据；另一个是科研论文引文信息数据集，该数据集共 500 条引文信息。这两个数据集是Seymore和McCallum等人在“Core论文检索项目”中建立的。实验中，对英文论文头部信息数据集，随机抽取其中的 700 篇作为训练语料，剩下的 235 篇作为测试语料。对引文信息数据集，随机把其中的 400 条作为训练语料，剩下的 100 条作为测试语料。

中文实验数据是从中国期刊网（China National Knowledge Infrastructure，简称 CNKI）全文数据库和引文数据库中检索计算机有关方面的论文整理标注而成，这些论文来源于计算机类的中文核心期刊。其中科研论文头部信息数据 600 篇，引文数据 1500 条。实验中，对中文论文头部信息数据集，随机抽取其中的 500 篇作为训练语料，剩下的 100 篇作为测试语料。对论文引文信息数据集，随机把其中的 1200 篇作为训练语料，剩下的 300 篇作为测试语料。

为了取得更好的实验效果，在训练集上训练 CRFs 模型时，采用 10 重交叉验证（10-fold cross validation）的方法来获取最优的模型。即训练集被随机划分为 10 个不相交的组，对模型训练 10 次，每次留出一组作为验证集，其他 9 组作为训练集用于调整模型参数，验证集用于评价推广误差，最终得到的模型有最低的推广误差。

2.5.2 实验结果及分析

2.5.2.1 不同语言数据集上的抽取结果比较

实验时，首先在公认的测试数据集，即 2.5.1 节提到卡内基梅隆大学的两个英文数据集上对本章提出的基于CRFs的方法进行实验，验证该方法的可行性。为了对中英

[⑤]该英文训练测试语料来源于：http://www.cs.cmu.edu/~kseymore/ie.html

文两种不同语言的科研论文头部信息的抽取性能进行比较，本章在对英文语料和中文语料进行头部信息抽取时选择都有的信息域进行评测比较。这些信息域包括：论文标题（Title）、作者（Author）、隶属单位（Affiliation）、地址（Address）、邮编（Zip Code）、摘要（Abstract）。各信息域具体抽取结果如表 2.4 所示。

表 2.4 中文和英文科研论文头部信息抽取结果

信息域	中文语料			英文语料		
	P	R	F	P	R	F
Title	0.971	0.918	0.944	0.982	0.921	0.951
Author	0.986	0.946	0.966	0.956	0.932	0.944
Affiliation	0.977	0.918	0.947	0.957	0.927	0.942
Address	0.976	0.936	0.956	0.961	0.902	0.931
Zip Code	0.996	0.979	0.987	0.997	0.976	0.986
Abstract	0.984	1.000	0.992	0.991	1.000	0.995

从表 2.4 可以看出，利用CRFs对科研论文头部信息抽取在中英文上性能相差不大，平均F值分别得到 96.5%和 95.6%，就英文而言略高于文献 [62]中提出的基于SVM的抽取性能（平均F值 95.1%）。其中作者和地址信息域在中文上的综合指标F值还要高于英文语料两个多百分点，这与中文姓名、地址这些信息域短而紧凑有关。

同样，在对中文语料和英文语料的引文信息进行抽取时，选择作者（Author）、标题（Title）、期刊或论文集（Journal）、发表年（Year）、卷（Volume）、页码（Pages）共 6 个信息域进行评测比较。图 2.5 比较了基于 CRFs 的中英文论文引文信息抽取结果的综合指标 F 值，从图 2.5 可以直观地看到：利用 CRFs 模型对科研论文引文信息抽取在中英文两种语言语料上性能很接近。

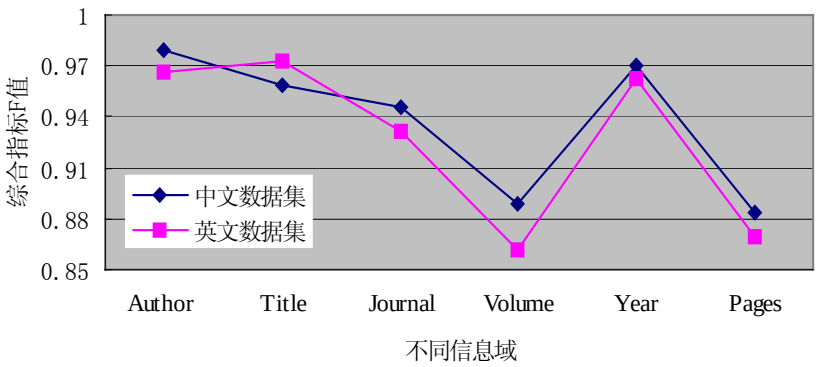

图 2.5 不同语言数据集上引文信息抽取结果比较

在不同语言数据集上的实验结果表明，本章提出的基于 CRFs 的科研论文信息抽取方法是可行的，在中英文两种数据集上抽取性能差别不大。

2.5.2.2 CRFs vs. HMM

本章重点比较了基于 CRFs 和 HMM 两种不同语言模型从中文计算机科研论文中抽取头部信息和引文信息的性能，其中 CRFs 模型中使用了前面用特征选择算法筛选出的全部四类特征。对中文论文头部信息的抽取固定在标题（Title）、作者（Author）、隶属单位(Affiliation)、单位地址(Address)、邮政编码(Zip Code)、摘要(Abstract)、关键词（Keywords）、中图分类号（CLC Number）、文献标识码（Document Code）共 9 个信息域。表 2.5 给出了中文科研论文头部信息抽取的结果。

表 2.5 中文科研论文头部信息抽取结果

信息域	CRFs			HMM		
	P	R	F	P	R	F
Title	0.971	0.918	0.944	0.915	0.848	0.880
Author	0.986	0.946	0.966	0.922	0.835	0.875
Affiliation	0.977	0.918	0.947	0.921	0.837	0.877
Address	0.976	0.936	0.956	0.931	0.851	0.889
Zip Code	0.996	0.979	0.987	0.958	0.916	0.937
Abstract	0.984	1.000	0.992	0.979	0.988	0.984
Keywords	0.969	0.937	0.953	0.929	0.905	0.917
CLC number	0.983	0.963	0.973	0.967	0.939	0.953
Document code	0.997	0.996	0.997	0.978	0.985	0.985

对中文论文引文信息的抽取固定在作者（Author）、标题（Title）、期刊或论文集（Journal）、发表年（Year）、卷（Volume）、页码（Page）共 6 个信息域。表 2.6 给出了中文科研论文引文信息抽取的结果。

表 2.6 中文科研论文引文信息抽取结果

信息域	CRFs			HMM		
	P	R	F	P	R	F
Author	0.989	0.970	0.979	0.907	0.864	0.885
Title	0.984	0.936	0.959	0.898	0.826	0.860
Journal	0.957	0.935	0.946	0.871	0.845	0.858
Volume	0.962	0.826	0.889	0.876	0.868	0.872
Year	0.982	0.959	0.970	0.978	0.932	0.954
Pages	0.891	0.875	0.883	0.839	0.862	0.850

从表 2.5 和表 2.6 可以看出，对于从中文科研论文中抽取头部信息和引文信息的任务，基于 CRFs 模型的性能明显优于基于 HMM 模型的性能。从抽取结果可以看到不论是头部信息抽取还是引文信息抽取，基于 CRFs 的方法在所有信息域的综合指标 F

值都要高于 HMM 的。对于头部信息抽取而言，从表 2.5 可以看到 CRFs 和 HMM 两种模型对有特定标识符的信息域（例如，摘要、中图分类号等）的抽取性能差别不大，这主要是由于特定标识符在两种方法中起重要作用所致。而对其他信息域，基于 CRFs 的抽取性能要比 HMM 的高出很多，例如，标题信息域的综合指标 F 值从 88.0%提高到了 94.4%；作者信息域的综合指标 F 值从 87.5%提高到了 96.6%，还有作者隶属单位、地址等信息域都要高出 6 个百分点以上。通过分析真实语料，可以看到这些信息域都具有丰富的局部、版面、外部词典等特征，基于条件随机场的方法将这些信息域的上下文特征用于信息抽取中，大大提高了抽取性能，这进一步说明上下文特征对半结构化文本信息抽取意义重大。

2.5.2.3 分块对抽取性能的影响

本章提出的方法首先要依据回车、逗号、分号等分隔符对头部信息或引文信息进行信息块划分，然后在信息块的基础上进行信息域的抽取。实验中分别在基于分词和基于信息块不同粒度的基础上，进行了头部信息抽取的对比测试，由此得到不同的实验结果，图 2.6 是头部信息抽取中一些信息域基于块和基于词的综合指标 F 值的比较。其中，由于特定标签对摘要、中图分类号、文献标识码等信息域的抽取作用重大，分块对这些信息域的抽取作用提升有限，所以在图 2.6 中没有比较这几个信息域。

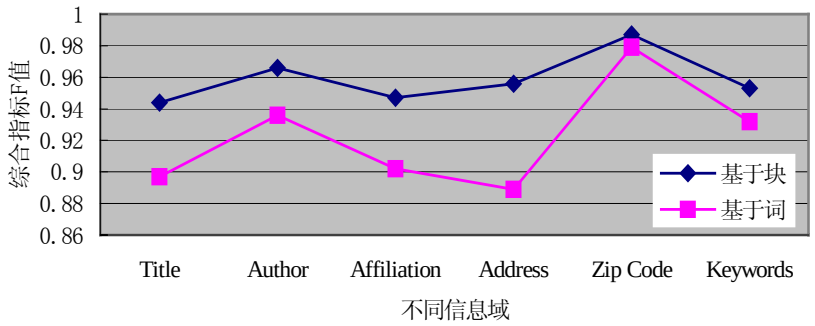

图 2.6 基于块和基于词的头部信息抽取结果比较

从图 2.6 可以看到，基于信息块的抽取方法的抽取结果明显比基于词的抽取结果要好。这主要因为依据回车、逗号、分号等分隔符划分的信息块一定属于一个信息域，而这些信息块经过分词之后，一般包含多个词语，一方面对多个词语的状态预测增加了出错的几率；另一方面分词经常将一些未登录词错误切分，特别是一些地名、专业术语等词语，而这些词语一旦被错误切分的话，对后面的状态预测影响很大。另外，

基于块的抽取方法在进行模型训练时也要快于基于词的方法。综合上面的实验结果及分析，可见在分块的基础上进行信息抽取可以大大提高抽取性能。

2.5.2.4 特征集的影响

为了比较不同的特征集对抽取性能的影响，在进行基于 CRFs 模型的头部信息抽取实验时，起先只使用筛选出的局部特征集，然后再加入版面特征、外部词典特征和状态转移特征，并观察不同特征集对抽取性能的影响，结果如表 2.7 所示。从表 2.7 可以看出，加入版面特征后使得头部信息抽取的平均 F 值从 91.9%提升到了 92.9%；加入词典特征后对抽取性能影响较大，平均 F 值从 91.9%提升到了 94.5%；加入状态转移特征后抽取性能也有一定提升。可见，筛选出的四类有效特征能够在很大程度上提高抽取的性能。

表 2.7 不同特征集时头部信息抽取结果

不同特征集	P 平均值	R 平均值	F 平均值
局部特征集	0.924	0.916	0.919
＋版面特征	0.939	0.922	0.929
＋词典特征	0.957	0.938	0.945
＋状态转移特征	0.937	0.918	0.928
全部有效特征集	**0.982**	**0.955**	**0.968**

2.6 小结

现有的方法一般将半结构化文本信息抽取看作序列数据标注问题，多采用 HMM 模型来实现。本章首先以科研论文的头部信息和引文信息为例分析了半结构化文本的特征，针对 HMM 模型不能充分利用对标注有用的上下文特征，提出了利用条件随机场模型进行半结构化文本信息抽取，并以从中文科研论文中抽取头部信息和引文信息为例进行了详细论述。科研论文是最常见、最重要的一类半结构化文本数据。随着科研论文资源的日益增多，人们迫切需要能够高效地进行科研论文的检索和统计，这使得准确抽取论文头部信息和引文信息非常必要，而基于条件随机场的科研论文头部信息和引文信息的自动抽取技术无疑能够大大提高工作效率。本章将条件随机场模型用于从中文科研论文中抽取头部信息和引文信息，实验结果表明，基于 CRFs 的中文科研论文信息抽取性能要明显优于基于 HMM 的性能。

第3章　事件抽取模式的自动获取

传统的信息抽取系统大多是基于模式匹配的，因此，如何自动获取抽取模式就成为信息抽取中的一个核心问题。本章提出了一种从未标注的中文文本中基于自扩展策略自动获取事件抽取模式的算法，该算法从少数几个种子抽取模式开始，通过一个增量迭代的过程发现新模式，每一轮迭代从三个层次对抽取模式进行扩展，然后采用类似于 TF/IDF 的评估方法对产生的候选模式进行评估，选择得分最高的几个模式并入到当前模式集。实验表明，该方法能较好地从未标注的中文文本中自动学习事件抽取模式，将这些模式用于事件信息抽取中取得了较好的抽取效果。

3.1 引言

作为一种实用的自然语言处理技术，文本信息抽取近几年来受到越来越多的重视。但是，到目前为止，信息抽取并未象信息检索一样被广泛应用。原因有多方面，其中最重要的原因有两个，一个是现有的信息抽取系统的性能还比较差，抽取精度还有待进一步提高；另一个是信息抽取系统的可移植性不好。由于一般的信息抽取系统是面向特定场景的，当一个成熟的信息抽取系统从一个场景移植到一个新的场景时，要耗费大量的人力物力财力，移植花费的时间也较长，有的甚至需要重新进行开发。这些因素都极大地制约着信息抽取技术的大规模应用。

而究竟又是什么制约着信息抽取系统快速地从一个场景向另一个场景移植呢？要搞清楚这个问题就需要明白信息抽取系统的机理。传统的信息抽取系统大多是基于模式匹配实现的[14][75][76]。即针对要抽取的特定类型事实事件（场景），构建相应的抽取模式库，也就是说抽取模式库是场景相关的。然后再用这些抽取模式去新的文档中匹配可能的事实事件，并将匹配的结果填入信息抽取模板中。当已有的信息抽取系统移植到一个新的场景时，一般都需要为新的场景构建相应的抽取模式库，而且抽取模式构建对用户的要求越高，系统的可移植性也就越差[125]。目前，采用的信息抽取方法中的大多数都能适应各种领域不同的信息抽取任务，唯独信息抽取模式是针对特定场景的，当转向一个新的信息抽取任务时，就必须重新创建一套模式。但是模式的手工创建不仅耗时费力，而且需要既熟悉信息抽取模式创建过程又精通应用领域的专

家，因而极大地限制了信息抽取的大规模应用。因此，如何快速地获取抽取模式对提高信息抽取系统的移植性至关重要，而研究通过机器学习方法来自动获取抽取模式无疑最有意义。

3.2 相关研究

为了自动获取信息抽取模式，人们先后在不同的信息抽取系统中采用过各种抽取模式获取方法，按照这些抽取模式获取系统所需要的用户辅助工作的不同和对用户工作量大小和技能要求高低的不同，可将这些系统分为五类：手工创建抽取模式的信息抽取系统，基于人工语料标注的抽取模式学习系统，基于人工语料分类的抽取模式学习系统，基于 WordNet/HowNet 和语料标注的抽取模式学习系统，基于种子模式的自扩展抽取模式获取系统。

3.2.1 手工创建抽取模式的信息抽取系统

这类系统有PLUM[78]、FASTUS[79]、GE NLTOOLSET[80] 、PROTEUS[81]等。这些系统主要出现在MUC评测会议的早期，它们的共同特点是采用手工的方式定制抽取模式或模式结构，系统的移植性最差，现在已经很少采用这种方式构建抽取模式。

3.2.2 基于人工语料标注的抽取模式学习系统

这些抽取模式学习系统进行模式学习的一般做法是：设计一种信息抽取模式表示方式；人工标注训练语料；使用机器学习方法从中学习相应的信息抽取模式。这类系统有AutoSlog[82]、PALKA[83]、CRYSTAL[84]、LIEP[85]、HASTEN[86]等，下面以AutoSlog为例来介绍该类抽取模式学习系统的基本原理。

AutoSlog 是 Univ. of Massachusetts 开发的世界上第一个采用机器学习方法进行模式获取的系统，能够从标注好的场景实例训练语料中学习抽取模式。AutoSlog 将一个信息抽取模式表示为一个概念节点（Concept Node）。该系统的核心部分是单个事件角色（事件要素）的模式学习算法，用学习出的抽取模式可以从文本中抽取单个事件角色，用来填充事件模板中单个槽。针对一个特定的信息抽取领域，该算法的输入除了标注后的实例语句外，还有一个系统预定义的与领域无关的语言模式集，其中包含了

13 个语言模式。该算法的输出是与每个实例语句对应的信息抽取模式。

3.2.3 基于人工语料分类的抽取模式学习系统

这类模式学习系统首先由人手工对训练语料进行领域相关与否的分类，然后根据这种粗浅的分类，从这些训练语料中学习出领域相关的抽取模式。典型的系统为AutoSlog-TS[87]，该系统是对AutoSlog的扩展。

作为 AutoSlog 的后继产品，AutoSlog-TS 的设计动机是为了比 AutoSlog 进一步减少用户在模式学习过程中的工作量并降低对用户的技能要求，而它的确做到了这一点。因为 AutoSlog-TS 在模式学习过程中，仅仅要求用户将文档分为领域相关和领域无关两类，然后就能从领域相关的文档中学习出抽取模式。这显然比 AutoSlog 中要求用户手工标注训练文本中出现的大量的事件实例要省时省力。

3.2.4 基于 WordNet/HowNet 和语料标注的抽取模式学习系统

这类系统的共同之处是在学习抽取模式时要用领域无关的概念层次知识库WordNet或HowNet作为支持。该类信息抽取模式学习系统的典型代表是TIMES[88]和GenPAM[5]。

TIMES 的功能是：在领域无关的概念层次知识库 WordNet 的支持下，用户通过一个 GUI 给出含有某类事件描述的语句（即本文所说的事件表述语句），系统调用部分句法分析功能对该语句进行部分句法分析，用户通过 GUI 指导系统从语法和语义两个层面对分析后的语句进行泛化，形成具有一定概括能力的事件信息抽取模式。TIMES 中的事件抽取模式由左右两部分组成，左边部分表示模式的触发条件，右边部分表示模式应采取的动作。

GenPAM 是中科院计算所的姜吉发提出的一种基于领域无关的概念知识库的事件抽取模式学习方法。该方法利用了领域无关的概念层次知识库 WordNet/HowNet 等的支持并能在模式学习的过程中同时实现词义消歧。GenPAM 在进行抽取模式学习的时候，不需要用户提供种子模式，也不需要用户进行语料分类，更不需要用户进行语料标注，只需用户进行信息抽取任务定义之后，系统就可以自动地从一个未经标注和分类的原始语料中学习出信息抽取模式，而且整个过程是在领域无关的概念知识库的支持之下实现的。

3.2.5 基于种子模式的自扩展抽取模式获取系统

这种信息抽取模式学习系统的典型代表是ExDisco[76][89]、Snowball[90][91][92]等。这两个系统获取抽取模式的机理都是基于种子模式的自扩展（bootstrapping）方法，下面以ExDisco为例叙述。ExDisco是纽约大学的Roman Yangarber等人 2000 年设计开发的。开发该系统的动机是为了比基于文本分类的信息抽取模式学习系统进一步减少用户的工作量并降低对用户技能的要求。对于一个特定的信息抽取领域，该系统只要求用户提供几个对场景来说有代表性的种子抽取模式，然后就能从一个未经领域相关性分类的文档集中学习出更多的抽取模式。下面从抽取模式的表示方式和学习过程两个方面进行详细介绍。

3.2.5.1 ExDisco 中抽取模式的表示方式

ExDisco 中采用二元组或三元组来表示信息抽取模式。例如，表 3.1 中列出的就是 ExDisco 在进行“职务变动”类事件的模式学习过程中用户给出的三个种子抽取模式，这三个模式对“职务变动”事件信息抽取任务有典型的代表性。

表 3.1 ExDisco 在“职务变动”场景中的三个种子模式

Subject	Verb	Direct Object
<C-Org>	<C-Appoint>	<C-Person>
<C-Org>	<C-Resign>	<C-Person>
<C-Person>	<C-Resign>	

其中，<C-Org>、<C-Person>、<C-Appoint>、<C-Resign>是四个概念语义类，<C-Org>和<C-Person>分别代表公司类和人员类的名词性概念语义类，通常是这两种语义类的命名实体，而<C-Appoint>和<C-Resign>是两个动词概念语义类，分别代表“任命”类和“解雇”类动词。

表 3.1 中的第一个模式是一个“任命”类事件模式，该模式是一个三元组，分别对应句子中的主语（Subject）、谓语动词（Verb）、宾语（Object），所以，这样的元组也称为 SVO 三元组。该模式的含义是：当符合各自语义约束的三个元组都出现在同一个句子中时，则说明该句子描述了一个“职务变动”类事件中的“任命”事件。同样的道理，第二个抽取模式是一个“辞职”类事件抽取模式，也是一个 SVO 三元组。该模式的含义是：当符合各自语义约束的三个语义类都出现在同一个句子中时，则说

明该句子描述了一个“职务变动”类事件中的“辞职”事件。由于自然语言十分复杂，同样一件事情可以有多种表述形式，所以，尽管第三个抽取模式也是一个“辞职”类事件模式，却是一个二元组。

3.2.5.2 ExDisco 的抽取模式学习过程

ExDisco 中从种子抽取模式通过自扩展的方法学习新的抽取模式的过程如下：

0）给定：

（a） 一个没有标注、没有分类的大文本集；

（b） 几个有代表性的种子抽取模式；

（c） 一个初始的概念语义类集合。

1）用当前的模式集来划分文档集，划分为和场景相关的文档和不相关的文档两个子集。

2）发现新的候选模式：

（a） 自动将相关文档中的每个句子转化为候选模式形式；

（b） 对这些候选模式打分并排序；

（c） 将分值高的候选模式加入种子模式集。

3）用户对新增加的抽取模式进行筛选，确定目前的抽取模式集。

4）返回步骤 1)，直到不再生成新的模式或达到某种循环条件为止。

3.3 面向中文文本的抽取模式获取难点

研究从中文文本中自动获取抽取模式，可以大大提升中文文本信息抽取系统的可移植性，对中文文本信息抽取的广泛应用意义重大。自动获取抽取模式又包含如下四个子问题：

（1）中文文本信息抽取中抽取模式的表示；

（2）抽取模式库的组织；

（3）如何从未标注的中文文本中自动产生候选抽取模式；

（4）候选模式的评估，即怎样的抽取模式是“好”的模式。

其中，问题（3）和问题（4）是自动获取抽取模式的关键所在。另外，中文以及中文文本所特有的一些特点，也使得从中文文本中自动学习信息抽取模式增加了一定

的难度。下面以“职务变动”场景为例给予说明：

（1）通过分析大量真实语料中的职务变动类事件的表述语句发现，对一类特定事件的描述，中文文本中可以用的动词个数要远大于英文文本中可以用的动词个数。

（2）中文的句式更灵活，变换更多，这使得一个意思的表达有更多的形式，也就是说，一类事件在中文文本中有更多的事件抽取模式，并且有许多模式出现的几率很小，但如果忽略的话，由这类模式匹配的事件就不能被正确抽取。

（3）在中文文本中，对一个事件的表述，常有主动语态和被动语态两种形式。例如，例句“山西人大常委会任命胡苏平为副省长。”对应的被动语态形式“胡苏平被任命为山西省副省长。”在真实文本中也经常出现。

3.4 基于自扩展策略的中文文本抽取模式自动获取

基于自扩展（bootstrapping）策略的机器学习方法是目前抽取模式自动获取的重要方法之一。一般来说，自扩展获取抽取模式无需用户提供大量的手工标注语料，只在初始阶段要求用户提供部分已知信息，利用这些信息以滚雪球的方式展开自扩展学习过程，具有启动速度快的优点，曾在词的歧义消解 [126]、通用概念发现 [127]、词典构建 [128][129][130]中得到较早的应用。后来又用于抽取模式的自动学习过程中，例如ExDisco[76][89]、Snowball[90][91][92]两个系统中都采用基于bootstrapping策略进行抽取模式的学习。自扩展方法的关键是寻找未知信息与已知信息之间的链接点，利用同一个信息由不同模式表达，同一模式又可以表达不同信息的特点，使得模式与信息成为相互的链接点；ExDisco[76][89]中利用抽取模式和文本文档的相关性、Riloff等 [128]利用语义类别与模式之间的相互关系，实现系统的自扩展。这些方法在很大程度上减少了人员的参与，提高了系统的可移植性。

综合上面的分析，针对中文文本中信息抽取模式获取的一些特点和难点，本章提出了一种从未标注的中文文本中基于自扩展策略自动获取事件抽取模式的算法，该算法无需用户提供大量的手工标注数据，只在初始阶段要求用户提供几个相关场景中典型的抽取模式作为种子模式。该算法从这几个种子抽取模式出发，通过一个增量迭代的过程发现新模式，在每一轮迭代中从三个不同的层次对抽取模式进行扩展，然后采用类似于 TF/IDF 的评估方法对产生的候选抽取模式进行排序，选择最合适的模式并入到当前模式集。

3.4.1 自扩展策略概要

自扩展策略是机器学习中的一个一般框架，用于提高从未标注数据集中进行机器学习的性能。ExDisco[76][89]和Snowball[90][91][92]中都采用了同样的学习策略。典型的自扩展过程是一个增量迭代的过程，在每一轮迭代中，用新产生的标注数据重新训练学习机，而新训练的学习机又可产生新的标注数据，这些新的标注数据被用于下一轮迭代，这样循环往复，直至收敛。本章采用基于自扩展的策略从未标注的中文文本中学习事件抽取模式，主要在划分相关文档集、产生候选抽取模式和评估候选抽取模式等步骤提出了与已有研究不同的方法，最终生成的事件抽取模式采用直接和间接两种方式进行验证。整个学习过程如图 3.1 所示，它是一个自扩展的框架。从图中可见，该框架包括以下这些步骤：

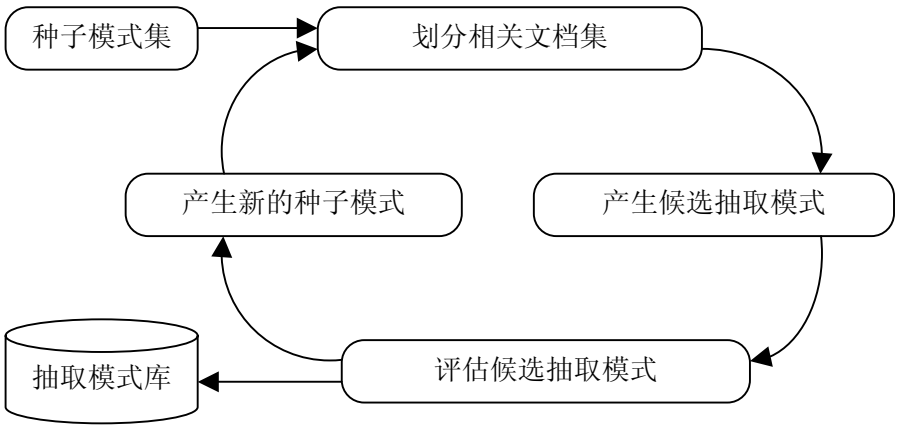

图 3.1 自扩展策略框架

（1）初始化：首先采用手工的方式选择几个事件抽取模式作为种子模式。例如，对于“职务变动”场景，手工选择的种子模式包括：“组织任命某人为某职位”和“某人出任某职位”等一些模式。

（2）迭代过程：

①划分相关文档集：系统根据当前的种子模式集将文档集划分为两部分：和场景相关的文档集和不相关文档集；

②生成新的抽取模式：该过程通过不同的策略从已有的抽取模式以及和场景相关文档集中扩展生成候选抽取模式；

③对候选抽取模式评估：依据一定的策略对候选抽取模式打分，依据分值再对候选抽取模式进行排序；

④产生新的种子模式集：将分值高的候选抽取模式加入种子模式集；

⑤返回步骤①

（3）构建抽取模式库：每一轮迭代将分值高的候选抽取模式并入当前种子模式集构成抽取模式库，经过一定次数的迭代或者系统收敛后，抽取模式库中就包含了用于该场景的最好抽取模式。

3.4.2 自动获取中文文本事件抽取模式

自然语言形式的文本中经常包含有大量的事件信息，中文文本事件信息抽取就是要从中文自由文本中抽取出特定类型的事件信息。传统的文本事件抽取系统大多是基于模式匹配来实现的。简单来说，就是用一组事件抽取模式，在新的文本中基于模式匹配的方法查找特定类型的事件。这些抽取模式是能够传递该类事件信息的句法形式。例如，如果希望从中文文本中抽取“职务变动”类的事件信息，就可以用“某组织任命某人为某职位”，或“某人出任某职位”这些事件模式在自由文本中匹配相应的事件并抽取相关的信息。

在进一步进行论述前，首先给出本章中事件抽取模式的表示方式。本章将一个事件抽取模式用一个四元组表示：{<C-Org> <C-Appoint> <C-Person> <C-Position>}，其中，C-Org 将仅仅匹配组织名称一类的命名实体；C-Appoint 表示“职务变动”类的动词词语（也有个别非动词的情况），可以是“任命”类词语、“辞职”类词语和“调职”类词语；C-Person、C-Position 分别匹配人员、职位类的命名实体或名词性词组。例如，如果一个事件抽取模式是四元组：{<C-Org> <任命> <C-Person> <C-Position>}，这样，这个抽取模式将会从中文文档中匹配如：“新浪网任命×××为公司副总裁”的一类事件。

