分类号：TP391.1　　单位代码：10110
学　　号：S20080434

中 北 大 学

硕 士 学 位 论 文

基于 CRF 的中文命名实体识别方法研究

硕士研究生　王峰
指导教师　王召巴 教授
学科专业　检测技术与自动化装置

年　月　日

图书分类号 TP391.1 密级 非密

UDC[注 1]

硕 士 学 位 论 文

基于 CRF 的中文命名实体识别方法研究

王峰

指导教师（姓名、职称） 王召巴 教授

申请学位级别 工学硕士

专业名称 检测技术与自动化装置

论文提交日期 年 月 日

论文答辩日期 年 月 日

学位授予日期 年 月 日

论文评阅人

答辩委员会主席

年 月 日

注 1：注明《国际十进分类法 UDC》的分类

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在指导教师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

论文作者签名：_____ 日期：_____

关于学位论文使用权的说明

本人完全了解中北大学有关保管、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：_____

基于 CRF 的中文命名实体识别方法研究

摘　要

作为文本信息中的基本信息元素，命名实体是正确理解文本的基础。命名实体识别就是将文本信息中规定的实体识别出来,它在自然语言处理中是一项基础性的工作，在信息抽取，机器翻译，自动问答等领域有着广泛的应用。本文以中科院网络科技监测平台建设为背景，采用条件随机域模型（CRF），研究中文命名实体的识别方法。

本文通过分析目前命名实体识别的研究现状，详细阐述了近些年来国内外命名实体识别的评测活动；在分析了马尔克夫模型和最大熵模型的基础上，确立了基于条件随机场模型的研究方案。

本文在条件随机场的预处理中，以字的方式作为输入标准，从字的角度来切割文本，以获得更多文本信息的上下文特征；在模型训练中，对不同的模板对文本进行了识别，得到了一个相对较为优化的训练模板，并在训练语料中加入词性的外部特征，通过实验表明，该方法可以弥补训练规模的不足，在一定程度上提高了实体的识别效果。

本文针对中科院网络科技监测平台建设的要求，利用 SIGHAN2006 MSRA 的语料库，通过对不同模板的测试，采用模式学习方法对不同词的粒度实体进行识别，自动识别出语料中的命名实体；通过对测试语料的识别，获得实体识别的详细信息，并与正确的人工标记结果进行比较，结果说明了采用 CRF 进行命名实体识别可以取得了不错的识别效果。

论文的研究成果为日后实现监测平台准确的进行实体识别打下基础。

关键词：条件随机场　命名实体　特征模板　标注集

Research on Chinese Named Entity Recognition Based on CRF

Abstract

Named Entity Recognition is to recognize specific entities in text. As the basic information unit of text, Named Entity is essential to the correct understanding of a text. Named Entity Recognition (NER) is a basic task in natural language processing research, which is widely used in machine translation, information extraction, automatic summarization and so on. So how to identify named entity has great theoretical and practical significance.

In this paper, firstly, it investigated and summarized the current status of the Name Entity Recognition. And then, it introduced the evaluation strategy for NER, which analyzed the current method of the Name Entity Recognition.

Detailed description of the conditional random field model, conditional random field is a statistical machine learning methods, it has good performance in labeling and fragmenting the sequence. Training in the model, we added the part of speech as the external characteristics of training data. The results show that the training corpus in the external features can make up for lack of training scale, to a certain extent, improved the entity recognition.

In this article, SIGHAN 2006 MSRA were our corpus. From our research, we could test template and word size from the Named Entity Recognition experience. Through the pattern learning style, it could recognize the named entity in the corpus. We could compare to the result with the true training corpus, which had been manually labeled. As a result, we could gain the result of the experience to analyze the validity and feasibility of the model.

Finally, from the result of the experiments ,the NER using CRF is feasible. And some comments about future works are made.

Keywords: Conditional Random Field; Named Entity Recognition; Characteristic Template; Tag-Set

目 录

1．绪论

1.1 本文的研究背景 1

1.2 命名实体的定义 3

1.3 国内外现状 3

1.4 命名实体识别的评测方法 5

1.5 研究的意义 6

1.6 组织结构 8

2．相关模型的算法

2.1 隐马尔科夫模型 10

2.1.1 离散马尔科夫过程 10

2.1.2 隐马尔科夫模型 10

2.1.3 隐马尔科夫模型的基本问题 12

2.1.4 隐马尔科夫模型的局限性 16

2.2 最大熵模型 16

2.2.1 最大熵模型的描述 16

2.2.2 最大熵模型的约束条件 17

2.2.3 最大熵模型求解 18

2.2.4 最大熵模型总结 19

2.3 条件随机场概率模型的形式 19

2.4 模型参数估计 21

2.5 本章小结 22

3．命名实体识别的思路和方法

3.1 命名实体识别的难点 23

3.2 命名实体识别的主要方法 25
3.3 命名实体识别方法结构的设计 27
3.4 本章小结 29
4．基于CRF的命名实体识别模型关键问题的解决方案
4.1 标记偏置问题 31
4.2 特征的选择 33
4.3 特征模板的定义 34
4.4 文本预处理 36
4.5 本章小结 37
5．CRF实验过程及结果分析
5.1 CRF模型实体识别流程 38
5.2 CRF 实验过程及结果 39
5.2.1 实验环境 39
5.2.2 不同模板的实现结果分析 39
5.2.3 加入词性标记实现的结果分析 42
5.3 实验分析 44
5.4 本章小结 45
结 语 46
参考文献 48
攻读硕士学位期间发表的论文 54
致谢 55

第一章　绪论

在信息高速发展的今天，人们需要将大量的非结构化的信息数据转换成为结构化的信息数据，才能有效地发现信息之间的内在的结构联系，从而为更进一步的知识体系的服务和拓展做准备。因此，怎样才能从非结构化的信息中抽取出真实世界中存在的具体或抽象的实体，来说明结构化的信息数据是值得深入研究和探讨的课题。本文通过吸取和借鉴国内外在信息抽取方面已取得的相关研究的成果，阐述了在非结构化文本中自动识别实体名的思路及相关技术的方法。

本章主要是从命名实体的研究背景入手，阐述了本文研究的目标及意义，并且简明的介绍了研究的过程、思路和组织的结构。

1.1 本文研究的背景

命名实体识别技术属于信息抽取技术的范畴，并且随着信息抽取技术进步而不断的向前发展。在当今时代，随着计算机的普及以及万维网的迅猛发展，大量的信息仅仅靠纸制的形式已经完全不能满足人们的需求，于是大量的信息被制成电子文档的形式出现在人们的面前或者存放在相关数据库中。在欧洲，很多机构正在致力于此项工作的研究，如欧洲科学技术与卓越描绘（Mapping of Excellent in Science and Technology，ME）[1]一直把机构的科研人员的数量、科研项目、科研水平等作为机构的结构数据进行研究。还有一些类似的前沿研究也取得了不错的研究成果，如研究趋势探测（Emerging Trend Detection ，ETD）[2]，研究领域描绘（Mapping of Research Specialty）[3]等。他们获得这些信息的途径主要是通过数据库。但是这类科研要素的获取量比较有限，而且揭示的也只是以文档的外部特征为主，而大量的非结构化的文本内的众多的语义信息未能得到利用。因此，如何从如此巨大的网络信息和科技信息中获取用户需要的信息（知识）是人工智能和Internet研究的一个主题，而其中关键的环节之一就是如何从非结构化的数字文本中自动识别结构化的信息数据。

为了从不同的信息源中获取用户不同层次和粒度的信息，于上个世纪的八十年代兴起，着力于解决非结构化数据中结构化信息的识别，人们发明了各种不同的信息获

取技术，得到数据间的关联关系，正是在这种背景下产生了信息抽取（Information Extraction，IE）[4]。总而言之，随着信息化的发展，旨在发现信息中隐含语义知识，集成，抽取数据资源中语义信息的IE抽取技术得到了更广泛的应用和研究。而命名实体识别（Name Entity Recognition，NER）是信息抽取中非常重要且不可或缺的一步。此外，近些年有许多研究者在NER的研究方面已经进行了许多的研究和实验，并取得了一些卓有成效的成绩，而他们对于该问题的兴趣源于该问题的广阔的应用领域和吸引力[5,6]。其中，作为这些研究中非常重要且必不可少的关键技术之一的命名实体识别技术，越来越受到人们的重视和关注，迄今为止它已发展成一个相对较为独立的研究领域。这些研究为本论文提供了一定的理论和实践的指导。

在国际上，早在 1987 年就由美国国防高级研究计划委员会（the Defense Advanced Research projects Agency，DARPA)资助的消息理解系列会议（Message Understanding Conference，MUC ），这个会议从开始于 1987 年，在长达十年的时间中总共举办了七届[7]会议。其中在 1995 年 9 月举行的第六届MUC会议中，正式的引入了命名实体评测任务[8,9]。在MUC的会议中命名实体被定义为：人们感兴趣的特定的专有名词和特定的数量词[10]。在 1997 年的MUC会议中命名实体任务的定义中，命名实体识别被指明包括了三个具体的任务：时间表示语、实体名、数量表达式的识别。其中实体名包括： 地名、人名、机构名；时间表示语包括：时间短语和日期短语；数字表示语包括：比例值和货币短语。在有关命名实体识别的最新的研究中，许多学者又根据不同的研究需求和应用实践又再次扩大了名实体的范围。

在MUC之后，从 2000 年至今，特别是经受了“911”之后的美国，出于对国土安全和国际恐怖主义的极度担忧的考虑，美国国家标准技术局[11]（NIST）开始资助了一系列的“自动内容抽取”（ACE）评测会议。先后文本理解会议（Document Understanding Conference ,DUC），自动内容抽取会议（Automatic Content Extraction，ACE）等信息抽取领域的国际评测会议，自动抽取新闻语料中出现的实体、关系、事件等内容作为了这些研究会议的主要目标。与同期的MUC相比较，目前的ACE评测主要采用基于误报（标准答案中没有而系统输出中有）为基础的一套评测体系和漏报（在标准答案中有而系统输出中却没有），并且对系统是否具有跨文档处理能力进行评测，而并不会针对某个特定的领域或场景。也在继续着命名实体的识别评测，并提供了相应的评测任

务。

在中文命名实体方面，我国对于这方面在不断地加大研究的力度，国家 863 命名实体识别评测小组研究并制定了《2004 年度命名实体识别评测大纲》121 大纲[12]，这个大纲详细地介绍了命名实体主要的评测任务: “命名实体的任务由三个子任务组成(命名实体、时间表达式、数字表达式)。被标注的表达式为命名实体(组织、人、地点)、时间(日期、时间)及数量。”

1.2 命名实体的定义

命名实体（Named Entity ，NE）作为一篇文章中的基本元素，包含了文章的许多内容，往往是正确理解这篇文章的基础，在不能阅读全文的情况下，命名实体识别可以作为一种快速了解该文章主要内容最为快捷的途径。就定义方面，狭义上说，命名实体是指现实世界中抽象或具体的实体，如人、组织、地点、公司等。通常用唯一的标识符（专有名词）表示，如人名、组织名、地点名、公司名等。广义上说，命名实体还可以包含数量、时间表达式等。总之，对命名实体而言，可以视具体的情况而定。而命名实体识别（Name Entity Recognition，NER）是指识别出文本中特定的实体。它对于中文信息处理来说是一项很有价值技术。

1.3 国内外现状

在国际上对于英文命名实体识别的研究比较早，而且在英文命名实体识别方面也取得了很大的成就，其中在MUC会议上的测试的F(评测度量的标准)值最优秀的可以达到 95%左右。Borthwick[13]在针对英文和日文的命名实体识别中利用了基于最大嫡模型（Maximum Entropy）的方法；Kiyotaka Uchimoto和Qing Ma在IREX竞赛中采用基于最大熵的转移规则的命名实体方法[14]。Roth实现了一个基于半指导的NE识别系统[15]则主要借助了句法分析信息帮助NE识别的条件。作者将NE识别问题细分为NE边界划分和NE分类两个不同的过程。第一阶段中从cnostiutneyc和dPenedneyc句法树中抽取候选NE实体并识别而获得训练语料主要是利用了句法分析信息帮助NE识别的条件。在第二阶段，作者则用EM算法进行参数的迭代计算和贝叶斯模型分类器进行分类。在利用

ACE语料训练和测试中，证明了constituency和dependency句法约束对NE分析器训练和抽取都是很有必要和可行的。Andrew McCallum对最大嫡马尔科夫模型和隐马尔科夫模型在英文名实体识别中的情况[16]进行了比较，并阐述了在文本处理中的隐马尔科夫模型的具体应用，隐马尔科夫模型作为一种有限的概率状态模型，其中包括两种参数:某一特定状态的观察概率(发射概率)和状态转移概率。

欧盟与 2008 第七框架计划项目OKKAM，就旨在探索一系列与实体相关的技术，为将互联网转换为以实体为节点的网络提供基础,其中就涉及了实体名称的规范问题；美国OCKC积极地与Illions大学[17]，马里兰州大学开展合作，解决命名实体的识别和排岐问题，为构建规范文档提供基础支撑；保加利亚的OntoText[18]公司在构建KIM语义知识库时也重点关注了实体名称问题；

与国外大规模的开展的实体识别项目相比。目前，中文实体识别的研究才开始不久，与英文实体相比，汉语命名实体的识别任务要复杂得多，主要表现在：汉语文本没有类似英语文本中空格之类的显示标示词边界的标示符,分词和命名实体相互影响。因此，在中文命名实体识别方面也将面临着更大挑战和难度。

