大连理工大学

硕士学位论文

中文命名实体识别的研究

姓名：丁卓冶

申请学位级别：硕士

专业：计算机应用技术

指导教师：黄德根

20081218

摘　要

中文命名实体识别是自然语言处理的基础任务，是机器翻译、信息检索、问答系统等技术的基础，研究并实现有效的中文命名实体识别方法是本文的主要研究内容。

本文主要采用基于机器学习的方法完成命名实体识别任务。

首先，通过分析中文人名、地名的特点，以抽取合适的特征；定义科学的特征模板，并建立了一种基于条件随机场(Conditional Random Fields, CRFs)的中文命名实体识别模型。通过对 CRFs 的识别结果进行分析，发现 CRFs 模型中给出的错误标记大都拥有较小的边缘概率，用边缘概率定位到 CRFs 模型中可能的错误标记，并分别引入了概率统计方法和边界模板的方法对这部分标记进行修正，以优化系统的识别效果。实验证明，这两种混合模型的识别效果明显好于单纯的 CRFs 方法。

另外，提出一种基于 Max-Margin Markov Networks 模型的地名识别方法。Max-Margin Markov Networks 模型将 Max-Margin 的思想应用于马尔可夫网络。它综合了支持向量机(Support Vector Machine, SVM)模型和无向图模型的优点。通过地名识别的实验证明，在相同的语料、特征和特征模板的条件下，基于 Max-Margin Markov Networks 模型的识别效果好于 CRFs 和 SVM 模型。

最后，提出了一种基于概率特征函数的 CRFs 模型。CRFs 模型是目前最优秀的机器学习模型之一，它定义的特征函数全部是 0、1 二值形式的，导致丢失一些有用的概率信息。本文在定义特征函数时融入了概率信息，以强化模型的学习能力，然后基于概率特征函数构造条件随机场。通过命名实体识别的实验证明，在相同的条件下，基于概率特征函数的 CRFs 比传统的 CRFs 具有更好的机器学习能力。

本文的研究成果可应用于其它自然语言处理任务中。

关键词：自然语言处理；命名实体识别；支持向量机；条件随机场

A Study on Chinese Named Entity Recognition

Abstract

Chinese Named Entity Recognition is a basic task of Natural Language Processing and also it is basis of some NLP tasks, such as machine translation, information retrieval, question answering and so on.

Firstly, a model based on CRFs is built to do NER task. CRFs model, one of the best machine learning models, is widely used in NLP area and has a good performance. CRFs, an undirected graphical model, can avoid bias problems belong to direct graphical models, and at the same time, it can consider the information between correlative nodes. Analyzing the results obtained by sole CRFs, we find that the errors mainly happened in labels with low marginal probabilities. Two methods, statistical method and boundary template method, are introduced to correct the errors. If the marginal probability is greater than the given threshold, the test sample is recognized by CRFs; otherwise, one of these two methods is used. Experimental results show that the two hybrid methods have better performance than the CRFs method.

Secondly, this paper introduces a new machine learning model, Max-Margin Markov Networks. It combines the advantages of both SVM and undirected graphical model. A novel method based on Max-Margin Markov Networks is presented in this paper to do Chinese location NER task and it obtains better results than CRFs and SVM.

Lastly, a kind of CRFs model based on probability feature functions is presented. Probability feature functions are defined to replace binary functions, as to improve machine learning ability of system. Then NER tasks are employed to test machine learning ability of this improved CRFs. Experimental results show that methods based on improved CRFs are better than CRFs method.

Our methods are expected to extend to other tasks of NLP area.

Key Words: Natural Language Processing; Named Entity Recognition; Support Vector Machine; Conditional Random Fields

大连理工大学学位论文独创性声明

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

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：中文命名实体识别的研究

作者签名：丁卓冶　　日期：2008年12月18日

大连理工大学学位论文版权使用授权书

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

学位论文题目：中文命名实体识别的研究

作者签名：丁卓冶　　日期：2008 年 12 月 18 日

导师签名：苏依拉　　日期：2008 年 12 月 18 日

1 绪论

1.1 研究背景与意义

随着互联网和信息产业的快速发展，大量信息以电子文档的形式出现在人们面前，人们迫切希望计算机能对网上出现的文本信息实现自动化处理。命名实体识别(Named Entity Recognition, NER)是信息处理技术中的关键基础技术。命名实体(Named Entity, NE)是文本中基本的信息单位，是文本中的固有名称、缩写及其他唯一标识，是正确理解文本的基础。狭义的讲，可以把命名实体分为人名、地名、组织名等。广义的讲，命名实体还可以包括时间表达式、数值表达式等，在各种应用领域，还可以根据具体的需要定义其他类型的命名实体，例如，在某个具体应用中，可能需要把住址、电子信箱、电话号码、会议名称等作为命名实体。

命名实体识别任务包括：(1) 发现命名实体，即判断一个文本串是否代表一个命名实体；(2) 标注命名实体，即将发现的命名实体标注为某一种具体类型。

命名实体识别属于文本信息处理的基础研究领域，它的研究成果将直接影响到文本信息自动化处理的深层次研究，它是以下多种自然语言处理技术的重要基础：

(1) 信息抽取

在信息抽取研究中，人们需要从文本中自动抽取出特定的事实信息，形成结构化数据。文本中通常会包括事件发生的时间、地点、参与人物等。命名实体是信息的主要载体，是实现信息抽取的第一步，也是信息抽取中最有实用价值的一项关键技术。

(2) 信息检索

在目前大规模知识库的情况下，信息检索过程对于准确率和相关度的要求要高于召回率，而提高准确率和改善相关度的一条重要途径就是以短语为索引词。索引的知识粒度越大，确定性越强，歧义性越小。有实验报告证明，命名实体的识别可以改善系统检索文档的相关度，并提高检索系统的召回率和准确率。