下面将给出基于自扩展的策略从未标注的中文文本中自动获取事件抽取模式的步骤。首先，原始的文档集需要经过词法分析和中文命名实体识别等预处理。然后，系统根据当前的抽取模式集来划分原始文档集，将原始文档集划分为两部分：一部分是和场景相关的文档，一部分是和场景不相关的文档。接着依据一定的方法从已有模式和相关文档集扩展出候选抽取模式。最后依据一定的策略给每个候选抽取模式打分，并选择得分最高的三个抽取模式加入抽取模式集，然后进入下一轮处理。下面详细论述文档预处理、文档划分、产生候选模式、候选模式评估四个阶段。

3.4.2.1 文档预处理

在进行抽取模式获取前，输入的原始文档首先要经过下面这些预处理步骤：句子切分、分词和词性标注、中文命名实体识别和浅层句法分析。

句子切分的功能是将输入的文档按照排版格式和标点符号分割为单个句子。目前的分割算法一般都是基于标点符号和版面格式进行的。例如，一个句子的结束符号一般为句号、问号、感叹号、后引号等等；一个段落通常以两个字符的空格开始；有时一个句子也以前引号开始等。

分词和词性标注是自然语言处理领域最常用的基本组件之一。本文采用中科院计算所开发的分词和词性标注模块：ICTCLAS。中文命名实体识别模块是自动识别文本中的命名实体，该模块采用我们实验室信息抽取小组开发的中文命名实体识别模块。本章涉及到的主要命名实体有：组织机构名、人名、职位名称、时间表达式等类型。

基于前面的分析可知，自扩展方法的关键是寻找未知信息与已知信息之间的链接点，利用同一个信息由不同模式表达，不同的模式之间存在语义类的交叉和链接点这一基本原则。在进行了中文命名实体识别之后，对其中包含有组织机构名、人名、职位名称这三类命名实体中的任意两类命名实体的句子进行浅层句法分析。因为这些句子有可能描述了特定类型事件，浅层句法分析将这个句子中的名词短语和动词短语等识别出来，供下一步分析利用。

3.4.2.2 文档划分

文档划分的功能就是将输入的原始文档集依据现有的抽取模式集进行划分，根据和场景的相关性大小划分为：场景相关文档集和不相关文档集。

首先我们来探讨对文档划分和候选模式评估起作用的内在机理是什么。对大量的中文自由文本中特定类型的事件实例分析、统计发现：①好的抽取模式出现在许多场景相关文档中，这是容易理解的，因为对特定类型的事件进行表述时，大家都会使用常用的句式进行叙述，而这些句式正是抽取模式的原型。由此可见一个抽取模式在相关文档中出现的次数可以用来衡量该模式的好坏。这里所说的好的抽取模式是指在一类特定事件表述中，人们经常用这种句式表述，所以这些抽取模式出现的次数较多。②如果一个抽取模式在不相关文档中出现的次数很多，则该抽取模式可能是一般的语言模式，而非该场景的抽取模式。③反过来，包含好的抽取模式说明该文档和特定事

件（场景）强烈相关。这三条原则可以作为制定文档划分和候选模式评估的评测标准的指导原则。

设当前的抽取模式集为：$P=\{p_1, p_2, p_3, \ldots\ldots, p_M\}$，共有 M 个抽取模式；输入的文档集为：$D=\{d_1, d_2, d_3, \ldots\ldots, d_N\}$，共有 N 篇文档。依据每篇文档的场景相关度对输入文档集进行划分，文档的场景相关度反映了该文档和场景相关的程度。一篇文档的场景相关度定义为抽取模式集中的抽取模式在该文档中出现的次数和与文档集 D 中的所有文档的平均抽取模式出现次数的比值。计算公式如下：

$$\mathrm{Re}levance\_rate = \frac{\sum_{i=1}^{M} tf_i}{\frac{\sum_{D}\sum_{i=1}^{M} tf_i}{N}} \qquad (3.1)$$

其中，tf_i 是抽取模式 p_i 在一篇文档中出现的次数，N 是文档集 D 中的文档数目。当一篇文档的相关度的值大于一定的阈值时，就认为该文档是相关文档。

3.4.2.3 产生候选模式

产生候选模式是抽取模式的自扩展阶段的开始，这个阶段本章从以下三个层次对已有的抽取模式进行扩展。

（1）基于抽取模式中动词同义的模式扩展。就是对目前抽取模式集中的抽取模式（未进行过动词同义模式扩展的那些）中的动词用同义词代替，产生新的抽取模式。这个层次的扩展主要基于这样一种思想 [125]：在同一场景中，同义词之间有一定的可替换性，利用这一特性可扩展语义的范围。通过该层次的抽取模式扩展，对中文文本中特定类型事件的表述中动词较多这一现象及难点进行了有针对的扩展。具体操作的时候采用通用性较强的《同义词词林》[131]进行已有抽取模式中的动词替换。《同义词词林》是一部对汉语词汇进行语义分类的词典，它根据汉语的特点和实用的原则，确定了以词义为主、兼顾词类、并充分注意题材集中的分类原则。它将各领域的词语分为大类、中类、小类 3 级，共有 12 个大类，94 个中类，1428 个小类，小类下面再以同义词原则划为词群，每一词群以一标题词立目，共有 3925 个标题词。

（2）主动语态和被动语态之间的相互扩展。这个层次的扩展主要基于这样一个分析结果（见 3.3 节）：在中文文本中，对一个事件的表述，常有主动语态和被动语态

两种形式。具体操作的时候我们将主动语态的模式转化成被动语态的模式。例如，将主动语态的抽取模式：<C-Org> <任命> <C-Person> <C-Position>转化成被动语态模式：<C-Person> <被任命（为）> <C-Org> <C-Position>；将被动语态的模式转化成主动语态的模式。例如，将被动语态的抽取模式：<C-Person> <被选举> <C-Org> <C-Position>转化成主动语态模式：<C-Org> <选举> <C-Person> <C-Position>。

（3）基于相同语义项从相关文档集中扩展抽取模式。这个层次的扩展主要基于这样一种思想：虽然中文文本中一类事件的不同表述形式对应了许多的事件抽取模式，但这些模式之间存在着大量的相同语义项。例如，抽取模式<C-Org> <任命> <C-Person> <C-Position>和<C-Person> <出任> <C-Org> <C-Position>中有三个相同的语义项：<C-Person> <C-Org> <C-Position>。在具体操作时，依据是否存在相同语义项抽取出和场景相关的句子，每个句子应该包含三个语义项中至少两个。例如，包含C-Person、C-Position 两个语义项。然后将所有相关句子转化为候选抽取模式。

在产生候选抽取模式阶段，不论从以上三个层次中的那个层次产生了一个候选模式，都要进行查重处理，如果这个候选模式已经存在，则这个候选模式将不再加入新的候选模式集。

3.4.2.4 候选模式评估

生成候选模式集后，系统依据类似于 TF/IDF 的得分策略来判断这些候选模式的“好坏”。TF/IDF（Term Frequency & Inversed Document Frequency）是词频与反比文档频的缩写，原本是计算文档特征权值的常用方法。其中，TF 用于计算词描述文档内容的能力，称作词频，是文本特征项在一篇文档中出现的频率；IDF 用于计算词区分文档的能力，称作反比文档频率。TF/IDF 算法基于这样的直觉知识：①词在文本中出现的次数越多，则该词对文本内容贡献越大；②词所出现的文本数越多，则该词的区分力越小。

本章借鉴TF/IDF的思想对候选抽取模式进行评估，这个方法也是基于 3.4.2.2 小节谈到的三个原则。其中TF表示抽取模式在相关文档中出现的频次；DF表示抽取模式在不相关文档中出现的文档数。本章对已有研究 [14][75]介绍的的评估候选模式方法进行了改进，计算候选模式p_i的得分公式如下：

$$Score_i = \frac{TF_i}{N_R} \times \log \frac{(N - \mathrm{N_R})}{DF_i} \qquad (3.2)$$

其中，TF_i 是抽取模式 p_i 在相关文档集中的文档里出现的次数，N_R 是相关文档集中的文档个数。DF_i 是不相关文档集包含抽取模式 p_i 的文档数，N 是输入文档集中的文档总数。最后，根据每个模式的得分多少，选择分值最高的三个模式并入抽取模式集作为当前的种子模式集。

3.5 基于抽取模式的中文文本事件抽取

通过前面的 3.4 节获取了事件抽取模式之后，就可以从新的文本文档中基于模式匹配进行事件信息抽取了，模式匹配是较成熟的技术，所以这一步实现较为容易。下面从抽取模式的匹配和事件模板的填充两个方面进行简要概述。

3.5.1 抽取模式的匹配

在计算机的研究领域中，文本的模式匹配是最早出现的计算机应用技术之一。在模式匹配发展的早期，出现了KMP[132]、BM[133]等模式匹配算法，但是这些算法用于中文文本的事件模式匹配不太合适。一方面，事件抽取模式是文本中句法到语义的映射，是一种能够传递某种特定类型事件信息的语言表达形式。这种模式反映了自然语言中某种语义约束或句式约束，比传统的文本模式要复杂的多。另一方面，由于汉语的复杂性，中文文本中同一语义类存在大量同义词替代现象，这也增加了事件模式匹配的难度。综合这些分析，本章的中文文本事件信息抽取中的模式匹配分为两步：概念语义类搜索和事件模式匹配。

概念语义类搜索主要从经过文本预处理的新文档中搜索当前所匹配的模式中存在的概念语义类。首先搜索模式中的动词概念语义类，在搜索时基于语义的词语相似度进行匹配，这里借鉴和使用了我们实验室的前期研究成果：基于语义的词语相似度计算 [134]。当在一个语句中搜索到该模式的动词概念语义类后，再在该语句中搜索模式中的名词性概念语义类，这些语义类一般对应到一个相应的命名实体或名词性词组。对这些概念语义类都要进行相应的标识，这些含有相应的概念语义类的句子就作为候选的语句继续进行下面的事件模式匹配。

在对候选语句进行事件模式匹配前，已经对这些语句进行了分词、词性标注以及场景相关类型的命名实体识别。具体的事件模式匹配过程如下：

（1）对候选语句过滤掉修饰性词语和中文停用词。

（2）生成候选语句对应的特征向量。即将动词性概念语义类以及其前后的相关类型命名实体或名词性短语对应的命名实体类型或语义类，生成该语句的特征向量：T_s＝{<CE_1> <CE_2> … <CV> … <CE_k>}。

（3）比较当前模式和候选语句特征向量中动词性概念语义类前后的实体类型或语义类是否一致，如果有两个命名实体类别或语义类匹配，则进行下一步。

（4）将当前模式对应的向量 T_p 与该候选语句生成的向量 T_s 用传统的余弦公式：

$$Sim(T_p,T_s)=\frac{\sum_{k=1}^{K}T_{pk}\times T_{sk}}{\sqrt{\sum_{k=1}^{K}(T_{pk})^2\sum_{k=1}^{K}(T_{sk})^2}} \qquad (3.3)$$

计算两者的相似度，当相似度大于一定阈值时，就认为该候选语句和当前模式匹配，是一个特定类型事件的表述语句。

3.5.2 事件模板的填充

基于模式匹配从新文本文档中找到一个特定类型的事件表述语句之后，依据当前的事件抽取模式和事件模板的对应关系，对该语句进行事件信息的抽取，并将相应的信息填充到事件模板中。这一步需要对事件表述语句中是否出现否定词、时态对事件性质的影响等问题进行处理，然后依据事件模式中的语义类和事件模板中的槽之间的对应关系将该事件表述语句表述的事件信息填充到相应的事件模板中。

3.6 实验结果与分析

本章的实验主要集中在评价利用自扩展策略从未标注的中文文本中自动获取的抽取模式的好坏及算法中几点改进措施的作用上，评价这些抽取模式的好坏并不容易。本章从两个方面来评价前面所提出的方法产生的抽取模式的好坏，一方面，在一个较小的文本集中用该方法自动获取抽取模式，并与手工从该文本集中创建的抽取模式进行比较，用这种方法来直接验证基于自扩展策略自动获取的事件抽取模式的好坏。另一方面，利用该方法在大的文本集中自动获取事件抽取模式，并将这些抽取模式用于新文本中的事件信息抽取，通过评测事件信息抽取的性能来间接验证获取的抽取模式的好坏。所有的实验都针对“职务变动”这一场景模板任务，该场景任务是抽

取中文文本中关于职务变动的事件信息，发现新闻报道中的职位变更信息等。该场景中涉及到的命名实体主要有人员、组织机构名、职位职务名称等。

实验中用到了 6 个种子模式，如表 3.2 所示，这些模式被用于“职务变动”类事件抽取任务。在这些模式中，C-Org 所在的位置应该为一个组织机构名，C-Person 所在的位置应该是一个人名，C-Position 所在的位置应该为一个职位职务名的命名实体。而“任命”、“出任”、“辞去”等职务变动类的动词在种子模式中表示一个动词性概念语义类，表示该位置应该为有相应语义的动词。

表 3.2 “职务变动”场景的种子模式

编号	种子模式	例句
1	<C-Org> <任命> <C-Person> <C-Position>	山西人大常委会任命胡苏平为副省长
2	<C-Person> <出任> <C-Org> <C-Position>	祁东风出任方正科技总裁一职
3	<C-Person> <被选举> <C-Org> <C-Position>	小王被选举为一班班长
4	<C-Person> <辞去> <C-Org> <C-Position>	魏新辞去方正科技董事长职务
5	<C-Org> <C-Position> <C-Person> <离职>	金蝶集团副总裁胡力离职
6	<C-Org> <辞掉> <C-Position> <C-Person>	实达公司辞掉副总裁吴庆生

从表 3.2 所示的 6 个种子模式开始，采用前面提到的基于自扩展策略的方法，从未标注的中文文本中学习“职务变动”类事件抽取模式。

3.6.1 实验数据集

本章实验用到的数据集有两类。一类是从发布“职务变动”新闻的站点获取的 5000 个相关网页，经过网页预处理得到 1367 篇有效文本，这类文本用于构建小的文本数据集，用于直接对获取的抽取模式的进行评价。另一类是《人民日报》1995 年全年的生语料，这些语料是纯文本格式的，其中包含了大量的职务变动类事件的新闻报道，这类文本用于构建大文本数据集，用于通过“职务变动”类事件的抽取来评价本章方法获取的模式的好坏。将任意 10 个月的生语料用来作为基于自扩展策略自动获取抽取模式的输入文本，另两个月份的语料作为间接验证获取的抽取模式的测试集。

3.6.2 实验结果

3.6.2.1 直接验证获取的抽取模式

这组实验用来直接验证本章所提出的自动获取事件抽取模式方法的性能。以表

3.2 中的 6 个模式作为种子模式，以 3.6.1 节给出的 1367 篇文本作为输入文档集。采用该方法自动获取的抽取模式共有 129 个。手工从 1367 篇文本中创建“职务变动”类事件的抽取模式，共创建了 93 个模式，并将这些模式分为：任命类、辞职类、调职类、其他四类。在自动获取的 129 个抽取模式中，有 62 个和手工创建的抽取模式精确匹配，占 66.7%；13 个基本匹配，精确匹配和基本匹配的抽取模式占到 80.6%。可见本章所提出的方法自动获取的事件抽取模式和人工构建的抽取模式吻合性较好。这 75 个抽取模式在四类模式中的分布如图 3.2 所示。

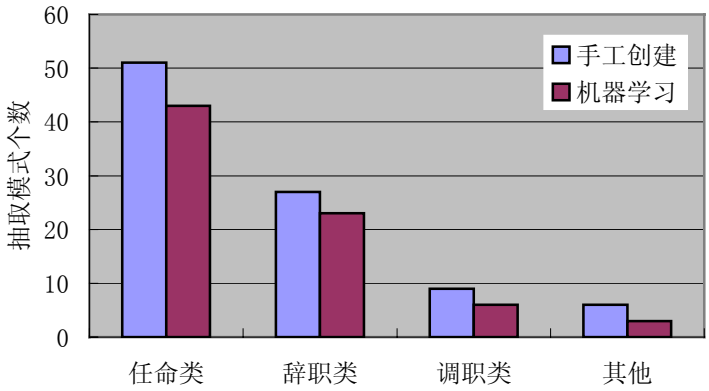

图 3.2 自动获取的抽取模式和手工创建模式的匹配情况

虽然基于自扩展策略从未标注文本中获取抽取模式的研究已经有许多，例如，ExDisco[76][89]和Snowball[90][91][92]等，其中ExDisco用于从文本中获取事件抽取模式，而Snowball系统用于从文本中学习实体关系模式。本章提出的基于自扩展策略的事件抽取模式获取方法以ExDisco中的方法为基础，并从如下三方面进行了改进：①从三个层次扩展抽取模式；②对文档集划分方法和候选抽取模式评估进行了改进；③不需要用户对新增加的抽取模式进行筛选。本组实验不仅直接验证了自动获取的抽取模式的吻合性，而且对这些改进所起的作用进行了验证。实验表明，本章从三个层次扩展抽取模式（详见 3.4.2.3），大大加快了抽取模式的自扩展过程，减少了自扩展迭代过程的次数。ExDisco及其一些改进方法 [75]仅仅将相关文档集中的语句或一些场景相关语句转换后来扩展抽取模式（类似于本章的第三个层次），产生抽取模式的迭代过程次数较多，速度较慢，花费的时间也较长。表 3.3 比较了几种不同的抽取模式扩展方式的性能，主要从最终获取的模式个数、精确匹配或基本匹配的模式个数、系统迭代次数、整个自扩展过程花费的时间等几个方面进行了比较。实验的平台为：DELL原装机；CPU为P4，2.4GHz；内存 512M；硬盘 80G；操作系统为Windows2000；输入文

档集为第一类实验数据：1367 篇网页正文。

表 3.3 不同扩展方式下抽取模式获取的性能比较

抽取模式的扩展方式	模式个数	匹配的模式数	迭代次数	整个过程时间（秒）
ExDisco 的方式	148	67	138	不能精确统计 [⑥]
仅仅用第 3 层次扩展	133	65	169	1036
第 1、3 层次扩展	126	72	85	309
第 2、3 层次扩展	117	68	106	515
三个层次共同扩展	129	75	52	196

从表 3.3 的实验结果可以得出如此结论，从三个层次扩展抽取模式，不仅大大减少了迭代次数，缩减了自扩展过程的时间，而且增加了获取模式的精度。特别是加入第一个层次的抽取模式扩展，即基于抽取模式中动词同义的模式扩展能够极大地提高抽取模式自扩展的性能，可以使迭代次数从仅使用第三层次扩展的 169 次减少到 85 次，耗费时间从 1036 秒缩减到 309 秒，匹配的模式个数也从 65 个增加到 72 个。

为了验证对划分文档集和候选抽取模式评估改进的作用，我们进行了对比实验，对本章所提出方法所产生的抽取模式的情况和 ExDisco 所产生的抽取模式的情况进行了比较，结果如表 3.4 所示。从表 3.4 可见，本章所改进的抽取模式评估方法获取的抽取模式有更好的性能。当然这也和划分文档集等其他辅助技术的改进有一定关系。

表 3.4 不同模式评估方法获取的抽取模式情况比较

抽取模式的评估方法	模式个数	精确匹配的模式数	基本匹配的模式数
ExDisco 的方法	148	56	11
本章的方法	129	62	13

另外，本章所提出的方法不再需要用户对新增加的抽取模式进行手工筛选，候选抽取模式的筛选完全自动进行，而且从表 3.4 所示的实验结果表明，该方法选取的事件抽取模式有较高的吻合性。

3.6.2.2 间接验证获取的抽取模式

间接验证抽取模式是将学习出的抽取模式用于从新文档中抽取事件信息，其中基于抽取模式从新文本文档中匹配“职务变动”类的事件采用 3.5 节介绍的匹配方法。实验中的训练语料和测试语料是 95 年的人民日报语料。实验时用不同的抽取模式集进行“职务变动”类事件信息抽取，这些抽取模式集是：只有 6 个种子模式、从种子

[⑥] 因为在 ExDisco 的模式扩展过程中用户对新增加的抽取模式要进行筛选，以确定目前的抽取模式集。

模式经过自扩展学习出的抽取模式、人工修订后的抽取模式集。实验结果如表 3.5 所示，其中的评测指标见第 1 章 1.3.3 节，综合指标F值中 β 取值为 1。实验结果表明，采用本章提出的方法自动获取的抽取模式能够较好地实现中文文本事件信息抽取，通过自扩展获取的事件抽取模式从新文本中匹配“职务变动”类事件信息，获得的综合指标F值得到 66.34%。这个结果要高于文献 [5]中，GenPAMS支持下的ICTIES进行的飞行事故事件的信息抽取实验的结果：综合指标F值为 63%。其中该实验中使用的实验数据为MUC-7 中提供的关于飞行事故的 200 篇新闻报道语料（英文）。

表 3.5 职务变动事件抽取结果

抽取模式集	模式个数	召回率 R (%)	准确率 P (%)	综合指标 F 值 (%)
Seed patterns	6	26.74	83.02	40.45
Seed + learned patterns	**89**	**61.26**	**72.34**	**66.34**
Manual patterns	80	62.16	72.89	67.10

综合直接验证和间接验证两种方法的验证结果，可以得出如下结论：对于从未标注的中文文本中自动获取事件抽取模式来说，本章所提出的基于自扩展策略的方法性能比已有的方法性能要好。一方面，采用自动获取的事件抽取模式进行中文自由文本事件信息抽取达到了较高的综合指标 F 值：66.34%；另一方面，本章的方法缩短了自动获取模式的时间和迭代次数，提高了获取的抽取模式的吻合性，并且进一步减少了用户的工作量，提高了系统的可移植性。

3.7 小结

自动获取抽取模式对于提升信息抽取系统的可移植性非常重要，它可以大大减少将一个信息抽取系统移植到一个新的场景时的工作量。本章提出了一种从未标注的中文文本中基于自扩展策略的事件抽取模式自动获取方法，该方法从少数几个种子抽取模式开始，通过一个增量迭代的过程发现新模式，每一轮迭代从三个层次对抽取模式进行扩展，然后采用类似于 TF/IDF 的评估方法对产生的候选模式进行评估，选择得分最高的三个模式并入到当前模式集。将该方法用于从中文自由文本中获取“职务变动”类事件抽取模式，实验结果表明，该方法获取的事件抽取模式能较好地实现中文文本事件信息抽取。

第4章 特定类型事件的探测与分类

事件探测和分类是基于触发词探测的事件信息抽取中的首要任务，对事件信息抽取的后继任务至关重要。事件探测采用基于触发词探测从自由文本中找出特定类型的候选事件，该阶段易于实现。传统的事件分类仅仅依据事件表述语句中的触发词，而忽略了触发词前后与事件表述密切相关的大量特征信息，致使分类效果不佳。为此提出了一种基于最大熵模型的事件分类方法，该方法能够综合事件表述语句中的触发词信息及各类上下文特征对事件进行分类。本章对其中的两个关键问题：参数估计、特征模板构建与特征选择进行了详细论述，采用 IIS 算法学习模型参数，使用增量选择方法选择有效特征。应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行了分类实验，结果表明，该方法有较好的分类效果。

4.1 引言

ACE评测会议[31]对事件探测与识别（Event Detection and Recognition, VDR）任务进行了详细规定，要求从文本文档中探测特定类型的事件，并将探测到的事件的相关信息识别出来，经过合并之后以统一的格式存储。ACE评测中的VDR任务支持中文和英文两种语言。基于触发词探测的事件信息抽取方法就是综合ACE评测对事件探测与识别的要求提出来的，该方法将事件抽取分为两步：第一步是特定类型事件的探测和事件的分类，主要探测特定类型事件的表述语句并确定事件的类别或子类别，这将是本章要解决的问题；第二步是从事件表述语句中抽取出事件要素及其语义角色并填充到预定义的事件模板中，这是本文第五章要解决的问题。

事件探测和事件分类是事件抽取的基础，事件探测旨在发现特定类型事件的表述语句，这些语句是进一步信息抽取的数据源。事件探测易于采用基于触发词探测的方法实现，本章将在 4.3 节简要介绍。而事件分类是用于确定该事件表述语句所叙述的事件类别，确定的事件类别正确与否对事件模板的选择以及究竟要抽取哪些事件要素来填充模板至关重要。传统的事件探测和事件分类主要依据触发词（trigger）来确定[54]，触发词是能够很好地概述出该事件中心意义的词。例如，职务变动事件中的“任命”、“辞去”等词语。基于触发词的事件探测和分类是将含有特定触发词的语

句作为候选事件语句并依据触发词对事件进行分类。例如，文本中的语句“方正集团董事长魏新辞去方正科技董事长职务”包含触发词“辞去”，就认为该语句是个离职类事件的表述语句。对大量中文文本中的事件表述语句研究发现：仅仅依据触发词就判定一个语句属于某类候选事件语句很容易出错。其一是有些包含触发词的语句并未表述相关事件；其二是一些词语在多个事件类别中充当触发词。所以简单地依据触发词来确定事件是不可取的，而触发词前后的上下文中包含了对事件类别确定有很大价值的各类特征。例如，触发词前后的一些特定类型的命名实体，一些用于表述某类特定事件的固定句式、短语结构和词语等。基于以上的分析，本章提出了一种基于最大熵模型的事件分类方法，将候选事件语句中的触发词及其上下文信息融合到最大熵模型中进行事件类别的判定，应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行了分类实验，结果表明，这种方法的分类效果明显比仅仅依据触发词的效果好。

4.2 最大熵模型及相关研究

4.2.1 最大熵理论

最大熵理论反映了自然界的一条基本原则：事物是约束和自由的统一体，并且在约束下事物总是争取最大自由度，即最大熵。因此，在已知条件下，熵最大的事物，最可能接近它的真实状态。可以应用最大熵理论来解决现实世界中的许多问题，这些问题都是在已知一些先验知识情况下，需要选择一个合适的模型对问题做出预测，无疑最大熵模型是解决此类问题的最合适模型。具体来说，对于一个事件，往往只了解它的部分情况，对于其它情况则一无所知。那么对这个事件建立模型时，对于已知的部分要尽量地拟合，使模型符合已知的情况。对于未知的情况，则保持均匀分布，即使该事件的熵最大。

最大熵模型是用来进行概率估计的。具体到事件分类的问题，所要解决的问题是判定一个事件表述语句的类别，也就是计算该事件表述语句属于各个类别的概率。例如处理一个类别数量为 5 的分类问题。假设已知包含词“任命”的语句属于职务变动事件类别的概率为 0.8，属于其它类别的概率未知。那么如果一个语句中出现了词语“任命”，就判定这个语句有 0.8 的可能性是属于职务变动类事件，属于其它类别的

概率均为 0.05。如果语句中没有出现“任命”，则这个语句属于各个类别的概率均为 0.2，这就是一个满足约束条件的最大熵模型。当然这里所举的例子比较简单，实际问题中，已知的先验知识往往会提供大量的约束条件，需要对约束条件进行一定的算法处理才能得到符合约束条件的模型。

4.2.2 最大熵模型的一般形式

对于分类问题，给定一些训练样本(x, y)，其中 x 表示上下文，y 表示问题的类别，可根据这些已知的样本构建一个能够对实际问题进行准确描述的统计模型 $p(y|x)$ 用来预测未知的事件。该模型的概率分布与训练语料中的经验概率分布应该相符，而训练样本中上下文信息与输出的经验概率分布 $\tilde{p}(x,y)$ 可由公式：

$$\tilde{p}(x,y)=C(x,y)\Big/\sum_{x,y}C(x,y)\text{，} \qquad (4.1)$$

计算，C(x, y)为(x, y)在训练样本中出现的次数。

最大熵原理表明，x, y 的正确分布应该是在满足已知条件（约束）的情况下，熵最大的分布。分布 p(x, y)的不确定性的数学度量是熵：

$$H(p)=-\sum_{x,y}p(x,y)\lg p(x,y)\text{，} \qquad (4.2)$$

条件分布 $p(y|x)$ 的不确定性的数学度量是条件熵：

$$H_t(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\lg p(y|x)\text{，} \qquad (4.3)$$

其中，$\tilde{p}$ 表示训练样本的概率分布。

如何表示已知的约束条件呢？研究者引入了特征函数的概念来表示已知的约束条件，特征函数一般情况下是一个二值函数 $f(x,y)\to\{0,1\}$，形式如下：

$$f(x,y)=\begin{cases}1 & \text{如果}(x,y)\text{满足某种约束}\\ 0 & \text{否则}\end{cases}\text{。} \qquad (4.4)$$

对于特征函数 f_i，它相对于经验概率分布 $\tilde{p}(x,y)$ 的期望值为：

$$E_{\tilde{p}}f_i=\sum_{x,y}\tilde{p}(x,y)f_i(x,y)\text{，} \qquad (4.5)$$

特征函数 f_i 相对于模型 $p(y|x)$ 的期望值为

$$E_p f_i=\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y)\text{，} \qquad (4.6)$$