新加坡大学提出的中文NE的识别方法[19] 主要是基于多重主体结构推理模型。中文命名实体的识别过程被分成了两个阶段。第一阶段它会使用NE的推理模型和贪心算法来提取文档中所有的待选命名实体并进行评价。接着对于命名实体过程中概率最大的过程作为一个多重主体协商问题来处理。这种方法能对很多复杂NE的识别，在准确率上有很大幅度的提高，但系统的迭代过程太多，计算量很大，导致运行效率不高，运行时间很长。

Yu[20]的识别系统利用语义标记、词性标注和命名实体列表等信息，构建了一个包含基于语言学识别模块和基于上下文识别模块识别系统。经过测试，系统的F值可以达 86.38%。但是由于对语言学的过度依赖性的特点，系统的可移植性不强。

谭鸿业通过基于转换的方法来识别地名[21]，取得不过的效果；李建华[22]利用了基于名用词和中文姓氏统计统计概率的方法对人名进行识别；黄德根则通过统计的方法来识别人名[23]；俞鸿魁等[24]通过构造如前缀、后缀、特征词等常用的机构名识别特征，利用不同角色标注，借助多层隐马尔可夫模型的工具，并由此对组织机构名的进行了测试并取得了不错的效果；吴雪军等[25]对两个标注器利用已标注好的组织机构名语料

进行了学习，并构造了两个不同的分类器分别由机构名上下文特征信和组织机构名的内部结构特征信息组成，最后从大规模未标注语料中利用两个标注器进行上下文特征信息和内部特征学习，从而取得机构名特征构造机构名识别知识库，并成功的将co-training机器学习方法应用到组织机构名的识别当中。

zhou[26]构建了模型系统是基于层叠条件随机场模型(CCRFs)。在本系统中，通过以观察值为条件的低层条件随机场模型为基础，用于一些基本命名实体的判别，当把最终结果传递给高层模型时，高层模型的输入变量将会包含数据和来自低层模型的识别观察值，这样就低层观察值为高层条件随机场模型提供了有效的决策支持，提高了复杂机构名的识别正确率，最后采用约束的forward-backward算法对识别结果进行可信度测试。通过在人民日报语料库上测试的结果F值达到了 92.89%。

而后来诞生一种新的模型，条件随机场（Conditional Random Field，CRF）[27]在命名实体识别方面也取得了不错的效果。CRF是一种以马尔科夫模型为基础的模型的统计方法，它着眼于字的角度，很好的利用了词以及词性的上下文信息的特征，很好利用了外部的特征，因为从字的角度出发理论上可以很好的避免碎片化的形成。而且从字角度出发来发现句子中词的特征，也更加符合中文以字为基础的特征，并且马尔科夫模型很好的解决了隐型马尔科夫模型和最大熵模型存在的一些缺点，如标记偏置等。而在自然语言处理领域马尔科夫模型也得到了很好的应用，如在英文的分词标注、英文短语名词的识别等领域都取得了较好的效果。与国际上的CRF学术研究相比，国内对于CRF模型方面的研究还比较少。

1.4 命名实体识别的评测方法

主要有两个评价指标来衡量命名实体识别的系统性能:准确率和召回率。准确率是系统正确识别的结果占所有识别结果的比例，召回率是系统正确识别的结果占所有正确结果的比例。两者的计算公式如下：

$$\text{准确率（P）}=\frac{\text{系统中正确标记的NE个数}}{\text{系统找到的NE总数}} \qquad (1.1)$$

$$召回率（R）=\frac{系统中正确标记的NE个数}{文档中的NE总数} \quad （1.2）$$

准确率是指抽取的信息中正确抽取的比例；召回率则是指正确抽取的信息占应抽取信息的比例。当比较两个不同信息抽取系统的性能时，一般使用这两个指标的综合值：F 度量，即

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

在上式中：P 为精度；R 为召回率；β 为对精度的偏重量。其中，β 决定了 R 和 P 的重要程度。若 β=1，则将 R 和 P 视为同等作用；若 β=2，则将 R 的重要度视为 P 的 2 倍；若 β=0.5，则将 P 的重要度视为 R 的两倍。一般取 β=1，本文中后期的评测都将取 β=1。

在中文命名实体方面，BAKEOFF-3[28-30]提供的命名实体评测分为简体文本和中文文本两类，简体语料有微软亚洲研究院（MSRC）和LDC（Linguistic Data Consortium）提供，繁体语料由香港城市大学（CITYU）提供，评测分为Open Track和Closed Track两种类型，共有 16 个系统参加了BAKEOFF-3 的命名实体识别的测试，表 1 提供了 6 个测试类别最好的测试性能指标。

表 1.1　BAKEOFF-3 命名实体识别六个评测任务中性能最好测试指标

数据来源	*测试类别*	*P*	*R*	**F**
MSRC（简）	Closed	0.8994	0.8420	0.8651
	Open	0.9220	0.9018	0.9118
LDC（简）	Closed	0.8026	0.7265	0.7627
	Open	0.7616	0.6621	0.7084
CITYU（繁）	Closed	0.9143	0.8676	0.8903
	Open	0.8692	0.7496	0.8051

1.5 研究意义

通过对背景以及国内外现状的介绍，可以发现从海量的数据中获取识别出中文命名实体，以满足对于知识数据挖掘的需求，为自然语言处理的其他方面提供知识支持，是整个 IE 抽取的基础。具体而言，该研究主要有以下几方面的重要意义。

（1）信息检索(Information Retrieval):命名实体是文本中信息中的主要载体，准确的分析和抽取文本中的命名实体，能够更准确的检索到相应的文档信息，从而使得信息检索的结果更加准确。

（2）开放域问答（Open-Domain Question Answering）：在许多开放域问答系统中，在很多情况下需要得到信息数据中的某个机构，地点，时间等具体的问题，而普通分词结果是很难满足要求的，而只能返回到原文档信息进行人工获取。而命名实体识别就能够很好的帮助开放域问答客服这种困难，他对文本中的这些实体信息进行了识别，使得系统能够获取更准确的答案。

（3）机器翻译（Machine Translation）：在机器翻译时，由于主要是通过词语对齐来进行翻译，在遇到机构名、地名等实体时，因此导致翻译的结果常常出现偏差，甚至有可能出现错误。但是通过加入实体识别之后，这样就能使机器翻译的时候的英汉对照能够达到短语级别，从而使翻译的语句更通顺，从而避免错误。

（4）按照MUC-7 的规定，在自然语言处理中，信息抽取过程主要包含 3 个不同的层次（由易到难）的任务：模板元素（Template Element ,TE）任务，模板元素主要的任务就是指提取文本中相关的命名实体，主要包括机构名，地名，人名的识别；模板关系（Template Relation，TR）任务，是指提取命名实体之间的各种关系，脚本模板（Scenario Template, ST）任务，提取出事件模型，包括事件中的各个属性，关系和实体[13]。而命名实体识别时上面所说的TE的关键技术，是整个信息抽取的基础，其准确率直接影响了接下来进行的TR和ST任务的质量。

本文以中科院网络科技监测平台建设课题为研究背景，在整个平台的建设中，命名实体识别作为自然语言处理的一个基础性工作在整个信息抽取占据着举足轻重的位置，它对于文本数据的知识获取，它是展开其他工作的基础，对于信息处理平台的构建，信息资源的检索以及整合等研究产生了深远的影响。因此，本文的研究任务大致

的可以分为以下 3 点：

（1）对中文命名实体识别的各种方法进行研究，对比找出最优或较优的方法进行实验，为今后进行进一步中文命名实体识别研究做好理论基础的准备。

（2）由于条件随机场作为目前学术界公认较优的机器学习方法之一，本研究将使用条件随机场模型去尝试着解决中文命名实体识别中存在的问题，通过构建一个基于条件随机场模型的识别系统，根据模型的训练和测试各方面的因素对系统进行设计和调优，最终完成一个能够识别中文命名实体的命名实体识别系统。

1.6 组织结构

本论文旨在探索自动，准确的识别非结构化文本信息中的人名，机构名和地名的技术研究方法，针对一些关键的问题提出适当的解决方案，并且通过相应的实验来验证该方法的有效性和可行性。因此，就本论文需要解决的问题，笔者在这里确定了几个需要突破的关键环节。

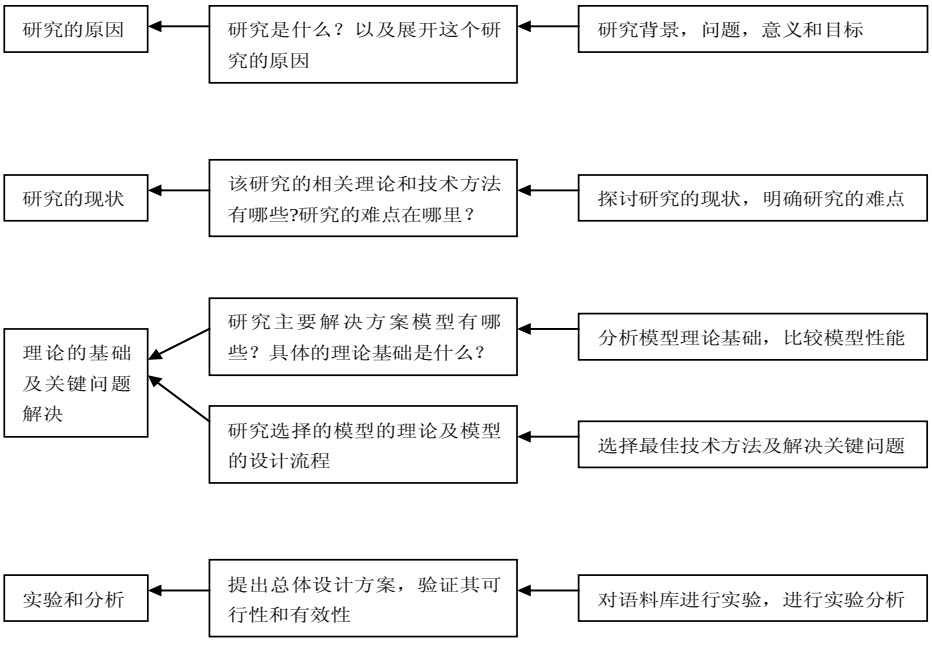

图 1.1 文本研究思路

与以上的研究思路相对应，本论文内容也主要是按照以上五个部分组织，详细阐述了研究的思路和过程，以下将根据这个思路介绍进行编写。

第一章 ，绪论。介绍命名实体的研究背景，介绍了国内外的研究现状以及命名实体的相关概念，并且阐述了本文研究的目标及意义。将本文的研究思路和组织结构的形式简要的描述。

第二章，相关模型的算法。本章主要介绍了现在比较常用的两种统计模型的算法：隐性马尔科夫模型，最大熵模型，条件随机场模型。隐性马尔科夫模型，最大熵模型，条件随机场模型是目前最常用的解决命名实体识别的三种模型方法，为后文的条件随机场模型的关键问题的解决方案和设计方案的提出打下理论基础。

第三章，命名实体识别的思路和方法。本章节简述了中文命名实体识别时的难点，并就研究领域比较主流的解决方法做了介绍，并提出自己对于本模型的设计思想，最后结合了 L-BFGS 的算法对 CRF 模型的模式学习流程的方式进行了解释。

第四章，基于 CRF 的命名实体识别模型关键问题的解决方案。针对条件随机场模型，介绍了模型创建时的关键问题解决方案。阐述了特征类型，标注等方面的知识，并对于条件随机场的优点做了一定的描述。

第五章，实验过程及结果分析。本章以第三、四章的理论和实践为基础。选取有效的模板，对实验结果进行了统计分析和计算，论证了条件随机场模型对中文命名实体识别的可行性。总结本论文的主要工作，研究成果，对实验中不足之处提出了改进意见，并对后续性的研究工作进行了展望。

第二章 相关模型的算法

在本文中将以统计的方法为主，对中文命名实体进行了识别研究。本章主要介绍了隐形马尔科夫模型、向前算法、Viterbi 解码算法、最大熵模型，比较了这几个模型之间的优缺点，这样有助于更好的理解模型，对日后的命名实体工作进行扩展打下坚实的理论基础。

2.1 隐马尔科夫模型

隐性马尔科夫模型的基本理论形成于上世纪 60 年代末期和 70 年代初期，是建立在俄国有机化学家 Markovnikov 提出的马尔科夫模型的基础上提出的。其数学本质是一个随机过程，在若干年之后重新受到欢迎。主要是基于它内在丰富的数学理论结构。因此在描述隐性马尔科夫定理之前，先简单介绍一下离散马尔科夫定理。

2.1.1 离散马尔科夫过程

设 Q 表示随机过程$\{q_t, t \in T\}$的状态空间，$\{S_1, S_2, S_3 ... S_n\}$表示有 N 个不同的随机过程的状态，$a_{ij}$表示状态$S_i$到$S_j$的转移概率。当随机过程满足：

$$P[q_t = S_i \mid q_{t-1} = S_j, q_{t-2} = S_k, ...] = P[q_t = S_i \mid q_{t-1} = S_j] \quad (2.1)$$

这个就是马尔科夫或无后效性指当前状态只依赖于过去有限的已出现状态历史，即过程“将来”的情况与“过去”的情况是无关的。可以用转移状态a_{ij}来表示

$$a_{ij} = P[q_t = S_i \mid q_{t-1} = S_j] \quad (2.2)$$

2.1.2 隐马尔科夫模型

隐马尔科夫模型[31]是一个双重马尔科夫随机过程，它是一个隐含的随机过程，这个序列本身对于观察者是不可见的，但是它可通过一个可见序列得到这个隐含的序列的概率，这其中就包括了状态转移概率的马尔科夫链和输出观测值的随机过程。由于其状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来，因此而称之为“隐”马尔科夫模型，即HMM。它包含隐马尔科夫模型主要有 5 个因素组成：

（1）N，表示马尔科夫模型的状态数。模型中的各个状态之间是相互连结的，任何状态能从其他状态到达，S表示模型中的状态$\{S_1,S_2,S_3...S_n\}$的集合，q_t 表示$\{q_t,t\in T\}$的状态空间。