(3) 机器翻译

在机器翻译系统中遇到机构名、地名等实体时，由于只能和词语对齐，因此常会使翻译结果不易理解，甚至出现错误。加入命名实体识别之后，它可以使机器翻译的英汉对照达到短语一级，从而使翻译语句更通顺更地道，减少错误。

(4) 问答系统

在开放域问答系统中，常常会遇到需要回答某个机构、地点、日期等具体问题，而通常的分词结果并不能满足需要，因此答案只能返回段落或篇章。将命名实体识别的技

术应用在其上，可以对文本中的上述信息做出更准确的分析，使问答系统给出更精确、更简洁的短语级的答案。

1.2 中文命名实体识别的特点与难点

1.2.1 中文命名实体识别的特点

中文姓名的特点主要有[1,2]：

(1) 中文姓名一般由二字到四字组成，第一字为姓氏字(复姓为前两字)，其后的一到两个字为名用字。

(2) 当今仍然使用、活跃的中文姓氏远没有某些姓氏典籍所列举的那么多，大概有1000多个。

(3) 姓氏分布很不均匀，但相对集中。“王、李、刘、张、陈”这5大姓就占了姓名样本数的29.1，前18个姓占50.3%，前181个姓占90.3%，前586个姓占98.6%，其余姓氏仅占不到1.5%。

(4) 某些姓氏可用作单字词，其中不乏高频单字词。常用的姓氏如“王、黄、马、高、于”等，不常用的姓氏如“从、那”等。

(5) 名字用字分布较姓氏用字要平缓、分散。共得到3679个名字用字，频率最高的前17个字的覆盖率为10.5%，前80个字为30.3%，前207个字为50.3%，前1122个字为90.4%。

(6) 名字用字涉及范围很广。从所属的词类看，不仅有实词，也有各类虚词。如副词“常、太、必、非、更、也、级、又、皆”等，介词“以、向、从、于、把”，连词“而、虽、且、与”等。从感情色彩看，多使用褒义字和中性字，但也出现了一些贬义字或不太文雅的字，如“狼、恶、悲、暴、虫”等。

(7) 某些汉字即可用作姓氏，又可用作名字用字。如“林、方、金、江、万、颜、童、柳”等。

上述各点，(1)，(3)和(5)赋予中文姓名具有统计意义上的可区别性，(4)和(6)使得部分姓名模糊，(7)则导致相邻候选姓名之间产生交叉歧义。

中文地名的特点主要有：

(1) 尽管在《中国地名录》有些地名没有被收录，但是，绝大部分没有收录的地名的用字都可以被地名用字库覆盖。

(2) 中文地名用词一方面比较自由、分散，地名录中共享汉字3685个。另一方面，中文地名用词又有相对集中的覆盖能力。

(3) 地名结尾经常有地名特征词出现。但地名特征词出现的情况比较复杂：既可以作为普通用词出现，并不表示真正的、具体的地名。又可以出现在地名其它位置或作为地名的前部词。既可以有一个地名特征词出现，又可以同时有多个地名特征词连着出现。这无疑增加了地名识别难度。

(4) 地名长度没有严格限制，不像中文姓名那样，长度在 2-4 个汉字之间。在真实文本中，经常会有地名简称出现。

(5) 地名用词的情况非常复杂。有一些词，如果在真实文本中出现，那么，作为地名用词的可能性非常大，如“峨嵋山”中的“峨嵋”。但是，可作单字词的汉字在地名中经常出现，如“西/直/门、马/家/塔”。每一个单字均为高频单字词，在真实文本中，作为地名出现的次数比较多，同时，作为非地名成分出现的次数也很多。

(6) 多字词可以在地名不同的位置出现，可以在地名首部出现，也可以在地名中部出现。同时，相对于单字词来讲，地名词典中的多字词统计的信息不充分，对多字词的判断也是地名识别中的一个难点。

(7) 地名有时同一些介词、动词、方位词之类的指示词出现，这些指示词对地名识别能起到标志作用，如“到北京、万家寨附近”。但有些指示词也可以作为地名组成部分在真实文本中出现，如“上甘岭、来复乡”。同时，这些词在文本中并不总是与地名同时出现，如“在此基础上、从计划到组织”。

(8) 经常多个地名通过一些连接词或者连接符号的连接一起出现，如“/吉林省/四平市/梨树县/梨树镇/霍家店村”。这样的地名多是表示行政地区的地名，这对地名识别来讲，是一个有利的信息。

(9) 真实语料中地名可能和其它词语发生冲突，首先，连续出现的地名自身发生冲突，如“重庆土兰寿县”。地名还可能成为其他命名实体的一部分，如“大连市机械厂”，地名作为机构名的一部分。地名用字还可能与其相邻字成词，如“海宁市长安邮电局”切分成“海宁/市长/安/邮电局”。

综上所述，(1)，(4)和(9)增加了地名识别难度，(3)和(7)可能使候选地名产生交叉歧义，(2)，(5)和(6)使部分地名边界模糊，(8)则有助于地名识别。

1.2.2 中文命名实体识别的难点

中文人名识别的难点在于如何正确确定出中文人名的左右边界。由于中文文本不含有西方语言的形态特征如大写字母等可以作为识别人名的依据，而且中文人名的结构复杂，表现形式多样，人名用字不仅可以自身成词，并能与相邻的字构成词。它们的识别存在以下问题：