这样，模型的约束可以表示为特征的期望等于训练数据特征的期望，即

$$E_p(f_i) = E_{\tilde{p}}(f_i)，\qquad (4.7)$$

那么，对于 k 个特征，求解的模型集合为：

$$C = \{p \mid E_p f_i = E_{\tilde{p}} f_i, i = \{1,...,k\}\}，\qquad (4.8)$$

而最好的模型是使其熵最大的模型 p^*：

$$p^* = \arg\max_{p \in C} H(p)。\qquad (4.9)$$

求解这个最优解的经典方法是拉格朗日乘子算法，可以证明 p^*满足下面的形式：

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x, y)\right]，\qquad (4.10)$$

其中，

$$Z(x) = \sum_{y} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x, y)\right]，\qquad (4.11)$$

k 为特征的数目，$Z(x)$ 为归一化因子，保证对所有 x，$\sum_{y} p(y \mid x) = 1$。参数 λ_i 指示了特征 f_i 对于模型的重要程度。通过在训练集上进行学习，求出 λ_i 的值，就得到了概率分布函数，完成了最大熵模型的构造，式（4.10）即为最大熵模型的一般形式。

4.2.3 相关研究

综上所述，最大熵模型是一种性能良好，且适应性、灵活性极好的统计模型，它可以从数据中提取各种相关或不相关的特征并进行综合处理。自从 1996 年Berger等人[124]首次将它应用于自然语言处理中后，近些年来，最大熵模型被广泛地应用于自然语言处理的各个领域中，包括命名实体识别[35]、文本切分[135]、词性标注[136]、汉语组块分析[137]、歧义消解[138]、文本分类[139][140]等。Nigam等[139]使用词频作为特征函数的值进行文本分类的研究。李荣陆等[140]使用分词和N-Gram两种中文文本特征生成方法用于文本分类研究，并比较了最大熵模型和Bayes、KNN、SVM三种常用的文本分类方法的性能。赵妍妍等[54]首先从训练语料中提取出一些触发词，然后采用《同义词词林（扩展版）》扩展这些触发词来构建触发词表，并探讨了基于触发词表和触发词表加机器学习方法来确定事件类别。本章的下面部分首先简要介绍中文文本中基于触发词的事件探测，据此发现特定类型事件表述语句。然后重点论述使用最大熵模型对基于触发词探测发现的事件表述语句进行子类别和类别判定，和文本分类相比事件

表述语句分类有如下特点：①分类的文本短，大部分都是一个完整的句子，最长的也不过 2～3 个完整的语句；②因为这些语句有可能是事件表述语句，所以语句中包含的信息量大；③这些要分类的事件表述语句已经进行了分词、词性标注及命名实体识别。针对这些特点本章所提出的基于最大熵的事件分类方法和一般的基于最大熵的文本分类方法也是不同的，主要表现在以下几个方面：①采用命名实体和分词相结合的特征生成方法；②对触发词进行了词频统计，统计结果也作为一类特征；③融合了触发词的特征和触发词上下文中的命名实体、短语等各种特征进行事件分类。

4.3 基于触发词的特定类型事件探测

特定类型事件探测旨在从自由文本中发现特定类型事件的候选事件表述语句，这些语句是进一步事件信息抽取的数据源。本文是从人民日报语料（包括人民日报 95 年全年生语料和 98 年 1 月份的熟语料）中探测职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件的候选事件表述语句。

4.3.1 触发词表的构建

基于触发词的特定类型事件探测是将包含有某类事件的触发词的语句作为该类事件的候选表述语句。所以，需要对每类事件构建一个触发词表用于事件探测，实验中用于进行“职务变动”等五类事件探测的触发词表采用手工的方式构建。因为对每类事件来说，其相应的触发词并不是太多。孙斌在他的博士学位论文 [56]中给出了 24 个“职务变动”方面的触发词。袁毓林 [96]在此基础上又增加了 11 个，给出了 35 个“职务变动”方面的触发词。本文在构建过程中参照真实语料，并借助于《现代汉语语法信息词典》[11]和《同义词词林》[131]，构建出的触发词表中的触发词数如表 4.1 所示。附录 1 给出了手工构建的“职务变动”类事件的触发词词表及相应例句。

表 4.1 手工构建的触发词表中触发词数目

事件类别	触发词数	触发词示例
职务变动	75	担任、当选、就任、任命、出任、辞职、罢免、撤销、调任
会见	38	会见、接见、会谈、会面、会晤、约见、拜会、晤面、召见
恐怖袭击	54	袭击、枪击、击毙、杀死、劫持、死亡、轰炸
法庭宣判	116	指控、控告、判决、定罪、逮捕、拘留、赦免、获释、监禁
自然灾害	47	夺去、爆发、发生、地震、天灾、引起

4.3.2 特定类型事件探测

从触发词表中取出一个触发词，然后在中文文本中探测，当探测到该触发词时，就认为该触发词所在的语句表述了一个特定事件，该语句就是一个候选事件语句。该语句的上下文范围有两种确定办法：①通常情况下，上下文的选取是基于核心词左右一定范围进行的，鲁松和白硕[141]对自然语言处理中词语的有效范围进行了定量研究，认为汉语核心词最近距离[－8，＋9]位置之间的上下文范围能包含 85%以上的信息量。就是说核心词之前的 8 个中文词语和之后的 9 个中文词语所包含的信息量超过了 85%，本文将触发词作为核心词。②触发词所在的完整语句。要确定完整语句就需要知道哪些符号是完整语句的结束符。一般来说，一个句子的结束符号有句号、问号、感叹号、后引号等。依据触发词前后最近的两个结束符就可以确定该语句的范围。

基于触发词探测并采用上面的后一种方法用程序从人民日报 1995 年全年的生语料中抽取出来的职务变动、会见、恐怖袭击、法庭宣判、自然灾害 5 类事件候选语句，共得到 27824 条语句，然后经过简单的人工筛选去除那些明显不是表述事件的语句后，剩下 8650 条语句，这些语句的分布情况如表 4.2 所示。

表 4.2 95 年人民日报语料中五类事件候选语句统计表

事件类别	候选语句数	典型候选语句
职务变动	2950	新华社北京 11 月 12 日电（记者李煦）日前举行的北京市十届人大常石会第 22 次会议，决定任命金人庆为北京市副市长。
会见	2150	中央军委副主席刘华清上将今天晚上在钓鱼台国宾馆会见了来访的泰国武装部队最高司令瓦达那差·武提西里上将一行。
恐怖袭击	450	正在撤离的维和部队六日遭袭击，两名士兵受伤。
法庭宣判	2530	长治市中级人员法院一审判决判处主犯于列海死刑，剥夺政治权利终身。
自然灾害	570	科学家们认为，冰岛发生的一次火山爆发可能导致该地气温降低和作物歉收，使人们大量外逃。
总数	8650	

4.4 基于最大熵模型的事件分类

使用最大熵模型进行事件分类首先要建立最大熵模型，其中的两个关键问题是：参数估计和特征选择。参数估计是从训练集学习每一个特征的权重参数，即求解向量

$\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_K\}$的过程。而特征选择是筛选出对最大熵模型有表征意义的特征，包括特征模板构建和依据特征模板进行有效特征的选择。

4.4.1 参数估计

参数估计就是从训练数据集学习权重参数向量Λ的过程。传统的办法是Della Pietra等人在GIS（Generalized Iterative Scaling Algorithm）算法[121]的基础上提出的IIS（Improved Iterative Scaling Algorithm）算法[122]。本章采用IIS算法学习模型参数，其过程如下：

算法 4.1：IIS 算法

输入：一个标注过的事件表述语句集和一组特征函数 $f_1, f_2, \ldots, f_n$

（1）对所有 $i\in\{1,2,\ldots,k\}$，初始化 $\lambda_i=0$；

（2）对所有 $i\in\{1,2,\ldots,k\}$，重复以下过程直到所有 λ_i 收敛：

①根据（4.10）式求解每个事件表述语句的 $p_\Delta(y|x)$；

②对每个参数 λ_i

a）求出下式的解 $\Delta\lambda_i$：

$$\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y)\exp(\Delta\lambda_i f^{\#}(x,y))=\tilde{p}(f_i), \qquad (4.12)$$

其中 $f^{\#}(x,y)\equiv\sum_{i=1}^{n}f_i(x,y)$；

b）根据 $\lambda_i=\lambda_i+\Delta\lambda_i$ 更新 λ_i；

输出：最优参数值Λ，最优模型

4.4.2 特征模板和特征选择

最大熵模型的主要优点是能够在同一个模型中集成不同的特征。所以，建立最大熵模型的另一个关键是如何针对特定的任务为模型选择合适的特征集，用简单的特征集表示复杂的语言现象。包括特征模板的构建和基于特征模板的特征生成和选择。

4.4.2.1 特征模板

特征模板的主要功能是定义上下文中某些特定位置的语言成分或信息对某类事件的出现概率是否有影响。由于本章是根据一个候选事件表述语句中的触发词及其前

后的上下文特征来确定该语句表述的事件类型，因此就由该语句中出现的语言成分及这些语言成分出现的位置来确定特征集合，即要考虑该语句中的触发词以及触发词前后的词、短语、命名实体所具有的特征，图 4.1 是可能的特征空间的图示。根据确定事件类别时影响因素的类别差异，具体定义特征空间为：

（1）触发词信息：触发词及其词性、词频等信息。其中对于触发词的词性，虽然大部分触发词都是动词，但是也有一部分其他词性的词语充当触发词。例如，名词或名词化的动词。而对触发词的词频信息用触发词的触发频和出现频两个值来表征。触发频是指训练集中包含该触发词且归属为特定类型事件的语句占所有包含该词的语句的比率，反映了该触发词出现的语句是特定类型事件表述语句的概率。出现频是训练语料中所有该事件类型的语句中，包含该触发词的比率，反映了特定类型事件表述语句中该触发词出现的概率。

（2）触发词上下文中的命名实体的信息：包括命名实体的类别、相对于触发词的位置等。

（3）句子中其他词或词组的词性标注、位置等信息。

（4）句子中的否定词信息。包括是否出现否定词、该否定词是否改变事件表述语句的意义等。ACE评测会议的事件探测与识别任务（VDR）中关于事件的属性进行了界定 [31]，认为事件的属性包括事件的类型（Type）、子类型（SubType）、语态（Modality）、极性（Polarity）、类属（Genericity）和时态（Tense）等。其中ACE07评测 [31]中给出的一些事件类型及子类型见附录 2。而该特征空间就是针对事件属性中的极性而设定的。

（5）句子中的时态信息，该特征空间是针对事件属性中的时态设定的。

（6）事件表述语句的整体或局部的简单句法结构信息。

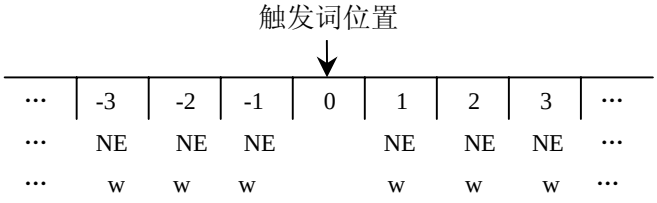

图 4.1 可能的特征空间

根据上面给出的特征空间，本章定义了最大熵模型中的特征模板，如表 4.3 所示，由于该表中每个模板只考虑一种因素，故称之为原子模板。原子模板也可以看作是当前上下文的一个特征函数。当特征函数取特定值时，则该模板被实例化，得到具体的

特征。当模板的取值确定后就可以产生一个特征，这个特征称为原子特征。

表 4.3 部分原子特征模板列表

原子特征模板	特征模板含义描述	原子特征模板	模板含义
Trigger	触发词	Word－1	触发词前后位置上的词
TriggerPOSTag	触发词词性标注	Word－2	
TriggerFreq	触发词触发频	Word－3	
TriggerTermFreq	触发词出现频	Word＋1	
NERType－1	触发词前后位置上的命名实体类型	Word＋2	
NERType－2		Word＋3	
NERType－3		WordPOSTag－1	触发词前后词的词性标注
NERType＋1		WordPOSTag－2	
NERType＋2		WordPOSTag－3	
NERType＋3		WordPOSTag＋1	
NP	名词短语	WordPOSTag＋2	
PP	介词短语	WordPOSTag＋3	
PrivativeAttribute	否定词属性	TenseAttribute	时态属性

由于语言现象十分复杂，仅仅用原子特征不足以表示上下文中的所有特征。通过对表 4.3 中各种原子模板进行组合，构成一些复合特征模板来表示更复杂的上下文环境，如表 4.4 所示。原子特征模板和各种复合特征模板共同构成了最大熵模型的所有特征模板，共有 30 多种模板类型。同样，对于复合特征模板，也是首先对各个原子模板通过实例化，对模板函数取值后，可能会确定一个事件类别，从而产生一个特征，称为复合特征。复合特征表示为二值特征函数的形式与原子特征相似，只是取值时需要满足的条件更多。

表 4.4 部分复合特征模板列表

复合特征模板	特征模板含义描述
NERType-1&NERType=Per	触发词前一个位置为命名实体，其类别为人（Person）
NERType+1&NERType=Per	触发词后一个位置为命名实体，其类别为人（Person）
Trigger-1 & NP	触发词前一个位置是名词短语（NP）
Trigger+1 & PP	触发词后一个位置是介词短语（PP）

4.4.2.2 特征选择

定义好特征模板后，要对训练语料中实际出现的特征进行特征选择，因为并不是所有的特征对模型都有贡献，太多的特征会增加模型训练的时间。最大熵方法不直接与特征选择联系起来，它仅仅提供了一种如何结合模型中多约束的巧妙方法。对于训

练数据，通过特征模板可以产生成千上万的特征，但并非所有特征都是有效的，有些特征和样本数据的多少有关系，在样本数据少的情况下，计算出的样本期望和真实期望并不一致，选择哪些特征是一个很关键的问题[142][143]。这个问题要通过特征选择算法加以解决，假定所有特征的集合是F，特征选择算法要从中选择一个有效特征集合S。常用的特征选择方法有基于频次的特征选择方法[139][144]（Count Cutoff Feature Selection, CCFS）和增量特征选择方法[124]（Incremental Feature Selection, IFS）。CCFS方法是给定一个阈值K，模型只考虑在训练集中出现的次数大于K次的特征。虽然该方法实施起来简单，但不能保证得到一个最小的特征集合。应用IFS算法时，每选一个特征需要对所有的候选特征调用IIS算法，对 λ 重新计算，还要利用训练数据对新模型的对数似然进行计算，以判断该特征是否使模型的对数似然增量最大，实现相对困难，但能够获得有效的特征集。所以在本章的研究中，作者采用IFS方法。

增量特征选择算法的基本思想是：设F是一个候选特征集合，其中只有一部分特征是语言模型的有效特征，有效特征子集设为S，特征选择就是从候选特征集F中选取有效特征子集S的过程。方法如下：开始设有效特征集S为空，然后不断地向S中增加候选特征，每次向S中增加的特征有训练数据决定。每增加一些特征需要对所有的候选特征调用IIS算法，对 λ 重新计算，还要利用训练数据对新模型的对数似然进行计算，以判断该特征是否使模型的对数似然增量最大，实现相对困难，但能够获得有效的特征集。具体算法如下：

算法4.2：增量特征选择算法

输入： $\tilde{p}(x), \tilde{p}(y \mid x)$ 和候选特征组 $\{f_i(x,y), 1 \le i \le N\}$

①初始化：$p^*(y \mid x)$ =均匀分布；

②计算选中的每一个特征的增益，取出增益最大的特征；

③计算增加了上述特征后的最大熵分布作为新的分布；

上述②，③步重复若干次，直到收敛。

输出：选出的一个特征：$\{f_i^*(x,y), 1 \le i \le N\}$ 和相应的 $p^*(y \mid x)$

算法4.2中，特征增益的计算如下：

$$G_{S,f} \equiv \Delta L(S,f) \equiv L(p_{S \cup f}) - L(p_S) \qquad (4.13)$$

计算特征增益可以通过使用先计算出 $p_{S \cup f}$, p_S，然后分别计算训练样本的对数似然。但这样计算量太大，不是一个可行的方案，所以采用下面近似计算的算法。

近似认为，加入一个特征后的模型仅依赖于原来的模型和参数 λ，即：

$$p_{S,f}^{\lambda} = \frac{1}{Z_{\lambda}(x)} p_s(y \mid x) e^{\lambda f(x,y)} \quad (4.14)$$

其中，$Z_{\lambda}(x)$为归一化因子。

$$G_{S,f}(\lambda) \equiv L(p_{S,f}^{\lambda}) - L(p_S^{\lambda}) = -\sum_x \tilde{p}(x) \log Z_{\lambda}(x) + \lambda \tilde{p}(f) \quad (4.15)$$

$$\sim \Delta L(S, f) \equiv \max G_{S,f}(\lambda) \quad (4.16)$$

$$\sim p_{S \cup f} \equiv \arg\max_{p_{S,f}^{\alpha}} G_{S,f}(\lambda) \quad (4.17)$$

为得到式 4.16 的特征增益，可以用式 4.15 求导等于 0，得到方程，用牛顿方法解方程来计算λ，代入式 4.15，求得每个特征的增益，找到增益最大的特征。

4.5 实验与分析

为验证本章提出的基于最大熵模型的事件分类方法，设计了多组实验进行比较。首先根据 4.4 节介绍的方法在 4.5.1 小节给出的训练数据集上对用于事件分类的最大熵模型进行参数估计和特征选择，构建最大熵模型。然后用构建好的最大熵模型对测试数据集中的 5 类事件语句进行分类实验。

4.5.1 实验数据集

训练数据集是用 4.3.2 节基于触发词探测从人民日报 1995 年全年的生语料中抽取出来的职务变动、会见、恐怖袭击、法庭宣判、自然灾害 5 类事件候选语句，共得到 8650 条候选语句。这些语句经过分词、词性标注、命名实体识别后、事件类别标注后作为训练数据集，用于构建最大熵模型。测试数据集是依据是否包含触发词从人民日报 1998 年 1 月的熟语料中抽取出来的上述 5 类事件的候选语句，这些标注过的语句仅仅经过命名实体识别后用于验证基于最大熵模型的事件分类方法。

4.5.2 实验结果及分析

本章使用国际上常用的文本分类评价指标：准确率和召回率的盈亏平衡点（precision/recall breakeven point）处分类器的微平均准确率（MicroP）[145]进行评测，从以下几个方面考察基于最大熵模型的事件分类的性能。

（1）分别对只使用分词、分词＋命名实体相结合这两种不同的语句特征生成方法下分类器的性能进行比较；

（2）选取不同数量特征时分类器的性能；

（3）触发词的触发频、出现频等特征对分类器性能的影响；

（4）基于最大熵模型的事件分类和仅根据触发词的分类性能的比较；

（5）与 Naïve Bayes、KNN 分类器的性能比较。

在事件表述语句的分类阶段，作者分别使用分词＋命名实体，只使用分词两种不同的方法来生成事件语句的特征。在进行测试时也基于这两种不同的方法生成特征，为了对使用不同特征生成方法、不同特征数目时基于最大熵模型分类器性能进行比较，在特征数目从 50 到 500 的情况下，对基于最大熵模型的事件分类方法进行了实验，实验结果如图 4.2 所示。

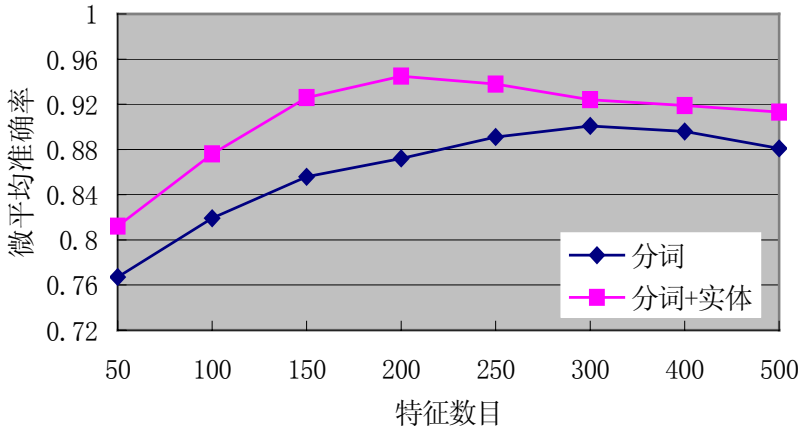

图 4.2 不同特征生成方法的性能比较

从图 4.2 可以得出如下结论：①使用分词＋命名实体的方法来生成语句特征的分类性能要优于只依据分词的方法；②随着特征数目的增加，分类准确率提高，当达到一定数目后，准确率不再升高，反而有所下降；③基于分词＋命名实体的语句特征生成方法用更少的特征能达到更好的分类性能。由此可见，事件表述语句中触发词前后的命名实体的类型、相对于触发词的位置等特征能够提升事件分类的性能。并且命名实体综合了多个词语的特性，用较少的特征就可以得到较高的分类精度。例如，图 4.2 中使用分词＋命名实体来生成事件语句的特征时，特征数在 200 时分类精度最高。而使用分词来生成事件语句的特征时，特征数在 300 时分类精度最高。

在对事件表述语句进行类别确定时，不仅触发词本身对判定事件类别非常重要，而且从大的语料集中统计出的触发词的一些词频信息对判定事件类别也至关重要。本

章进行事件分类时，用到了前面 4.4.2.1 节提到的触发词的触发频和出现频两个词频特征。触发频反映了该触发词出现的语句是特定类型事件表述语句的概率；出现频反映了特定类型事件表述语句中该触发词出现的概率。实验中我们以 4.3.2 节探测到的 5 类事件候选语句为对象，对职务变动、会见、恐怖袭击、法庭宣判、自然灾害 5 类事件的触发词的词频信息进行统计，并将这些统计信息作为特征用于基于最大熵模型的事件分类中。实验时比较了使用这些触发词的词频信息作为特征和不使用这些词频信息的分类效果。使用触发词的词频信息时分类的微平均准确率（MicroP）为 94.7%，而不使用这些特征时微平均准确率（MicroP）仅为 89.5%，相差了 5.2%。由此可见，触发词的词频统计信息对事件表述语句的正确分类有很大帮助。

为了对基于最大熵模型（ME）的事件语句分类方法和其他分类方法进行比较，选择了基于触发词的分类方法、Naïve Bayes 和 KNN，其中，Naïve Bayes 使用多项式模型，KNN 方法 K 值取 60。实验结果如表 4.5 所示（基于触发词的事件分类不受特征数目的影响）。

表 4.5 不同分类方法的性能比较

特征数目	Bayes (MicroP)	KNN (MicroP)	Trigger (MicroP)	ME (MicroP)
50	0.781	0.821	0.784	0.812
100	0.836	0.869		0.876
150	0.892	0.896		0.926
200	0.905	0.925		0.945
250	0.912	0.937		0.938
300	0.893	0.934		0.924
400	0.895	0.929		0.919
500	0.873	0.914		0.913

从表 4.5 可以得出如下结论：①基于最大熵模型的事件语句分类、Naïve Bayes 分类、KNN 分类都远远优于仅仅依据触发词的事件分类方法；②基于最大熵的事件语句分类性能优于 Naïve Bayes 分类，稍好于 KNN 分类方法。可见本章提出的基于最大熵模型的事件分类方法有最好的分类效果，分类的微平均准确率（MicroP）从仅仅依据于触发词的 78.4%提升到了 94.5%。

4.6 小结

发现事件表述语句并判定该语句表述的事件类型是文本事件信息抽取的首要任

务。本章基于触发词从中文文本中探测特定类型的事件，获取候选事件表述语句。然后使用基于最大熵模型的方法对事件进行分类，并且就模型参数估计、特征模板的构建和特征选择进行了详细论述，就两种不同的特征生成方法、特征数目对基于最大熵模型的事件分类器的性能影响进行了实验和分析，验证了触发词的词频信息在分类中的作用，并比较了几种不同的事件分类方法。实验结果表明，使用分词和命名实体相结合的特征生成方法，利用最大熵分类器对事件表述语句进行事件分类有很好的效果，比单纯使用触发词进行事件类别及子类别的确定准确率要高很多。

第5章 基于论元结构的事件要素及其角色识别

第 4 章详细论述了事件探测和事件分类方法，当从文本中探测到特定类型事件，并确定了相应的事件语句所表述的事件类别和子类别后，事件要素及其角色的识别就成为最重要的任务。论元结构是沟通认知与句法结构的桥梁，是语义和句法的接口，在现代句法学和语义学研究中有着相当重要的地位，对于确定句子含义和进行文本理解意义重大。本章提出一种基于触发词的论元结构，利用条件随机场模型来识别事件要素及其角色的方法。首先简要介绍了论元结构理论及其在事件信息抽取中的具体应用。然后利用事件表述语句中触发词的论元和要抽取的事件要素之间的对应关系，以浅层句法分析为基础，把短语或命名实体作为识别的基本单元，将条件随机场模型用于事件要素及其角色的识别。应用该方法对“职务变动”和“会见”两类事件的事件要素及其角色进行识别，分别获得了 77.3%和 74.2%的综合指标 F 值。

5.1 引言

在文本中出现的事件，无论是新闻报道，还是历史事件回放，或者是一般的事件表述，一般都包含着“时间、地点、人物、事件、原因”等要素。即事件发生的时间（When）、发生的地点（Where）、事件的主角（Who）、发生的事情（What）、发生的原因（Why）。因为相应的五个英文词汇都以 W 开头，所以也将这五要素简称为 5W。除了这 5 个事件要素之外，有时还应包括“结果”在内。从中文文本中进行事件信息抽取，就是要将某个事件的事件要素（event argument）识别出来，并确定这些要素在整个事件中充当的角色，然后填充到相应的事件模板中。也就是从文本中抽取该事件涉及到的人员、组织、时间、地点等信息，并将这些信息填充到事件模板的相应槽中。这就是文本事件信息抽取的第二步：事件要素及其语义角色的识别，其实质是确定候选事件表述语句中触发词和其他句法成分之间的语义关系，这些研究属于自然语言处理中的语义分析范畴。

对自然语言形式的句子进行正确的语义分析，一直是从事自然语言理解研究的学者们追求的主要目标。语义分析旨在让计算机能够根据句子的句法结构和句子中每个实词的词义推导出这个句子的意义。对中文文本事件信息抽取中的事件要素及其语义

角色的识别这一具体问题来说，就是要分析事件表述语句中的触发词和它所支配的句法成分（论元）间的语义关系，进而确定该事件的事件要素及其角色，并将这些信息填充到事件模板中。例如，对于“职务变动”的事件表述语句：“1 月 1 日，巴西新总统任命球王贝利为体育部长”，计算机经过一系列处理和语义分析后，能从这个句子中抽取出该事件的相关信息。如图 5.1 所示（注：①图中未标出时间状语；②由于介词大多数从动词演变而来，大部分介词还保留有动词的用法[11]，所以有人也把例句中的“为”看作动词）。类似的例句如：“中央军委任命孔英为陕西省军区政委”。

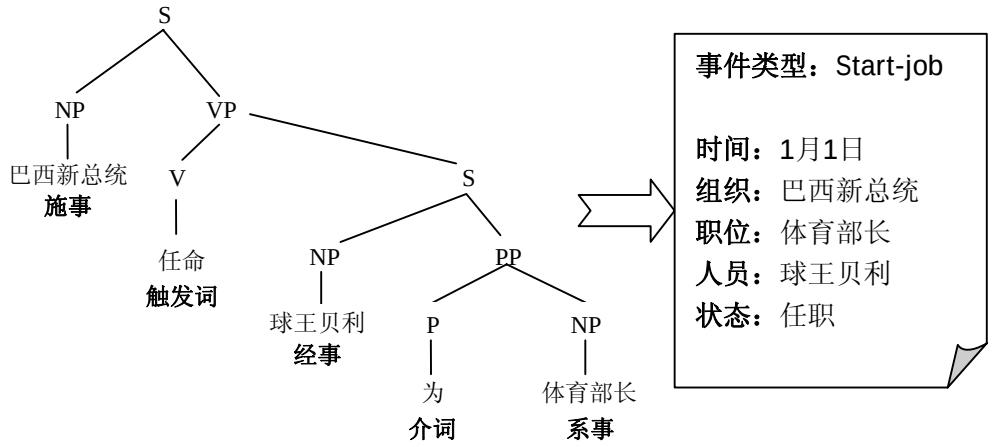

图 5.1 从“触发词”的论元结构抽取事件信息

从事件表述语句中识别事件要素并判定其语义角色是文本事件信息抽取中难度最大的问题，其中涉及到自然语言处理中许多核心问题，特别是动词的论元结构（argument structure）分析和语义角色标注（Semantic Role Labeling,SRL）。动词的论元结构分析旨在研究动词及其所支配的论元之间所存在的各种语义关系，对文本事件信息抽取有很大帮助作用。袁毓林 [95]曾指出：动词的论元结构可以传递到事件模板中，动词的论元最终将成为填入事件模板中的信息条目。例如，图 5.1 的例句中，触发词“任命”所支配的核心论元有：“巴西新总统”、“球王贝利”和“体育部长”，另外还有一个辅助论元：“1 月 1 日”。从句法结构层次上看这些论元分别属于：主语、直接宾语、间接宾语、状语等句法成分；从语义层次上看这些论元分别属于：施事、经事、系事、时间等语义角色。通过这些分析，就能将触发词的论元和事件模板中的槽对应起来。例如施事指自主动作、行为的发出者，该例句中指发出“任命”的人或组织。袁毓林 [94]对现代汉语动词常见的 17 种论元角色及其语义特征进行了详细论述。语义角色标注是根据一个句子中的动词（谓词）与相关的各类短语等句子成分之间的语义关系而赋予这些句子成分的语义角色信息。动词在特定的句式中有固定