（2）O，表示每个状态的观察值，M 为每个状态对应的可能观察符号，即输出可观察的字符个数。观察值对应于模型系统的实际输出，记这些观察值为：$V=\{v_1,v_2,...v_m\}$。

（3）$A=\{a_{ij}\}$，状态转移概率矩阵，其中$a_{ij}=P(q_t=S_i\,|\,q_{t-1}=S_j)$，$1\le i,j\le N$。$a_{ij}$表示从状态 i 转移 j 的概率。当状态$S_i$ 经过一步到达S_j 时，a_{ij}满足：$a_{ij}\ge 0,\forall i,j$；且$\sum a_{ij}=1,\forall i$。

（4）$B=\{b_j(k)\}$，观察值概率分布矩阵，其中$b_j(k)$表示在状态下，t 时刻满足的概率，即$b_j(k)=P$（在 t 时刻出现$v_k\,|\,q_t=s_i$），$1\le j\le N,1\le k\le M$。$b_j(k)$满足：$b_j(k)\ge 0,\forall j,k$；且$\sum_k b_j(k)=1,\forall j$。

（5）$\pi=\{\pi_i\}$，初始状态分布矢量，其中$\pi_i=P(q_1=S_i),1\le i\le N$，即在t=1 时刻处于状态$S_i$的概率。$\pi_i$满足：$\sum_i \pi_i=1$。

首先从最简单的离散 Markov 过程入手，可以发现，一个 HMM 是一组有限的状态，Markov 随机过程具有如下的性质：在任意时刻，从当前状态转移到下一个状态的概率与当前状态之前的那些状态没有关系，以下是一个 HMM 的组成示意图。

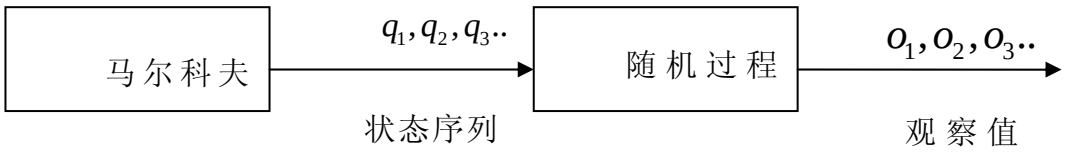

图 2.1 HMM 组成示意图

所以，这里可以用 $\lambda=(A,B,\pi)$ 来表示一组完整的隐性马尔科夫模型。给定 $\lambda=(A,B,\pi)$，观察序列 $O=O_1O_2O_3...O_t$ 可以由下列的步骤产生：

（1） 根据初始状态概率分布令 $\pi=\pi_i$，选择一个初始状态 $q_1=s_i$；

（2） 设 t=1；

（3） 根据状态 S_i 的输出 $b_i(k)$,输出O_t=V_k。

（4） 根据状态转移概率分布 a_{ij}，转移到新的状态、

（5） 设 t=t+1。如果 t<T，重复（3）（4），否则结束。

2.1.3 隐形马尔科夫模型的基本问题

1 预测问题

如何从观测序列 $O=O_1O_2O_3...O_t$ 以及隐形马尔科夫模型 $\lambda=(A,B,\pi)$ 快速的计算出观测序列的概率 $p(O|\lambda)$ 是首相要解决的问题。对于一个观测序列来说最简单的方法就是观察序列长度为T，从而枚举观察序列O所有可能的输出状态。假设状态数为N，那么就会导致计算量为 $2T*N^T$，这种方法会耗费大量的资源，因此不可行。目前主要采用forward-backward的方法解决优化此类问题。

首先，这里定义在时间t时刻的输出字符是 $O=O_1O_2O_3...O_t$ 并且位于状态S_i的概率是 $a(i)=P(O_1O_2...O_t,q_t=s_i|\lambda)$，先前算法主要是通过前一状态的值以及概率分布计算出后一状态的值，从而求得 $p(o|\lambda)$。其中 $p(o|\lambda)=\sum_{i=1}^{N}a_T(i)$ 表示的是所有状态 q_t 下观察到序列 $O=O_1O_2O_3...O_t$ 的概率。其次马尔科夫存在 3 个假设：

（1） t 时刻的状态只依赖于 t-1 时刻的状态；

（2） t 时刻所生成的值只依赖于 t 时刻的状态；

（3） 状态与具体的时间无关；

根据以上条件，这里可以用动态规划推导出向前变量 $a_t(i)$，即该算法通过将时间 t+1 的向前变量表示为在时间 t 时刻向前变量 $a_t(1),a_t(2),...,a_t(N)$ 值递归的方式的出来，向前算法的推导步骤如下：

（1） 初始化：

$$a_1(i) = \pi_i b_i \tag{2.3}$$

（2） 归纳计算：

$$a_j(t+1) = P(O_1 O_2 ... O_{t+1}, q_{t+1} = s_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_1 O_2 ... O_t, q_t = s_i, q_{t+1} = s_j \mid \lambda) P(O_{t+1} \mid q_{t+1} = s_j, \lambda)$$

$$= [\sum_{i=1}^{N} a_i(t) a_{ij}] b_j(O_{t+1}) \tag{2.4}$$

$$= P(o \mid \lambda) a_{ij} b_j(O_{t+1})$$

由以上可知forward算法统计 $P(O \mid \lambda)$ 的算法复杂度时，需要得到每个 $a_i(j)$ 的和，而 $1 \le j \le N, 1 \le t \le T$ ，所以复杂度为N^2*T，远小于直接计算所用的复杂度。

同理根据forward的算法，可以很轻松的推导出backward的。同样，这里假设在时间t状态为s_t的条件下，变量 $\beta_t(i)$ 为：

$$\beta_t(i) = P(O_{t+1} O_{t+2} .... O_T \mid q_t = s_i, \lambda) \tag{2.5}$$

通过递归可以得到：

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(i) \tag{2.6}$$

综上所述，以上可以用一个图示更好的解释以上算法模型：

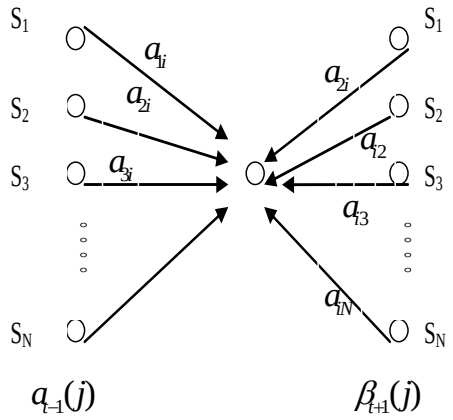

图 2.2 先前向后算法模型

2 解码问题

给定观察向量序列 $O=O_1O_2O_3...O_t$ 和隐性马尔科夫模型 $\lambda=(A,B,\pi)$，如何选择一定意义下的最优解。一种理解是该状态序列中的每一个状态都单独具有最大的可能性，用向前向后变量来表达：

$$\gamma_i(i)=P(q_t=s_i\mid o,\lambda)=\frac{P(q_t=s_i,o\mid\lambda)}{P(o\mid\lambda)} \tag{2.7}$$

上式中：$P(q_t=s_i,o\mid\lambda)=a_t(i)\beta_t(i)$ 为 HMM 的输出 o，节点的概率为: $P(o\mid\lambda)=\sum_{i=1}^{N}P(q_t=s_i,O\mid\lambda)=\sum_{i=1}^{N}a_t(i)\beta_t(i)$ 。因此 $\gamma_i(i)$ 可以表示为：

$$\gamma_i(i)=\frac{a_t(i)\beta_t(i)}{\sum_{i=1}^{N}a_t(i)\beta_t(i)} \tag{2.8}$$

有了 $\gamma_i(i)$ 之后，问题就变成了求最大值的问题：

$$q_t=\arg\max_{1\le i\le N}[\gamma_i(i)] \tag{2.9}$$

另一种理解是在给定的模型 λ 和观察序列O的条件下求出现此观察序列最大的状态序列。Viterbi算法就是运用这种思想实现的最优状态序列的算法。利用viterbi算法主

要是为了找到一个概率最大的路径Q=$q_1q_2\ldots q_T$，复杂度为NT^2，可以定义$\delta_t(i)$为t时刻状态$q_t=s_i$时刻的最大概率当前序列和前t个时刻观察序列的联合概率：

$$\delta_t(i)=\max_{q_1q_2\ldots q_T} P(q_1q_2..q_t=s_i,o_1o_2...o_t\mid\lambda) \qquad (2.10)$$

由t时刻状态$q_t=s_i$时的最优状态序列和前t个观察序列的联合概率$\delta_t(i)$可以递推得到$t+1$时刻状态$q_t=s_i$时的最优状态序列和前t个观察序列的联合概率可递推得到$t+1$时刻状态$q_{t+1}=s_j$时的最优状态序列和前$t+1$个观察序列的联合概率 $\delta_{t+1}(j)$:

$$\delta_{t+1}(j)=[\max_{1\le i\le N}\delta_t(i)a_{ij}]b_j(O_{t+1}) \qquad (2.11)$$

这里可以用数组$\varphi_t(j)$保存 t 时刻的各个状态及处于最优路径的前一个状态的索引值，viterbi 算法步骤如下：

（1） 初始化：

$$\delta_1(i)=\pi_i b_i(o_1) \qquad (2.12)$$

（2） 递归得到：

$$\delta_{t+1}(j)=[\max_{1\le i\le N}\delta_t(i)a_{ij}b_j(O_{t+1})] \quad 2\le t\le T,1\le j\le N \qquad (2.13a)$$

$$\varphi_{t+1}(j)=\arg\max_{1\le i\le N}[\delta_t(i)a_{ij}b_j(O_{t+1})] \quad 2\le t\le T,1\le j\le N \qquad (2.13b)$$

（3）最大概率P的最优路径和最优路径状态序列的最后一个概率状态记为q_t，

$$P=\max_{1\le i\le N}[\delta_T(i)] \qquad (2.14a)$$

$$P=\arg\max_{1\le i\le N}[\delta_T(i)] \qquad (2.14b)$$

（4） 路径（状态序列）回溯

$$q_t=\varphi_{t+1}(q_{t+1}) \quad t=T-1,T-2,\ldots1 \qquad (2.15)$$

虽然 Viterbi 算法和 forward 计算过程很相似，但其区别在于 viterbi 算法是根据前一个的状态序列得到使当前的状态具有最大概率的那个状态，而 forward 算法是得到的是前一个的状态序列计算转移到当前状态的的概率和。

3 学习问题

学习问题归根结底给定一个观察序列的参数估计问题，如何根据最大似然估计来

求模型的参数值，HMM 模型一般采用的是 EM 方法进行参数估计。其基本思想是：初始时随机地给模型的参数赋值，得到该模型 λ_0，由 λ_0 可以得到模型中隐变量的期望值。然后可以从 λ_0 得到从某一状态转移到另一个状态的期望次数。以期望的次数作为实际的次数，便可以得到模型参数重新估计，由此可以得到模型 λ_1，然后再循环，直到模型收敛于最大释然估计。

2.1.4 隐形马尔科夫模型的局限性

隐性马尔科夫模型是属于生成模型所要求的严格的独立假设,即它定义了一个在给定序列 $O = O_1O_2O_3...O_t$ 和标记序列Q=$q_1q_2…q_t$ 符合贝叶斯公式

$$P(O|Q)P(Q) = P(O,Q) = P(Q|O)P(O) \tag{2.16}$$

定义了标记序列和观察序列的联合分布以后,则就假设所有可能的观察序列必须都是能够枚举，如果长距离依赖存在在观察序列中，枚举所有可能的观察是不现实的。他的独立性假设，在实际上，数据的序列不可能完全的成为一个独立的单元。下面所要介绍的最大熵和条件随机场模型就是克服了这种问题。

2.2 最大熵模型

最大熵模型（Maximum Entropy Models,MEMs）是基于最大熵理论的统计模型，广泛应用于自然语言处理当中。最大熵的概率分布服从已知道的不完整信息的约束，如果训练数据包含的信息用特征函数的集合表示，那么最大熵的值在分布最一致的时候熵能得到最大值，一切事物都能够达到最大值在一定的客观条件下。如果在有一系列限制条件下的选择一种分布，但在这种条件下又不能确定某种分布是唯一的，这就是最大熵分布的适用范围。

2.2.1 最大熵模型描述

在自然语言处理中有很多问题都可以归结为统计分类问题，例如对于中文人名和地名自动识别任务，用有限集Y表示一个候选词分到某个类别看成的一个事件，该词的上下文在语料中可以看成事件发生的环境并用集合X表示，则在给定的上下文信息

$x \in X$，计算输出的词x是属于Y的哪一类的概率，即计算 $y \in Y$ 的条件概率P（y|x），模型的输入为人工标注后的训练数据样本集D={(x_1,y_1),(x_2,y_2)…(x_t,y_t)}。这里可以从训练样子本中归纳出随机变量x和y的联合经验概率分布 $\tilde{P}(x,y)$：

$$\tilde{P}(x_i,y_j)=\frac{count(x_i,y_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}count(x_i,y_j)} \tag{2.17}$$

这个方法存在最大问题就是处理稀疏数据时候，如果数据量有限，或者在此次测试数据中的正好这个二元数组的值没有出现，因此肯定不能就此认为该二元组发生的概率为零。最大熵模型就是用来解决这种稀疏事件的，它原则上使未知事件的概率分布尽可能的分布均匀，即得到最大熵的分布。具体的来说根据 Shannon 的定义，熵的计算公式为：

$$H(X)=-\sum_{x}p(x)\log_2 p(x) \tag{2.18}$$

其中熵有如下特性：

$$0\le H(X)\le \log|X| \tag{2.19}$$

|x|表示离散分布时的随机变量的个数。当 X 为确定值，即没有变化的可能时 H（x）=0 成立，由条件：$\sum_{x\in X}p(x)=1$，对熵求极值，当随机变量 X 服从均匀分布时，H（x）=log|x|成立，即熵最大。