(1) 缺乏明显的特征标志。英文命名实体大多首字母大写，因此易于识别，而中文命名实体不具有明显的标志，增加了识别的难度。

(2) 分词影响命名实体的识别。分词的错误有时会导致命名实体的边界错误。

(3) 不同种类的命名实体间存在歧义问题，主要包括边界歧义和分类歧义。边界歧义是指根据命名实体边界的不同，可以有不同的识别结果；分类歧义是指一个命名实体，可以标为几种不同的实体类型。

(4) 大部分命名实体是未登录词。汉语词汇是个开放的集合，不可能将所有的词都放入词库。

1.3 国内外研究现状

近几年来随着计算机信息检索技术的不断发展，中文命名实体识别已成为学术界研究的热点课题，国内外很多学者和专家进行了深入的研究。根据查阅的文献，目前中文命名实体识别的方法主要有：基于规则的方法[3-6]、基于统计的方法[7-20]、规则和统计相结合的方法[21,22]等。

(1) 基于规则的方法

在中文命名实体识别的早期研究中，大多采用人工总结各种判定规则，然后通过规则匹配的方法识别各种类型的命名实体。规则方法主要是利用两种信息：命名实体用字分类和限制性成分。即：分析过程中，当扫描到具有明显特征的命名实体用字时，开始触发命名实体的识别过程，并采集命名实体前后相关的成分，对命名实体的前后位置进行限制。此外文献[3]采用基于转换的错误驱动的方法来获取识别地名的上下文有关规则，然后应用这些规则对当前标注结果进行转换来实现中文地名的识别。小规模测试的结果表明，其准确率可以高达97%。

基于规则的系统，通过分析命名实体的内部和外部特征，人工构造规则模板实现命名实体的识别。基于规则的命名实体识别方法在小规则测试效果较好，速度快。但是，规则方法的存在一些缺点在于：

① 无论是人工总结出判定规则，还是收集规模巨大的命名实体库与真实语料库，都对语言知识要求较高，需要很大的人力物力。

② 一旦增加新特征的命名实体，或移植到其它语言就必须对以前的规则重新修订，增加新规则，规则方法很难扩展。

③ 规则虽然可以保证很高的准确率，但是覆盖范围都是有限的，对于覆盖集之外的命名实体就完全无能为力。

④ 规则较多时还会引起规则之间的冲突。

因此，目前中文命名实体识别的主流技术就是采用统计模型，以及统计和规则相结合的方法。

(2) 基于统计的方法

中文命名实体识别系统中采用的统计模型主要有：隐马尔可夫模型[7,8](Hidden Markov Model, HMM)、最大熵模型[9](Maximum Entropy Model, ME)、决策树[10](Decision Tree)、boosting[11,12]、支持向量机[13-15](Support Vector Machine, SVM)以及传统的概率统计方法[16,17]、条件随机场(Conditional Random Fields, CRFs)[18-20]。文献[7]提出了一种基于角色标注的命名实体识别方法，首先采用 Viterbi 算法对切词结果进行角色标注，然后在此基础上进行模式最大匹配，最终实现中国人名的识别。文献[10]采用决策树的方法，首先把命名实体识别问题看成一种分类问题，然后用决策树的方法来解决这个分类问题。从语料库及现代汉语语素数据库中共统计出六类知识，用这些知识作为属性构建了训练集，最后生成了决策树。文献[14]采用支持向量机方法进行中国人名和组织机构名的自动识别。文献[15]提出了支持向量机与概率统计结合的混合模型进行命名实体识别，取得了较好的效果。文献[16]采用传统的概率统计的方法对中国人名进行识别，针对姓名语料库来训练某个字作为姓名组成部分的概率值，并用它们来计算某个候选字段作为姓名的概率，其中概率值大于一定阈值的字段为识别出的中国人名。文献[18]提出并实现了一种基于 CRFs 的中国人名识别方法，并取得了较好的识别精度。

(3) 规则和统计相结合的方法

目前一些系统将统计与规则结合起来，它采用统计方法对命名实体进行识别，利用规则机制对其进行校正过滤。例如，文献[21]使用从大规模真实文本语料库得到的统计信息，通过计算人名的构词可信度和接续可信度并结合规则对中国人名进行识别。文献[22]针对有特征词的中文地名进行了研究，并实现了以统计为主、规则为辅的有特征词的中文地名识别系统，该系统使用从大规模地名词典和真实文本语料库得到的统计信息以及针对地名特点总结出来的规则，通过计算地名的构词可信度和接续可信度从而识别中文地名。

(4) 存在的问题

早期的规则系统面对大规模真实文本束手无策的原因在于语言学家编写的有限的规则不能够全面、准确地描写输入符号串到输出符号串的映射。在这种情况下，统计语言模型以及统计和规则相结合的模型成为了当前主流技术。

但是这些解决方案仍然存在一些不足：

① 命名实体的候选字段大都选取切分后的单字碎片，这样内部成词以及上下文成词的命名实体很难召回。

② 机器学习方法在某些样本上的表现很差，导致整体的识别效果不高。

③ 基于机器学习的方法由于泛化不够，导致召回率偏低；另外，机器学习模型的学习能力有限，成为识别效果提高的瓶颈。

1.4 本文的工作

本文主要研究中文命名实体识别问题，重点放在对人名、地名这两种命名实体识别的研究上。首先，应用 CRFs 模型作为机器学习的方法，选取合适的特征，完成命名实体识别任务，并定位出 CRFs 模型中错误机会较大的标记，引入其它方法对这部分标记进行修正，以优化识别结果；之后，考虑尝试其它机器学习模型，期望获得更好的识别效果。

本文的主要工作如下：

(1) 基于 CRFs 模型进行命名实体识别任务。条件随机场是一种判别无向图模型，它继承了最大熵模型的优点，具有较强的机器学习能力，被广泛应用于自然语言处理领域。