的语义角色，这些角色表示动词所涉及的主体、客体或动作、行为、状态、所处的场所、发生的时间、借助的工具等。在自然语言理解的诸多应用领域，SRL有着广泛的应用[97]。例如：问答系统、信息抽取、机器翻译、文本数据挖掘、词汇语义消歧等。

5.2 相关研究

动词的论元结构理论在国内外语言学界已经有大量研究，但是将这一理论应用到自然语言处理领域的研究较少，应用到文本信息抽取的研究就更少。徐烈炯和沈阳[146]认为汉语中的论元结构理论和相关的配价理论（Valence Theory）主要借鉴了法国和德国的配价理论和题元理论（Thematic Theory）。国内的论元结构理论和配价理论研究主要集中在：基本理论研究[59][146][147][148][149]、论元的特征分析[94][150]、汉语语义资源分析[95][151]等。北京大学的袁毓林教授[95][96][152]结合一些具体的实例分析了论元结构在信息抽取中所起的作用。

语义角色标注是近几年自然语言处理领域中的研究热点之一，国内外已经进行了一些卓有成效的研究和实验。目前人们大多采用统计学习的方法解决语义角色标注问题。基本的思想是把句子中连续的句子成分作为标注的基本单元，然后根据一定的语言学知识列出该单元的各种特征，并与该单元的语义角色类型组成学习的实例，最后使用某种统计学习方法对这些实例进行自动的学习，以便对新的实例进行预测。Gildea等人[97]是第一个使用统计的方法进行语义角色标注研究的；Thompson等人[98]使用产生式模型（generative model）进行语义角色标注；刘挺等[100][101]使用最大熵分类器来实现语义角色标注；车万翔等[102]使用混合回旋树核函数进行语义角色标注；Chen等人[103]使用决策树C4.5算法进行语义角色标注的实验；Pradhan等[104]将支持向量机（SVM）用于语义角色标注。

使用统计学习的方法进行语义角色标注，离不开标注好语义角色的语料资源，英语中较为知名的有FrameNet[105]和PropBank[106]两种。其中，U.C.Berkeley开发的FrameNet以框架语义为标注的理论基础，它试图描述一个词汇单元（称为谓词或目标动词，可以是动词、部分名词及形容词）的框架及其关系。FrameNet中包含了 1,462 个词汇单元、800 多个框架、1000 多种语义元素和 49,013 个标注好的句子。PropBank是UPenn在Penn TreeBank句法分析的基础上标注的浅层语义信息。PropBank中定义了 50 多个语义角色，其中核心的语义角色为Arg0-5 六种，Arg0 通常表示动作的施事，

Arg1 通常表示动作的影响，即受事，等等。其余的语义角色为附加语义角色，使用ArgM表示，如ArgM-LOC表示地点，ArgM-TMP表示时间等等。图 5.2 是PropBank中对一个句子的标注实例。对于汉语的语义角色标注资源，Chinese PropBank是UPenn基于Chinese Penn TreeBank标注的汉语浅层语义标注资源⑦；台湾科学研究院的中文句法结构树库（Sinica TreeBank）[107]，利用 74 种语义角色，在句法关系链上添加了语义标签；山西大学的刘开瑛等人⑧正在构建的汉语框架语义知识库（Chinese FrameNet，简称CFN）是一个以框架语义学为理论基础、以真实语料为事实依据的语义词典，其资源用语义Web标记语言描述。

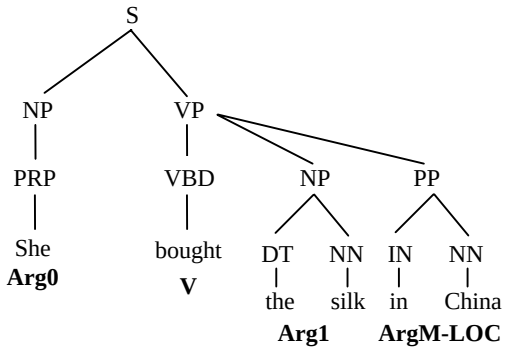

图 5.2 PropBank 中语义角色标注示例

综合以上分析可见，事件要素及其角色识别问题的实质是对触发词的论元进行语义角色标注。本章的下面部分首先简要介绍论元结构理论及其在文本信息抽取中的作用，并分析了触发词的论元和事件模板中的槽之间的对应关系，在此基础上提出了一种基于条件随机场（Conditional Random Fields，CRFs）的事件要素及其语义角色识别方法。CRFs是Lafferty等人[108]于 2001 年提出的一种用于序列数据标注的条件概率模型，是一种判定性模型（Discriminative Model）。CRFs通过定义标记序列和观察序列的条件概率P(S|O)来预测最可能的标记序列。近几年来，CRFs已经被成功地应用到许多自然语言处理领域，例如，词语切分及词性标注[108][109]、浅层句法分析[110][111]、组块分析和短语识别[112][113]、命名实体识别[114][115][116][117]、信息抽取[118][119]等。本章采用L-BFGS（Limited-memory BFGS）算法[120]对模型参数进行估计，并选择基于句法成分的、基于谓词的、句法成分-谓词关系、语义四类特征作为模型特征集。为了验证本章提出的方法的性能，将该方法用于对“职务变动”和“会见”两类事件表述语句中事件要素及其角色识别。实验结果表明，该方法有较好的识别性能。

5.3 论元结构理论

近十多年来，在汉语学界进行了许多配价理论和论元结构的研究，并取得了一些

⑦ 详见 http://www.cis.upenn.edu/~Chinese/
⑧ 见刘开瑛在中文信息学会二十五周年学术研讨会上的讲稿“汉语框架语义知识库构建工程介绍”。

研究成果[94][146][147][148][149][150][151]。这些研究有两方面作用[94][95][96][146]，一方面可以发现施事、受事等语义角色跟主语、宾语等句法成分之间的映射关系，加深我们对汉语的结构面貌的全面认识；另一方面可以为计算机处理中文信息提供比较充分的语义方面的知识资源，满足信息抽取、自动问答系统、机器翻译等涉及语义信息处理技术的需求。 所以说，汉语学界的论元结构及相关的配价理论的研究对中文文本信息抽取，乃至整个中文信息处理都有重大意义。

5.3.1 论元结构的基本概念及含义

下面首先引入几个和论元结构相关的理论概念，参照文献[94][146][148]给出下列术语及其定义：

（1）论元（argument）：指带有论旨角色的名词短语。论元原本是逻辑学的术语。谓词逻辑把简单命题分解为“论元”和“谓词”，“论元”表示命题所涉及的客体事物，“谓词”表示客体之间的关系或与客体事物有关的行为、动作、特性等。一个谓词同一个或多个论元组成一个论元结构。论元结构把命题抽象为谓词和论元，恰好反映了语言表达中陈述与指称两方面的基本内容，于是语言学家从中发现了新的语义描述视角，以论元结构所概括的语言事实、语义现象为基础，构建了以论元结构来描述某些语言现象的研究模式。尽管对论元的理解不尽相同（汉语中的译文也不同，如：谓元、题元、论旨等），但论元理论所蕴涵的语言实质却为大多数学者认同。

（2）论旨角色（thematic role）：由谓词根据其与相关的名词短语之间的语义关系而指派给这些名词短语的语义角色。目前公认的论旨角色有施事者（agent）、感受者（experencer）、受惠者（benefactive）、客体（theme）、使役者（cause/causer）等。论旨角色这一概念的产生及运用反映出语言学家试图透过表层的语法关系，如主语、宾语与述语的结构关系，更深入地了解述语与论元角色之间的语义关系，以及这种语义关系对语法的影响。

（3）论元结构（argument structure）：一个词项的论元结构就是该词项所能拥有的一组已经标有论旨角色名称的论元。

5.3.2 配价理论及论元结构的研究内容

汉语的配价研究及论元结构的研究经常是交织在一起的，配价理论及论元结构研

究的许多内容也是重叠的。

5.3.2.1 配价理论的研究内容

“配价”这一概念来自化学。化学中提出的“价”（valence，亦称“原子价”或“化合价”）的概念为了说明在分子结构中各元素原子数目间的比例关系。最早把化学中的“价”明确引入语法研究中的是法国语言学家特思尼耶尔（Lucien Tesnière）[59]。语言学中引进“价”的概念，为的是说明一个动词能支配多少个属于不同语义角色的名词词组。汉语中的配价理论研究的基本问题 [59][146]如下：

（1）配价理论旨在研究句子中词与词之间的句法关联。因为动词是句子的核心，所以重点研究动词与由名词性词语形成的论元之间的关联。

（2）动词所关联的论元的多少就决定了动词的配价数目。

（3）与动词所关联并能决定动词配价数的论元是指在句子里位于动词前后的主语、宾语等名词性成分。

（4）不同类别的动词可以支配的名词性成分个数和性质都不相同。

利用动词与不同性质名词之间的配价关系来研究、解释某些语法现象，这种研究、分析手段，就称之为“配价分析法”，或简称为“配价分析”；由此而形成的语法理论就称为“配价理论”。

5.3.2.2 论元结构的研究内容

徐烈炯等 [146]认为目前汉语学界进行的“配价”研究所关注的关于名词和动词的一些语法现象，其实在欧洲的“配价理论”中涉及的并不太多，反而在“题元理论”研究中倒是有过很多比较深入的讨论。所以说汉语中论元结构理论和配价理论是交织在一起的。袁毓林 [94]认为汉语中论元结构的研究应该吸收国内外配价理论、格语法、生成语法、论元结构研究的有关成果，并根据计算机处理中文信息的实际需要来确定汉语动词的论元结构的研究内容。认为汉语论元结构研究的主要内容应该包括：

（1）论元属性：指的是动词所能关联的论元的数目，即确定每一个动词能支配多少个必用论元、多少个可用论元。这方面的研究可以参考配价理论的研究成果。

（2）论旨属性：指的是各个论元的论旨角色，即标定这些论元在语义上的功能，即论元的语义角色。这方面的内容可以参考格语法的研究成果。

（3）语法特征：指的是描写这些论元在句法上的功能和所受到的句法约束。语法特征包括句法功能和范畴特征两个方面，前者指各个论元在句子中各自可以充当什么样的句法成分（如主语、宾语、状语等），后者指各个论元通常由什么样的词类范畴来实现（如施事、受事通常由名词性成分实现，致事通常由名词或动词性成分来实现，场所、源点、终点通常由处所性成分来实现），这方面的内容可以参考论元结构的研究成果。

（4）语义特征：指的是描写这些论元的动态的语义特征和静态的语义特征。前者指的是各个论元在述谓结构中表现出来的施动性、受动性等语义特点，后者指实现不同的论旨角色的词语在语义上受到的约束（比如：施事、与事通常由指人名词来实现，受事则既可以由指人名词来实现、也可以由指物名词来实现），对于这方面的内容，可以参考词汇语义学和论元结构理论等的研究成果。

（5）配位方式：描写动词及其论元的句法配置方式，指的是依存于同一个动词的各个论元在句子中的共现和选择限制，即怎样构成一个或几个相关的句式，这方面的内容可以参考配价理论和论元结构理论的研究成果。

5.3.3 论元结构在事件信息抽取中的应用

论元结构是沟通认知与句法结构的桥梁，是语义和句法的接口。论元结构理论在现代句法学和语义学研究中有着相当重要的地位，对于确定句子含义和进行文本理解意义重大。在许多深层次的自然语言处理问题上动词论元结构分析都非常重要。例如，问答系统、信息抽取、篇章理解、自动文摘等。下面具体分析论元结构在文本事件信息抽取中的作用。

（1）论元结构理论有助于对事件表述语句的句法结构进行分析。当我们探测到特定类型事件的候选表述语句之后，利用触发词（一般为动词）的论元结构分析可以为事件表述语句的句法结构分析提供很大帮助。

（2）论元结构理论有助于对事件表述语句中的事件要素及其角色进行识别。“论元”不是纯句法的成分性概念，而是在句法概念中加入语义内容，是句法成分分析的一种扩展和语义成分分析的一种抽象，或者说是句法和语义的一种接口。这有助于对事件表述语句进行浅层语义分析，例如，发现语句中的“触发词”的论元及其语义角色，而这些论元及其语义角色恰好对应了事件要素及其角色。

（3）论元结构理论有较强的解释力。论元结构是认知结构向语义层的投射，又继而投射到句法层次，而在人的认知结构中，动作、行为与时间、地点、原因、方式、目的、工具、始点、终点等因素密切相关，在一定语境中，任一因素都有可能成为对动作、行为施加影响的必要因素。

5.3.4 论元结构和事件模板的对应

在文本事件信息抽取中，事件模板起到把要抽取的信息内容类型化和结构化的作用。比如，用户所关心的一个职务变动事件中的五个信息项目：谁、什么时候、什么组织、什么职务、事件状态（任职还是离职），可以表示为职务变动事件模板中的五个模板元素。这样，与某种特定类型事件相关的模板就是一个事件模板，模板中的槽就是事件的要素。如果把一个事件模板看作是一个句子的语义的某种抽象化表示，那么模板元素之间的关系就是动词的意义，各个模板元素就是动词所支配的论元。因此，袁毓林 [95]指出：动词的论元结构可以传递到事件模板中，动词的论元最终将成为填入事件模板中的信息项目。对于中文文本事件信息抽取来说，触发词的论元将和事件模板中的大部分槽相对应。下面通过分析几个例子进一步感性认识触发词所支配的论元和事件模板中的槽之间的对应关系。

（1）触发词“任命”的论元结构中的论元：

Arg0：进行任命的人或组织

Arg1：被任命的人

Arg2：职位或职务

ArgM-TMP：时间

例句 1：{1 月 1 日}$_{ArgM-TMP}$，[巴西新总统]$_{Arg0}$ **任命**[球王贝利]$_{Arg1}$ 为[体育部长]$_{Arg2}$。

例句 2：[中央军委]$_{Arg0}$ **任命**[孔英]$_{Arg1}$ 为[陕西省军区政委]$_{Arg2}$。

（2）撤职类触发词的论元结构中的论元：

Arg0：发布撤职的人或组织

Arg1：被撤职的人

Arg2：职位或职务

ArgM-CAU：原因

例句 1：[田凤山]$_{Arg1}${因违纪}$_{ArgM-CAU}$ 被[国务院]$_{Arg0}$ **免去**[国土资源部部长职务]$_{Arg2}$。

例句 2：{借丧事敛财}$_{ArgM-CAU}$，[天津市人事局长]$_{Arg2}$ [王志平]$_{Arg1}$ 被**撤职**。

例句 3：[咸阳市人大常委会]$_{Arg0}$ 决定**撤消**[张定会]$_{Arg1}$ [副市长职务]$_{Arg2}$。

5.4 利用条件随机场识别事件要素及其角色

在以上分析的基础上，本节利用条件随机场来识别事件表述语句中的事件要素及其语义角色，该方法以浅层句法分析为基础，把短语或命名实体作为识别标注的基本单元，将条件随机场用于事件表述语句中事件要素及其语义角色的识别。该方法的关键在于模型参数估计和特征选择。具体应用中采用L-BFGS算法学习模型参数，并选择基于句法成分的、基于谓词的、句法成分-谓词关系、语义四类特征作为模型特征集。关于CRFs的基础知识和参数估计部分可以参考第 2 章的 2.3 节，这里不再赘述，下面仅对不同部分进行阐述。

5.4.1 利用 CRFs 识别事件要素及其角色的机理

中文文本中特定类型事件的事件要素及其角色识别的实质就是识别出事件表述语句中触发词所支配的论元的语义角色，即语义角色标注。本章所处理的事件表述语句已经确定了触发词，并进行了分词和词性标注、浅层句法分析和命名实体识别。所以本节所进行的事件要素及其语义角色的识别，是建立在浅层句法分析的基础上，将短语或命名实体作为标注的基本单元，并将该问题看成一个序列数据标注问题，利用条件随机场来实现。这样，识别事件表述语句中的事件要素及其语义角色就变成了如此一个问题：首先，事件触发词所支配的事件要素（论元）被确定；然后，基于 CRFs 模型利用上下文的各种特征对每一个事件要素标记一个合适的语义角色。这样根据识别出的事件要素和相应的语义角色就可以填充相应的事件模板。要完成这两个任务，需要首先确定在应用 CRFs 模型识别事件要素及其角色时所使用的特征集。

5.4.2 识别事件要素及其语义角色的特征集

要利用条件随机场模型进行事件要素及其语义角色的识别的关键是如何针对特定的任务为模型选择合适的特征集，用简单的特征集表示复杂的语言现象。一些实验[117]已经显示了特征选择和归纳对CRFs的性能有明显影响。针对事件要素及其角色识

别这一任务，在详细分析输入文本的上下文信息的基础上，参照文献[94][104][153]对语义角色标注及汉语语义角色的特征分析，本章将特征分成四类：基于句法成分的特征、基于谓词的特征、句法成分-谓词关系特征和语义特征。

5.4.2.1 基于句法成分的特征

基于句法成分的特征是指事件表述语句中一个句法成分中所包含的短语、命名实体、词语所具有的类别、词性等特征以及该句法成分前后的标注单元所具有的特征，这些特征是最基本的特征。本章用到的部分基于句法成分的特征如表 5.1 所示。

表 5.1 基于句法成分的特征列表

No.	特征	特征含义描述
F1	PhraseType	句法成分的短语类别
F2	NEType	句法成分包含的命名实体及类别
F3	HeadWord	句法成分包含的中心词
F4	POS	句法成分的词性标注
F5	FirstWord	句法成分中的第一个词
F6	LastWord	句法成分中的最后一个词
F7	Prepositions	句法成分的前置词
F8	PreviousUnit	句子中该句法成分的前一个标注单元
F9	NextUnit	句子中该句法成分的后一个标注单元
F10	WordNumber	句法成分中词的数目

5.4.2.2 基于谓词的特征

基于谓词的特征主要指待处理的事件表述语句中谓词（触发词）的类型、位置、词义等所具有的特征。本章用到的基于谓词的特征如表 5.2 所示。

表 5.2 基于谓词的特征列表

No.	特征	特征含义描述
F11	PredicatePOS	谓词词性标注
F12	PredicatePostion	谓词的位置
F13	PredicateVoice	谓词的语态
F14	PredicateSense	谓词的词义

5.4.2.3 句法成分-谓词关系特征

句法成分和谓词之间有各种各样的外在关系，如句法成分相对于谓词的位置，谓

词到句法成分的路径等。例如，图 5.3 示意了图 5.2 中谓词“bought”到句法成分“She”所在 NP 的路径为：VBD↑VP↑S↓NP。符号↑表示在句法分析树中向上移动，即由孩子结点向父结点移动，而符号↓表示向下移动。这类特征如表 5.3 所示。

表 5.3 句法成分-谓词关系特征列表

No.	特征	特征含义描述
F15	Path	句法树中，从谓词到句法成分的句法路径。
F16	PathLength	句法成分到谓词的路径长度
F17	Position	句法成分和谓词的位置关系

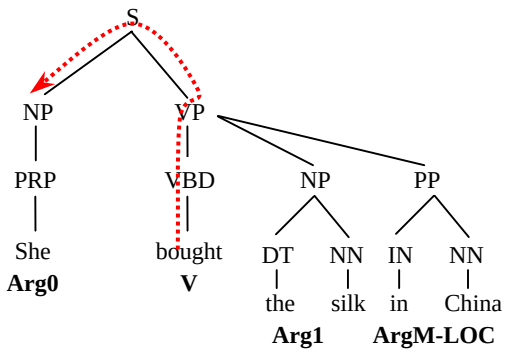

图 5.3 谓词到句法成分的路径示意

5.4.2.4 语义特征

语义特征是语义学中的概念，指的是该句法成分及其上下文在意义上所具有的特征，这些特征对确定语义角色有重要作用。可以通过分析某类句法成分处于特定位置时所具有的共同的语义特征，来解释、说明哪些词语或命名实体可以出现在这些位置，充当相应的角色。这类特征如表 5.4 所示。

表 5.4 语义特征列表

No.	特征	特征含义描述
F18	Independent	句法成分先于谓词所表示的事件独立存在
F19	Causation	该句法成分所指的事物施行某个动作、或造成某个事件或状态
F20	StateChange	使事物或事件状态发生变化
F21	Affected	所指事物承受某一动词的动作、行为的影响

5.4.3 语义角色标注的一般过程

从训练语料构建好条件随机场模型之后，接下来就可以采用 CRFs 来进行语义角

色标注了。进行语义角色标注的基本单元可以是句法成分、短语、命名实体等。在图5.2的句法分析树中，每个非终端结点，如NP、PP等都是句法成分，一般认为每个语义角色是与某一句法成分相对应的。如在图5.2中，Arg0对应一个NP，ArgM-LOC对应一个PP等。然而，这种深层句法分析的结果很难自动得到，尤其对汉语而言，深层次的句法分析还不成熟。所以，本章的事件要素语义角色标注是建立在浅层句法分析基础之上，待标注的基本单元是短语或命名实体等。本章实验中采用了中国科学院计算所张浩的概率句法分析器。

语义角色标注一般分为四个阶段。第一个阶段首先确定目标动词，这里的目标动词是触发词；第二个阶段是过滤掉不可能成为语义角色的成分；第三个阶段是识别目标动词可支配的短语或命名实体（触发词所支配的论元）的边界；第四个阶段是确定触发词每个论元（事件要素）的语义角色。

5.4.4 事件要素及其语义角色的识别

事件信息抽取的过程一般包括两个阶段：第一个阶段是事件的探测和分类，该问题已经在第四章进行了详细论述。第二阶段是从候选事件语句中抽取事件的事件要素，在此阶段就需要对与触发词相关联的短语或命名实体进行语义角色标注。

本章利用CRFs模型对“职务变动”类事件和“会见”类事件进行事件要素及其语义角色的识别，即“职务变动”事件信息抽取和“会见”事件信息抽取。所谓“职务变动”事件信息抽取就是从相关的文本中抽取“职务变动”类事件信息。也就是从自然语言形式的文本中发现“职务变动”类事件并抽取出该事件的要素的过程。该类事件的事件要素包括：事件发生的时间、人员、组织机构、职位等。“会见”事件信息抽取就是从相关的文本中抽取“会见”类事件信息。也就是从自然语言形式的文本中发现“会见”类事件并抽取出该事件的要素的过程。该类事件的事件要素包括：会见的时间、参与者、会见地点、持续时间等。深入分析这两类事件的大量表述语句之后，确定对“职务变动”类事件标注的语义角色包括八个：下达任命的人（Arg0-PER）、下达任命的组织（Arg0-ORG）、被任命者（Arg1-StartJobPER）、离任者(Arg1-LeftJobPER)、职位（Arg2-Position）、事件发生的时间（ArgM-TMP）、事件发生的地点（ArgM-LOC）、事件发生的原因（ArgM-CAU）。确定对“会见”类事件标注的语义角色包括六个：施行会见的人（Arg0-PER）、被会见者（Arg1-PER）、事件

发生的地点（ArgM-LOC）、事件发生时间（ArgM-TMP）、事件发生的原因（ArgM-CAU）、方式（ArgM-MNR）。

5.5 实验结果与分析

为验证本章提出的基于触发词的论元结构，利用条件随机场进行事件要素及其语义角色识别的方法，设计了多组实验。实验时分别对“职务变动”和“会见”两类事件在各自的训练集和测试集上进行相应要素及其角色的识别。另外，在实验中还探讨了不同大小训练集和不同特征集对事件要素及其角色识别的影响。

5.5.1 实验数据集

实验采用的训练数据集是基于触发词探测从人民日报 1995 年全年的生语料中抽取出来的职务变动、会见两类事件表述语句，共得到 5100 条语句。详细情况可参看第 4 章的 4.3.2 节，后又经过第四章的事件类别确定后，得到 2650 条职务变动方面的事件表述语句和 1850 条会见方面的事件表述语句。这些语句经过分词、词性标注、命名实体识别、浅层句法分析、语义角色标注后作为训练数据集，用于训练CRFs模型。测试数据集是依据是否包含触发词从人民日报 1998 年 1 月的熟语料中抽取出来的上述两类事件的表述语句，这些经过词性标注的语句经过人工筛选、命名实体识别、浅层句法分析、语义角色标注后用于验证本章提出的事件要素及其角色识别方法。

5.5.2 性能评估

在对事件要素及其语义角色识别性能进行评估时，采用了三个评测指标：准确率（P）、召回率（R）、综合指标 F 值（F）。准确率表示在识别的全部事件要素及其语义角色中，正确的所占的比值。召回率指语句中所有的事件要素及其角色，正确识别出的语义角色所占的比值。综合指标 F 值则同时考虑 P 和 R。三个指标计算公式如下：

$$P=\frac{\text{准确识别的事件要素及角色数}}{\text{标记的所有事件要素及角色数}}, \qquad (5.1)$$

$$R=\frac{\text{准确识别的事件要素及角色数}}{\text{应该标注的事件要素角色数}}, \qquad (5.2)$$

$$F=\frac{(\beta^2+1)PR}{(\beta^2P+R)}, \qquad (5.3)$$

其中在式 5.3 中，β 决定对 P 侧重还是对 R 侧重，通常设定为 1、2 或 1/2。本章 β 取值为 1，即对二者一样重视。

5.5.3 两类事件语句的事件要素及其语义角色识别结果

首先依据本章提出的方法对“职务变动”和“会见”两类事件表述语句进行了事件要素及其语义角色识别的实验，其中 CRFs 模型中使用了前面提到的全部四类特征集。识别结果如表 5.5 所示。

表 5.5 “职务变动”和“会见”两类事件事件要素及语义角色识别结果

语义角色	P	R	F	语义角色	P	R	F
Arg0-PER	0.847	0.826	0.836	Arg0-PER	0.835	0.809	0.821
Arg0-ORG	0.819	0.802	0.810	Arg1-PER	0.816	0.782	0.798
Arg1-StartJobPER	0.823	0.768	0.794	ArgM-LOC	0.796	0.758	0.776
Arg1-LeftJobPER	0.725	0.769	0.746	ArgM-TMP	0.750	0.696	0.722
Arg2-Position	0.836	0.758	0.795	ArgM-CAU	0.645	0.622	0.633
ArgM-LOC	0.760	0.735	0.747	ArgM-MNR	0.708	0.693	0.700
ArgM-TMP	0.815	0.706	0.757				
ArgM-CAU	0.702	0.695	0.698				
平均性能	0.791	0.757	0.773	平均性能	0.758	0.727	0.742

由于文本事件信息抽取中事件要素及其语义角色的研究相对较少，并且ACE系列评测会议在以前的评测中（2005 年以前），对没有参加评测的单位和个人，不公开ACE评测的结果和相关技术资料。只是在最近一届评测（2007 年评测）之后，在其网站上公开了ACE2007 的评测结果，但是从参加的单位及任务列表来看，涉及事件探测与识别任务（VDR）的只有BBN Technologies一家，而且是在英文语料上进行。并且ACE评测有自己独特的一套评测体系，这使得本组实验的结果和BBN的结果没有可比性。国内参加ACE2007 评测的虽然有 6 家，但是大家的研究大都集中在中文命名实体、实体提及与探测（EMD）、时间表达式的识别等任务上。好在我们可以从另一个侧面来验证本组实验的结果，即从语义角色标注来对实验结果做出评判。近几年，语义角色标注成为国内外NLP领域研究的热点问题，国内哈尔滨工业大学的刘挺教授、车万翔等对该问题进行了深入的研究，他们开发的语义角色标注系统 [101]在CoNLL2005 评测提供的测试数据上评测结果的综合指标F（β 取值为 1）值为 75.6%。而事件要素及其角色识别问题的实质是对事件表述语句中的触发词的论元进行语义角色标注。从表 5.5 可以看到两类事件的事件要素及其语义角色识别的综合指标F平均值分别达到了

77.3%和 74.2%。这个结果表明，本章提出的事件要素及其语义角色识别方法有较好的性能。

5.5.4 不同大小训练集对性能的影响

一般情况下，训练语料集的大小会对事件要素及其语义角色识别的性能产生影响，因此，我们在“职务变动”类事件的实验中进行了几组不同的实验，每组实验使用了不同数量的训练语句来训练条件随机场模型，然后再利用相应的模型在测试集上进行事件要素及其语义角色的识别，用于估计多大的训练语料集对该方法是较为合适的。实验结果如图 5.4 所示。

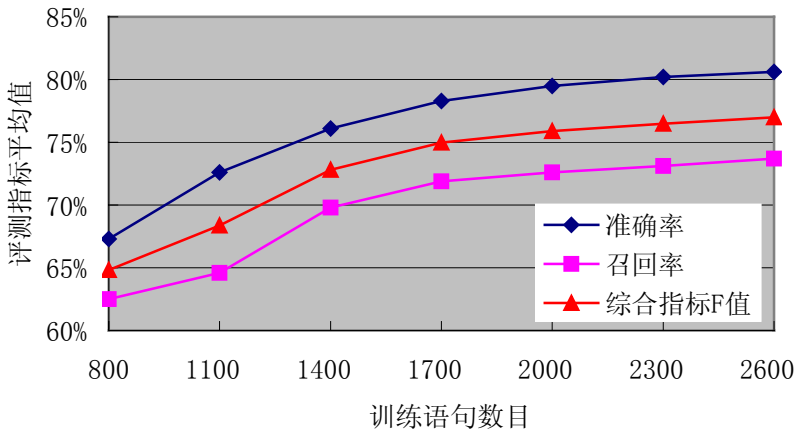

图 5.4 不同大小的训练集对实验结果的影响

从图 5.4 可见，随着训练语句数目的增加，事件要素及其语义角色识别的评测指标：准确率、召回率和综合指标 F 值的平均值不断增加，可见，大的训练语料集有利于性能的提高。其中，在训练语句从 800 句提高到 1700 句的过程中，增幅比较大。从 1700 句增加到 2600 句时，增幅放缓，提高的幅度非常小。所以，在条件允许的情况下，要尽量使用较多的训练语料，以达到最好的性能，不过标注训练语料需要耗费大量人力物力财力，本章的实验中使用的训练语料已经能够得到较高的性能。