2.2.2 最大熵模型的约束条件及原则

最大熵模型是利用条件约束的最优化问题来获得相应的概率分布的最终解。而通过引入特征方程来表示这种训练样本中的数据的约束，用特征函数 f(x,y)来表示，例如，在分句中可以有这样的特征函数

$$f(x,y)=\begin{cases}1 & x\text{和y满足的前提和动作}\\ 0 & otherwise\end{cases} \tag{2.20}$$

设 $\tilde{P}(f)$ 是现对于经验分布 $\tilde{P}(x,y)$ 的特征函数 f 的数学期望，称为经验期望，p(f)是相对于模型确定的概率分布 p(x,y)的数学期望，称为模型期望。其中约束由模型得到的特征函数的数学期望等于由训练样本得到的特征函数的经验期望，

即 P(f)= $\tilde{p}(f)$，这里称之为模型约束条件。因此，以上可以得到下面的等式：

$$\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{2.21}$$

最大熵模型的原则，可以根据p（y|x）来定义。P表示所有条件分布的集合C，设有n个训练样本集（x_1,y_1),$(x_2,y_2)\ldots(x_n,y_n)$，其中x_i是由k个属性特征构成的样本向量，y_i为类标记。在满足上面所叙述的模型条件约束的前提下，随机事件的不确定性可以用条件熵来衡量，其定义如下：

$$H(p)=-\sum p(x)p(y|x)\log p(y|x) \tag{2.22}$$

对于任意跟定约束集 C，能找到唯一的 $p^*\in C$ 使得 H(p)取得最大值，如何找到 p* 是一个约束优化的问题。其中 P（y|x）具有以下的形式：

$$p(y|x)=\frac{1}{Z（x）}\exp\sum_{i=1}^{n}\lambda_i f_i(x,y) \tag{2.23}$$

$$Z(x)=\sum_{y}\exp(\sum_{i=1}^{n}\lambda_i\ f_i(x,y)) \tag{2.24}$$

其中，λ_i 是参数，在上面的式子中表示为特征函数的权值。在 x 确定时 Z(x)为归一化的范化常数。

2.2.3 最大熵模型求解

从以上关于最大熵原理的叙述，这里发现最大熵的目标主要是通过计算特征的权值的估算，来产生模型。目前，用得比较多的方法是 IIS（Improved Iterative Scaling）算法，它是在 GIS 算法上做出了改进，降低了求解算法的约束条件，增加了算法的适用性。

IIS 的算法如下：

（1） 初始化 $\lambda_i=0$，并用公式 $p(y|x)=\frac{1}{Z(x)}\exp\sum_{i=1}^{n}\lambda_i f_i(x,y)$ 计算相应的 $p(y|x)$。

（2） 将 $\Delta\lambda_i$ 带入到 $\sum_{x,y}\tilde{p}(x)p(y|x)\exp(\Delta\lambda_i f^*(x,y))=\tilde{p}(f_i)$ 其中：

$$f^*(x,y)=\sum_{i=1}^{n}f_i(x,y)$$

(2) 更新计算 $\lambda_i = \lambda_i + \Delta\lambda_i$

(3) 若 λ_i 没有收敛，返回步骤（2）；

最后得到 $\lambda^* =< \lambda_1, \lambda_2, ..., \lambda_n >$

2.2.3 最大熵模型总结

最大熵是条件模型中的代表模型。它没有产生式模型中对观察序列设置非常严格的假设条件，它可以灵活的设置约束条件。选择不同的特征可以构建出不同的特征模型，扩大了模型的适用性，使得统计模型实现了面向问题的思想。最大熵可以调节模型对未知数据的适应度主要是通过控制约束条件的多少。另外，不需要列出所有的可能的观察序列。自动解决了统计模型中参数平滑的问题。

经典的 ME 模型适用于任何的分类问题，即如何通过给定的样本找到其最优解，在文本分类中有很多应用。但是在应用过程中会碰到这样一类问题，即输入的样本不再是单个的，而是一段连续的样本序列，例如汉语分成，词性标注，命名实体识别等问题。当遇到此类问题时，由于每个随机变量非独立，因此在最大熵中引入了马尔科夫模型，即最大熵马尔科夫模型。

2.3 条件随机场概率模型的形式

Lafferty[32]对CRF势函数的选择是以最大熵模型作为了一个基准。它定义了一个势函数：

$$\phi_{y_c}(y_c) = \exp(\sum_k \lambda_k f_k(c, y \mid c, x) \tag{2.25}$$

其中 c 表示无向图中 Y 的最大全连通环中的节点的索引，C 是无向图中所有的全连通环，$\phi_{y_c}(y_c)$ 是一个建立在无向图中的 Y 的最大全连通环至上的势函数，它要求此函数必须为正实数。y|c 个布尔 表示第c个全连通环中的节点对应的随机变量，f_k是一个布尔型的特征函数，则 p(y|x)定于为：

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{c \in C} \sum_k \lambda_k f_k(c, y_c, x) \tag{2.26}$$

其中 Z(x)是归一化因子：

$$Z(x)=\exp(\sum_{c\in C}\sum_{k}\lambda_k f_k(c,y_c,x) \tag{2.27}$$

在一阶链式结构的图 G=（V,E）中，最大的全连通环是它的边的 E，它仅包含两个节点。对于一个最大的全连通环中的无向边$e=(v_{i-1},v_i)$，势函数的一般表达式可以扩展为：

$$\phi_{y_c}(y_c)=\exp(\sum_{k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{k}u_k g_k(y_i,x,i)) \tag{2.28}$$

其中$t_k(y_{i-1},y_i,x,i)$表示整个观察序列和相应标记序列在 i-1 和 i 时刻的特征，他表示状态转移函数。而$g_k(y_i,x,i)$表示状态函数，是在 i 时刻标记的特征序列和整个观察序列。所以，联合概率的表达形式为：

$$p(y\mid x)=\frac{1}{Z(x)}\exp(\sum_{i}\sum_{k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i}\sum_{k}u_k g_k(y_i,x,i)) \tag{2.29}$$

通过分别为上式定义一个特征函数，状态值以及对应的观察值，可以为 CRFs 建立 HMM 属性。在函数定义之前，必须先把观察序列的真实特征 b（x,i）的集合构造出来，这个集合同时体现了模型分布和训练数据的经验分布的特征。

$$b(x,i)=\begin{cases}1 & \text{If}\quad \text{i 的位置的观察值是}\\ 0 & \text{Otherwise}\end{cases} \tag{2.30}$$

以上式子的模型是只要获得 i 时刻的观察值 b（x，i）的真实特征，与对应的标注结果相结合，便能得到模型的特征函数集。如果前一个状态和当前状态（转移函数）或当前状态（状态函数）值是已经规定好的，则所有的特征函数都将会是实数值。例如：

$$t_k(y_{i-1},y_i,x,i)=\begin{cases}b(x,i) & \text{If}\quad y_{i-1}=\text{LOC},y_i=\text{N}\\ o & \text{otherwise}\end{cases} \tag{2.31}$$

2.4 模型参数估计

CRF 模型的训练主要有两个阶段，第一阶段是建立相应的特征函数，第二阶段是估算模型的参数，也就是对数据的权重进行估计。CRF 的参数估计主要有两种方法：贝叶斯估计（Bayes estimation）和极大似然估计（maximum likelihood estimated，MLE）。

其中 MLE 的方法应用地更加广泛，这里将介绍 MLE 估计 CRF 模型。

假设给定的训练集 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ 上，训练集上的数据与联合经验分布 $\tilde{p}(x,y)$ 同分布且是独立分布的。则利用条件模型 p（y | x，λ)得到训练数据的相似度函数为：

$$L(\lambda)=\prod_{x,y}p(x\mid y)^{\tilde{p}(x|y)} \tag{2.32}$$

MLE 就是利用似然估计的方法求得使 $L(\lambda)$ 最大时的参数 λ 的值，使得模型分布接近于经验分布，从而可以得到条件概率 $p(y\mid x,\lambda)$ 的对数似然函数的形式：

$$L(\lambda)=\log\prod_{x,y}p(x\mid y)^{\tilde{p}(x|y)}=\sum_{x,y}\tilde{p}(x\mid y)\log p(x\mid y) \tag{2.33}$$

其中设定归一化因子是：

$$Z(x)=\exp[\sum_{j}\lambda_j f_j(y(k),x(k)] \tag{2.34}$$

已知条件概率 $p(y\mid x,\lambda)$ 公式为：

$$p(y\mid x,\lambda)=\frac{1}{z(x)}\exp[\sum_{j}\lambda_j f_j(y(k),x(k)] \tag{2.35}$$

对于 CRF 模型，所有的训练数据都是独立的，因此训练集上所有序列的 $p(x^{(k)}\mid y^{(k)},\lambda)$ 的乘积就是相似度，对于使得 $L(\lambda)$ 最大时的参数值时 $L(\lambda)$ 的对数形式可以表示为：

$$L(\lambda)=\sum_{k}[\log\frac{1}{z(x^{(k)})}+\sum_{j}\lambda_j f_j(y(k),x(k)]$$

此函数是个凹函数，因此 $L(\lambda)$ 会存在会收敛于全局的最大值。这里可以先求出经验概率和模型的得到的概率的序列期望：

$$E_{\tilde{p}}[f_k]=\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n}fk(y_i,x_{,i})=E_{\tilde{p}}\sum_{i=1}^{n}f_k(y_i,x_{,i}) \tag{2.36}$$

$$E_p[f_k]=\sum_{x,y}\tilde{p}(x,y)p(y\mid x,\lambda)\sum_{i=1}^{n+1}f_k(y_i,x_{,i})=E_p\,p(y\mid x,\lambda)\sum_{i=1}^{n}f_k(y_i,x_{,i}) \tag{2.37}$$

然后对 $L(\lambda)$ 求得 λ_j 一阶偏导数。

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = \sum_j \tilde{p}(x,y)[f_j(y(k),x(k)] - \sum_{x,y} \tilde{p}(x,y) \frac{\exp[\sum_j \lambda_j f_j(y(k),x(k)]}{z(x)} \sum_{i=1}^{n+1} f_k(y_i, x_{,i})$$

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = E_{\tilde{p}}[f_k] - E_p[f_k] \tag{2.38}$$

通过梯度为零就可以将这个一阶方程取零也就可以得到最大熵的约束条件，其实 CRF 模型也是符合最大熵的原理的。但是极大似然估计求解必须采用一种迭代的梯度的方法，利用常规的方法求解一般是得不到一个封闭的解，因而。本次试验中的模型是使用的 L-BFGS 算法，他是一种近似的二阶算法。此种方法相比如最大熵时所采用的训练模型，将要快很多。它主要是通过引入罚函数来解决这个问题，这样原问题，就变为:

$$L(\lambda) = L(\lambda) - \frac{\sum_j \lambda_j^2}{2\sigma^2} + const \tag{2.39}$$

其一阶导数可以变为：

$$\frac{\partial L(\lambda)}{\partial \lambda_j} = \frac{\partial L(\lambda)}{\partial \lambda_j} - \frac{\lambda_j}{\sigma^2} \tag{2.40}$$

数据稀疏现象指的是基于统计技术的知识获取的方法中一般方法中，在语料库的规模比较小的时候，大多数的词或邻接词以及各个属性的搭配在语料中会出现很少或者是不出现的现象，这样就会造成特征值为空的现象，对于计算起到很大负面作用。而上式中可以发现原来的一阶导数等式多了一个因子：$\frac{\lambda_j}{\sigma^2}$，对于数据稀疏的现象的控制，主要是通过上面的因子 σ 起到一种平滑作用。

2.5 本章小结

本章首先对两种常用的统计模型进行了介绍：HMM 模型和最大熵模型，他们是常用的两种命名实体识别模型。这两种模型的是基于条件的模型，有各自的优缺点，是条件随机场模型的理论基础，接着对于 CRF 模型的理论进行了详细的介绍，对于下一章更好的实现条件随机场中的关键问题打下坚实的理论基础。

第三章　命名实体的识别的思路和方法

前面一章详细的介绍了条件随机场模型的原理。本文的研究任务主要是集中在机构名，人名和地名的研究，而对其他的命名实体将不做深入研究。因此本章的即针对以上实体的识别方法展开探究，并为进一步的分析出本论文研究中的关键问题提供了解决方法和思路。

3.1 命名实体识别的难点

命名实体识别就是判断一个字符串是否代表着一个命名实体，并确定其类别。但是在实际环境中随着语境的不同，可能会出现大量不同的意思，使得这些实体出现不可辨别的情况，这也就导致命名实体的确切含义，只能根据具体的应用来确定。例如：一串数字在不同的地方，可能代表的是汽车牌号，邮编，电话号码。

总而言之，文本中时常出现的大量命名实体，而信息抽取想要从文本中得到事件或关系信息，其中就必然涉及到事件角色的对象，而这些对象大部分是命名实体。如果想要从这些文本中提出命名实体识别，无论是英语还是汉语都不是一件简单的事，其难点可以概括于以下几点：

（1）命名实体是一个很大的开放性的类，不可能用穷举的方法收集。例如：全世界人的姓名，想把它穷举出来是不可能的。如果想编写一步完整的字典显然是不可能的，而且随着信息社会的到来，新词将会层出不穷，这就注定了字典或列表只能以辅助的形式出现在命名实体识别的过程中。