(2) 通过边缘概率较准确的定位出 CRFs 模型中错误机会较大的标记，并分别引入边界模板方法和概率统计方法对这些标记进行修正，以优化系统的识别效果。

(3) 基于 Max-Margin Markov Networks 模型进行地名实体任务。Max-Margin Markov Networks 是一种优秀的机器学习模型，它综合了 SVM 与无向图模型的优点，具有较强的机器学习能力。

(4) 提出一种基于概率特征函数的 CRFs 模型。它在定义特征函数时，通过融入了概率信息，以提高模型的学习能力，然后基于概率特征函数构造条件随机场。通过命名实体识别任务来验证此改进模型的学习能力。

2 条件随机场

条件随机场(Conditional Random Fields, CRFs)是由 Lafferty 等人于 2001 年提出[23]，其模型思想主要来源于最大熵模型。它可以看成是一个无向图模型或马尔可夫随机场，是一种用来标记和切分序列化数据的统计框架模型。目前，条件随机场被应用于解决分词[24,25]、命名实体识别[26]、浅层语法分析[27,28]等自然语言处理任务，取得了较好的效果。根据条件随机场的特性以及它在序列标注上的良好表现，本文将命名实体的识别问题转换为标注问题，并使用 CRFs 来解决命名实体识别的问题。

2.1 判别无向图模型

无向图模型的优点在于其没有隐马尔可夫模型那样严格的独立性假设，同时克服了最大熵马尔可夫模型的标记偏置问题。

无向图模型或称为马尔可夫随机场是一个非循环图$G=(V,E)$，其中V是图中节点集合，E是V间的无向边集合。节点V表示一组连续或者分散的随机变量。由以上分析我们知道，在有向图模型$G^d=(V^d,E^d)$中，随机变量间的联合概率分布被表示为公式(2.1)的形式：

$$p(v_1^d,v_2^d,...,v_n^d)=\prod_{i=1}^{n}p(V_i^d\,|V_{\pi i}^d) \tag{2.1}$$

其中$V_{\pi i}^d$是V_i^d的所有的父节点集合。

不同于有向图模型，马尔可夫随机场的无向性很难确保每个节点在给定它的邻节点的条件下的条件概率和以图中其他节点为条件的条件概率一致。由于这个原因，马尔可夫随机场中的联合概率并不是用条件概率参数化表示的，而是定义为由一组条件独立的局部函数的乘积形式。

参数化表示一个无向图的第一步，即找出每个局部函数所作用的那部分节点。为了找出这些节点，我们引入了无向图中条件独立的概念。A，B，C表示的是三个不相交的索引子集，如果节点集V_B将V_A和V_C分隔开，那么就可以认为节点集V_A所表示的随机变量，在给定节点集V_B所表示的随机变量的条件下，是条件独立于V_C所表示的随机变量的。即对于一个无向图来说，在给定V_B的条件下，V_A条件独立于V_C，从V_A中的任意一个节点V_i到V_C中的任意一个节点V_j之间的路径，至少经过V_B中的一个节点V_k。为了确定局部函数起作用的那部分节点，根据无向图模型G的条件独立属性，我们可以发现若G中两个节点V_i和V_j之间不存在边，则意味着在给定图中所有其他的节点的条件下，这

两个节点一定是条件独立的。因此，当选择局部函数时，我们必须保证能够通过合理分解，使V_i和V_j不出现在相同的局部函数中。

实现这个分解要求的最简单的方法，就是使得每个局部函数所作用的那部分节点，在 G 中形成一个全连接的节点子集或团(clique)。这就确保了没有一个局部函数是作用在任何一对没有直接连接的节点上的，并且如果两个节点同时出现在一个团中，则在这两个节点所在的团上定义一个局部函数来建立这样的依赖。更进一步地精练局部函数的概念，我们把每个局部函数都建立在一个最大团，或者说一个不可能再被扩展以包含其他的节点同时保持全连接的团上。因此，最简单的局部函数集就是那些在图 G 中最大团 C 所包含的节点集上定义的函数集。这些局部函数$\psi Vc(v_c)$被称为势函数，并且是严格正实数的函数形式。然而，一组正实数函数的乘积并不能保证满足概率公理。因此，为了满足概率公理，并确保乘积确实是 G 中节点所表示的随机变量的联合概率分布，我们定义了一个归一化因子 Z，形式如下：

$$Z = \sum_{v_1,\ldots,v_c} \prod_{c \in C} \psi Vc(v_c) \tag{2.2}$$

其中 C 是 G 中所有的最大团的集合。根据 Hammersley-Clifford 定理，我们得到联合分布的计算公式(2.3)：

$$p(v_i,\ldots v_n) = \frac{1}{Z} \prod_{c \in C} \psi Vc(v_c) \tag{2.3}$$

虽然无向图模型中随机变量的联合分布可写成势函数的乘积形式，但需要说明的是，一个孤立的势函数并没有直接的概率表示，而只是表示在随机变量结构上的约束条件而已，即该函数是在这些随机变量上定义的。

2.2 条件随机场(CRFs)模型

2.2.1 条件随机场的无向图结构

CRFs 是无向图模型的一种形式，在给定将要标记的观测序列的情况下，无向图模型可以被用来在标记序列上定义一个联合概率分布。假设 x，y 分别表示需要标记的观察序列和它对用的标记序列的联合分布随机变量，条件随机场(x,y)就是一个以观测序列 x 为全局条件的无向图模型。

通常，我们定义$G=(V,E)$是一个无向图，$y=\{y_v \mid v \in V\}$。即 V 中的每个结点对应着一个随机变量所表示的标记序列的成分Y_v。因而，整个图和与图相关的分布类别以 x 为条件，所以与 G 相关的联合分布的类别的形式是$P(y_1,\ldots,y_n \mid x)$，这里 y 和 x 分别是类