5.5.5 特征集的影响

为了比较不同的特征集对事件要素角色标注性能的影响，在进行“职务变动”类事件的事件要素及其语义角色识别时，起先只使用基于句法成分的特征集，然后再加入谓词特征、句法成分-谓词关系特征和语义特征，并观察不同特征集对标注性能的

影响，结果如表 5.6 所示。从表 5.6 可以看出，加入谓词特征后对标注性能影响并不大，平均 F 值从 72.5%提升到了 73.1%；加入句法成分−谓词关系特征后对标注性能影响较大，平均 F 值从 72.5%提升到了 74.8%；加入语义特征性能也提升不小，平均 F 值从 72.5%提升到了 74.5%。

表 5.6 不同特征集的事件要素及其语义角色识别结果

不同特征集	P 平均值	R 平均值	F 平均值
句法成分特征集	0.734	0.716	0.725
＋谓词特征	0.739	0.724	0.731
＋句法成分-谓词关系特征	0.757	0.740	0.748
＋语义特征	0.756	0.736	0.745
全部特征集	0.791	0.757	0.773

5.6 小结

本章首先简要介绍了动词论元结构理论的基本概念，以及汉语中配价理论和论元结构理论研究的内容，并分析了动词论元结构理论在文本信息抽取中的应用。提出了一种基于触发词的论元结构，利用条件随机场模型来识别事件要素及其角色的方法。该方法利用事件表述语句中触发词的论元和要抽取的事件要素之间的对应关系，以浅层句法分析为基础，把短语或命名实体作为识别的基本单元，选择基于句法成分的、基于谓词的、句法成分−谓词关系、语义四类特征作为模型特征集，将条件随机场模型用于事件要素及其语义角色的识别。应用该方法对“职务变动”和“会见”两类事件的事件要素及其语义角色进行识别，在各自的测试集上分别获得了 77.3%和 74.2%的综合指标 F 值。

第6章 基于隐马尔可夫模型的文本事件信息抽取

文本信息抽取是处理海量文本数据的重要手段，事件信息抽取是信息抽取研究中最具挑战性的任务之一。本章针对一些简单的事件表述语句，提出了一种基于隐马尔可夫模型的中文文本事件抽取方法，该方法首先通过触发词探测从文本中发现特定类型的候选事件语句，然后利用隐马尔可夫模型从这些语句中抽取每个候选事件的事件要素，为每一类事件要素构建一个独立的隐马尔可夫模型用于该类事件要素的抽取，构建模型的关键是模型结构的优化和参数估计，本章采用随机优化的方法从最简单的模型获取最优的模型结构，采用最大似然估计学习模型参数，并采用 Good-Turing 平滑方法对参数进行平滑处理。最后将该方法用于“职务变动”和“自然灾害”两类事件信息抽取，结果表明，该方法能较好地实现简单事件的信息抽取,对表述复杂的事件要素的抽取效果较差。

6.1 引言

事件信息抽取是信息抽取研究中最具挑战性的任务之一。本文的第 3 章探讨了事件抽取模式的自动获取以及基于模式匹配的事件信息抽取；第 4 章和第 5 章探讨了基于触发词探测的文本事件信息抽取。下面仍然以“职务变动”事件为例，来探讨基于统计概率模型的事件信息抽取，主要针对一些句式简单的事件语句，用统计概率模型来对某类事件要素的上下文建模，从而实现该类事件要素的抽取。例如：

（1）栗晓峰 1993 年 4 月 10 日**出任**中国女排主教练。

（2）西曼先生于 1997 年加入 SAP 中国公司，**担任**大中国区总裁。

（3）法律系讲师林瑞莲昨天**当选**工人党主席。

（4）12 月 26 日，在中共安阳市第九届委员会第一次全会上，靳绥东**当选**安阳市委书记。

（5）1993 年 4 月 6 日，孔繁森同志到阿里**就任**地委书记。

（6）1 月 1 日，球王贝利被巴西新总统**任命**为体育部长。

（7）12 月 28 日方正科技发布公告称，方正集团董事长魏新**辞去**方正科技董事长职务。

（8）原 CA 中国公司总经理吴沛殷女士已于 6 月**辞职**。

（9）大连市第十二届人民代表大会第一次会议 1 月 10 日**选举**薄熙来为大连市市长。

（10）1996 年初，李长水**担任**了市公安局局长、党委书记。

分析上面列出的 10 个“职务变动”类事件的表述语句，可以总结出这些语句有如下特点：

（1）句式简单。所有这些例句在表述职务变动事件时所使用的句式比较简单。

（2）事件表述语句短。一般都是用一个语句就将整个事件表述清楚。

（3）事件要素全。在这些事件表述语句中，“职务变动”类事件的要素：时间、人员、相关组织、职位等一般都出现。

针对这些特点，借鉴第 2 章中半结构化文本信息抽取的一些方法，本章采用概率统计语言模型来实现这些简单事件表述语句中的事件信息的抽取。

6.2 相关研究

国内外对文本信息抽取已经进行了大量的研究和实验，概括起来主要有三类方法。第一类是基于抽取模式的文本事件抽取，其核心是抽取模式的自动获取。本文的第 3 章深入探讨了从未标注的中文文本中自动获取事件抽取模式。抽取模式是进行信息抽取时所使用的匹配模式，是能够传递待抽取事件信息的句法形式。例如：“职务变动”类事件的抽取模式有：①某公司总裁×××辞职；②公司任命×××为总经理。第二类是基于触发词探测的文本事件抽取 [54]。触发词是能够很好地表述出某类事件中心意义的词。例如，职务变动事件中的“任命”、“担任”、“辞职”等词语。本文的第 4 章论述了事件的探测与分类，提出了基于最大熵模型的事件分类方法。第 5 章在第 4 章的基础上深入探讨了如何识别事件表述语句中的事件要素及其语义角色。第三类是基于概率统计模型的文本信息抽取，这方面的研究已经有很多。Seymore[67]用一个隐马尔可夫模型（Hidden Markov Model，HMM）对计算机科研论文头部信息的所有域进行抽取。张玲等人 [65]采用基于符号特征提取的HMM结构学习方法，对科研论文的引文信息进行抽取。林亚平等人 [69]用基于最大熵的HMM对科研论文头部信息和学术报告信息进行抽取。Freitag[70]使用随机优化技术动态选择最合适的HMM模型结构，并从四类文档集：研讨会公告、公司合并报道、招聘告示、会议征文公告中抽取相应的

信息。虽然将统计模型用于文本信息抽取的研究很多，但这些研究和本章所述的文本事件抽取是有区别的，所处理的文本对象特征不同。上面这些研究中待抽取的信息域都可以看成一个非常紧凑的序列，并且有丰富的版面结构，属于半结构化文本之列，而自由文本中事件的表述往往并不具备这些特征，需要抽取的事件要素域是分散的、稀疏的，有的待抽取域甚至距离事件表述中心[⑨]有一定的距离。

经过对大量事件表述语句的分析研究，本章提出了一种新的中文文本事件信息抽取方法。该方法将后两类方法结合起来，首先在中文文本中通过触发词探测找到候选事件语句，然后在触发词前后一定范围内的上下文中利用 HMM 模型来抽取该事件的各个要素，在进行抽取时为每类事件要素构建一个独立的 HMM 模型，构建模型时对每一个模型在训练数据集上进行了结构优化和参数估计。从中文文本中抽取“职务变动”和“自然灾害”两类事件的实验表明，该方法对一些简单的事件表述语句效果很好，但并不适合于表述过于分散的复杂事件的信息抽取。

6.3 隐马尔可夫模型简介

隐马尔可夫模型是一种用参数表示，用于描述随机过程统计特性的概率模型，它是在马尔可夫模型基础上发展起来的。早在 20 世纪的 60 年代末，HMM的基本理论就由Baum等人建立起来，并由卡内基梅隆大学（CMU）的Baker和IBM的Jelinek等人将其应用到语音识别之中，取得了很大的成功 [154]。1983 年以后，Bell实验室的Rabiner等人发表了一系列系统介绍HMM模型理论和应用的文章使马尔可夫模型和HMM得到了广泛应用，并应用到许多新的领域，例如，自然语言处理领域的词性标注 [155][156][157]、命名实体识别 [39]、文本信息抽取 [65][66]以及生物信息学中的基因序列分析 [158]等。

马尔可夫模型中，每一个状态代表一个可观察的事件，这限制了模型的适用范围。由于实际问题更为复杂，观察到的事件并不是与状态一一对应，而是状态的随机函数，为了模拟这些问题就产生了隐马尔可夫模型。它是一个双重随机过程，其中之一是马尔可夫链，这是基本随机过程，它描述状态的转移，该过程是不可观察（隐蔽）的。另一个随机过程是隐蔽的状态转移过程的随机函数，描述状态和观察值之间的统计对应关系。这样，站在观察者的角度，只能看到观察值，不像马尔可夫模型中的观察值和状态一一对应，因此，不能直接看到状态，而是通过一个随机过程（观察值序列）

⑨ 可以简单地将事件表述中心看作是触发词所在的位置。

去感知状态的存在及其特性，因而称之为隐马尔可夫模型。

综上所述，HMM 可以看成是能够随机进行状态转移并输出符号的有限状态自动机，它通过定义观察序列和状态序列的联合概率对随机生成过程进行建模。每一个观察序列可以看成是由一个状态转移序列生成。状态转移过程是依据初始状态概率分布随机选择一个初始状态开始，输出一个观察值后再根据状态转移概率矩阵随机转移到下一状态，该状态输出一个观察值后再转移到下一状态，直到到达某一预先指定的结束状态为止。图 6.1 是一阶隐马尔可夫模型的状态转移过程并输出观测序列的图示。

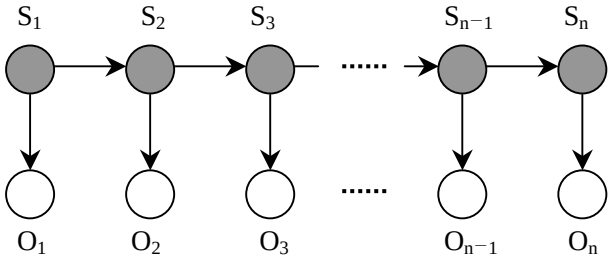

图 6.1 一阶 HMM 模型的图形结构

一个 HMM 有五个组成部分，记为一个五元组{S,O,Π,A,B}，其中：

（1）S 是模型的状态集，设模型共有 N 个状态，记为 $S=\{s_1, s_2, \ldots, s_N\}$。

（2）O 是模型中状态输出符号的集合，符号数为 M，符号集记为 $O=\{o_1, o_2, \ldots, o_M\}$。

（3）Π 是初始状态概率分布，记为 $\Pi=\{\pi_1, \pi_2, \cdots\cdots, \pi_N\}$，其中 π_i 是状态 s_i 作为初始状态的概率。

（4）A 是状态转移概率矩阵，记为 $A=\{a_{ij}\}$，$1\leqslant i\leqslant N, 1\leqslant j\leqslant N$。其中 a_{ij} 是从状态 s_i 转移到状态 s_j 的概率。

（5）B 是符号输出概率矩阵，记为 $B=\{b_{ik}\}$，$1\leqslant i\leqslant N$，$1\leqslant k\leqslant M$。其中 b_{ik} 是状态 s_i 输出 o_k 的概率。

在上述给定的模型框架下，要用 HMM 解决实际问题，首先需要解决评估、解码、训练三个基本问题：

（1）评估问题。给定一个观察序列 $O=O_1O_2\cdots\cdots O_T$ 和模型 $\lambda=\{\Pi, A, B\}$，如何高效率地计算概率 $P(O|\lambda)$，也就是在给定模型 λ 的情况下观察序列 O 的概率。

（2）解码问题。给定一个观察序列 $O=O_1O_2\cdots\cdots O_T$ 和模型 $\lambda=\{\Pi, A, B\}$，如何快速地选择在一定意义下“最优”的状态序列 $Q=q_1q_2\cdots\cdots q_T$，使得该状态序列“最好地解释”观察序列。

（3）训练问题。给定一个观察序列 $O=O_1O_2\cdots\cdots O_T$，以及可能的模型空间(不同的

模型具有不同的模型参数)，如何来估计模型参数，也就是说，如何调节模型λ＝{Π，A，B}的参数，使得P(O|λ)最大。

对上面的三个问题而言，评估问题用于判断最佳模型；解码问题用于寻找最有可能生成观察序列的状态序列，常用韦特比算法进行解决；训练问题用于从已有数据中估计模型的参数。常采用最大似然估计算法（对于已标记的训练集）或Baum-Welch算法（对于未标记的训练集）解决。

6.4 基于 HMM 的中文文本事件抽取

设定要从职务变动相关的文本中抽取“职务变动”类事件信息。也就是从文本中发现“职务变动”类事件并抽取出该事件的相关信息，即事件要素，该类事件的事件要素包括：事件发生的时间、涉及人员、组织机构、职位等。图 6.2 所示的文本中包含一个辞职事件。

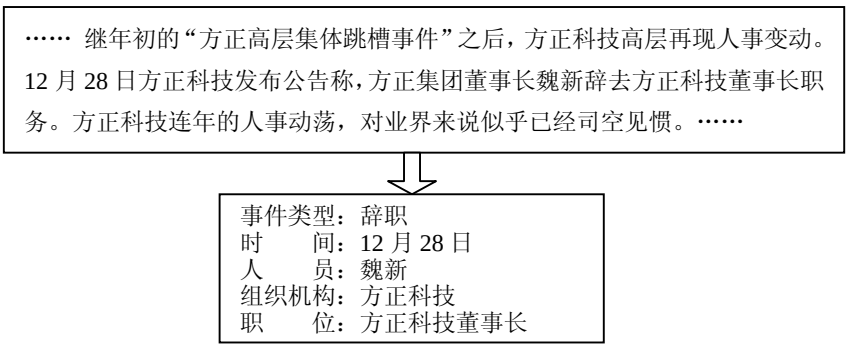

图 6.2 中文文本事件抽取示例

本章提出的事件抽取方法分两个阶段：第一个阶段是通过触发词探测发现候选事件所在的语句，这部分内容已经在第 4 章的 4.3 节进行了论述。第二阶段是利用HMMs从候选事件语句中抽取事件要素。在第一阶段，依据手工构建的触发词表，当在文本中探测到触发词时，就认定该触发词所在的语句表述了一个特定事件，该语句就是一个候选事件语句。该语句的上下文范围有两种确定办法：①通常情况下，上下文的选取是基于核心词左右一定范围进行的，鲁松、白硕 [141]对自然语言处理中词语的有效范围进行了定量研究，认为汉语核心词最近距离[－8，＋9]位置之间的上下文范围能包含 85％以上的信息量。就是说核心词之前的 8 个中文词语和之后的 9 个中文词语的信息含量超过了 85％，本章将触发词作为核心词。②触发词所在的完整语句。要确定触发词所在的完整语句就需要知道那些符号是完整语句的结束符。一般来说，一个句

子的结束符号有句号、问号、感叹号、后引号等。依据触发词前后最近的两个结束符就可以确定该语句的范围。最后对候选事件语句进行预处理，包括分词、词性标注、过滤停用词等。

第二阶段的机理可以描述为：为每类待抽取的事件要素构建一个HMM模型，例如：职务变动事件抽取中，对时间、人物、组织机构、职位分别构建四个独立的HMM模型。候选事件语句中的词语作为这些HMM模型中状态的输出符号，如果模型给定，那么事件抽取过程就是搜索最可能创建词语序列的状态序列。这个问题可以由Viterbi算法通过动态规划解决。用于事件信息抽取的HMM模型结构应该能反映待抽取的事件要素的内容和它的上下文特征。为了实现正确的事件抽取，用于对待抽取的事件要素及其上下文进行建模的HMM模型一般引入四种类型的状态[70][71]：

(1)目标状态：可分为多个状态，用于对目标短语进行建模，例如对要抽取的“组织机构”事件要素建模。

(2)前缀状态：前缀包含一个或多个状态，这些状态被连接成字符串，一个前缀状态仅仅转移到位于该字符串中的下一个状态，或者如果它是该字符串的最后一个状态，则它转移到一个或多个目标状态。也就是说前缀状态可以有一个或多个状态，位于目标状态之前。

(3)后缀状态：后缀状态在结构上类似于前缀。如果状态序列一旦从目标状态集离开，应该经过一个或多个后缀状态。即后缀状态可以有一个或多个状态，位于目标状态之后。

(4)背景状态：背景状态主要是对没有被其它类型状态建模的任何文本建模。一个背景状态仅能转移到它自身或前缀的开始状态，同时它的输入转换仅能来自它自身或后缀的尾端结点。图 6.3 示意了用于事件抽取的两个 HMM 模型结构，其中上面的模型是用于事件要素抽取的最简单 HMM 模型，每类状态都仅有一个状态；下面那个是一个较为复杂的 HMM 模型，包含有四个目标状态、三个前缀状态、三个后缀状态和一个背景状态。

综上所述，本章利用 HMMs 对中文文本事件信息进行抽取有如下特点：（1）每个 HMM 仅仅抽取一类事件要素（例如“职位”或“组织机构”），当多类事件要素被抽取时，则为每类事件要素建立一个独立的 HMM。（2）模型中的状态并不是全连接的，一些转移结构的约束提高了事件要素抽取的准确性。（3）应用该方法进行事件要素抽取不需要预先进行命名实体识别。

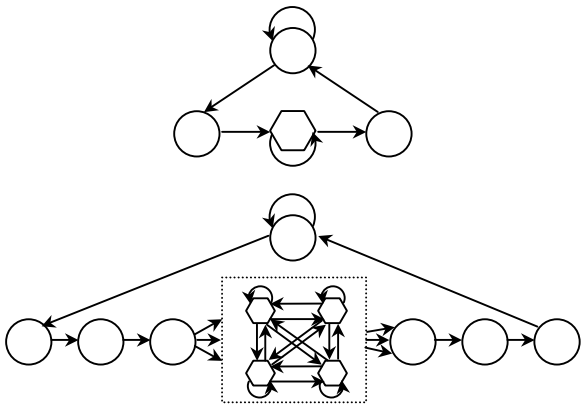

图 6.3 HMM 模型结构的两个示例（圆的结点表示
非目标状态，六边形结点表示目标状态）

6.5 HMM 模型的构建

利用HMMs进行文本事件信息抽取时，首先对不同类型的事件要素建立相应的HMM模型，其中最重要的两个问题[67][70][71]是模型结构优化和模型参数估计。

6.5.1 模型结构优化

6.5.1.1 为什么要优化模型结构？

HMM的模型结构是体现文本上下文关系的重要部分，是对事件要素上下文建模的关键。一些学者通过对数据的分析手工确定HMM的模型结构[73]。但是，人工确定的结构不仅难以推广到大规模的数据集中，而且人对数据的先验知识往往与真实的数据不相符。因此，很多学者研究了HMM结构的自动学习方法[67][70]。

要建立用于事件信息抽取的 HMM 模型，需要对每类事件要素从训练数据集学习得到一个优化的模型结构，以便用这个优化的模型结构对该类事件要素及其上下文进行建模。虽然每个模型包含四类状态，并且这些状态之间的转移也有一定的约束，但用于抽取每类事件要素的模型究竟应该包含几个前缀状态、几个后缀状态、几个目标状态会更符合真实语料，却需要从训练数据集学习确定。图 6.3 所示的最简单模型显然不符合真实情况，例如：事件语句“*原国务院新闻办公室主任赵启正同志出任人民大学新闻学院院长一职*。”中事件要素人物：*赵启正*前面的“*原国务院新闻办公室主*

任”在这里由前缀状态输出，显然这不是一个前缀状态可以完成的。所以，从训练语料中学习 HMM 模型结构对准确抽取该类事件要素非常必要。

6.5.1.2 优化模型结构的方法

目前常用的模型结构优化方法有两种：状态合并方法（State Merging）[67]和随机优化方法（Stochastic Optimization）[70]。其中状态合并方法从最大化最细化的全连通模型开始，使用横向合并（Neighbor-merging）和纵向合并（V-merging）进行状态合并。横向合并将紧相邻的几个同类状态合并为一个状态，合并后的状态通过自转移来模拟多个同类状态的转移。纵向合并将那些从同一个状态转移而来或要转移到同一个状态去的且具有相同类标签的状态合并为一个状态。经过这些合并后，最终就得到优化的HMM模型结构。而随机优化方法是从一个具有最少状态的简单模型开始，利用爬山法在可能的结构空间内，通过循环调节数据集来对模型结构进行优化。

在本章进行模型结构优化时采用了和文献 [70]类似的方法。首先从最简单的模型结构开始（如图 6.3 中上图所示），该结构含有最少的状态，即仅包含一个目标状态、一个前缀状态、一个后缀状态和一个背景状态。对该模型利用爬山法在可能的结构空间中进行优化，在每一轮对现有模型逐个施以一组操作中的每一个操作并在数据集上验证新产生的结构，选择其中最优的一个结构作为下一轮的起始模型结构。对现有模型进行的操作包括以下七个：

（1）增加一个前缀状态：在最后一个前缀状态和目标状态之间增加一个前缀状态，原先位于最后的前缀状态转移到新增加的前缀状态，而新增加的前缀状态则转移到目标状态。

（2）复制一个前缀状态：将某一个前缀状态连同它的转移情况复制，使原先的前缀状态和新复制的前缀状态有相同的连通性。

（3）增加一个后缀状态：在最前一个后缀状态和目标状态之间增加一个后缀状态，目标状态转移到新增加的后缀状态，而新增加的状态转移到原先最前的后缀状态。

（4）复制一个后缀状态：类似于复制一个前缀状态的操作。

（5）增加一个目标状态：类似于增加一个前缀状态的操作，所不同的是目标状态可以自转移。

（6）复制一个目标状态：类似于复制一个前缀状态的操作。

（7）增加一个背景状态：在背景状态字符串中增加一个背景状态。

在每一轮，经过七个操作中的每一个操作会产生一个新的模型结构，将这些新的模型结构作为候选结构，并在一个标注好的测试集上进行验证，将得分最高的结构作为下一轮循环的起始模型结构，直到最后找到一个最优的模型结构或循环终止。

6.5.2 参数估计

6.5.2.1 模型参数估计

为每类事件要素构建独立的 HMM 模型，当模型结构确定后，就可以从标注好的训练语料中用最大似然估计学习模型的参数。计算模型的初始状态概率、状态转移概率和输出概率的公式如下：

$$\pi_i = \frac{C(X_1 = s_i)}{\sum_{j=1}^{N} C(X_1 = s_j)} \quad ，1\leqslant i\leqslant N \qquad (6.1)$$

$C(X_1=s_i)$是训练语料中，以 s_i 为初始状态的序列个数。

$$a_{ij} = \frac{C_{i,j}}{\sum_{k=1}^{N} C_{i,k}} \quad ，1\leqslant i，j\leqslant N \qquad (6.2)$$

$C_{i,j}$ 是训练序列中，从状态 s_i 转移到状态 s_j 的次数。

$$b_{ik} = \frac{C_{i,k}}{\sum_{j=1}^{M} C_{i,j}} \quad ，1\leqslant i\leqslant N，1\leqslant j\leqslant M \qquad (6.3)$$

$C_{i,k}$ 是训练序列中，从状态 s_i 输出词语 o_k 的次数。另外由于标注的训练语料数量有限及数据稀疏现象的影响，所以对这些参数需要进行平滑处理，参数的平滑处理的详细论述见下一小节。

6.5.2.2 数据稀疏问题及平滑处理

在参数求解过程中，由于训练语料规模的限制，数据稀疏问题是经常必须考虑的问题。所谓数据稀疏就是在整个训练语料中，有许多样本的出现概率很低甚至为零，它将会导致经过统计学习得到的参数是不可信的或是不充分的。与自然语言文本中较大的词语特征维数相比，HMM 使用的训练数据常常过于稀疏，从而导致计算输出概

率时缺乏足够的训练数据。为了解决这一问题，人们通常使用平滑方法计算在训练语料中没有出现的词语的输出概率。这些方法的基本思想都是从已出现词语的输出概率中分配一些给未出现的词语。

为了解决HMM模型中常见的数据稀疏问题，并且保证模型的简洁性，本章使用了Good-Turing平滑方法[159]来计算在训练样本中没有出现的词语的输出概率。Good-Turing平滑方法的基本思想是：对训练语料中观察到的样本进行概率调整，将调整出来的概率和按照一定的规则分配给在训练语料中未观察到的样本上，从而消除零概率估计。

如果设 n_r 表示训练语料中出现 r 次的词语（样本）的个数，训练语料的规模用 N 表示，则：

$$N = \sum_r r \times n_r \qquad (6.4)$$

按照最大似然估计方法，出现 r 次的词语的出现概率应为 $P = r/N$。为了克服零概率，可对词的出现频次 r 作适当调整，以 r^* 代替 r。如果取调整频次：

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \qquad (6.5)$$

则得到的概率估计称为 Good-Turing 概率估计，记为 P_{GT}，可由下式计算：

$$P_{GT} = \frac{r^*}{N} = \frac{(r+1)\times n_{r+1}}{N \times n_r} \qquad (6.6)$$

这样，样本中所有概率之和为：

$$\sum_{r>0} n_r \times P_r = 1 - \frac{n_1}{N} < 1。 \qquad (6.7)$$

也就是说，为了保证限制条件 $\sum P = 1$，有 n_1/N 的剩余概率量可分配给所有的未出现的样本（r＝0 的词语）。

6.6 实验结果与分析

为了验证本章提出的基于 HMMs 的事件信息抽取方法的可行性，本章在给出的实验语料上进行了抽取“职务变动”类事件要素的实验，同时为了探讨该方法用于复杂事件表述语句中事件要素的抽取效果，将该方法用于人民日报语料中“自然灾害”类事件要素的抽取实验。最后探讨了上下文范围和模型结构对抽取性能的影响。

6.6.1 触发词表构建

实验中用于进行“职务变动”和“自然灾害”两类事件探测的触发词表采用手工的方式构建。因为对每类事件来说，其相应的触发词并不是很多。详细过程可参见第4章的4.3.1节，构建出的两类触发词表分别包含了75个职务变动类事件触发词（见附录一）和47个自然灾害类事件触发词。

6.6.2 训练和测试数据集

本章实验用到的数据集有两类。一类是《人民日报》1995年全年的生语料，这些语料是纯文本格式的，其中包含了大量的职务变动类事件和少量的自然灾害类事件的新闻报道，用这类实验数据进行基于HMMs的“职务变动”和“自然灾害”两类事件的事件要素抽取的可行性验证。将任意10个月的生语料用第4章4.3.2节的方法将两类事件表述语句抽取出来，进行分词、标注后作为训练数据集。另两个月份的语料作为测试集。另一类数据集是从发布“职务变动”新闻的站点上抓取的5000个相关网页，经过网页预处理得到1367篇有效文本，这些文本产生的实验语料用作比较不同上下文范围和模型结构对事件要素抽取性能的影响。随机抽取出其中的1000篇作为训练集，并将这1000篇中的职务变动事件语句抽取出来并进行了分词、标注等，这些标注后的语句作为HMM模型结构学习的训练集和参数估计的训练集，剩下的367篇文本用作测试集。

6.6.3 抽取性能评估

对抽取性能进行评估时，采用了类似于第1章1.3.3节的3个评测指标：准确率（P）、召回率（R）、综合指标F值（F）。准确率表示在抽取出的全部事件要素中，正确的所占的比值。召回率指正确识别的事件要素占所有应该抽取出的事件要素的比值。计算公式如下：

$$P=\frac{\text{准确抽取的事件要素数}}{\text{抽取出的所有事件要素}}, \qquad (6.8)$$

$$R=\frac{\text{准确抽取的事件要素数}}{\text{应该返回的事件要素数}}, \qquad (6.9)$$

实际评估一个系统时，应同时考虑 P 和 R，但同时要比较两个数值，很难做到一目了然。所以常采用综合两个值进行评价的办法，综合指标 F 值就是其中一种。计算公式如下：

$$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}, \qquad (6.10)$$

其中，β 决定对 P 侧重还是对 R 侧重，通常设定为 1、2 或 1/2。本章 β 取值为 1。

6.6.4 职务变动和自然灾害两类事件抽取结果

本组实验首先在人民日报实验语料集上进行了“职务变动”类事件要素的抽取实验，将实验数据统计综合，各事件要素域抽取结果如表 6.1 所示，其中候选事件语句范围采用 6.4 节提到的第二种方法确定，即触发词所在的完整语句。