（2）世界各地区都没有对于命名实体有严格的命名规范，也就导致了在不同的地方，不同文化中的实体是千差万别。

（3）不同的领域和场景下，命名实体的外延有差异，外在表达形式也是各种各样。

（4）在许多文档中因为自身的风格，领域不同，使得电子文档差异性增加，也加大了识别难度，而文档中的标点，空格，文字编码等不同，同样也会对实体识别产生影响。

总而言之，以上这些是所有实体识别必须都要面对的问题，而不同语言之间的命

名实体识别的难度也不一样。其中，在中文命名实体识别方面比英文识别方面的难度显然就要大很多。例如英文单词之间是以空格隔开、专有名词的开头一般都回大写，所以对于英文实体识别的主要任务可以概括为两个：（1）识别实体短语，（2）对实体进行分类，用来识别这个实体表示的是人名、地名、机构名还是其它实体。但如果语种是汉语，情况就会复杂很多，本文总结中文命名实体识别的实体主要有以下困难：

（1）缺乏明显的特征标志。在中文中命名实体之间是没有明确的区分的特征标志，例如：在英文的实体中一般情况下会大写。而且中文不像英文以空格作为划分词的标志，这就给分词的时候也会带了影响，造成标记错误，这样使得后期实体识别不准确。这样就需要提取更多的上下文的特征，同时加大了命名实体识别的难度；

（2）汉语自身相比英语组织非常复杂，没有设定严格的规定。例如，在有些组织机构名中的长度可能比较短，只有 2 个词，但是有些确有 8 个词之多，而且并没有可以遵循的规律。

（3）表达形式的多样性，中文博大精深，语义场景性非常强，这一点在机构名的识别中表现得特别明显。例如，在今天的院足球比赛中，计算机以 2:1 的比分战胜了法学。从这个句子中，可以发现“计算机”和“法学”都不是机构名，但是从语义上分析，这个两个词分别代表着是“计算机学院”和“法学院”的意思，应该作为机构名被识别处理。而在另外的场景下，例如，在大学生选择专业时，计算机和法学都是现在的非常热门的专业。在这里，计算机和法学是作为两个学科的名字，可见在中文命名实体识别中，对于语义和上下文的分析是非常重要的。

（4）中文中常常出现一些缩略语的形式，也加大了实体识别的难度。例如，“太原理工大学”的缩写形式为“太理工”。

（5）中文中实体互相嵌套的情况也是经常遇到的。例如，“巴拿马总督米雷娅—莫斯科索”，其中“米雷娅—莫斯科索”正确的识别应该是人名实体，但是“莫斯科”这个地名，嵌套在人名当中，这样就会很可能导致系统判别时出现错误，这样也就给实体识别带来了困难。

（6）普通名词和实体名之间会出现歧义性。例如，“高峰”既可以作为普通词，也可以作为人名。

（7）由于中文命名实体研究的比较晚，使得现在能够应用于命名实体识别的大规

模标注语料还很少，因此，一方面对于应用于中文命名实体识别的大型标注语料库显得非常急切，同时，对于不依赖于大型标注语料库的命名实体识别的算法研究也就具有更加重要的意义。

（8）一些专有名词出现过一次以后，有些以后将用简写的格式出现，在英文中出现的简写，一般会以大写字母的形式出现，而在中文中就不会有特别明显的标志性特征，由于命名实体包含了文本中重要的信息，命名实体识别是信息抽取研究中最有意义的研究内容之一。另外，文本中频繁出现的命名实体，也是制约分词精度提高的最主要原因。其识别的好坏将直接影响分词精度以及其后的词性标注和句法分析的精度，命名实体作为信息抽取中最关键，也是最基础的一环，有着相当重要的地位。

3.2 命名实体识别的主要方法

目前，主要有三种比较优秀的命名实体识别方法：基于规则的方法、基于统计的方法、规则和统计相结合的方法。

一般来说，基于规则的方法在某一次的特定测试中，要优于基于统计的方法的结果 2%左右,这主要是因为规则的制定大部分是本领域的专家所提供，对于各自领域的实体规则研究得还算比较完备，但是当语料的来源是不同的交叉学科的时候，由于各个学科对命名实体的认识是不一样，这样在再次移植到其他的语料中时，往往效果会不尽人意，这也就是用规则方法进行命名实体识别时移植性差的缺点。而基于统计的方法只需要少部分人工标注的训练语料进行学习，在标注语料时并不一定要广博的语言学知识，由于语料小，因此不需要耗费大量的人力和物力，而大部分知识的获取是通过机器完成，客观性比较强，但是对于上下文的特征的发现显然不如规则那样准确，可能出现偏差。综合分析以上两种方法的优缺点，将两种方法结合起来，在规则中加入统计的方法，或者在统计中制定适当的规则，这类方法即规则与统计相结合的方法。这种方法在使用大量的语料进行训练和学习的时候，尽可能多地通过利用语言和规则方面的知识，从而为有效进行命名实体识别研究提供进一步的支持。

基于规则的命名实体识别主要是通过规则和词典的方法相结合来进行识别，一般习惯把常用的人名、地名、组织机构名等专有名词作为基础加入词典，通过规则的方

法对于词典中没有出现过的名词进行识别和标记。以下是一个简单的命名实体规则，如：

组织名→{[人名][组织名][地名][核心名]}*[组织类型]<指示词>

人名→<姓氏><名字>

地名→<名字部分>*<指示词>

在英文中实体本身的语言特征有些会比较的明显，可以结合上下文，发现这些实体往往包含固定的特征。例如，在组织名中实体经常以 INC 和 COMMITTEE 做结尾。

一个规则并不能一次性写好的，很多时候甚至依赖规则制定者的灵感，经常需要反复的调试。常用的做法是利用一个测试语料做规则检查，体标记出来其中不能正确识别的命名实体，然后分析原因，再修改规则，如此不断反复，直到最后获得一个满意的准确率。

总而言之，基于规则的命名实体识别的方法，主要是通过人工的方法分析了实体的外部和内部特征，编写了相应的模板和规则进行命名实体的识别。在特定的环境下基于规则的方法，对于特定的领域的测试的效果较好。但规则的制定在于制定规则的人对该领域的了解程度，这对规则制定人的专业知识的要求很高。而且当该系统应用于新的领域时很可能出现新的问题，这时可能又得重新制定规则，这样做的适用性不是很好，而且浪费大量的金钱和人力。另外，不同领域的规则一起应用时，很有可能引起规则之间的冲突。

由于规则获取实体的方法自身瓶颈的原因，越来越多的人开始关注基于统计的方法了。基于统计的方法只是利用了很小的一部分语料，对于大规模语料库依赖性不强，而较小的语料则基本满足它的要求。总之，基于统计的统计方法的优点在于，对于语言的依赖性较小，可移植性能强。目前使用比较多的统计的方法有：支持向量机的方法（Support Vector Method，SVM）[33]、隐性马尔科夫模型、最大熵（Maximum Entropy ME）[34,35]、条件随机场（Conditional Random Field，CRF）[36]等。在CoNLL-2003 的[37]会议上，所有参赛的 16 个系统全部都是采用统计的方法，其中最大熵的模型有 5 个系统[38-42]，隐性马尔科夫模型有 4 个系统适用[43-46]，还有其他的一些相关模型在使用，但是可以发现这个以马尔科夫模型为基础的系统在逐渐成为主流。

3.3 命名实体识别方法结构设计

根据以上的结论，本文将采用以马尔科夫模型为基础的理论模型进行实验，这个统计模型需要实现的基本步骤主要包括：

（1） 统计模型：

在任何一个统计学习的方法中都会选取相应的模型作为统计学习的模型，这些先后应用于自然语言处理中的模型，例如，支持向量机的方法、隐性马尔科夫模型、最大熵、条件随机场等数学的模型在命名实体识别中，在不同的阶段都取得了不错的结果。

（2） 特征的提取

统计的模型大致可以认为是模式识别中分类问题的一种特例，而特征的提取作为模式识别中关键，有着重要的意义，例如：可以选取的特征一般可以选取词性，词位置值等。

（3） 标注

利用机器学习方法获取相应的统计模型并对测试语料进行统一标注，测试序列将会得到相对应的标注结果。

在此次的测试语料的训练数据中，本实验将实体标注为 3 种命名实体：person(PER),organizations(ORG),locations(LOC)。当一个命名实体中含有多个命名实体时，这里只会识别最外层的实体，而内嵌的命名实体将不会做考虑。本次试验中采用的标记方式是由 Ramshaw 和 Marcus 提出的 IOB 模式，即 B（begin，开始）、I（internal ，内部）和 O(other，其他)。其中，B-XXX 标注的是用于命名实体 XXX 词首的字，当有两个类型为 XXX 的实体直接相邻时，则用 B-XXX 来表示另一个实体的开始。而 I-XXX 标注的是用于命名实体 XXX 非词首的字。定义了 7 种标记的集合，记为 L={ B-ORG，I-ORG，B-PER，I-PER，B-LOC，I-LOC， O}。在这个集合中每个标记的意思分别为：机构名开始，机构名的内部，人名的开始，人名的内部，地名的开始，地名内部，其他。为了更好的演示，将在选取训练数据中的一段标注实例作为说明：

近 O/年 O/来 O/国 O/家 O/给 O/北 B-LOC/图 I-LOC/的 O/行 O/政 O/拨 O/款 O/还 O/是 O/逐 O/年 O/递 O/增 O/的 O/，O/

这里可以发现这种对于中文文本采取的院子切分的方法可以看出是以字为切分标准的，其中“/…/”表示在每个字必须是单独的占一行，这里所谓的字，必须有一个统一的标准。字，它不仅仅是传统意义上的汉字。它包括非常多的非汉字的字符，在这里都把他们统称为字，并且他们单独的作为一个切分单元独立作为一行，一般来说，汉字还是构成切分标准中最多的字。通过最已有的方法做的大量研究中，发现在中文处理中，以字作为切分的方法具有很多优势。

1、该方法在算法实现上更加简洁，性能更优。如：在命名实体识别系统中，对比计算基于词模型，基于句子模型时建模，在字一级建模的算法性能是最好的。

2、他能够平衡的对待未登录词和词表词的识别问题。语料中的未登录词和词表词都是由统一的字标注算法来实现的。没有必要特意的为未登录词设计新模块。一个概率模型的生成主要是利用字根据预定义的特征进行词位的特征学习，再进行标注，学习的过程。最后再根据字与字之间的结合的特征的相似度，得到一个词的标注结果。在整个过程中，是不需要考虑未登录词和词表词的因素，期间的一切处理过程都是在字一级上完成的。

3、之前做的大多数系统都是在词一级，对字一级的命名实体研究还不多。而字一级的研究不会出现由于分词不当而造成歧义现象，如果在词一级的处理不当，对于后期的命名实体的识别会起到很大的干扰作用，从而造成实体识别的准确率下降。

将语料库按照上述的标注体系的定义，以字的形式进行标记，生成所需要的的训练文件。

（4） 训练模型

利用机器学习到的统计模型对测试语料进行标注，得到测试序列对应的标注结果。在获得特征参数后，这里需要对数据进行 CRF 模式的学习。模型训练流程的步骤如下：

1、模型利用动态规划的算法获取数据中的特征函数的模型期望和经验期望。

2、由模型的平滑因子σ，特征的模型期望以及经验期望，可以获得$\frac{\partial L(\lambda)}{\partial \lambda_j}$，通过 L-BFGS 算法模块进行训练并且获得修正后的参数λ。如果迭代中止的条件被满足则退出模式学习，否则转到（1）。

模型的输入，输出如下：

输入：特征函数集 F，平滑因子σ；

输出：特征函数集 F 及其参数集合；

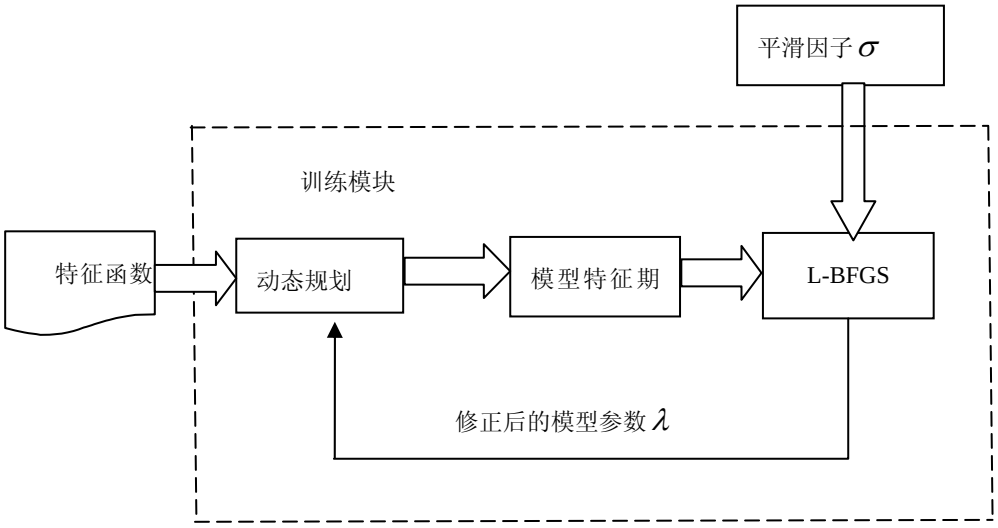

图 3.3CRF 模式学习示意图

（5） 后处理

测试后的序列进行经后续的处理，这样就可得到命名实体识别的具体结果。

在这些统计方法中，马尔科夫模型是基础的系统，现阶段来说取得了最佳的成果。这种基于语言模型的实体识别，都是先利用内部信息得到候选实体，再通过迭代计算的方法产生概率，并利用解码算法得到最优的路径，而且它将分词过程与命名实体识别过程合并进行，基于单字的识别，能有效避免由于分词错误而产生的专名识别错误。