别序列和观测序列。如果每个随机变量 y_v 满足关于 G 的马尔可夫属性，给定 x 和 y_v 以外的所有随机变量 $y_{(u|u\neq v,\{u,v\}\in V)}$，则随机变量 y_v 的概率式为：

$$P(y_v|x, y_u, u \neq v)=P(y_v|x, y_u, u \sim v) \tag{2.4}$$

其中 $u \sim v$ 表示 u 与 v 在图 G 中相邻，那么，(x,y) 就是一个条件随机场。

理论上，如果图 G 表示了将要建模的标记序列之间的条件依存关系，则它的结构可以是任意的。但是当用于序列标记任务建模时，所遇到的最简单和最通用的图结构是这样的：与 y 的元素相对应的结点形成了一个简单的一阶链(First-order Chain)。我们将这种条件随机场称为线性链条件随机场(Linear-chain CRFs)，如图 2.1 所示。表示 y 的随机变量只是图 G 的一部分，这是因为我们希望定义一个概率分布 $P(x,y)$。另外，x 的元素间并不存在任何图结构，这是因为我们只是将观察序列作为条件，所以并不对 x 做任何的独立假设。

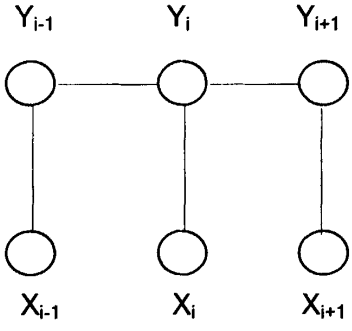

图 2.1 线性 CRFs 的模型结构

Fig. 2.1 The model structure of a linear-chain CRFs

2.2.2 条件随机场的势函数表示

CRFs的图结构可以被用来将其联合分布分解为一个归一化(Normalized)的势函数的乘积，势函数来自条件独立的概念，是严格非负的、实数值函数，这里 y_v 是 y 的元素。每一个势函数涉及的 G 中的顶点表示随机变量的一个子集。根据无向图模型的条件独立定义，如果 G 中两个顶点之间没有边，则意味着两个顶点表示的随机变量独立于 G 中其它给定的顶点。因而势函数必须保证可以将联合概率分解，以至于条件独立的随机变量不会出现在相同的势函数中。最容易满足这个要求的方法是保证每一个势函数作用于随机变量的一个集合上，而这些变量对应的顶点形成了一个最大的全通环(Clique)。这

确保了势函数所涉及的任何随机变量对，其顶点是直接联系的，如果两个顶点在一个全通环(Clique)中一起出现，则这种关系就明确表示出来了。在链结构的 CRFs 下，每一个势函数作用于相邻的标记变量 y_i 和 y_{i+1} 对。

尽管无向图模型中随机变量的联合分布可写成势函数的乘积，需要指出的是一个孤立的势函数并没有直接的概率意义，而是表示了定义这个势函数所涉及的随机变量的结构上的约束而已。这反过来也影响了全局结构的概率，即一个概率大的全局结构较概率小的全局结构更能满足这些约束条件。

在给定观测序列 $\boldsymbol{x}$ 的情况下，Lafferty 等定义了标记序列 $\boldsymbol{y}$ 的概率是势函数(Potential Function)乘积的一个归一化形式，其中每个因子形式如公式(2.5)所示：

$$\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_k \mu_k S_k(y_i, \boldsymbol{x}, i)) \tag{2.5}$$

这里 $t_j(y_{i-1}, y_i, \boldsymbol{x}, i)$ 是关于整个观测序列和位置 i 以及 i-1 标记的特征函数，$S_k(y_i, \boldsymbol{x}, i)$ 是关于位置 i 的标记和观测序列的状态特征函数，这里参数 λ_j 和 μ_k 是特征权重，可从训练语料中估计得到。

当定义特征函数时，可以构造了观测序列的实数值特征 $b(\boldsymbol{x}, i)$ 集合来描述训练数据的经验分布特征，这些特征与模型具有相同的分布。下面是一个例子：

$$b(\boldsymbol{x}, i) = \begin{cases} 1 & \text{if 位置 } i \text{ 的观测值为“市”} \\ 0 & \text{其它} \end{cases}$$

每个特征函数表示一个实数值的观测特征 $b(\boldsymbol{x}, i)$，如果当前状态(状态函数)或前一个状态和当前状态(转移函数)具有特定的值，则所有的特征函数都是实数值的。例如下面的转移函数。

$$t_j(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} b(\boldsymbol{x}, i) & \text{if } y_{i-1} = \text{动词}, y_i = \text{名词} \\ 0 & \text{其它} \end{cases}$$

在后面的描述中，我们用公式(2.6)来表示状态函数：

$$s_k(y_i, \boldsymbol{x}, i) = s_k(y_{i-1}, y_i, \boldsymbol{x}, i). \tag{2.6}$$

且

$$F_j(\boldsymbol{y}, \boldsymbol{x}) = \sum_i^n f_j(y_{i-1}, y_i, \boldsymbol{x}, i) \tag{2.7}$$

特征函数 $f_j(y_{i-1}, y_i, \boldsymbol{x}, i)$ 是一个状态特征函数 $S_k(y_i, \boldsymbol{x}, i)$ 或者是一个转移特征函数 $t_j(y_{i-1}, y_i, \boldsymbol{x}, i)$。