表 6.1 “职务变动”类事件要素抽取结果

事件要素	准确率（P）	召回率（R）	综合指标 F 值
人员	0.862	0.689	0.766
组织机构	0.786	0.653	0.713
职位	0.846	0.724	0.780
时间	0.681	0.547	0.607
总体性能	0.795	0.648	0.714

从表 6.1 可以看到，本章所提出的基于 HMMs 的方法在“职务变动”这一场景中的综合指标 F 值达到了 71.4%，将这一结果和别的系统或方法进行比较，发现本章的方法的在“职务变动”这一场景中的抽取性能要略优于其他方法。例如，比较表 3.5 给出了基于模式匹配的“职务变动”事件抽取性能，可见本章的方法的整体抽取性能比基于模式匹配的最好的测试结果（手工修订抽取模式）要好。由此可见，本章提出的这种方法用于自由文本事件信息抽取是可行的，但是这种方法需要进行训练语料的标注，而且对抽取的事件表述有一定限制。另一方面，对有些事件要素域，该方法的抽取性能并不高，如“时间”要素，分析发现这主要因为时间要素较其他事件要素域的边界要模糊的多，边界判断出错的概率较大。

由于“自然灾害”类事件的表述一般过于分散，不像“职务变动”类事件的表述相对集中。为了探讨基于 HMMs 的事件要素抽取方法在这些类型事件表述语句中的抽取性能，在人民日报语料上进行了自然灾害类事件信息抽取的实验。抽取结果如表 6.2 所示。从抽取结果可见，本章的方法在复杂的事件表述语句中效果并不理想，在“自

然灾害”这一场景中的综合指标 F 值仅为 52.2%。

表 6.2 “自然灾害”类事件要素抽取结果

事件要素	准确率（P）	召回率（R）	综合指标 F 值
灾害类型	0.629	0.568	0.597
发生地点	0.567	0.473	0.516
发生时间	0.576	0.497	0.536
死伤人数	0.482	0.405	0.440
总体性能	0.564	0.486	0.522

为了进一步解释“职务变动”和“自然灾害”这两类事件抽取结果的差别，我们从这两类事件表述语句中分别随机抽取 300 条语句进行统计，发现一个语句中完整表述一个“职务变动”类事件的占 36%，两个语句中完整表述的占 79%。而对“自然灾害”类事件而言这两个值分别是 6%和 23%。可见，本章所提出的事件信息抽取方法适合于表述相对紧凑的简单事件语句，对表述过于分散的复杂事件语句效果较差。

6.6.5 上下文范围和模型结构对抽取性能的影响

候选事件表述语句的范围对抽取性能也有一定的影响，在实验中对 6.4 节介绍的两种确定候选事件语句范围的办法进行了比较。①距触发词距离[－8，＋9]位置之间的上下文，记为D_Trigger。②触发词前后最近的两个句号之间的语句，记为Full_Stop。两种办法在职务变动场景的四类事件要素上的抽取性能比较见图 6.4。从图 6.4 中可以看出，方法②确定的语句范围下抽取性能要好于方法①确定的候选语句范围，“职务变动”类事件的四个要素域的抽取结果方法②的性能都要好于方法①。这主要由于方法②确定的事件表述语句一般是完整的语句，而方法①则会造成多余词或缺少某些词语的现象，给抽取模型增加噪声。

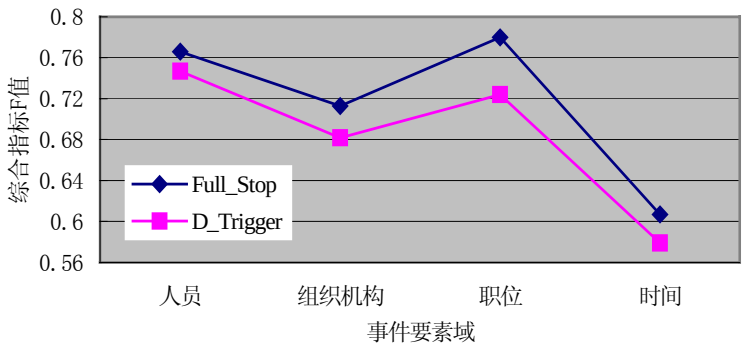

图 6.4 不同上下文范围的抽取结果比较

另外实验中还对比了最简单的模型结构和经过优化后的模型结构在“职务变动”场景中的事件要素抽取性能，如图 6.5 所示。从图 6.5 可见：优化模型结构的抽取性能要明显好于最简单模型结构的抽取性能，优化模型结构的所有事件要素域的抽取结果都高于最简模型结构的结果，有的事件要素甚至要高出 10 个百分点。由此可见对用于抽取事件要素的 HMM 模型结构的优化非常必要。

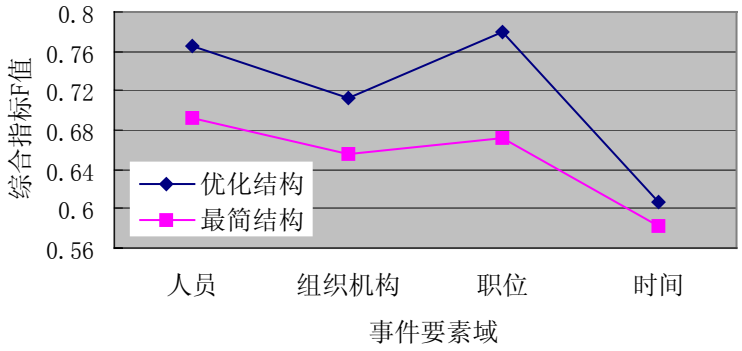

图 6.5 最简单模型与优化模型抽取结果比较

6.7 小结

事件信息抽取是信息抽取研究领域最具挑战性的任务之一，它的相关技术及抽取性能制约着信息抽取更广泛更深入的应用。针对一些简单的事件表述语句，本章提出了一种新的中文文本事件抽取方法，该方法首先通过触发词探测发现候选事件语句，然后对事件的每一类事件要素构建一个独立的 HMM 模型去抽取该要素，实验结果表明，该方法对简单的事件表述语句有较好的抽取性能。

第7章 面向 Web 的信息抽取原型系统构建

面对互联网时代庞杂无序的海量信息，智能高效地处理及深层次综合利用 Web 信息有着重大的社会价值和现实意义。本章以实验室的在研项目“特定领域网络信息提取服务系统”为背景，以前面章节的技术方法作支持，在分析研究已有的几个典型信息抽取系统基础上，设计并实现了一套面向 Web 的文本信息抽取原型系统：WebIE。本章首先给出该系统的设计目标和体系结构，并简要介绍了各模块的功能，然后给出了事件信息抽取模块的详细设计，最后对该系统的整体性能进行了评测。

7.1 引言

随着计算机的普及和互联网的飞速发展，网上信息呈现爆炸式的增长态势。同样在我国，伴随着我国互联网络的发展，越来越多的人开始使用互联网。据中国互联网信息中心（CNNIC）统计 [160]，截止 2006 年 6 月，我国网民人数达到 1.23 亿人，上网计算机数达到 5940 万台；中国网站总数达到了 843,000 个，网页总数达到 44.7 亿页。本次调查结果还显示，网络已经成为网民获取信息的主要途径之一，网民选择的比例为 82.6%（多项选择），如图 7.1 所示 [160]，已经远远超过了其他途径，如电视 64.5%，报纸 57.9%。由此可以看出，对于许多人来说，网络已经成为其获取信息的最主要途径，其次才是大众化的电视，然后是纸质的平面媒体，最后是广播。

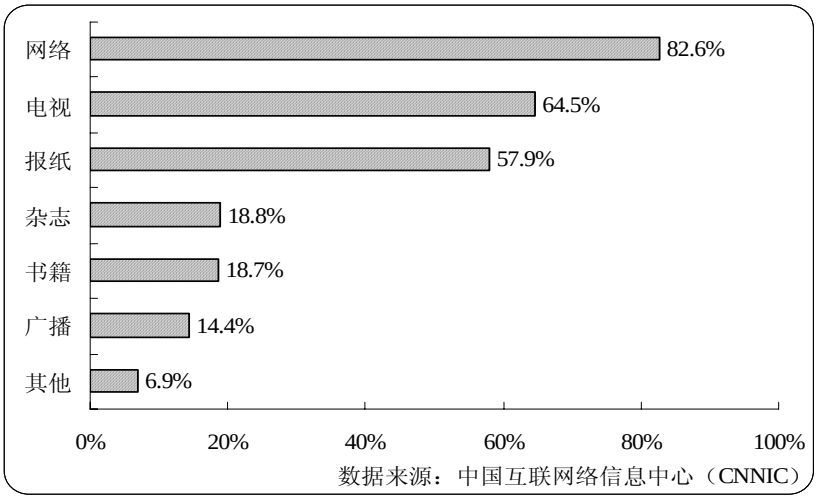

图 7.1 网民获取信息的主要途径

而且在互联网这个分布式信息空间中蕴涵着具有巨大潜在价值的信息，如何从这个海量数据源中发现有用的信息或者知识，成为急待解决的问题。网上信息有多种媒体形式，但其中占主要地位的还是文本信息。统计显示，网上信息 80％以上的形式是文本。例如，新闻报道、评论文章、产品介绍、统计报告等都是文本形式。如何从这些文本数据中抽取和发掘出有用的信息和知识已成为一个日趋重要的课题。本章以实验室的在研项目“特定领域网络信息提取服务系统”为背景，以前面几章介绍的文本信息抽取技术作支持，设计并实现了一个面向 Web 的文本信息抽取原型系统：WebIE（Web Information Extraction）。

7.2 已有的信息抽取系统简介

在信息抽取的研究历程中，出现了许多信息抽取系统。例如，国外的PLUM[78]、FASTUS[79]、GE NLTOOLSET[80]、PROTEUS[81]、AutoSlog[82]、PALKA[83]、CRYSTAL[84]、LIEP[85]、HASTEN[86]、AutoSlog-TS[87]、TIMES[88]、ExDisco[76][89]、Snowball[90][91][92]、GATE[161]，等等。国内的GenPAM[5]、InfoX[56]等。虽然这些系统大多还不是一套完整的信息抽取系统，有的只是实现了信息抽取中的一部分功能，但是也有许多值得借鉴的地方。在这一节里，我们对国内外一些典型的、相对而言比较完整的信息抽取系统进行简要概述，重点集中在这些的系统实现的功能及机理、模块组成等部分。

7.2.1 国外的一些典型信息抽取系统

国外的信息抽取系统大都是基于模式匹配的，为了进行抽取模式的学习，人们先后在不同的信息抽取系统中采用过各种抽取模式获取方法。下面简要概述两个典型的信息抽取系统：PROTEUS[81]和TIMES[88]。

7.2.1.1 纽约大学的 PROTEUS

PROTEUS[81]是美国纽约大学研发的一个信息抽取系统，该系统的核心抽取引擎包括了七个模块：（1）词法分析（Lexical Analysis）；（2）命名实体识别（Name Recognition）；（3）浅层句法分析（Partial Syntactical Analysis）；（4）事件抽取（Event Extraction）；（5）指代消解（Reference Resolution）；（6）篇章分析（Discourse Inference）；

（7）模板输出（Output Generation）。该系统的体系结构[81]如图 7.2 所示。

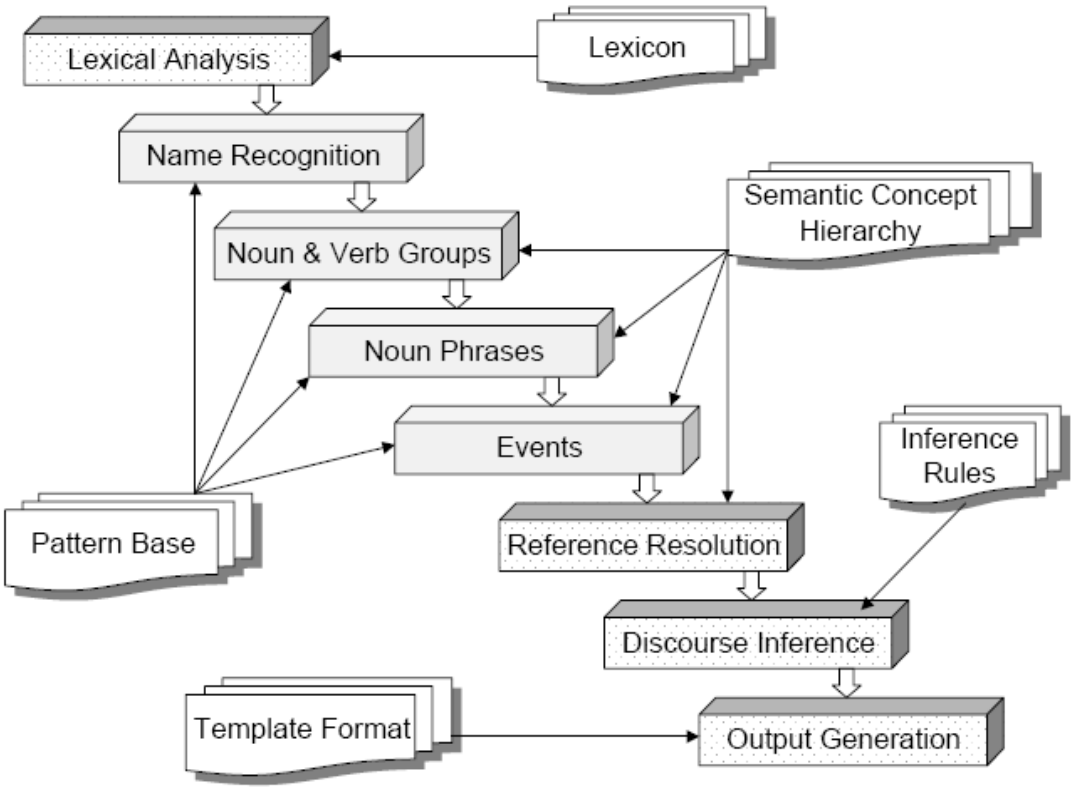

图 7.2 纽约大学 PROTEUS 信息抽取系统体系结构

词法分析模块首先将文档分割为句子，然后基于词典（Lexicon）进一步将句子切分成词并进行词性标注。命名实体识别模块识别文本中的命名实体，该模块在识别命名实体的过程中要使用模式库中相关的命名实体识别模式。浅层句法分析主要包括对名词、动词词组（Noun & Verb Groups）和名词短语（Noun Phrases）的分析，这些分析也要利用模式库中相关的语言模式，并且要用到语义概念层次知识库（Semantic Concept Hierarchy）。事件抽取模块利用模式库中的事件抽取模式从文本中匹配候选的事件语句，在事件抽取过程中也需要语义概念网络知识库的支持。然后对候选的事件语句进行指代消解和篇章分析，并进行事件的合并最终通过模板输出模块将抽取到的事件信息输出。其中 PROTEUS 系统中的模式库中的模式采用手工的方式构建。

7.2.1.2 TIMES

TIMES[88]是一个基于WordNet和语料标注的信息抽取系统，该系统的功能是：在领域无关的概念层次知识库WordNet的支持下，用户通过一个图形用户界面给出含有某类事件表述的语句，系统调用部分句法分析功能模块对该语句进行浅层句法分析，

并通过界面来指导系统从语法和语义两个层次对分析后的语句进行泛化，产生具有一定概括能力的事件信息抽取模式。然后再利用这些模式从新的文本数据中抽取特定类型的事件。其中TIMES系统进行模式学习的基本流程如下：

（1）用户通过图形用户界面选择一个含有特定类型事件的表述语句；

（2）系统对该语句进行分词、词性标注、命名实体识别和浅层句法分析，并以每个短语最后的一个词作为该短语的中心词；

（3）用户通过系统界面来指导系统将相关的名词短语与其所能充当的事件角色关联起来；

（4）用户对某些有歧义的中心词进行词义消歧；

（5）系统记录用户的相关操作并形成相应的特例模式；

（6）系统从语法和语义两个方面对形成的特例模式进行泛化，形成一个泛化模式。其中语法泛化的途径是：去除某特例模式中某些元素，然后改变剩余元素之间的先后顺序。语义泛化的途径是：将特例模式中某些元素用它们各自的上位概念来代替。某个概念的上位概念来自概念层次知识库 WordNet。

7.2.2 国内的信息抽取系统

国内开展信息抽取的研究起步比较晚。台湾大学在信息抽取方面的研究相对较早，它的资讯工程研究所的自然语言处理实验室从 90 年代初就开始了信息抽取的研究，并在MUC-7 提交了一个TE系统参加了测试，测试了中文命名实体（人名、地名、时间等名词性短语）的识别，取得了与英文命名实体识别系统相近的性能 [28]。东北大学从上世纪 90 年代末就开始了信息抽取方面的研究，朱靖波、姚天顺等 [162]构造了一个从文本中抽取计算机设备名称、用途、生产厂家等信息的中文信息抽取的模型系统。上海交通大学的李芳 [57]等实现了一个多语种投资信息抽取实验系统。北京邮电大学的李蕾等 [58]实现了一个财经新闻领域的中文信息抽取实验系统，并探索了信息抽取技术在移动信息服务中的应用。

北京大学的孙斌在定义信息的继承-归纳这一思想的基础上，开发了一个中文信息提取原型系统InfoX[56]，它支持灵活的信息描述机制，用户可以由预定义的信息描述工具来定义新的信息提取任务。在该系统中孙斌 [56]提出了一种基于“状态跃迁链”的职务变动事件抽取技术，称为ST-算法(State-Transition Algorithm)。其处理过程类似

于原子物理中费米子(Fermion)的能级跃迁。一个费米子具有唯一的一个状态，只有该状态为空时才允许其它费米子占据。费米子在能级上相继的跃迁构成一个事件链，每个跃迁对应两个能级，二者满足相继的条件，即二者的配对表示了一个事件。这个算法的步骤如下：

（1）构造状态集 S ：一个状态是标识(label)-域的 n 元组，s =def (l1 : B1, …, ln : Bn);

（2）识别跃迁算法：t(si , sj) 画出跃迁图，每条边连接的两个状态具有至少一个相同的域；以某个（或几个）相同的域为准，使相继的状态配成跃迁状态对；

（3）为所有未配对的状态构造隐含状态，即“隐状态”，其某些域具有 UNKNOWN 值。

（4）输出：所有配对的状态对应两个相继的事件；配对的状态-隐状态对应一个事件和一个“隐事件”（通常略去）。

该系统选取北大计算语言所加工出来的 98 年“人民日报”语料进行测试。通过人工阅读头两个月的语料（约 17.5MB文本），从中找出了 70 多个职务变动事件，其中“任职”45 个，“离职”16 个，“调职”11 个。实验结果 [56]如表 7.1 所示：

表 7.1 InfoX 抽取结果

事件	实际数目	提取数目	正确数目	Recall	Precision
Start_job	45	38	19	42%	50%
Leave_job	16	15	7	44%	47%
Change_job	11	8	6	54%	75%
Total	72	61	33	45%	54%

7.3 WebIE 的设计目标和体系结构

7.3.1 WebIE 的设计目标

面向 Web 的文本信息抽取原型系统：WebIE 的设计目标是能够从互联网上的海量文本数据中自动抽取出特定类型的事实事件信息，并将这些信息转化为结构化的数据存储到一个数据库中供用户查询或进一步分析利用。例如，可以从 Web 上抽取丰富的事实信息：企事业单位的联系信息，某类产品的型号、属性、价格等信息，房屋租赁信息，工作招聘信息等。或者从 Web 上抽取特定类型的事件信息：国家及各省市行政部门的职务变动信息，突发事件信息，自然灾害信息等。该原型系统能根据用户的需

求预先设定要抽取的事实或事件类型，并产生相应的事实事件模板，然后自动地从Web上搜索并爬取相关的网页，接着通过网页预处理模块将网页正文取出，再利用文本信息抽取技术从中抽取特定的事实事件信息，最后通过用户容易接受的方式显示输出或存储到数据库中。

7.3.2 WebIE 的体系结构

在探讨该原型系统的体系结构之前，首先看看信息抽取系统的主要组成模块和一般过程。Hobbs[163]提出信息抽取系统一般应有文本分块、预处理、过滤、预分析、分析、片断组合、语义解释、词汇消歧、共指消解、模板生成十个模块组成；Cowie和Lehnert[21]认为典型的信息抽取系统有六个模块组成：过滤、词语切分与标注、语义标注、浅层分析、指代消解、模板生成。

面向Web的文本信息抽取的一般过程如图7.3所示，有以下六个步骤：（1）网络爬虫模块从Web获取领域相关的网页；（2）网页预处理模块抽取网页正文；（3）文本的分词、词性标注及命名实体识别；（4）实体关系抽取模块抽取事实信息；（5）事件信息抽取模块抽取特定类型的事件信息；（6）模板填充和结果输出。

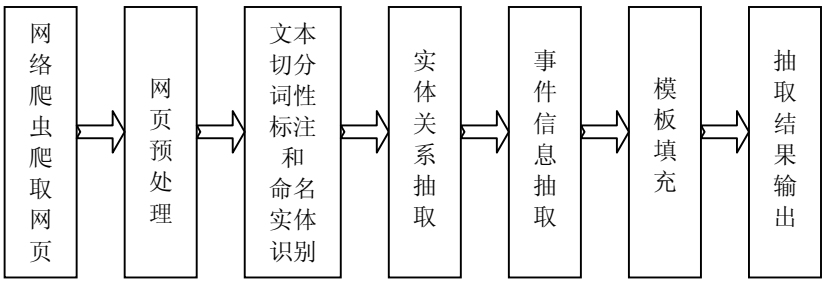

图7.3 Web信息抽取的一般过程

本章设计的WebIE系统的体系结构如图7.4所示。该系统主要由以下四大模块组成：图形用户接口（GUI）模块、数据获取模块、文本粗加工模块、信息抽取模块。

7.3.2.1 GUI 模块

GUI模块是为了方便用户操作而提供的一个图形界面，用户通过该界面可以实现对系统的设置、维护、查询等各种操作。这些操作主要包括网络爬虫参数的设置、知识库的维护、数据表的维护和手工修订、模板创建与维护、用户查询及结果显示等。

网络爬虫参数设置主要用于设置网络爬虫的种子站点、爬取网页时的链接深度、

网页爬取的主题词或领域关键词等。知识库包含了文本信息抽取过程需要的一些资源，包括：通用词典、抽取模式库、触发词表、受限领域本体库、和通用知识库等。可以通过 GUI 模块对这些知识库进行维护。在 WebIE 的一系列操作中会产生许多数据表，这些数据表存放在数据库中，通过 GUI 可以浏览、甚至修订这些数据表。

通过 WebIE 原型系统的一系列处理后，抽取出的特定类型的事实事件信息要保存到数据库中，用于保存这些事实事件信息的结构就是模板。用于保存事件信息的称为事件模板，而事实信息一般对应于事件模板中的一个或几个槽。在 GUI 模块中可以根据要抽取的特定事件类型创建相应的事件模板，在事件模板创建之后还可以进行修改维护等操作。WebIE 抽取出的特定类型事件信息就填充到事件模板的相应槽中，最后填充好的模板将作为一个整体存入数据库，供用户查询或进一步分析利用。用户可以通过 GUI 界面查询并获取系统的抽取结果。

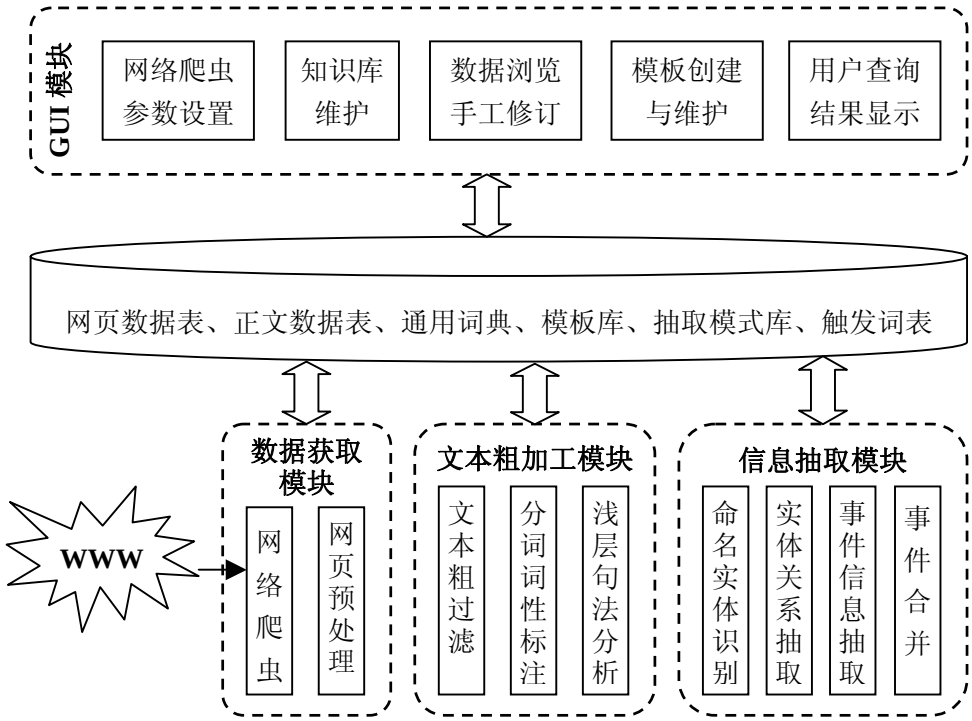

图 7.4 WebIE 原型系统的体系结构

7.3.2.2 数据获取模块

数据获取模块用于从互联网上爬取可能包含有特定类型事件信息的网页，并将网页进行预处理抽取网页正文，为下面的进一步处理准备数据源。该模块包括网络爬虫和网页预处理两个子模块。

网络爬虫子模块根据设置好的参数从相关的 Web 站点抓取网页。其中的参数之一是种子站点，该参数用于决定从那些站点抓取网页。例如，为获取“职务变动”方面的新闻报道网页，设置的种子站点有：http://cn.news.yahoo.com/cn/cn_renshibiandong/（中国人事变动要闻_雅虎专栏）；http://tech.sina.com.cn/focus/exec_change/index.shtml（IT 业界重要人事变动网）；http://unn.people.com.cn/GB/22220/58514/（人民网地方人事变动专题）；http://china.dayoo.com/gb/node/node_114.shtml（大洋网_中国新闻_人事变动）。另外该模块还可以根据提供的一些主题词或关键词对抓取的网页进行粗略的过滤，以保证获取的网页大部分是和主题、场景相关的。

网页预处理子模块的主要功能是去除网页中对文本信息抽取没有用的信息，将网页的正文抽取出来。由于互联网上出现的大部分网页已经不再是单一的文本信息，而是包含有大量的图片、flash 小动画、超链接等多种媒体的信息，并且页面结构也多种多样，所以首先需要对抓取的网页进行预处理。过滤掉其中的非文本数据，仅将大块的文本数据保留下来，并保存为自由文本的格式，供下一步文本信息抽取使用。

7.3.2.3 文本粗加工模块

文本粗加工模块包括三个子模块：文本粗过滤、分词及词性标注、浅层句法分析。分别叙述如下：文本粗过滤子模块主要实现对一篇或一段中文文本进行停用词过滤、领域不相关的句子过滤等操作。分词及词性标注子模块实现对一篇文档、一段文本或一个句子的词语切分和词性标注。浅层句法分析子模块能够对一段中文文本或一个中文语句进行浅层句法分析，识别其中的名词短语、动词短语及相应的句法成分。

7.3.2.4 信息抽取模块

信息抽取模块是整个系统中的核心模块，主要包括命名实体识别、实体关系抽取、事件信息抽取、事件合并四个子模块。命名实体识别子模块主要实现中文文本中的命名实体的识别，识别出文本中的人名、地名、机构名、时间表达式、数值表达式等命名实体。实体关系抽取子模块主要是对文本中的命名实体间的一些特定类型关系进行识别，该模块的识别方式是基于模式匹配技术的。事件信息抽取子模块就是采用前面章节中提出的方法从中文文本中抽取特定类型的事件信息，然后经过事件合并子模块将相同的或有嵌套关系的事件合并之后，再将这些信息填充到事件模板中。

7.4 事件信息抽取模块详细设计

在本章设计和实现的面向 Web 的文本信息抽取原型系统中，本文作者主要负责和事件信息抽取相关的一些模块的设计和实现工作，主要包括事件模板的创建与维护、事件信息抽取、事件合并等模块的设计和实现。下面以采用基于模式匹配的事件信息抽取方法对这些模块进行详细论述。

7.4.1 事件信息抽取模块层次结构

本章将文本事件信息抽取任务详细划分为事件模板的创建与维护、事件模式的获取与维护、文本事件信息抽取三个子任务，这三个子任务又可以进一步细分。划分后的事件信息抽取模块层次结构如图 7.5 所示。