3.4 本章小结

本章节简述了中文命名实体识别时的难点，并就研究领域比较主流的解决方法做了介绍，并提出自己对于本模型的设计思想，最后结合了 L-BFGS 的算法对 CRF 模型的模式学习流程的方式进行了解释。

基于以上对命名实体识别方法的设计，希望最终能够达成以下目标：以 CRF 模型

为实验模型和理论向导，在现有的语料库的基础上，针对 CRF 中特征函数集的生成以及模型训练，进行测试和调优。为实现这些任务所需做的主要工作如下：

1、对中文命名实体识别的各种方法进行理论学习，对比找出最优或较优的方法进行深入分析和学习，为今后进行进一步中文命名实体识别打好理论基础。

2 由于条件随机场是目前自然语言处理领域认为较优的一种机器学习的方法，本研究主要将利用条件随机场模型解决中文命名实体识别任务中存在的困难。在本研究中，将构建一个基于条件随机场模型的命名实体识别系统，对系统从训练到测试等各方面性能进行测试和分析，按照 CRF 模型原理以及现在主流命名实体识别的方法，提出了本实验将会进行的实验方法和思路，最后完成一个相对完整的中文命名实体识别的条件随机场模型系统设计步骤。

第四章 CRF 的命名实体识别模型关键问题的解决方案

条件随机场模型（Conditional Random Fields，CRF）是一种优秀的统计机器学习的方法，在自然语言处理领域被广泛的应用，他在序列标注，分词和命名实体的识别取得了很好的了效果，本章将就基于 CRF 的命名实体识别的模型实验原理进行阐述。

4.1 标记偏置问题

基于马尔科夫模型的统计中，最大熵模型是表现的比较优秀的一种。最大熵把一个状态链的条件转化为多个单状态链的乘积,其原理还是马尔科夫链。它同时结合了 ME 和 HMM 的优点，利用这个模型时，只需要给命名实体的选择最佳的特征值，而这些特征值的使用则不需要考虑，这样使得特征值的选择具有多样灵活性。但是又由于 MEMM 是基于 next-state 分类器进行特征的分类，它也就存在了一个标记偏置的问题，这个问题可以用一个简单的例子图来说明。

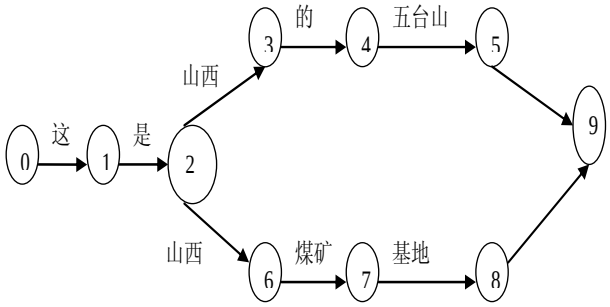

图 4.1　标记偏置实例

图4.1表示了两个个简单的固定的待标注的句子，他的具体表示两个独立的句子为：

（1） 这是山西的五台山。

（2） 这是山西煤矿基地。

从图中可以发现这里有两条不同的路径，假设的观察到的序列是“这是山西的五台山”，在 MEMM 作分析时，在第一步遇到的路径是 123,126 这是，这时从初始状态

到现阶段的路线都是匹配的，因而在第一步中 123 和 126 的传递概率相等。当继续向下分析，观察到了 4 和 7 分别作为 3 和 6 的唯一标记序列的后续词，但是 4 和 7 状态都只有一个输出连接，状态 4 在训练语料中经常出现，而状态 7 几乎很少出现。但是状态 4 和 7 没有其他的选择，只能把现在所有的状态都传递给后续的唯一路径，这样就导致了最终的输出序列将会是分叉处那个转移概率较大的路径的值，也就是第一条路径，从而忽略概率较小的那条路径的值，也就导致了和观察值序列无关，而训练语料中标记的分布概率起到了决定性的因素。最大熵马尔科夫模型等这种基于判别式的马尔科夫模型都可能因为忽略该词上下文的特征而出现标记偏置问题，只是机械将其最大的概率的标记作为最终的序列输出。因此，在最大熵马尔科夫模型中，标记偏置严重影响了最终识别的准确率。

为了解决标记偏置的问题，又保留 MEMM 等条件概率模型的优点。Lafferty 等人于 2001 年提出了条件随机场（Conditional Random Fields ，CRF）模型，模型的最主要的思想依旧是来源于 MEMM，但是 MEMM 对于给定了当前状态的被用作了下一状态的条件概率使用，这也是导致标记偏置的根本原因，而 CRF 模型是一个无向图，上一状态的概率没有作为下一状态的概率指数进行估算，而是对整个观察序列的联合概率进行联合概率的计算，这样的标记序列可以信赖观察序列中非独立，相互作用的特征，并且给不同的特征赋予不同的权值，从决定特征的重要度。

条件随机场是一种以给定输入序列为条件来预测输出序列的概率的无向图模型，当给定一组需要标记的观察序列的条件时，可以用它来预测一个待标记序列的联和概率分布。假定$X=\{x_1,x_2,x_3,\ldots x_T\}$表示被观察的序列的集合，$Y=\{y_1,y_2,y_3,\ldots y_T\}$表示被预测的标注状态集合。

一个无向图G=（V，E）其中V表示顶点集合，E表示边集合。接着把标记S用作顶点索引，即$Y=\{Y_v \mid v\in V\}$即V中的每个节点对应于一个随机变量所表示标记序列的变量Y_v。当每个随机变量Y_v都具有马尔科夫性的时候，那么（X，Y）称为一个马尔科夫条件随机场，随机变量S_V的概率为：

$$p(Y_v \mid X, Y_u, u\neq v, \{u,v\}\in V) \tag{4.1}$$

根据图论，图 G 的结构可以是任意的，他意味着标记序列中独立性的条件。因此，

可以用一个最普遍和最简单的一阶链结构作为例子，在该结构中的每个节点对应于 S 中的元素，如图所示：

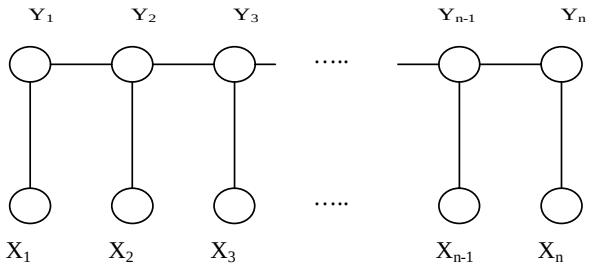

图 4.2 线链 CRF 的图形结构

在无向图G=(V,E)中，S集合中的随机变量是图G中的一部分，对于输入的观察序列O ，元素之间没有图的关系，这是因为在p（S_v|O）中只是将观察序列作为条件，并不对O做任何的独立假设，因此很好解决了标记偏置的问题。

4.2 特征的选择

在获取训练文本之后，接下的工作就是获取特征函数集。这部分的内容直接关系着 CRF 模型识别出来的实体正确率和召回率的关键，下文将是 CRF 模型中特征函数获取的方法。

特征函数是 CRF 的一个很重要的概念，首先根据上一章的定义从训练语料中获取特征 $f_j(y_{i-1}, y_i, x, i)$，然后再给这个特征赋一个权重，即模型参数 λ_i。

$f_j(y_{i-1}, y_i, x, i)$ 是转移特征函数和特征函数的统一形式表示，每个特征函数取值为一个观察特征 $b(x,i)$，并以后文的语料库中的特征实例“乡长邵华就”这半句进行举例说明，例如：

$$f_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x,i) & \text{If} \quad y_{i-1}=N, y_i=\text{B-PER} \\ o & \text{Otherwise} \end{cases} \tag{4.2}$$

$$b(x,i)=\begin{cases}1 & \text{If}\quad \text{i 的位置的观察值是}\\ 0 & \text{Otherwise}\end{cases} \tag{4.3}$$

当在i时刻获得观察值的特征$b(x,i)$ 的时候，只要再结合上文的标注结果，就能够获得一个特征函数的集合。而对于这个观察值“邵”，也不会局限于只是获得一个单元的信息量，这样得到特征显然会很稀疏。因此，首先会考虑将特征的窗口扩大到 2，即考虑前后两个单元的信息，那么对于 i 这个位置的字，将获得，i-1/i ，i/ i+1 的特征信息，这里的特征的取值就不会局限于单个的字，而且有利于获取更多的特征信息。那么它的这一时刻真实的特征函数形式为：$\mathrm{f}\left(\mathrm{y}_{\mathrm{i}-1}=N,\mathrm{y}_{\mathrm{i}}=B-PER,\mathrm{x}_{\text{-1}}=\text{“长”},\ i=3\right)$，表示成二值函数为：

$$\mathrm{f}_{\mathrm{j}}=\begin{cases}1 & \text{If}\quad \text{前一观察值为“长”，前一标注为“N”，当前标注为“B-PER”}\\ 0 & \text{Otherwise}\end{cases} \tag{4.4}$$

上述例子实例化的介绍了一种窗口为 2 时的一种简单的情况，这样可以将特征函数归为一个的直观表示的形式：$\left\{\mathrm{y}_{\mathrm{i}-1}=\mathrm{Label},\mathrm{y}_{\mathrm{i}}=\mathrm{Label},\mathrm{x}_{\mathrm{w}}=\text{“content”}\right\}$，其中$\mathrm{Label}\in\left\{\text{B-PER，I-PER，B-LOC,I-LOC,B-ORG,I-ORG,O}\right\}$，$\mathrm{y}_{\mathrm{i}}$表示 当前的标记，$\mathrm{y}_{\text{i-1}}$表示下一个标记，w 表示考察的位置，Content 表示 w 位置的内容。下文中将按上述实例化信息进行实验。

4.3 特征模板的定义

CRF 模型的最大优点就是不仅能综合的使用字、词等的上文信息，而且还能综合利用各种外部信息。特征模板的作用就是按照模板上的规定给 CRF 模型提供一个统一的特征函数的生成模式，模板是对上下文中的特定信息和信息的特定位置的综合考虑。在实验中，通过对模板的编辑和选择来调整获取的特征函数集合，因此，在本次实验中，由于特征模板将直接关系着特征函数的形成。因此，模板的选择需要多次调优，才能得到最佳的模板，本章将会首先尝试最简单的特征模板并得到一个基本的测试数据，然后再尝试不同的模板以求得到最佳的特征模板。

在模板中，把只考虑一种因素的简单模板叫做原子模板。针对下文的特征制定的

模板如下：

x[-2,0]，x[-1,0]，x[0,0]，x[1,0]，x[2,0]，中括号内的第一个数字表当前所选取的字符的位置，第二个数字表示考虑的上下文特征的种类。这里对 CRF 模板进行进一步的说明：

```
# Unigram
U02:%x[-2,0]
U03:%x[-1,0]
U04:%x[0,0]
U05:%x[1,0]
U06:%x[2,0]
U10:%x[-2,0]/%x[-1,0]
U11:%x[-1,0]/%x[0,0]
U12:%x[0,0]/%x[1,0]
# Bigram
B
```

模板的输入格式为%X[row,col],用于表示输入数据的一个确定的token位置，它可以包含很多列，token的定义可根据具体的任务，如词、词性等。一个token的序列可构成一个sentence，sentence之间用一个空行间隔其中，row确定与当前的token的相对行数。col用于确定特征的绝对列数。如下表：

表 4.1 例子中的上下文的标记

	col 0	col 2	
row -2	乡	N	
row -1	长	N	
row 0	邵	B-PER	当前行
row 1	华	I-PER	
row 2	就	N	

表 4.2 例子中的上下文的特征

模板	指代的特征
U01：%x[-2,0]	乡
U01：%x[-1,0]	长
U02：%x[0,0]	邵
U03：%x[1,0]	华
U04：%x[2,0]	就
U05:%x[-2,0]/%x[-1,0]	乡/长
U06:%x[-1,0]/%x[0,0]	长/邵
U07:%x[0,0]/%x[1,0]	邵/华

以上图表形象的描述了CRF模板的训练特征，其中以字母U开头，Unigram template。当模板前加上U之后，CRF生成一个特征函数集合，一个模型生成的特征函数的个数总数为L*N，其中L是输出的类别数，N是根据给定的template扩展出的独立串的数目。第二种特征模板以B开头，即Bigram template 。它用于描述Bigram特征。系统将自动产生当前输出token与前一个输出token的组合。产生的可区分的特征的总数是L*L*N，其中L是输出类别数，N是这个模板产生的unique feature数。

表 4.3 例子中的特征函数和状态函数

转移特征函数	状态特征函数
$\{y_{i-1}=O, y_i=\text{B-PER}, x_{-2}=\text{“乡”}\}$	$\{y_{i-1}=\#, y_i=\text{B-PER}, x_{-2}=\text{“乡”}\}$
$\{y_{i-1}=O, y_i=\text{B-PER}, x_{-1}=\text{“长”}\}$	$\{y_{i-1}=\#, y_i=\text{B-PER}, x_{-1}=\text{“长”}\}$
$\{y_{i-1}=O, y_i=\text{B-PER}, x_{0}=\text{“邵”}\}$	$\{y_{i-1}=\#, y_i=\text{B-PER}, x_{0}=\text{“邵”}\}$
$\{y_{i-1}=O, y_i=\text{B-PER}, x_{1}=\text{“华”}\}$	$\{y_{i-1}=\#, y_i=\text{B-PER}, x_{1}=\text{“华”}\}$
$\{y_{i-1}=O, y_i=\text{B-PER}, x_{2}=\text{“就”}\}$	$\{y_{i-1}=\#, y_i=\text{B-PER}, x_{2}=\text{“就”}\}$

从表4.3的信息中，可以轻松的得到上下文特征的信息，上下文的特征也称为字面特征，主要考虑的是输入序列本身的观察值，根据上面的例子的标注，可以的得到一个上下文的示例。而本文将根据这些上下文的信息，自行的设计模板，通过测试不同模板对上文的信息获取情况，选择在系统性能和算法结果上最优化的模板，进行测试。