因此对于一个给定观测序列 x，其对应的标记序列 y 的概率为公式(2.8)所示：

$$P(y|x,\lambda)=\frac{1}{Z(x)}\exp(\sum_j \lambda_j F_j(y,x)) \tag{2.8}$$

$Z(\mathbf{x})$ 是归一化因子(Normalization Factor)其形式如公式(2.9)所示：

$$Z(x)=\sum_y \exp(\sum_j \lambda_j F_j(y,x)) \tag{2.9}$$

现在用 CRFs 建立了 $P(y|x)$ 的统计模型，求解序列标记任务就是求得 y^* 满足 $P(y|x)$ 最大，$Z(x)$ 与 y 无关，所以 y^* 为公式(2.10)所示：

$$\begin{aligned} y^* &= \arg\max_y P(y|x) \\ &= \arg\max_y \frac{1}{Z(x)}\exp(\sum_j \lambda_j F_j(y,x)) \\ &= \arg\max_y \sum_j \lambda_j F_j(y,x) \end{aligned} \tag{2.10}$$

使用 Viterbi 等动态优化方法，即可求出最优解 y^*。

2.2.3 条件随机场的参数估计

建立 CRFs 模型的主要任务就是从样本数据中估计得到特征权值 λ。CRFs 参数估计可以使用最大似然估计(Maximum Likelihood Estimation, MLE)和贝叶斯估计(Bayes Estimation)。下面主要介绍用最大似然估计 CRFs 的模型参数。

在训练集 $T=\{\langle x^k, y^k\rangle\}$ 中，最大似然参数估计就是假设 $P(y|x,\lambda)$ 为 λ 的函数，使 $P(y|x,\lambda)$ 的对数值最大的 λ 为估计值，其似然值为公式(2.11)所示，其最大值为公式(2.12)所示：

$$\begin{aligned} L_\Lambda &= \sum_T \log P(y^k|x^k,\lambda) \\ &= \sum_T \log\frac{1}{Z(x^k)}\exp(\sum_j \lambda_j F_j(y^k,x^k)) \\ &= \sum_T \left(\sum_j \lambda_j F_j(y^k,x^k) - \log(Z(x^k))\right) \end{aligned} \tag{2.11}$$

$$\Lambda^* = \arg\max_\lambda \sum_T \log P(y^k|x^k,\lambda) \tag{2.12}$$

由于 L_Λ 为凸函数，导数为零的最值点。故对 λ 求导，则偏导数公式为(2.13)所示：

$$\frac{\partial L_\Lambda}{\partial \lambda_j} = \sum_i \left(\sum_j F_j(y^k, x^k) - E_{P(y|x^k)}\left[F_k(y, x^k)\right] \right) \tag{2.13}$$

可简写为：

$$\frac{\partial L_\Lambda}{\partial \lambda_j} = O_j - E_j = 0 \tag{2.14}$$

公式(2.14)中，O_j为λ_j在训练集T中出现的频率，$E_j = \sum_i E_{P(y|x^k)}\left[F_k(y, x^k)\right]$是$\lambda_j$在模型分布中的特征期望。$E_j$如果直接计算需要很大的计算量，可以使用动态规划的方法求解，如向前-向后(Forward-Backward)算法，我们将在第 2.2.4 节详细介绍动态规划策略。

如果直接使用最大似然估计，可能会发生过度学习问题，可以通过引入罚函数的方法解决这一问题。例如使用惩罚项$\frac{\sum_j \lambda_j^2}{2\sigma^2}$，则原问题变为式(2.15)：

$$L_\Lambda' = L_\Lambda - \frac{\sum_j \lambda_j^2}{2\sigma^2} + const \tag{2.15}$$

其导数变为公式(2.16)：

$$\frac{\partial L_\Lambda'}{\partial L_\Lambda} = \frac{\partial L_\Lambda}{\partial \lambda_j} - \frac{\lambda_j}{\sigma^2} \tag{2.16}$$

于是λ的参数估计问题可以用最优化方法解决。可以使用 GIS，IIS 等迭代方法，本文的实现使用 L-BFGS(Limited-memory Broyden-Fletcher-Goldfarb-Shanno)算法[29]。

2.2.4 动态规划方法

对于一个链式结构的 CRFs，可以为每个句子添加开始状态标记和结束状态标记来标记序列，y_0和y_{n+1}：分别表示开始标记和结束标记，给定一个观测序列x，标记序列y的概率$P(y|x,\lambda)$可以使用矩阵进行有效的计算。

设ψ是标记的字母表，y和y'是来自这个字母表的标记，我们定义了$n+1$个矩阵的集合$\{M_i(x)|i=1,...,n+1\}$，这里每个$M_i(x)$都是一个$|\psi \times \psi|$的矩阵，矩阵元素形式如公式(2.17)所示：

$$M_i(y', y|x) = \exp\left(\sum_j \lambda_j f_j(y', y, x, i) \right) \tag{2.17}$$

给定观测序列 x，没有归一化的标记序列 y 的条件概率可以表示为 $n+1$ 个矩阵的元素的乘积，如公式(2.18)所示：

$$P(y \mid x,\lambda)=\frac{1}{Z(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i \mid x) \tag{2.18}$$

相似的，观察序列 x 的归一化因子 $Z(x)$ 可以通过使用 Closed Semirings 方法从 $M_i(x)$ 矩阵中计算得到，该代数结构是一个处理图中的路径问题的一般框架。$Z(x)$ 的值由从开始位置到结束位置的 $M_i(x)$ 矩阵的乘积给定。其形式如式(2.19)所示：

$$Z(x)=\left[\prod_{i=1}^{n+1}M_i(x)\right]_{start,end} \tag{2.19}$$

因此只要求出 $M_i(x)$ 就可计算出 $Z(x)$ 的值。

在参数估计过程中，无论是使用迭代收敛还是基于导数的方法，为计算出极大似然的参数，对于训练数据中的每个观测序列 x^k，都必须有效地计算出每个与 CRFs 模型分布相关的特征函数的期望 $E_{P(y \mid x^k)}\left[F_k(y,x^k)\right]$，如公式(2.20)所示：

$$E_{P(y \mid x^k)}\left[F_j(y,x^k)\right]=\sum_{y}P(y=y^* \mid x,\lambda)F_j(y^*,x^k) \tag{2.20}$$