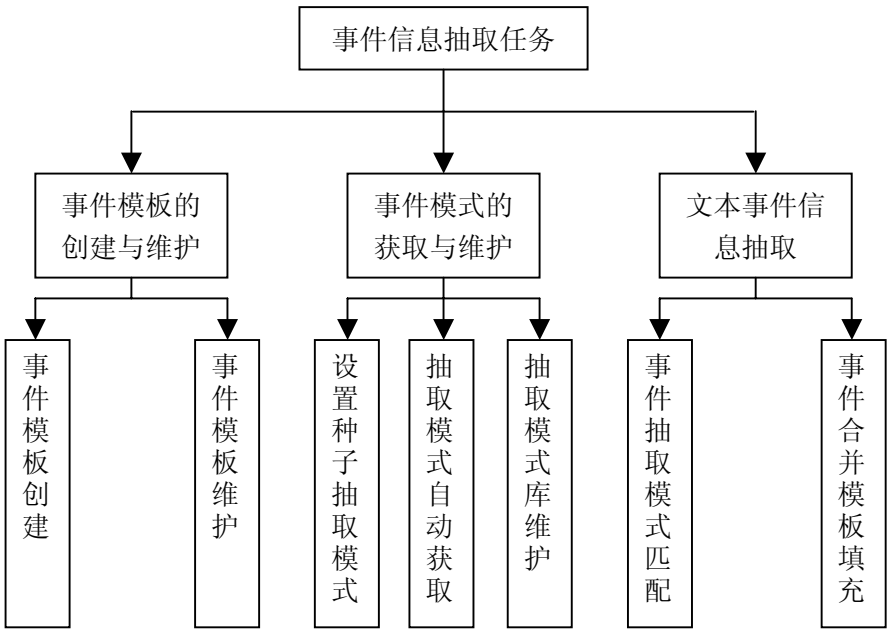

图 7.5 事件信息抽取模块的层次结构

事件模板的创建与维护主要用于根据一定的条件对事件填充模板进行创建，并能够编辑、修订、浏览这些模板，可以通过 GUI 模块提供的界面来调用该模块，完成相应的操作。该部分又可以进一步细分为事件模板创建和事件模板维护两个子任务。事件模式的获取与维护是事件信息抽取中的最重要的模块，该模块采用第 3 章提出的方法从未标注的中文文本中自动获取事件模式。又可以进一步细分为种子模式设置、抽取模式自动获取、抽取模式库维护三个子任务。文本事件信息抽取模块是基于模式匹

配的技术从新的文本中抽取特定类型事件，并经过事件合并处理，然后填充到事件模板中的过程。又可以细分为事件抽取模式匹配和事件合并及模板填充两个子任务。

7.4.2 模块详细设计

本小节从三个方面来介绍事件信息抽取模块的详细设计，这三个方面是：该模块的三个子模块（事件模板的创建与维护、事件模式的获取与维护、文本事件信息抽取）的关系、事件抽取模式的自动获取、基于模式匹配的事件信息抽取。

7.4.2.1 三个子模块的关系

事件模板的创建与维护、事件模式的获取与维护、文本事件信息抽取这三个子模块是事件信息抽取中三个部分，分别处理不同层面的事务。事件模板给出了特定类型事件信息的存放结构，将事件信息以结构化的形式存储并呈现给普通用户；事件模式确定了在自然语言形式的文本中特定类型的事件所表现出的句法结构上的规则；事件信息抽取利用学习出的事件模式在文本中进行匹配，将获取的事件信息填充到事件模板中。这三个子模块的关系如图 7.6 所示。

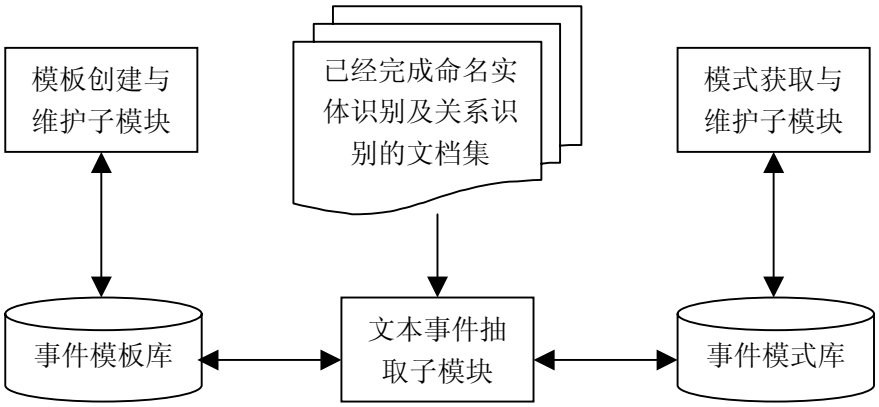

图 7.6 三个子模块的关系

7.4.2.2 事件抽取模式的自动获取

事件抽取模式的自动获取是事件信息抽取的关键所在。WebIE 原型系统中采用第 3 章提出的基于自扩展策略的事件模式获取方法，从几个典型的种子抽取模式开始，通过一个增量迭代的过程发现新的抽取模式，每一轮迭代用类似于 TF/IDF 的评估策略对产生的候选抽取模式进行打分，选择最合适的候选模式并入到种子模式集。然后

再开始下一轮迭代。处理过程如图 7.7 所示，两个半圆虚线表示的是迭代的过程。

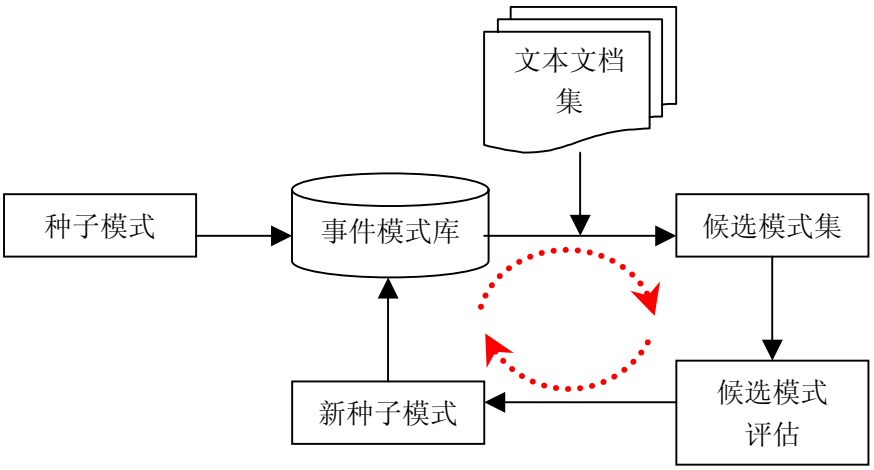

图 7.7 基于自扩展策略的模式自动获取过程

7.4.2.3 基于模式匹配的事件信息抽取

基于抽取模式的中文文本事件信息抽取的处理过程如图 7.8 所示。要处理的文档集已经经过命名实体识别和实体关系抽取，然后调用文本粗加工模块中的一些功能子模块对这些文本进行处理，主要要将句子切分。然后采用第 3 章 3.5.1 节的事件模式匹配算法来产生候选事件，再对候选事件进行评估，选择可信度大的候选事件作为最终的事件，进行事件合并后填充到事件模板中。

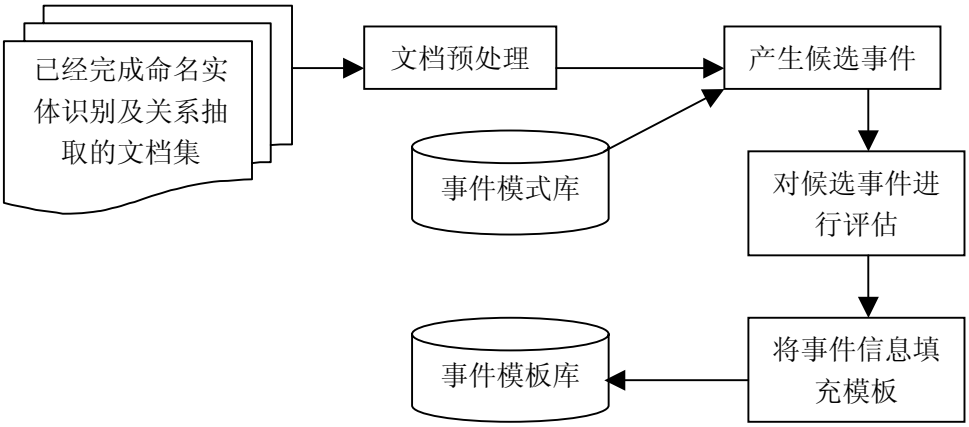

图 7.8 基于模式匹配的事件信息抽取过程

7.5 WebIE 原型系统的性能评测

下面首先给出系统评测的目的和评测指标，然后对面向 Web 的文本信息抽取原型

系统：WebIE 的性能进行简要概述。

7.5.1 系统评测目的和评测指标

建立该原型系统有两方面目的，一方面是要对文本信息抽取的技术和方法进行检验，特别是对所建立的语言模型和提出的机器学习方法提供一个测试的平台；另一方面建立一套完整的信息抽取系统也有重大的现实意义，可以用于中文文本信息处理的方方面面，并且从系统的层次上验证所提出的设计思想正确与否。从这两个方面来说，建立原型系统，可以达到如下目的：

（1）评定系统的设计思想和技术实现；

（2）评定文本信息抽取中各种语言模型与算法的质量和效率，包括这些方法的可行性与性能；

（3）指导文本信息抽取的进一步研究，完善提出的语言模型和机器学习方法，为设计和实现更完善的信息抽取系统打下坚实的基础。

在对整个系统和部分模块进行评测时采用了第 1 章 1.3.3 节给出的评测指标。

7.5.2 原型系统的性能

在 1995 年全年人民日报语料和 1998 年 1 月份标注语料，以及从互联网上获取的大量语料上，本文对 WebIE 原型系统和前面章节提出的语言模型和机器学习方法进行了性能评测，评测结果显示其设计目标基本实现，前面多数章节的局部评测大都建立在该平台之上，各部分详细的评测数据可以参看前面章节。

7.6 小结

目前的互联网是一个巨大的、分布广泛的、全球性的信息资源中心，它包含了许多有价值的但又不容易获取的信息，如何利用计算机自动地抽取这些有用信息具有重要研究价值。本章以实验室的在研项目“特定领域网络信息提取服务系统”为背景，设计并实现了一套面向 Web 的文本信息抽取原型系统：WebIE。对该系统的设计目标、体系结构、各模块的主要功能进行了概述，并详细介绍了事件信息抽取模块的功能及其详细设计。最后对该原型系统进行了评测，基本上实现了预期的目标。

第8章 结束语

8.1 研究工作总结

文本信息抽取是当前自然语言处理领域研究的热点问题，有着非常重要的现实意义和广阔的应用前景。本文研究工作以实验室的在研项目“特定领域网络信息提取服务系统”为背景，是文本信息抽取关键技术研究的一部分，也是其中最具挑战性的部分。由于自然语言的复杂性，当前文本信息抽取系统的实现在技术上还有很多难点问题需要解决，本文对半结构化文本信息抽取和自由文本事件信息抽取中的语言模型和机器学习方法进行了深入研究。在半结构化文本信息抽取方面，提出了一种基于条件随机场的信息抽取方法；在抽取模式自动获取方面，提出了一种基于自扩展策略的事件模式自动获取方法；在事件探测和分类方面，提出了一种基于最大熵模型的中文文本事件分类方法；在事件要素及其角色识别方面，提出了一种基于触发词的论元结构，利用条件随机场模型来识别事件要素及其语义角色的方法；对于中文文本中一些简单事件的信息抽取，提出了一种基于隐马尔可夫模型的事件要素抽取方法。具体说来，本文主要的创新工作如下：

（1）针对隐马尔可夫模型不能充分利用对抽取有用的上下文特征，提出了一种基于条件随机场的半结构化文本信息抽取方法。该方法能充分利用半结构化文本中的版面结构等特征，并针对中文文本的特点，在进行信息抽取时先利用分隔符、特定标识符进行信息块划分，在信息块的基础上利用条件随机场模型进行信息抽取。对中文科研论文的头部信息和引文信息抽取的实验结果表明，该方法抽取性能要明显优于基于隐马尔可夫模型的方法。

（2）为了提高信息抽取系统的可移植性，提出了一种从未标注的中文文本基于自扩展策略自动获取事件抽取模式的算法，该算法从少数几个种子抽取模式出发，通过一个增量迭代的过程发现新模式，每一轮迭代从三个层次对抽取模式进行扩展，然后采用类似于 TF/IDF 的评估方法对产生的候选模式进行评估，选择得分最高的几个模式并入到当前模式集。实验表明，该方法能较好地从未标注的中文文本中自动学习抽取模式，将由该方法获取的相应的模式用于“职务变动”类事件信息抽取中，得到了综合指标 F 值为 66.3%的抽取效果。

（3）提出了一种基于最大熵模型的事件分类方法，该方法能够综合事件表述语句中的触发词信息及触发词前后的命名实体、短语等各类特征对事件表述语句进行分类。应用该方法对人民日报语料中的职务变动、会见、恐怖袭击、法庭宣判、自然灾害五类事件进行的分类实验表明，该方法的分类效果明显优于传统的分类方法。

（4）提出一种基于触发词的论元结构的事件要素及其角色识别方法。该方法利用事件表述语句中触发词的论元和要抽取的事件要素之间的对应关系，以浅层句法分析为基础，把短语或命名实体作为识别的基本单元，选择基于句法成分的、基于谓词的、句法成分-谓词关系、语义四类特征作为模型特征集，将条件随机场模型用于事件要素及其角色的识别。应用该方法对“职务变动”和“会见”两类事件的事件要素和角色进行识别，在各自的测试集上分别获得了 77.3%和 74.2%的综合指标 F 值。

（5）针对中文文本中一些简单事件的表述特点，提出了一种基于隐马尔可夫模型的中文文本事件抽取方法，该方法利用 HMM 模型从事件表述语句中抽取每个候选事件的事件要素，为每一类事件要素构建一个独立的隐马尔可夫模型用于该类事件要素的抽取。构建 HMM 模型时采用随机优化的方法从训练语料中自动学习模型结构。在人民日报和从互联网获取的网页文本两类语料中进行了实验，结果表明，该方法能较好地实现中文文本中简单事件的信息抽取。

8.2 下一步研究设想

本文的研究工作是信息抽取研究中的一部分，主要研究和探讨了文本信息抽取中的语言模型和机器学习方法。当一个研究课题暂时告一段落，人们要思量下一步该如何去做的时候，无非要从两个方面做更多的努力，即一方面结合更多的研究和实验，对现有的理论和方法进行检验并向纵深挖掘；一方面在现有的研究成果基础上，探索如何开辟更广阔的研究空间。本课题的研究也不例外，信息抽取系统的研究是一个非常浩大的工程，涉及到很多学科。本文仅在几个小的方面做了一些尝试，真正面向实用的信息抽取系统的研究还有大量的工作要做。就本论文所做的一些工作而言，也还有很多地方需要进一步的深入研究：

（1）对于半结构化文本信息抽取来说，还有两方面问题急需解决，一方面条件随机场模型的参数训练和特征选择归纳还存在速度慢的问题，需要进一步研究能够提升参数估计和特征选择归纳速度的算法；另一方面，需要更全面的分析半结构化文本

的特征，探索性能更好、普适性更高的半结构化文本信息抽取方法。

（2）针对中文文本的句法分析，还没有广泛地应用到文本信息抽取中。进一步研究句法分析，特别是浅层句法分析在文本事件信息抽取中的作用有重大意义，对提高事件抽取模式的可信度、事件表述语句分类的精度、事件要素及其语义角色识别的准确度都将有很大帮助。

（3）进一步将动词论元结构分析、语义角色的相关特征分析同自由文本事件信息抽取结合起来进行研究，对提升事件要素及其角色识别的准确度将会有很大帮助。

（4）文本事件信息抽取是信息抽取中最具挑战性的任务之一。进一步分析本文提出的基于模式匹配的文本事件抽取、基于触发词探测的事件抽取、基于 HMM 的文本事件抽取三类方法的实质，探索将这些方法的优势结合起来，进一步扩展思路，更深入地进行文本事件信息抽取的研究也需要今后逐步展开。

由于作者水平有限，论文中肯定存在不足或不妥之处，真诚希望各位老师和同学批评指正。

参考文献

[1] Appelt D E. Introduction to information extraction [J]. AI Communications, 1999:12(3):161-172.

[2] 孙斌. 信息提取技术概述（上）[J]. 术语标准化与信息技术, 2002,(3):28-32.

[3] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用,2003,39(10):1-5.

[4] Turmo J, Ageno A, Català N. Adaptive informaiton extraction [J]. ACM Computing Surveys, 2006,38(2):1-47.

[5] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 北京:中国科学院计算技术研究所，2004.

[6] Atkinson-Abutridy John, Mellish C, Aitken S. Combining information extraction with genetic algorithms for text mining [J]. IEEE Intelligent Systems, 2004,(3):22-30.

[7] 于海滨，秦兵，刘挺，等. 命名实体识别和指代消解在文摘系统中的应用[J]. 计算机应用研究, 2006,(4):180-182.

[8] Gregg D G, Walczak S. Exploiting the information Web [J]. IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 2007, 37(1):109-125.

[9] 于中华，张容，唐常杰，等. 基于前后文词形特征的生物医学文献句子边界识别[J]. 小型微型计算机系统, 2006,27(1):180-184.

[10] 许嘉璐. 现状和设想——试论中文信息处理与现代汉语研究[J]. 中文信息学报, 2001, 15(2):1-8.

[11] 俞士汶，朱学锋，王惠，等. 现代汉语语法信息词典详解[M]. 北京:清华大学出版社，1998.

[12] 俞士汶. 语言信息处理研究的意义与方法[N]. 中国计算机报，1991, (18):3.

[13] Muslea I. Extraction patterns for information extraction tasks: a survey [A]. In Proceedings of the AAAI’99 Workshop on Machine Learning for Information Extraction [C]. 1999. 1-6.

[14] Kiyoshi Sudo. Unsupervised discovery of extraction patterns for information extraction [D]. Ph.D. dissertation, New York University, 2004.

[15] Agichtein E. Extracting relations from large text collections [D]. Ph.D. dissertation, Columbia University, 2005.

[16] 周俊生，戴新宇，尹存燕，等. 自然语言信息抽取中的机器学习方法研究[J]. 计算机科学，2005,32(3):186-189.

[17] Freitag D. Machine learning for information extraction in informal domains [D]. PhD dissertation, Carnegie Mellon University, 1998.

[18] Freitag D. Machine learning for information extraction in informal domains [J]. Mahine Learning, 2000,39(3):169-202.

[19] Bagga A. Coreference, cross-document coreference, and information extraction methodologies [D]. Ph.D. dissertation, Duke University, 1998.

[20] Gaizauskas R, Wilks Y. Information extraction: Beyond document retrieval [J]. Computational Linguistics and Chinese Language Processing, 1998,3(2):17-60.

[21] Cowie J, Lenhert W. Information extraction [J]. Communications of the ACM, 1996,39(1): 80-91.

[22] Riloff E. Information extraction as a stepping stone toward story understanding [A]. Ram A, Moorman K. In Understanding Language Understanding: Computational Models of Reading [C]. MIT Press, 1999. 435-460.

[23] Kosala R, Blockeel H. Web mining research: A survey [J]. ACM SIGKDD, 2000,2(1):1-15.

[24] Hsu C N, Chang C C. Finite-state transducers for semi-structured data extraction form the Web [J]. Journal of Information Systems, 1998,23(8):521-538.

[25] Sager N. Natural language information processing: A computer grammar of English and its applications [J]. Reading, Massachusetts: Addison Wesley, 1981.

[26] Dejong G. An overview of the FRUMP system [A]. In Proceedings of Strategies for Natural Language Processing [C]. Lawrence Erbaum, 1982. 149-176.

[27] Lytinen S, Gershman A. ATRANS: Automatic processing of money transfer messages [A]. In Proceedings of the 5th National Conference of the American Association for Artificial Intelligence [C]. IEEE Computer Society Press, 1993. 93-99.

[28] Elaine Marsh, Dennis Perzanowski.MUC-7 Evaluation of IE Technology: Overview of Results. 1998.

[29] Nancy Chinchor. MUC-7 Information Extraction Task Definition V 5.1. 1998.

[30] Aone C, Halverson L, Hampton T, et al. SRA: description of the IE^2 system used for MUC-7 [A]. In Proceedings of MUC-7 [C]. 1998.

[31] 美国国家标准技术局（NIST）. The ACE 2007 (ACE07) evaluation plan [EB/OL]. http://www.nist.gov/speech/tests/ace/ace07/doc/ace07-evalplan.v1.3a.pdf.2007-2-22/2007-4-1.

[32] 孙茂松，黄昌宁，高海燕. 中文姓名的自动辨识[J]. 中文信息学报, 1995, 9(2):16-27.

[33] 刘秉伟，黄萱菁，郭以昆. 基于统计方法的中文姓名识别[J]. 中文信息学报, 2000, 14(3):16-24.

[34] 郑家恒，李鑫，谭红叶. 基于语料库的中文姓名识别方法研究[J]. 中文信息学报, 2000, 14(1):7-12.

[35] 蔡晓白，樊孝忠. 疾病命名短语识别的最大熵方法[J]. 北京理工大学学报，2006, 26(6):517-520.

[36] 王宁，葛瑞芳，苑春法，等. 中文金融新闻中公司名的识别[J]. 中文信息学报, 2002, 16(2): 1-6.

[37] Zhang Y M, Zhou J F. A trainable method for extracting Chinese entity names and their relations [A]. In Proceedings of the 2nd Chinese Language Processing Workshop [C]. Hong Kong, 2000.

[38] 黄德根，岳广玲. 基于统计的中文地名识别[J]. 中文信息学报, 1995, 17(2):36-41.

[39] 郑家恒，张辉. 基于 HMM 的中国组织机构名自动识别[J]. 计算机应用，2002, 22(11):1-2.

[40] 庄明，老松杨，吴玲达. 一种统计和词性相结合的命名实体发现方法[J]. 计算机应用, 2004, 24(1):22-24.

[41] 刘非凡，赵军，吕碧波，等. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2006,20(1):7-13.

[42] Deng Bo, Fan Xiaozhong, Yang Ligong. Chinese entity relation extraction based on word semantics [J]. Journal of Computational Information Systems, 2005,(3):913-920.

[43] 邓擘，樊孝忠. 基于向量空间模型的自动关系提取[J]. 计算机科学，2006, 33(9A):12-14.

[44] 车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2):1-6.

[45]姜吉发,王树西. 一种自举的二元关系和二元关系模式获取方法[J]. 中文信息学报, 2005,19(2):71-77.

[46]张素香，李蕾，秦颖，等. 基于 Bootstrapping 的中文实体关系自动生成[J]. 微电子学与计算机, 2006,23(12):15-18.

[47]王厚峰,何婷婷. 汉语中人称代词的消解研究[J]. 计算机学报， 2001,24(2):136-143.

[48]张威，周昌乐. 汉语语篇理解中元指代消解初步[J]. 软件学报，2002,13(4):732-738.

[49]钱伟，郭以昆，周雅倩，等. 基于最大熵模型的英文名词短语指代消解[J]. 计算机研究与发展，2003,40(9):1337-1343.

[50]王晓斌，周昌乐. 基于语篇表述理论的汉语人称代词的消解研究[J]. 厦门大学学报（自然科学版），2004,43(1):31-35.

[51]王厚峰，梅铮. 鲁棒性的汉语人称代词消解[J]. 软件学报，2005,16(5):700-707.

[52]姜吉发. 一种跨语句汉语事件信息抽取方法[J]. 计算机工程,2005,31(2):27-29.

[53]姜吉发. 一种事件信息抽取模式获取方法[J]. 计算机工程,2005,31(15):96-98.

[54]赵妍妍，王啸吟，秦兵，等. 中文事件抽取中事件类别的自动识别[A]. 第三届学生计算语言学研讨会论文集[C]. 2006. 240-245.

[55]廖乐健，曹元大，李新颖. 基于 ontology 的信息抽取[J]. 计算机工程与应用，2002,38(23):110-113.

[56]孙斌. 继承-归纳关系及其在对象系统和信息提取技术中的应用[D]. 北京:北京大学，2000.

[57]李芳，盛焕烨，张冬茉. 多语种投资信息抽取系统的实现[J]. 上海交通大学学报，2004, 38(1):21-25.

[58]李蕾，周延泉，王菁华. 基于全信息的中文信息抽取系统及应用[J]. 北京邮电大学学报，2005,28(6):48-51.

[59]陆俭明. 现代汉语语法研究教程[M]. 北京:北京大学出版社. 2005.

[60]张华平，刘群. 基于 N-最短路径的中文词语粗分模型[J]. 中文信息学报, 2002，16(5):1-7.

[61]李向阳,张亚非. 基于语义标注的信息抽取[J]. 解放军理工大学学报(自然科学版),

2004,5(4):39-43.

[62] Han H, Giles C, Manavoglu E, et al. Automatic document metadata extraction using support vector machines [A]. In Proceedings of Joint Conference on Digital Libraries [C]. Houston: IEEE Press, 2003. 37-48.

[63] Hobbs J R, Appelt D, Bear J, et al. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text [A]. Finite State Devices for Nature Language Processing [C]. MIT Press, Cambridge, 1996.

[64] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [A]. In Proceedings of the IEEE [C]. 1989,77(2).

[65] 张玲,黄铁军,高文. 基于隐马尔可夫模型的引文信息提取[J]. 计算机工程, 2003,29(20):33-34.

[66] 刘云中,林亚平,陈治平. 基于隐马尔可夫模型的文本信息抽取[J]. 系统仿真学报, 2004,16(3):507-510.

[67] Seymore K, McCallum A, Rosenfeld R. Learning hidden Markov model structure for information extraction [A]. In Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction [C]. Orlando: AAAI Press,1999. 37-42.

[68] Yin P, Zhang M, Deng Z H, et al. Metadata Extraction from Bibliographies Using Bigram HMM [A]. In Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL2004) [C]. Shanghai, published by LNCS (Lecture Notes in Computer Science), Springer-Verlag. Dec.2004. 310–319.

[69] 林亚平,刘云中,周顺先,等. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报, 2005,33(2):236-240.

[70] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization [A]. In Proceedings of the 18th Conference on Artificial Intelligence [C]. Edmonton: AAAI Press, 2000. 584-589.

[71] Freitag D, McCallum A. Information extraction with HMM and shrinkage [A]. In Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction [C]. Orlando: AAAI Press, 1999. 31-36.

[72] 于琨. 互联网半结构化信息抽取研究[D]. 合肥:中国科学技术大学, 2005.

[73] Ray S, Craven M. Representing sentence structure in hidden Markov models for

information extraction [A]. IJCAI-2001 [C]. 2001.

[74] Riloff E. Automatically Generating Extraction Patterns from Untagged Text [A]. In Proceedings of Thirteenth National Conference on Artificial Intelligence (AAAI-96) [C]. 1996. 1044-1049.

[75] Kiyoshi Sudo, Satoshi Sekine, Grishman R. Automatic Pattern Acquisition for Japanese Information Extraction [A]. In Proceedings of the Human Language Technology Conference (HLT2001) [C]. 2001.

[76] Yangarber R. Scenario Customization for Information Extraction [D], Ph.D.Thesis, New York University, January 2001.

[77] 马彦波，张蕾. 一种创建事件模式的新方法[J]. 微机发展，2005,15(1):20-23.

[78] Ayuso D, Boisen S, Fox H, et al. BBN PLUM: Description of the PLUM system as used for MUC-4 [A]. In Proceedings of the 4th Message Understanding Conference (MUC-4) [C]. 1992. 169-176.

[79] Hobbs J R, Appelt D, Tyson M, et al. SRI International: Description of the FASTUS system used for MUC-4 [A]. In Proceedings of the 4th Message Understanding Conference (MUC-4) [C]. 1992. 268-275.

[80] Krupka G, Jacobs P, Rau L, et al. GE NLTOOLSET: Description of the system as used for MUC-4 [A]. In Proceedings of the 4th Message Understanding Conference (MUC-4) [C]. 1992. 177-185.

[81] Yangarber R, Grishman R. NYU: Description of the Proteus/PET system as used for MUC-7 ST [A]. In Proceedings of the 7th Message Understanding Conference: MUC-7 [C]. Washington, 1998.

[82] Riloff E. Automatically Constructing a Dictionary for Information Extraction Tasks [A]. In Proceedings of the Eleventh National Conference on Artificial Intelligence [C]. 1993. 811-816.

[83] Kim J, Moldovan D. Acquisition of linguistic patterns for knowledge-based information extraction [J]. IEEE Transactions on Knowledge and Data Engineering, 1995,7(5):713-724.

[84] Soderland S, Fisher D, Aseltine J, et al. CRYSTAL: Inducing a conceptual dictionary [A]. In Proceedings of the Fourteenth International Joint Conf on Artificial Intelligence

[C]. 1995. 1314-1321.

[85] Huffman S. Learning information extraction patterns from examples [A]. IJCAI-95 Workshop on new approaches to learning for natural language processing [C]. 1995. 127-142.

[86] Krupka G. Description of the SRA system as used for MUC-6 [A]. In Proceedings of the 6th Message Understanding Conference (MUC-6) [C]. 1995. 221-235.

[87] Riloff E, Shoen J. Automatically acquiring conceptual answer patterns without an annotated corpus [A]. In Proceedings of the Third Workshop on Very Large Corpora [C]. 1995. 148-161.

[88] Chai J Y. Learning and generalization in the creation of information extraction systems [D]. PhD thesis, Duke University, 1998.

[89] Yangarber R, Grishman R, Tapanainen P, et al. Automatic acquisition of domain knowledge for information extraction [A]. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000) [C]. Saarbrucken, Germany, 2000.

[90] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections [A]. In Proceedings of the ACM International Conference on Digital Libraries [C]. 2000. 85-94.