4.4 文本预处理

CRF 的训练数据和测试数据有严格的要求，首先它是由很多块组成的，每个块是一个句子，块与块之家必须有空行隔开，其次，训练数据和测试数据要求必须每个字是一个行，这同时也决定了模板是基于字符的，而不是基于句子或词的。这样，在对模板切分时，可以避免出现词级别的切分错误。

4.5 本章小结

本章介绍了条件随机场的图形结构，它是一个用于标记和分割数据的无向图模型，解决了标记偏置的问题。接着本文详细描述了本实验对文本特征函数的表示、分类和选择，提出了对于特征函数模板的设计思想，本文将会自行设计模板，通过测试不同模板对上文的信息获取情况，选择在系统性能和算法结果上最优化的模板。并且指明了文本预处理的方式等这些在命名实体识别过程中需要解决的关键性问题的解决方案。

第五章 实验过程及结果分析

本章详细阐述了基于条件随机场的模型中文命名识别系统的实现。通过对不同的模板的比较，这里选择了最合适的模板作为 CRF 模型的，用来提高命名实体识别时的准确率，召回率和 F 值，然后再加入词性标记属性，规则进一步对中文命名实体的性能做进一步的实验。

5.1CRF 模型实体识别流程

这个系统是建立在第 3 章所说的 CRF 模型的基础上的，本文通过对具有不同的属性特征的 CRF 模型进行测试，生成不同的 CRF 测设模型，然后再把同一个测试数据带入到模型中，最后得到命名实体识别的情况，具体的流程图如下所示：

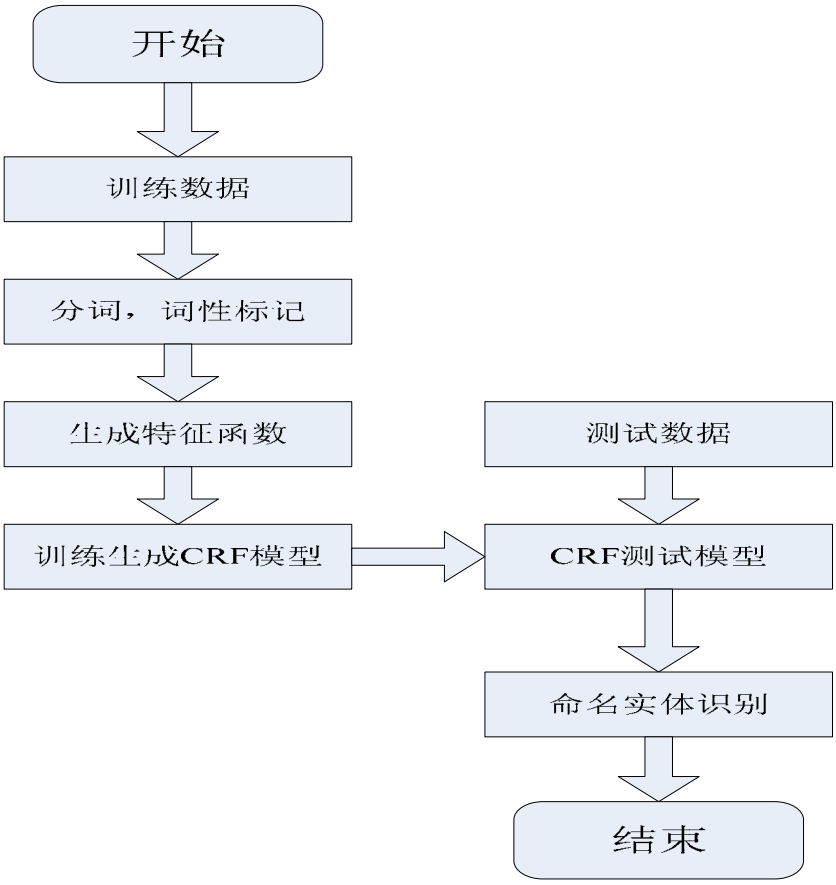

图 5.1 实验流程图

5.2 CRF 实验过程及结果

条件随机场模型的训练和测试这里使用了 CRF++的开源工具。CRF++实现了 CRFs 的模型，并且很好了封装了各种算法，此工具在序列数据的分割以及标注方面具有很强的实用性，在自然语言处理的很多方面都取得不错的成果，如信息抽取，命名实体识别，文本分类等方面。

CRF++的训练一个特征模板文件，训练数据文件，测试数据文件，文本格式需按照上文所介绍的格式生成。本文将会采用 SIGHAN2006 MSRA 语料库作为训练和测试语料，该语料库的训练语料共有 172070 个字符，在测试预料中共有 172601 个字符。

5.2.1 实验环境

下面列举将是本实验进行时的软件环境和硬件环境。

硬件环境： 电脑 CPU 3.0GHz 双核

物理内存 3G

虚拟内存 6G（CRF++主要是占用的虚拟内存）

软件环境： 测试语料为 SIGHAN2006 MSRA 语料库。

操作系统 WindowsXP SP3

CRF 模型软件 CRF++5.3

系统初始运行占用内存 653M

5.2.2 不同模板的实验结果分析

通过简单的模板对 CRF 的识别出来的命名实体进行测试，其中采用的评测工具是 CoNLL-2003 所用的 conlleval.pl。

```
# Unigram
U03:%x[-1,0]
U04:%x[0,0]
U05:%x[1,0]
```

U09:%x[-1,0]/%x[0,0]

U10:%x[0,0]/%x[1,0]

Bigram

B

测试结果为：

表 5.1 简单模板测试命名实体实验结果

命名实体	精确度	召回率	F 值
人名	89.76%	79.83%	84.51
机构名	80.03%	68.34%	73.73
地名	91.31%	71.20%	80.01

此次测试所用的训练时间为 1000s 的数量级，从模板中，可以发现 template 使用了当前字符和前后的各一个字符共 3 个字即：%x[-1,0]，%x[0,0]，%x[-1,0]，一共包含了 3 个单字特征，同时对应的二元上下文特征是：x[-1,0]/%x[0,0]，当前的字符和前一个字符组合的二元上下文特征，%x[0,0]/%x[1,0]表示当前的字符和后一个字符组合的二元上下文的特征。从测试的结果可以发现 LOC,PER 的测试值已经达到了 80%，而 ORG 的测试值为 73.73%理想的测试值的差距较大，为了提高测试的数据精确度，扩大了字的窗口的长度，以便获得跟多的特征值，以下便是引入的新的模板和测试数据，当字的窗口长度扩展到 2 和 3 时，他的测试的数量级将会大幅增加。

Unigram

U02:%x[-2,0]

U03:%x[-1,0]

U04:%x[0,0]

U05:%x[1,0]

U06:%x[2,0]

U10:%x[-2,0]/%x[-1,0]

U11:%x[-1,0]/%x[0,0]

U12:%x[0,0]/%x[1,0]

U13:%x[1,0]/%x[2,0]

Bigram

B

字的窗口为 2 测试结果为：

表 5.2 1000s 数量级测试命名实体实验结果

命名实体	精确度	召回率	F 值
人名	91.18%	81.49%	86.06
机构名	84.35%	79.07%	81.63
地名	94.37%	74.22%	83.09

通过实验的数据，测试时间的数量级为 10000s，训练时间大幅增加，达到了 10 倍以上，这主要是因为这里增加了 2 个单字的一元特征函数和 2 个二元特征函数，但是可以测试的结果是可以接受的，LOC，ORG 和 PER 的测试指标都有不同程度的上升。因此，继续修改模板，继续扩大字的窗口的长度，使得字的窗口长度为 3：

```
# Unigram
U01:%x[-3,0]
U02:%x[-2,0]
U03:%x[-1,0]
U04:%x[0,0]
U05:%x[1,0]
U06:%x[2,0]
U07:%x[3,0]
U08:%x[-2,0]/%x[-1,0]
U09:%x[-1,0]/%x[0,0]
U10:%x[0,0]/%x[1,0]
U11:%x[1,0]/%x[2,0]
 # Bigram
 B
```

测试结果如下：

表 5.3 10000s 数量级测试命名实体实验结果

命名实体	精确度	召回率	F 值
人名	91.37%	80.18%	86.06
机构名	84.64%	75.72%	79.68
地名	94.39%	73.85%	82.86

测试的数据的时间达到了 100000s，训练的时间非常的长，我在这里只加入了字的一元特征函数的获取，并没有增加二元特征函数的获取。这里发现测试数据的准确率反而下降了，这里可以发现一味的增加特征函数的数量会使系统承受更大的负荷，而测试数据的准确度并不会因为这个特征函数数量的增加而上升，反而出现了一定程度的下降，可见，只有选取适当的窗口长度才能得到最佳的测试结果，因此，在本次后面的测试中均会选取单字窗口的长度为 2。而在多元特征方面，这里将会引入下面的三元特征的共现模板，来进行特征，因此，只需要特征为 2 的模板上加入：

```
# Unigram
U01:%x[-3,0]
U02:%x[-2,0]
U03:%x[-1,0]
U04:%x[0,0]
U05:%x[1,0]
U06:%x[2,0]
U07:%x[3,0]
U10:%x[-2,0]/%x[-1,0]
U11:%x[-1,0]/%x[0,0]
U12:%x[0,0]/%x[1,0]
U13:%x[1,0]/%x[2,0]
U14:%x[-1,0]/%x[0,0]/%x[1,0]
```

所得的测试结果如下：

表 5.4 100000s 数量级测试命名实体实验结果

命名实体	精确度	召回率	F 值
人名	91.37%;	79.62%	85.10
地名	76.53%	78.84%	77.67
机构名	94.52%	72.45%	82.02

同样测试的结果的准确率并不会增加，此特征的加入只是引入了冗余的信息，反而影响了效果，因此不再试图增加更过的三元共现特征。

5.2.3 加入词性标记实现的结果分析

通过之前的实验，已经看到了字的特征遭遇了性能瓶颈，需要突破原始数据文件受限的信息框架，如果只是从字的角度着眼，已经不够，为了获取更多的非冗余的特征信息，而文本中的词性的前后依赖对于提取实体也具有启发作用。因此，这里选择词性来作为观察序列的外部特征。从语言学的角度来看，实体是专有名词，它在一个句子中承当的语法角色也相对是比较固定的，词性在一定程度含有实体相对固定的意义。在很多情况下，人们可以在对实体没有具体的认知的时候，人们常常根据在这句子中的词性，可以大致的判断是否这个词汇可以作为命名实体。这里利用了分词的特征和、、中科院的分词器 ICTCLAS30 来帮助提供分词和词性特征，由模板提供的分词后的特征如下表：

表 5.5　加入词性的上下文标记

	col 0	col 1	col 2	
row -2	党	B-nt	B-ORG	
row -1	中	I-nt	I-ORG	
row 0	央	I-nt	I-ORG	当前行
row 1	机	B-n	N	
row 2	关	I-n	N	

而人名，地名，机构名使用的词性作为外部特征制定的原子模板如下：

x[-2,1]、x[-1,1]、x[0,1]、x[1,1]、x[2,1]。括号内第二个数字 1 表示单字情况下的外部特征。括号内第二个数字是 2 时表示的是 2-Gram 情况的外部特征，因此得到的新的数据文件模板的格式如下例所示：

```
# Unigram
U01:%x[-2,0]
U02:%x[-1,0]
U03:%x[0,0]
U04:%x[1,0]
U05:%x[2,0]
U06:%x[-2,0]/%x[-1,0]
U07:%x[-1,0]/%x[0,0]
```

U08:%x[0,0]/%x[1,0]
U09:%x[1,0]/%x[2,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]
U20:%x[-1,0]/%x[-1,1]
U21:%x[0,0]/%x[0,1]
U22:%x[-1,0]/%x[-1,1]
Bigram
B

测试的结果是：

表 5.6 加入词性后测试命名实体实验结果

命名实体	精确度	召回率	F 值
人名	92.92%;	88.52%	90.67
机构名	93.44%	83.85%	88.38
地名	92.33%	92.09%	92.21

从以上系统的实验结果，在加入词性的分析以后，词性作为一种上文特征，由于它是基于语言的特性来标注的，在整体上对于模型获取更多的语言的内在规则是非常有用的。因此，这里可以发现实现结果都有不同程度的上升，可以发现 ORG 的 F 超过了 85%,，LOC 和 PER 的测试值比较的好分别到了 90%。

5.3 实验分析

在本次实验中，这里就实验中出现的错误进行了统计和分析，通过对不同的错误信息进行总结，可以得出了以下几点原因：

（1）由于分词的粒度过大或错误分词造成的负面影响，将词作为特征的根基不如直接以字来得好。因此实验中期又切换回到以字为根基的特征模版上。词性的加入使得实验结果的各项指标都得到了显著的提高，将分词的信息作为辅助特征也在一定程度要提高了指标，但是由于 ICTCLAS30 自身在分词方面也不能做到百分百的准确，所以对于后期的实验也会造成影响。

（2）机构名的识别效果差有很大原因是其复合性，一个复合机构名常嵌套着简单的人名和地名实体。可以考虑将这种嵌套关系作为一种特征，以引导标注器对复合机构名进行识别。而另一种在机构命中的识别中出现了大量的常规的名称和动词，这些都会降低识别准确性。如：“中 B-ORG/国 I-ORG /政 I-ORG /府 I-ORG /代 I-ORG /表 I-ORG /团 I-ORG/”被识别为“中 B-LOC/国 I-ORG /政 N/府 N/代-ORG/表 I-ORG/团 I-ORG ”,这种类似的嵌套输入造成了机构名识别的准确率不高。