对式(2.19)直接计算的开销十分巨大，若标记序列 x^k 有 n 个元素，则 y^k 对应元素为 $n^{|y|}$ 个，因此，通常使用类似 HMM 中的向前向后算法解决这个问题。

我们改写公式(2.20)的右边为：

$$\sum_{i=1}^{n}\sum_{y',y}P(y_{i-1}=y',y_i=y \mid x,\lambda)f_j(y',y,x) \tag{2.21}$$

接下来，便可使用动态规划方法计算 $P(y_{i-1}=y',y_i=y \mid x,\lambda)$，我们定义向前-向后向量 $\alpha_i(x)$ 和 $\beta_i(x)$ 为分别为公式(2.22)与公式(2.23)所示：

$$\alpha_0(y \mid x)=\begin{cases}1 & if\ y=start\\ 0 & 其它\end{cases} \tag{2.22}$$

$$\beta_{n+1}(y \mid x)=\begin{cases}1 & if\ y=end\\ 0 & 其它\end{cases} \tag{2.23}$$

其递归定义分别为公式(2.24)与公式(2.25)所示：

$$\alpha_i(x)^T=\alpha_{i-1}(x)^T M_i(x) \tag{2.24}$$

$$\beta_i(\boldsymbol{x}) = \boldsymbol{M}_{i+1}(\boldsymbol{x})\beta_{i+1}(\boldsymbol{x}) \tag{2.25}$$

在给定观察序列 $\boldsymbol{x}$ 的条件下，$y_i = y$ 的概率，我们称为边缘概率，可如下给定：

$$P(y_i = y \mid \boldsymbol{x}, \lambda) = \frac{\alpha_i(y \mid \boldsymbol{x})\beta_i(y \mid \boldsymbol{x})}{Z(\boldsymbol{x})} \tag{2.26}$$

同理，在给定观察序列 $\boldsymbol{x}$ 的条件下，$y_{i-1} = y'$ 和 $y_i = y$ 的概率，可如下给定：

$$P(y_{i-1} = y', y_i = y \mid \boldsymbol{x}, \lambda) = \frac{\alpha_{i-1}(y' \mid \boldsymbol{x})\boldsymbol{M}_i(y', y \mid \boldsymbol{x})\beta_i(y \mid \boldsymbol{x})}{Z(\boldsymbol{x})} \tag{2.27}$$

我们将公式(2.27)代入公式(2.21)便可使用动态规划的方法，有效计算特征期望，从而能够使用机器学习的方法计算得到模型特征权值 λ。

3 Max-Margin Markov Networks

Max-Margin Markov Networks(M^3Net)是一种优秀的机器学习的模型，Ben Taskar 等人于 2003 年提出[30]。M^3Net 的原理是将 Max-Margin 的思想应用于马尔可夫网络上。Max-Margin(最大边缘)思想是尽量增大正确标注与其它非正确标注之间的差距从而增加标注正确的可能性。一些机器学习的方法建立在 Max-Margin 思想的基础上[31-34]。我们熟悉的支持向量机(SVM)就是来源于 Max-Margin 思想，SVM 思想是最大化两类超平面的间隔来增加分类的可信度。M^3Net 将 Max-Margin 建立在马尔可夫网络上，可以方便考虑相邻节点之间的联系，可以说 M^3Net 模型综合了 SVM 和无向图模型二者的优点，是一种优秀的统计模型。

M^3Net 模型的构建很大程度来源于多类 SVM，所以本章先回顾 SVM 的一些原理。

3.1 支持向量机

支持向量机(Support Vector Machine，简称为 SVM)是九十年代中期由 Vladimir N. Vapnik 等人根据统计学习理论提出的一种新的机器学习方法，它体现了结构风险最小化的思想和方法，是统计学理论中很新的内容，具有较强的学习能力和泛化性能，能够较好地解决小样本、高维数、非线性和局部极小等问题，SVM 的出现经历如下阶段：

1992 年，Boser，Guyon 和 Vapnik 提出了最优边界分类器的概念[35]，被认为是支持向量机的最初原型。

1995 年，Vapnik 首次完整地提出了基于统计学习理论的支持向量机方法[36]。

1997 年，Vapnik，Gokowich 和 Smola 提出了基于支持向量机方法的回归估计方法(Support Vector Regression, SVR)和信号处理方法。

3.1.1 最大间隔分类超平面

支持向量机最初是针对线性可分情况下的二类模式分类问题而提出的。给定观测样本集 $S=\{(x_1,y_1),...(x_l,y_l)\}$，其中，$\chi \subset R^n$ 称为输入空间或输入特征空间，$y_l \in \{-1,+1\}$ 是样本的类标记。分类的目的就是寻找一个分割超平面将正负两类样本完全分开，如图 3.1 所示。

设 $\zeta=\{w.x+b=0, w\in R^n, b\in R\}$ 是所有能够对 S 完全正确分类(经验风险为 0)的超平面的集合，其中，“.”是内积运算符。“完全正确分类”的意义是：任意一个由法向量 w 和常数 b 确定的分类超平面 H，它对样本集 S 的分类结果为：

$$\begin{cases} w.x_i + b \geq 0, \text{ 若 } y_i=+1 \\ w.x_i + b \leq 0, \text{ 若 } y_i=-1, \end{cases} \tag{3.1}$$