[91] Agichtein E, Gravano L, Pavel J, et al. Snowball: A prototype system for extracting relations from plain-text collections [A]. In Proceedings of the ACM SIGMOD Conference [C]. 2001.

[92] Agichtein E. Extracting Relations from Large Text Collections [D]. Ph.D.Thesis, Columbia University, 2005.

[93] 郑家恒, 王兴义, 李飞. 信息抽取模式自动生成方法的研究[J]. 中文信息学报, 2004, 18(1):48-54.

[94] 袁毓林. 论元角色的层级关系和语义特征[J]. 世界汉语教学, 2002,(2):10-22.

[95] 袁毓林. 信息抽取的语义知识资源研究[J]. 中文信息学报, 2002,16(5):8-14.

[96] 袁毓林. 用动词的论元结构跟事件模板相匹配[J]. 中文信息学报,2005,19(5):37-43.

[97] Gildea D. Jurafsky D. Automatic labeling of semantic roles [J]. Computational Linguistics, 2002,28(3):245-288.

[98] Thompson CA, Levy R, Manning CD. A generative model for semantic role labeling [A]. In Proceedings of ECML-2003 [C]. LNAI 2837, Springer Berlin Heidelberg, 2003. 397-408.

[99] 刘怀军，车万翔，刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报，2007, 21(1):79-84.

[100] Liu T, Che W X, Li S, et al. Semantic role llabeling system using maximum entropy classifier [A]. In Proceedings of CoNLL-2005 [C]. Ann Arbor, Michigan, 2005. 189-192.

[101] 刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007, 18(3):565-573.

[102] Che W X, Zhang M, Liu T, et al. A hybrid convolution tree kernel for semantic role labeling [A]. In Proceedings of ACL-2006 [C]. 2006.

[103] Chen J, Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments [A]. In Proceedings of EMNLP-2003 [C]. Sapporo, Japan, 2003.

[104] Pradhan S, Hacioglu k, Krugler V, et al, Support vector learning for semantic argument classification [J]. Machine Learning, 2005.

[105] Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Project [A]. In Proceedings of ACL & Coling-1998 [C]. 1998. 86-90.

[106] Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles [J]. Computational Linguistics, 2005,31(1):71-105.

[107] You J M, Chen K J. Automatic semantic role assignment for a tree structure[A]. In Proceedings of 3rd ACL SIGHAN Workshop [C]. 2004.

[108] Lafferty J, Pereira F, McCallum A. Conditional random fields: probabilistic models for segmenting and labeling sequence data [A]. In: Proceedings of 18th International Conference on Machine Learning [C]. 2001. 282-289.

[109] Peng F C, Feng F F, McCallum A. Chinese segmentation and new word detection using conditional random fields [A]. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004) [C]. 2004. 562-568.

[110] Sha F, Pereira F. Shallow parsing with conditional random fields [A]. In Proceedings of Human Language Technology-NAACL [C]. Edmonton, Canada, 2003.

[111] Tan Y M, Yao T S, C Q, et al. Applying conditional random fields to Chinese shallow parsing [A]. In Proceedings of CICLing 2005 [C]. LNCS 3406, 2005. 167-176.

[112] 冯冲，陈肇雄，黄河燕，等. 基于条件随机域的复杂最长名词短语识别[J]. 小型微型计算机系统，2006,27(6):1134-1139.

[113] Lee Y H, Kim M Y, Lee J H. Chunking using conditional random fields in Korean texts [A]. In Proceedings of IJCNLP 2005 [C]. LNAI 3651, 2005. 155-164.

[114] 周俊生，戴新宇，尹存燕，等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006,34(5):804-809.

[115] Feng Y Y, Sun L, Zhang J L. Early results for Chinese named entity recognition using conditional random fields model, HMM and maximum entropy [A]. In Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering [C]. 2005. 549-552.

[116] Settles B. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets [A]. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications [C]. 2004.

[117] McCallum A. Efficiently inducting features of conditional random fields [A]. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence [C]. 2003. 403-410.

[118] Pinto D, McCallum A, Wei X, et al. Table extraction using conditional random fields [A]. In Proceedings of ACM SIGIR'03 [C]. 2003. 235-242.

[119] Zhu J, Nie Z Q, Wen J R, et al. 2D conditional random fields for Web information extraction [A]. In Proceedings of the 22nd International Conference on Machine Learning [C]. Bonn, Germany, 2005.

[120] Byrd R H, Nocedal J, Schnabel R B. Representations of quasi-Newton matrices and their use in limited memory methods [J]. Mathematical Programming, 1994, 63(2):129-156.

[121] Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models [J]. Annals of Mathematical Statistics,1972,43(5):1470-1480.

[122] Della Pietra S, Della Pietra V, Lafferty J. Inducing features of random fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997,19(4):380-393.

[123] Sha F, Pereira F. Shallow parsing with conditional random fields [A]. In Proceedings of Human Language Technology of NAACL [C]. 2003. 213-220.

[124] Berger A L, Della Pietra S A, Della Pietra V J. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996,22(1):39-71.

[125] 李向阳，张亚非. 一种基于自举原理的语义模式自动获取方法[J]. 微电子学与计算机，2005,22(2):188-192.

[126] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods [A]. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics [C]. Cambridge, MA, 1995. 189-196.

[127] Strzalkowski T, Wang J. A self-learning universal concept spotter [A]. In Proceedings of the 16th International Conference on Computational Linguistics [C]. Copenhagen, 1996. 931-936.

[128] Riloff E, Jones R. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping [A]. In Proceedings of Sixteenth National Conference on Artificial Intelligence (AAAI-99) [C]. Orlando, Florida. 1999. 474-479.

[129] Riloff E, Shepherd J. A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction [J]. Natural Language Engineering, 2002,5(2):147-156.

[130] Thelen M, Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts [A]. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002) [C]. 2002. 214-221.

[131] 梅家驹，竺一鸣，高蕴琦，等. 同义词词林[M]. 上海：上海辞书出版社，1996.

[132] Knuth D E, Morris J H, Pratt V R. Fast pattern matching in strings [J]. SIAM J. Comput. 1977:6(1):323-350.

[133] Boyer R S, Moore J S. A fast string searching algorithm [J]. Comm. ACM, 1977,20(10):762-772.

[134] 夏天. 中文信息处理中的相似度计算研究与应用[D]. 北京：北京理工大学，2005.

[135] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation [J]. Machine Learning,1999,34(1-3):177-210.

[136] 赵岩，王晓龙，刘秉权，等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展, 2006,43(2):268-274.

[137]李素建，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003, 26(12):1722-1727.

[138]张锋，樊孝忠. 基于最大熵模型的交集型切分歧义消解[J]. 北京理工大学学报, 2005, 25(7):590-593.

[139]Nigam K, Lafferty J, McCallum A. Using Maximum Entropy for Text Classification [A]. In Proceedings of the IJCAI99 Workshop on Information Filtering [C]. Stockholm, Sweden, 1999.

[140]李荣陆，王建会，陈晓云，等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展，2005,42(1):94-101.

[141]鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报，2001,24(7):742-747.

[142]Ratnaparkhi A, Roukos S, Ward R T. A maximum entropy model for parsing[A]. In Proceedings of the International Conference on Spoken Language Processing [C]. Yokohama, Japan, 1994. 803-806.

[143]Reynar J C, Ratnaparkhi A. A maximum entropy approach to identifying sentence boundaries[A]. In Proceedings of the 5th Conference on Applied Natural Language Processing [C]. Washington D.C, 1997. 16-19.

[144]Ratnaparkhi A. Maximum entropy models for natural language ambiguity resolution [D]. Pomsy Lvania: University of Pennsylvania, 1998.

[145]Yang Y. An evaluation of statistical approaches to text categorization [J]. Information Retrieval, 1999, 1(1):76-88.

[146]徐烈炯，沈阳. 题元理论与汉语配价问题[J]. 当代语言学，1998,(3):1-21.

[147]徐烈炯. 语义学[M]. 北京:语文出版社，1990.

[148]顾阳. 论元结构理论介绍[J]. 国外语言学，1994,(1).

[149]沈阳，郑定欧. 现代汉语配价语法研究[M]. 北京:北京大学出版社，1995.

[150]袁毓林. 一套汉语动词论元角色的语法指标[J]. 世界汉语教学，2003,(3):24-36.

[151]詹卫东. 基于配价的汉语语义词典[J]. 语言文字应用，2000,(1):37-43.

[152]袁毓林. 用逻辑和篇章知识来约束模板匹配[J]. 中文信息学报,2005,19(4):39-45.

[153]Moreda P, Palomar M. The role of verb sense disambiguation in semantic role labeling [A]. In Proceedings of FinTAL-2006 [C]. LNAI 4139, Springer Berlin

Heidelberg, 2006. 684-695.

[154]谢锦辉. 隐 Markov 模型（HMM）及其在语音处理中的应用[M]. 武汉:华中理工大学出版社,1995:1-46.

[155]张孝飞，陈肇雄，黄河燕,等. 词性标注中生词处理算法研究[J]. 中文信息学报，2003,17(5):1-5.

[156]胡春静，韩兆强. 基于隐马尔可夫模型（HMM）的词性标注的应用研究[J]. 计算机工程与应用，2002,38(6):61-64.

[157]刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004,41(8):1421-1429.

[158]吴晓明，宋长新，王波,等. 隐马尔可夫模型用于蛋白质序列分析[J]. 生物医学工程学，2002,19(3):455-458.

[159]Gale W A. Good-turing smoothing without tears [J]. Journal of Quantitative Linguistics, 1995, 2(3):217-237.

[160]中国互联网络信息中心. 中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/uploadfiles/doc/2006/7/19/103601.doc. 2006-7-19/2007-4-1.

[161]Cunningham H, Maynard D, Bontcheva K, et al. GATE: A framework and graphical development environment for robust NLP tools and applications [A]. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics [C]. 2002.

[162]朱靖波，姚天顺. 中文信息自动抽取[J]. 东北大学学报(自然科学版)，1998,19(1):52-54.

[163]Hobbs J R. The generic information extraction system [A]. In Proceedings of the 5th Message Understanding Conference (MUC-5) [C]. Morgan Kaufman, 1993. 87-91.

博士学位论文支撑课题

1．教育部博士点基金——受限领域自动问答系统研究（项目编号：20050007023）
2．北京理工大学自然语言处理实验室与扬州万方电子技术有限责任公司合作课题“特定领域网络信息提取服务系统”
3．北京理工大学自然语言处理实验室与金图信息科技有限公司合作课题“中华助考网”

攻读博士学位期间发表的论文

1．Jiangde Yu, Xiaozhong Fan. Automatic Acquisition of Extraction Patterns for Chinese Information Extraction. Journal of Computational Information Systems, 2007, 3(4):1607-1614.（EI Compendex 收录，Accession number: 072610674448）

2．于江德，樊孝忠，尹继豪. 基于条件随机场的中文科研论文信息抽取. 华南理工大学学报（自然科学版），2007, 35(9). （EI Compendex 刊源）

3．Jiangde Yu, Xiaozhong Fan. Metadata Extraction from Chinese Research Papers Based on Conditional Random Fields. In Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, 2007. （EI Compendex, ISTP）

4．Yu Jiangde, Fan Xiaozhong, Pang Wenbo, Yu Zhengtao. Semantic Role Labeling Based on Conditional Random Fields. 东南大学学报(英文版), 2007,23(3).（EI Compendex 刊源）

5．于江德，樊孝忠，庞文博. 事件信息抽取中语义角色标注研究. 计算机科学. （已录用）

6．于江德，樊孝忠，顾益军，尹继豪. 信息抽取中领域本体的设计和实现. 电子科技大学学报（自然科学版）. （已录用），（EI Page One 刊源）

7．于江德，樊孝忠，汪涛，顾益军. 本体论在 Web 信息检索中的应用. 微电子学与计算机，2006, 23(4):160-163.

8．于江德，樊孝忠，尹继豪. 基于 Ultra Search 的桌面搜索设计与实现. 广西师范大学学报，2007, 25(2):218-221.

9．于江德，樊孝忠，尹继豪，顾益军. 基于隐马尔可夫模型的中文科研论文信息抽取. 计算机工程，2007,33(19).

10．于江德，樊孝忠，尹继豪. 隐马尔可夫模型在自然语言处理中的应用. 计算机工程与设计. （已录用）

11．汪涛，樊孝忠，于江德，邓擘. 基于链接分析的搜索结果聚类. 计算机科学，2005, 32(9A):328-330.

12．尹继豪，樊孝忠，于江德. 基于类语言模型的中文机构名称自动识别. 计算机科学, 2006, 33(11): 212-214.

13. 顾益军，樊孝忠，黄维金，于江德. 一种文本讨论线索的自动获取方法. 华南理工大学学报（自然科学版）， 2004, 32(S1):96-98. （EI Compendex 已收录，Accession number: 05048803589）

14. 尹继豪，樊孝忠，赵攀超，于江德. 基于组块分析技术的中文机构名称识别. 哈尔滨工程大学学报, 2006, 27(7S): 466-470. （EI Compendex 已收录，Accession number: 064310200814）

15. Jihao Yin, Xiaozhong Fan, Kaixuan Zhang, Jiangde Yu. Chinese organization name recognition using chunk analysis. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. Wuhan, China, Nov. 2006, 347-353.

附录 1：中文文本中职务变动类事件触发词表

场景	子类	序号	触发词	例　　句
职务变动	任命类触发词	1	补选	鉴于安启元同志已调动工作，不再担任中央纪委常委职务，全会**补选**祁培文同志为中央纪委常务委员会委员。
		2	充任	那就是现任全国政协副主席、中华全国科学技术协会主席、著名物理学家朱光亚教授，在板门店的『停战谈判』中，隐藏其教授身份，隐姓埋名地**充任**谈判代表的翻译，前后达一年半之久！
		3	出任	● 周小川出任中国人民银行行长。 ● 44 岁的栗晓峰自 1993 年 4 月 10 日**出任**中国女排主教练，到目前已一年零八个月。
		4	担当	郎平接手中国女排后，选择的突破口是二传，并颇具胆识地启用了头脑好用的云南队何琦**担当**主力二传，她平时练得最多的人也是何琦。
		5	担起	1988 年 8 月，原县委书记调到市委工作，王耀平**担起**了县委书记的重任。
		6	担任	西曼先生于 1997 年加入 SAP 中国公司，**担任**大中国区总裁。
		7	当	刘沈明原在福建省海洋渔业公司**当**车间主任，下岗 5 年多，他到处打工。
		8	当选	● 法律系讲师林瑞莲昨天**当选**工人党主席。 ● 12 月 26 日，在中共安阳市第九届委员会第一次全会上，靳绥东**当选**安阳市委书记。 ● 刘鹏**当选**为全国青联主席。
		9	到任	新**到任**的公司董事长兼总经理赵乃志对企业的前途充满信心，他说：“改制后，企业从事资产运营更方便了。
		10	到职	1955 年组织决定他由山西省军区司令员调任北京军区后勤部政委，办完交接，刚**到职**接任的新司令员因患病不能坚持工作，当时又没有合适的人选，组织又决定紫峰同志继续主持山西省军区的工作，他不计个人得失，愉快地服从组织安排，继续把山西省军区的工作搞得很好，受到北京军区领导的赞扬。
		11	供职	德尔诺夫舍克毕业后曾在建筑联合企业、银行界任职，担任过地方银行行长，其后又在前南斯拉夫驻外使馆**供职**。
		12	获任	原中联办主任姜恩柱**获任**人大外事委副主任委员。
		13	继任	伊斯兰圣战组织的文传通告还宣布，42 岁的拉马丹·阿卜杜拉已**继任**为该组织新的领导人。
		14	兼任	用友总裁何经华今天离职，王文京将继续**兼任**。
		15	接任	11 月 2 日上午 10 时，新浪科技从有关人士处获得证实，何经华即将离任，总裁一职仍将由 王文京**接任**。

场景	子类	序号	触发词	例　　句
		16	接替	金蝶公司目前也已经确定了胡力的**接替**者，该人士为金蝶公司副总裁、华东总经理章勇。
		17	就职	韩国新任总理李寿成 12 月 18 日正式**就职**，他表示将推进改革。
		18	就任	1993 年 4 月 9 日，孔繁森同志到阿里**就任**地委书记的第三天，我被组织上安排到他身边当通信员。
		19	加入	西曼先生于 1997 年加入 SAP 中国公司，**担任**大中国区总裁。
		20	晋升	上届政府主管工业、科技和对外经济关系的副总理梅特**晋升**为第一副总理。
		21	历任	林坤华，43 岁，当过兵，上过大学，**历任**乡镇党委书记、县委组织部长，1993 年任柘荣县委书记。
		22	连任	治理国家政绩显著，是梅内姆总统这次能够连选**连任**的主要原因。
		23	聘	沈阳市体委主任张家祥介绍说，由于沈阳队主教练张增群未能完成升级指标，将不再**聘**其为主教练。
		24	聘请	江苏省毛泽东诗词研究会成立大会近日在镇江市召开。研究会**聘请**杜平同志为总顾问，匡亚明同志任名誉会长。
		25	聘任	1994 年 3 月，董事会又**聘任**傅守法担任总经理，任期四年。
		26	聘为	谭盾在西方音乐界享有一定的声誉，今年还被 B B C 苏格兰交响乐团**聘为**留团作曲家兼副指挥。
		27	任	● 原北大副校长陈章良**任**中国农大校长。 ● 邢云，1952 年生，大学温和。历任内蒙古伊克昭盟副盟长、盟委副书记、盟长，1996 年 10 月起**任**盟委书记、盟人大工委主任。
		28	任命	球王贝利被巴西新总统**任命**为体育部长。
		29	任职	陈小红已经到卓越网曾经的股东之一老虎基金下属的老虎科技(Tiger Technology)**任职**，主要负责在中国开拓新的投资业务。
		30	上任	今年 1 月**上任**的美国财政部长鲁宾多次表示，美国政府致力于维持美元坚挺。
		31	升为	邱娥国的职务虽已**升为**分管户籍、外勤的副所长。
		32	受聘	德国慕尼黑市长**受聘**南开大学客座教授
		33	胜选	选前民调显示巴勒斯坦解放组织主席阿巴斯笃定**胜选**。
		34	提拔	1984 年，他被**提拔**为商丘地区人民银行副行长，不久，他又挑起了商丘地区工商银行行长的重担。
		35	提名	以色列工党政治局昨晚一致**提名**佩雷斯为工党主席和总理候选人。
		36	提升	梅杰首相**提升**乔治·扬为交通大臣，黑格为威尔士事务大臣，霍格为农业大臣，福赛恩为苏格兰事务大臣。
		37	推荐	转业退伍军人在河北涿州市华北新型建材机械厂深受欢迎。复员军人李文山被职工**推荐**担任厂长。

场景	子类	序号	触发词	例　　句
		38	推举	2 月 7 日晚，波兰执政联盟两党领导人一致**推举**现议长约瑟夫·奥列克西为政府总理候选人。
		39	推任	由于多年以来苏蒂夫在法国高教科研部负责仪器仪表方面的工作，所以被**推任**为法中仪器仪表业合作委员会的法方主席。
		40	委派	1990 年 1 月，受组织**委派**，34 岁的田德营出任该厂党总支书记兼厂长。
		41	委任	抗日同盟军组建，冯将军**委任**席液池为第 3 2 军（骑兵军）军长，
		42	新任	巴西**新任**总统费尔南多·恩里克·卡多佐今天在巴西国会宣誓就职。
		43	选举	1994 年 6 月，中国工商银行行长张肖女士被**选举**为储协副主席。
		44	选为	施罗德被德国联邦议院**选为**新任德国总理任期 4 年
		45	走马上任	1989 年底，时任北京市丰台区委教育部长的他，**走马上任**中国抗战馆馆长。
		46	主持工作	两年前已在这个经济增长强劲的大镇**主持工作**的副镇长则是原华中理工大学计算机自动控制专业副教授、51 岁的周冠雄。
	离职类触发词	47	罢免	7 月 4 日，北京市宣武区人民代表大会常务委员会举行第 13 次会议，依法**罢免**了王宝森的北京市第十届人民代表大会代表资格。
		48	撤除	新华社在会后不久就宣布，中共中央已经**撤除**张文康在卫生部的党职。
		49	撤消	在大跃进中，景晓村同志反对浮夸，敢讲真话，结果被错误地戴上右倾机会主义的帽子，被**撤消**德阳工业区党委副书记、二重厂党委书记职务。
		50	撤销	中共天津市委近日作出决定，并经中央纪委审定报党中央批准，**撤销**王志平的天津市委委员、市委组织部副部长、市人事局党组书记职务。
		51	撤职	借丧事敛财，用公款吃喝，天津市人事局长王志平被**撤职**。
		52	辞掉	林格尔还表明无意再返新加坡并**辞掉**了在国立大学的教职。
		53	辞去	12 月 28 日方正科技(600601)发布公告称，方正集团董事长魏新**辞去**方正科技董事长职务，由西南合成董事长方中华接任；
		54	辞职	● 惠普 CEO 卡莉昨天**辞职**。 ● 在 10 月 12 日上午 CA 公司举行的新存储软件 BrightStor11.1 的发布会上，CA 中国公司市场总监尹婉智证实，原 CA 中国公司总经理吴沛殷女士已于 6 月**辞职**。
		55	革职	9 月 9 日，内务部长差瓦立下令将差罗和梭蓬两警中将**革职**查办。
		56	解除	9 月 7 日，铁道通信信息有限责任公司发生重大人事变动。原总经理彭朋经董事会召开临时会议，被**解除**职务，新任铁通公司总经理由乔金洲担任。
		57	解雇	白宫首席大厨称因无法满足第一夫人要求被**解雇**。

场景	子类	序号	触发词	例　　句
		58	解聘	为了维护富乐公司每一个投资者的合法权益，在 1993 年 4 月 3 日的富乐公司董事会上，多数董事认为，杨惠安的行为违反了公司章程，也违背了国家有关法规，不再适宜担任总经理职务，决定**解聘**其在富乐公司的职务，由燃气集团总公司调回重新安排工作。
		59	解职	原俄中央银行行长格拉先科 1994 年 10 月因卢布暴跌事件被**解职**，由帕拉莫诺娃任代行长。
		60	开除	1995 年 3 月，经市委常委会讨论并报省委批准，决定**开除**石全志党籍。
		61	离任	11 月 2 日上午 10 时，新浪科技从有关人士处获得证实，何经华即将**离任**，总裁一职仍将由 王文京接任。
		62	离职	用友总裁何经华今天**离职**，王文京将继续兼任。
		63	免除	上个月，萨达姆总统又**免除**了其堂弟阿里·哈桑·马吉德的国防部长的职务。
		64	免去	田凤山因违纪被**免去**国土资源部部长职务。
		65	免职	人大常委会通过一批**免职**与任命名单。
		66	去职	现年 65 岁的英国外交大臣赫德在英国外交部对新闻界发表了一个简短的**去职**声明。
		67	任满	今天上午 11 时，法国举行总统职权交接仪式，**任满**总统密特朗将国家最高权力交给新当选总统希拉克，希拉克正式入主爱丽舍宫，开始其 7 年的总统任期。
		68	退下	熊复从《红旗》杂志总编辑岗位上**退下**来之后，有了空闲，大发填词之兴，出了三本词集。
		69	退了	一九七六年，欧珠从乡长的岗位上**退了**下来。
		70	退居二线	北京时间 6 月 21 日消息，据《华尔街日报》报道，继上周盖茨宣布**退居二线**不久，本周微软即曝出高层离职新闻。
		71	下台	1989 年 2 月，统治巴拉圭达 35 年之久的独裁者斯特罗斯纳在军事政变中被赶**下台**，发动军事政变的罗德里格斯将军自任临时总统。
	其他触发词	72	任免	国务院最近**任免**一批国家工作人员。
		73	调任	1994 年 3 月，朱志东**调任**铜山县供销社主任后，他一心一意为农民服务的意识更加强烈了。
		74	调动	鉴于安启元同志已**调动**工作，不再担任中央纪委常委职务，全会补选祁培文同志为中央纪委常务委员会委员。
		75	调入	何长工 1952 年 8 月**调入**地质部，此前曾任重工业部副部长、代部长。

附录 2：ACE2007 评测中给出的事件类型及子类型

事件类型	事件子类型	例　　句
生命	出生	一名出生仅 7 天的女婴被重复接种了卡介苗，家人又急又怕。
	结婚	杨振宁与翁帆在汕头结婚。
	离婚	离婚女向前夫计宠物探视权。
	伤害	美国一核潜艇触礁，20 人受伤。
	死亡	阿拉法特侄子首次披露阿翁可能死于谋杀。
运动	运输	深圳地铁首日开通运送乘客 11 万人，创下记录。
交易	变换所有权	秘报严厉批评美国向台湾出售先进武器。
	交换钱	安锋集团当年向“中华开发”申请资金借贷。
商业	创办	云南大学法学院大三的男生陈俊耕在 2004 年这个暑假成立了自己的公司。
	合并	中国水利电力对外公司并入中国水利投资公司。
	破产	北京“王麻子”剪刀宣布破产清算工作展开。
	关闭	4 年打拼现金耗尽, 一家著名游戏公司关门。
冲突	攻击	1999 年 5 月 7 日北约以 5 枚导弹袭击中国驻南联盟使馆。
	示威	大约有一万人参加了示威活动。
接触	会面	伊拉克前总统萨达姆 16 日首次获准与自己的辩护律师会面。
	致电与信件	外交部长李肇星 20 日打电话给中国驻伊拉克大使杨洪林。
人事	任命	德国慕尼黑市长受聘南开大学客座教授。
	离任	德国胡玛纳公司解雇四名豆奶粉事件责任人。
	提名	美国总统布什 16 日正式提名国家安全事务助理赖斯为新一届政府的国务卿。
	选举	杀人嫌疑犯被选为巴西一城市市长。
司法	逮捕与监禁	山东省人大常委会表决许可逮捕“下跪”副市长。
	释放与假释	美国释放东突分子。
	审判与听证	五角大楼计划将至少 20 名涉嫌虐囚的美兵送上军事法庭接受审判。
	指控	网易前高管被指控违规交易。
	起诉	邹雪正式起诉赵薇要求经济赔偿并公开赔礼道歉。
	定罪	李少民被判间谍罪驱逐出境。
	宣判	谢霆锋被判 240 小时社会服务。
	罚款	切拉因吐出“天价口水” 罚金超过 7000 美元。
	处决	罪有应得刘涌昨日在锦州市被执行死刑。
	引渡	联合公约开始生效加拿大有义务引渡赖昌星回国
	宣告无罪	企业家被疑私藏枪支入狱 500 天宣告无罪获释。
	上诉	27 日，原告方不服东京高院作出的维持一审判决的裁定，将此案上诉至最高法院。
	赦免	巴基斯坦军方表示不会赦免中国人质事件主谋。

致　谢

首先要衷心感谢我的导师樊孝忠教授。本论文从选题起就得到了樊老师细心的指导和耐心的讲解，在论文工作的整个过程中，我从思想上、学术上都得到了樊老师的悉心帮助和指导。樊老师渊博的知识，对实质问题敏锐的洞察力，使我克服了研究过程中遇到的一个个困难；他孜孜不倦的工作态度和严谨求实的工作作风，潜移默化地影响了我；他宽容热情的待人态度，对学生体贴入微的关怀和帮助，都使我受益匪浅。同时，樊老师为我创造了宽松和谐的研究环境，也为我的研究工作提供了非常好的客观条件。在此，向樊老师表示由衷的敬意和感谢！

感谢实验室已经毕业的李宏乔、李良富、顾益军、许云、张锋、夏天、刘林、骆正华、刘祥瑞、郭庆琳、余正涛、康海燕、汤世平、林培光等博士，在我攻读硕士学位和博士学位的六年时间里，在和你们一起学习和讨论中，从你们那里学到了许多知识，为我的课题研究奠定了基础，向你们致以最诚挚的谢意！

感谢信息抽取小组的邓擘、尹继豪、庞文博、蒋效宇、玉德俊、陈晓阳、王金帅同学，一起学习、讨论、出游的日子是多么令人难忘，共同奋斗的每时每刻将永远留在我的记忆之中。特别是庞文博和陈晓阳在程序辅助编写方面进行了辛苦的工作，在此向你们表示由衷的感谢！

感谢实验室的陈康、刘杰、许进忠、贾可亮、傅继彬、毛金涛等在读的同学们，大家在一起相处和讨论的日子里，让我获益不少。

感谢同宿舍的翟岩龙、王彦杰同学，我们在一起度过了许多美好的时光，从大运村到新 2 号，我们一起走过。还有同班的任小金、张昱、刘法旺、谭励、贺巧艳等同学，不论一起学习英语还是探讨问题，或是一起品尝美味，那些时光都将永远留在我的记忆深处。

感谢我远在家乡的亲人们，是你们对我的关心、支持与帮助，才使得我能够全身心的投入到学习和研究之中，取得今天的成绩。特别是我的父母和岳父岳母，你们虽然年纪已大，但却为支持我的学业付出了许多。在此，祝你们身体健康！

感谢我的妻子李学钰女士，是她担负起了照看与养育孩子的责任，是她的任劳任怨的付出才使我有了今天的成绩，她和快四岁的女儿是我动力的源泉。

感谢那些不曾相识，却有幸拜读你们作品的所有研究人员，你们的真知灼见为我的研究工作指明了方向。衷心感谢为评阅本论文而付出辛勤劳动的各位专家学者。