（3）对于地名和人名的识别中有些词，在上下文特征不明显时很难被正确识别，如“韩”,“美”在很多地方就是被错误的是识别，或者是漏掉了。

（4）通过对不同的模板的测试可以发现对于不同模板的引入，系统的性能影响很大，在今后进行开放式的测试时，很有可能系统由于超过负载，对于平台的建设会造成极大性能影响，因此，这里通过不同实验选取了一个性能和结果都较优的特征模板作为实验模板是非常有必要的。

综上所述，这里发现对于命名实体识别，其中 ORG 的召回率还是很低的，这是阻碍笔者进一步提升 F 值的一个瓶颈。通过引入外部特征，来提高命名实体识别的 F 值，如加入词表，但是由于语料和时间有限，没有加入这方面的东西，在以后的工作研究中，可以针对以上出现的错误原因，做进一步更深入的研究。

5.4 本章小结

本章通过对比实验中，展示了 CRF 不同模板下的系统的召回率，准确率及系统的

F 值的具体数据，自行设计了模板并对测试出来的数据的性能和测试结果进行了对比，这里得到了一个在系统性能和识别结果都相对优化的模板，并提出了在语料中加入了词性的分析，词性作为语言独有的一种特征，对于实体识别是非常有帮助的，最后再进行了分析，得到了一个相对满意的值，但是对于机构名的识别仍没有达到最好的程度，因此提出了需要改进的地方，在今后的工作中将会就现阶段出现的问题做进一步的加强。

结 语

命名实体的识别在自然语言处理的各个领域有着广泛的应用，是对文本进行处理的基础性工作。本文就如何利用条件随机场对命名实体进行识别，提出了条件随机场模型中加入词性进行机器学习的训练方法，实现了一个对地名、人名和组织机构名识别任务的基于条件随机场模型的实验系统。本文在实现该系统的条件下，主要完成了以下几点工作：

（1）采用 CRF 模型对 MSRA 的语料测试。条件随机场作为一种相对优秀的机器学习的方法，在实体识别方面取得不错的效果。本文首先从理论方面着手，对 CRF 模型的模型推导、训练算法、标注方式等进行了详细的阐述，为今后进一步进行扩展研究打下坚实的理论基础。

（2）在语料库上通过 ICTCLAS30 进行词性切分，提出了将词性以特征列的形式加入特征模板中，词性作为语言的一种独有的特征，在一定程度上可以使得命名实体识别时能够获取更多的有效的特征值，发挥条件随机场获取上下文信息强的优势，对于提高实验效果是非常有用。

（3）在训练中，本文通过对不同大小的特征窗口进行实验，对比在不同的特征窗口的时间所需要的时间和正确率，这样在相对能够接受的时间内，选择一种相对优秀的特征模板进行实验，在耗费相同计算成本的条件下，获取最优化的系统性能，同时也为今后部署在监测平台上得到最优化的性能做准备，并论证了在封闭的测试环境下，部署在监测平台上的可行性。

本文虽然取得以上阶段性的研究成果，也基本上完成了预期的目标，但受研究时间所限，本文还是存在着不足之处。在以后的研究中，还有以下几个方面需要进一步的完善：

（1）为了得到更多的外部特征，在特征选取方面加入词典，以求得到更多的有效特征。

（2）在后续处理中，尝试着加入一些有用的规则，用规则和统计相结合的方法，以达到提高命名实体结果。

（3）本文仅是对 CRF 进行命名实体方法的研究，没有应用与字符规模很大的语料库进行研究，而且最后得到的 F 值，并没有达到 90%。

在以后的工作中，选取更多的外部特征和模板进行研究，在在后续处理中加入一定规则，以提高测试的 F 值。由于本实验是基于大型网络科技监测项目的研究型课题，面对的是大型的非结构化的网页数据，因此在开放性测试的要求很高，所以本实验今后努力将其运用到开放性的实验环境中去。

参考文献

[1] COMMUNICATION FROM THE COMMISSION TO THE COUNCIL , THE EUROPEANPARLIAMENT,THE ECOMOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OFTHE REGIONS-Towards a European research area [EB/OL].[2009-05-24]. http://cordis.europa.eu/documents/documentlibrary/C001190EN'.pdf

[2] April Kontostathis et al. A Survey of Emerging Trend Detection in Textual Data Mining. A Comprehensive Survey of Text Mining[M].Springer-Verlag,2003.pp 1-44

[3] S. A. Morris, B. Van der Veer Martens. Mapping research specialties. Annual Review of Information Science and Technology[J], Volume 42: 213-295

[4] Workshop on High-level Information Extraction[EB/OL].[2010-8-30] http://en.wikipedia.org/wiki/Information_extraction

[5] 赵健. 条件概率模型研究及其在中文名实体识别中的应用[D]. 哈尔滨工业大学..2006:3~5

[6] G.. R. Krupka, K. Hausman. IsoQuest Inc. Description of the NetOwl (TM) Extractor System as Used for MUC-7[C]. In Proceedings of the Seventh Message Understanding Conference.1998

[7] 李保利. 信息抽取研究综述[J]. 计算机工程与应用.2003,4:1~5

[8] R. Grishman, B. Sundheim, Message Understanding Conference-6: A Brief History[C]. In Proceedings of the 16h International Conference on Computational Linguistics (COLING-96). 1996, 8

[9] N. Chinchor, E. Marsh, MUC-7 Information Extraction Task Definition(version 5.1)[C].In Proceedings of the Seventh Message Understanding Conference. 1998

[10] 庄明. 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, 1

[11] ACE[EB/OL]: http://www.nist.org. speech/tests/ace/

[12] 863 命名实体识别评测组. 2004 年命名实体评测大纲[S]. http://www.863data.com.cn

[13] A. Borthwick. Maximum Entropy Approach to Named Eniity Recognition[J]. PhD Dissertation, NewYorkUniversity.1999:18-25.

[14] Kiyotaka Uehimoto，QingMa，etal. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules[C]. In Proceedings of the IREX workshop，1999.

[15] Roth, D. Learning To Resolve Natural Language Ambiguities: A Unified Approach [M]. Menlo Park，California: AAAI Press.1998:806-813.

[16] McCallum. A, Freitag.D, Pereira.F. Maximum Entropy Makov Models For Information Extraction And Segmentation[C]. Stanford, California. 2000.

[17] Veselin Stoyanov, Claire Cardie, Ellen Riloff.Nathan Gilbert. In: Proceedings of the 47 th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2008(8): 656-664.

[18] OntoText . [EB/OL].[2010-5-30]. http://www.ontotext.com/ .

[19] Shiren Ye, Tatseng Chua, Liu Jimin. An Agent-based Approach to Chinese Named Entity Recognition[J]. Coing, Taipei, 203:1149~1155.

[20] Shihong Yu, Shuanhu Bai and Paul Wu. Description of the Kent Ridge Digital Lbas System Used for MUC-7. Proceedings of the Seventh Message Understanding Conefrence，Washington. DC，USA，1998.

[21] Li Jianhua, Wang Xiaolong. An Effective Method on Automatic Identification of Chinese Name[J]. High Technology Letters. 2000,10(2):46-49.

[22] Tan Hongye, Zheng Jiaheng, Liu Kaiying. Research on Method of Automatic Recognition of Chinese Place Name Based on Transformation[j]. Journal of Software 2001.

[23] 黄德根，杨元生，王省，张艳丽，钟万. 基于统计方法的中文姓名识别[J].中文信息学报. 2001(2):31-37.

[24] 俞鸿魁，张华平，刘群，等. 基于层叠隐马尔科夫模型中文命名实体识别[J]. 通信学报 2006,27(2):87-94.

[25] 吴雪军，朱靖波. 基于统计和规则的中文人名识别[C]. 第一届学生计算语言研讨会(SWCL2003)，2002.97-102.

[26] Zhou Junsheng, He Liang, Dai Xinyu, et al. Chinese Named Entity Recognition with a Multi-Phase Model[C].In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney Australia, 2006.213-216.

[27] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In Proceedings ICML 2001:282~289.

[28] Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff[C]. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing，Sapporo，Japan，2003.95-104

[29] Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff[C]. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Republic of Korea, 2005.88-97

[30] A. Borthwick. Maximum Entropy Approach to Named Eniity Recognition[J]. PhD Dissertation, NewYorkUniversity.1999:18-25.

[31] R. Lawrence, Babiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[C].Proceedings of the IEEE.1989,(77),No.2

[32] Clifford. Markov random fields in Statistics [A]. In: Grimmett G, Welsh D, Disorder in physical systems, Vol. in Honour of J.M.Hammersley 70th Birthday [M],London: Oxford University Press,1990,19-32.

[33] H. Isozaki, H. Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition[C]. Proceedings of the 19th International Conference on Computational Linguistics (COLING-02),2002:390~396

[34] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman. Exploiting Diverse Knowledge Sources via Maximum Entropy[C].Proceedings of the 6th Workshop on Very Large Corpora(VLC-98),1998:152~160

[35] K. Uchimoto, M. Murada, Q. Ma, H. Ozaku and H. Isahara. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules[C]. Proceedings of the 38th Annual Meeting of the Association for Computational

Linguistics(ACL-00),2000:326~335

[36] Andrew McCallum, Wei Li, Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons[C]. Proceedings of CoNLL-2003,Edmonton, Canada, 2003:188-191

[37] Erik F. Tjong Kim Sang, Fien De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition[C]. Proceedings of CoNLL-2003,Edmonton, Canada,2003:142-147

[38] Oliver Bender, Franz Josef Och and Hermann Ney, Maximum Entropy Models for Named Entity Recognition[C]. Proceedings of CoNLL-2003, Edmonton, Canada, 2003:148-151

[39] Hai Leong Chieu, Hwee Tou Ng, Named Entity Recognition with a Maximum Entropy Approach[C]. Proceedings of CoNLL-2003, Edmonton, Canada, 2003: 160-163

[40] James R.Curran and Stephen Clark,Language Independent NER using a Maximum Entropy Tagger[C]. In Proceedings of CoNLL-2003, Edmonton, Canada, 2003:164-167

[41] Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang, Named Entity Recognition through Classifier Combination[C]. Proceedings of CoNLL-2003, Edmonton, Canada, 2003:168-171

[42] Dan Klein, Joseph Smarr, Huy Nguyen and Christopher D. Manning, Named Entity Recognition with Character-Level Models[C].Proceedings of CoNLL-2003, Edmonton, Canada, 2003:180-183

[43] James Mayfield, Paul McNamee and Christine Piatko, Named Entity Recognition using Hundreds of Thousands of Features[C]. Proceedings of CoNLL-2003,Edmonton, Canada, 2003:184-187

[44] Casey Whitelaw, Jon Patrick, Named Entity Recognition Using a Character-based Probabilistic Approach[C]. Proceedings of CoNLL-2003, Edmonton, Canada,2003:196-199

[45] Tong Zhang, David Johnson, A Robust Risk Minimization based Named Entity Recognition System[C]. Proceedings of CoNLL-2003, Edmonton, Canada, 2003: 204-207

[46] Xavier Carreras, Lluís Màrquez, and Lluís Padró, Learning a Perceptron-Based Named Entity Chunker via Online Recognition Feedback[C]. Proceedings of CoNLL-2003, Edmonton, Canada, 2003:156-159

攻读硕士学位期间发表的论文

1 王峰，王召巴，陈友兴，丁战阳. 聚合物混合状态的超声检测及在线监控.塑料科技（已录用）.

2 徐德山，张智雄，王峰，邢美凤. 上下文分析与统计特征相结合的英文术语抽取研. 现代图书情报技术.（已录用）.

致　谢

首先，特别感谢我中科院的实习导师张智雄老师。张老师在知识平台长期建设领域有着深厚的科研基础，并一直保持着持续学习、潜心思考的科研态度。在中科院实习期间，我的每一点点滴滴的进步都离不开张老师的悉心培养。从论文的选题、构思，再到论文的写作、修改，每一个过程都凝聚着张老师的心血，张老师一丝不苟的工作作风等优秀品质使我终生受益。

特别感谢我的导师王召巴老师。如果没有王老师对学生的理解和帮助，我想我是很难顺利完成中科院的学习过程的，在最后论文的提交和修改方面，非常感谢王老师能够如此细致的对我论文提出修改意见，帮助我对论文进行了进一步的修改和完善，王老师对于学生的宽容以及严谨求实的治学态度，对我今后的在做人以及工作方面都受益匪浅。

特别感谢我们实验组的陈友兴老师、金永老师和赵霞老师。在学校期间，无论是生活、学习还是工作，我都得到了你们无私的帮助，在实验过程中给出了很多非常好的建议，使得实验顺利完成并能顺利的完成了学校的课程和任务。

特别感谢李春旺老师、乐小虬老师、孙秀娟老师等在信息系统部所为我提供的帮助。在这里，我得到了最好的学习环境。感谢赵华铭老师在云存储技术方面给予我的指导；感谢李宇老师对我的论文思路提出的宝贵建议；感谢周强老师在 Java 技术上对我的悉心指教；感谢我中科院的师兄徐健、许德山、杨代庆、邹益民，学姐吴思竹、邢美凤、洪娜，刘建华，以及同学高建秀，王科，钱力，袁国华，管仲，谢靖。你们不仅在我的论文写作过程中提供了力所能及的帮助，而且在生活中与我互助互勉，一起熬过最困难的日子。

感谢我的同学郭惠平、程擂、范玉磊、李正、马娟娟、刘卉芳、李宏等等，你们陪我渡过了一段难忘的研究生生活。在我在中科院学习的时候，你们总能把学校的事情第一时间转达给我，让我时刻和学校保持联系。

最后，我特别感谢我的父母，是你们对我无私和不求回报的支持，才让我一路平稳的走到现在。

感谢所有帮助过我的老师、同学、朋友、亲人以及本论文的引文作者们！