在所有的超平面中，最大间隔分类器要寻找的是一个最优超平面(Optimal Hyperplane)。这个最优超平面是指满足两类的分类间隔(Margin)最大的超平面。分类间隔被定义为：每类距离超平面最近的样本到超平面的距离之和。

此分类间隔可以经过如下的计算得到：设 H 为最优超平面，在 H 两侧分别作一个经过距离 H 最近的样本并且平行与 H 的超平面，记为 H_1 和 H_2。这两个超平面的表达式分别为：H_1：$y = w.x + b = 1$，H_2：$y = w.x + b = -1$。

显然，超平面 $H: y = w.x + b = 0$ 仍然属于 ζ。我们把超平面 H_1 和 H_2 之间的距离称为 H 的“分类间隔△”，并将 H_1 和 H_2 称为 H 的“间隔超平面”或者“间隔边界”。容易计算，$\Delta = 2/\|w\| = d^+ + d^-$。

所谓的“最大间隔分类超平面”就是在正确分类所有学习样本(即满足约束条件 $y_i(w.x_i + b) \geq 1$ 的前提下)，使得分类间隔△取最大值的超平面，例如，图 3.1 中所示的平面 H。

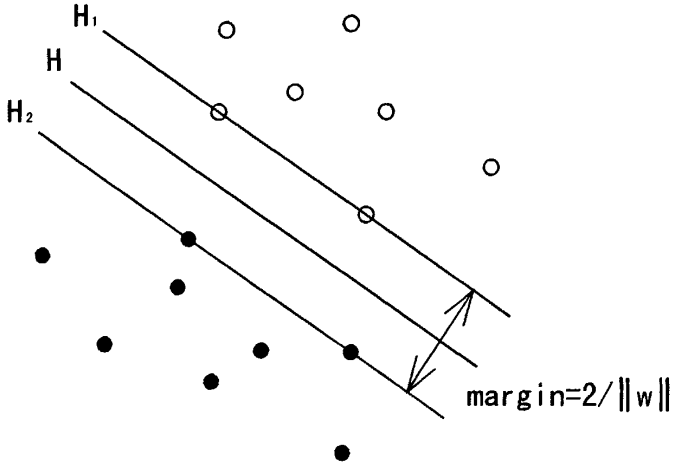

图 3.1　线性可分的分类超平面

Fig. 3.1　Sketch chart of SVM in the case of linear separable

3.1.2 支持向量机

(1) 线性的情况

依据前一小节的讨论，为了求解线性可分问题的最大间隔超平面，需要在满足约束 $y_i(w.x_i+b)\geq 1$ 的前提下最大化间隔△，等价于如下的优化问题：

$$\min_{w} \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w.x_i+b)\geq 1,\ i=1,...,l \tag{3.2}$$

这是一个典型的线性约束的凸二次规划问题，它唯一确定了最大间隔分类超平面。它的 Lagrange 函数是：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i(w.x_i+b)-1) \tag{3.3}$$

其中，$\alpha_i\geq 0$ 是每个样本对应的 Lagrange 乘子。将函数 $L(w,b,\alpha)$ 关于 w，b 求其极小值，由极值条件 $\nabla_b L(w,b,\alpha)=0$ 和 $\nabla_w L(w,b,\alpha)=0$ 得到：

$$\sum_{i=1}^{l} y_i\alpha_i=0 \tag{3.4}$$

$$w=\sum_{i=1}^{l}\alpha_i y_i x_i \tag{3.5}$$

将公式(3.4)和公式(3.5)代入 Lagrange 函数 $L(w,b,\alpha)$，并考虑 wolfe 对偶性质，得到优化问题(3.2)的对偶问题，如公式(3.6)所示：

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j\alpha_i\alpha_j(x_i.x_j)+\sum_{i=1}^{l}\alpha_i$$
$$s.t.\begin{cases}\sum_{i=1}^{l} y_i\alpha_i=0\\ \alpha_i\geq 0,\ i=1,...,l\end{cases} \tag{3.6}$$

可见，对偶问题仍然是线性约束的凸二次优化，存在唯一的最优解 α^*。

根据约束优化问题的 Karush-Kuhn-Tucker(KKT)条件，优化(3.6)取最优解 α^* 时应该满足如下的条件：

$$\alpha_i^*(y_i(w^*.x_i+b^*)-1)=0,\ i=1,2,...,l \tag{3.7}$$

从图 3.1 中可以看出，由于只有少部分观测样本 x_i 满足 $y_i(w^*.x_i+b^*)=1$，它们对应的 Lagrange 乘子 $\alpha_i^*>0$，而剩余的样本满足 $\alpha_i^*=0$。我们称解 α^* 的这种性质为“稀疏性”。

我们把$\alpha_i^* > 0$的观测样本称为“支持向量”，它们位于间隔边界H_1或H_2上。结合公式(3.5)和公式(3.7)可知，w^*和b^*均由支持向量决定。因此，最大间隔超平面$w^*.x_i + b^* = 0$完全由支持向量决定，而与剩余的观测样本无关。

这时，可以得到如下的最优决策函数或者分类器：

$$f(x) = \mathrm{sgn}(w^*.x + b^*) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i y_i (x.x_i) + b^*) \tag{3.8}$$

Vapnik把公式(3.8)称为“线性硬间隔支持向量机[37]”，而公式(3.2)和公式(3.6)分别称为它的原始优化问题和对偶优化问题。

另外，当样本线性不可分时，由于不存在使得分类间隔△取正值的超平面，严格要求所有样本被正确分类的硬间隔方法是行不通的。换句话说，必须适当松弛公式(3.1)中的约束条件。我们通过引入松弛变量$\xi_i \geq 0$, $i=1,...l$，可以得到“软化”的新约束条件：

$$y_i(w.x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i=1,...,l \tag{3.9}$$

显然，当ξ_i充分大时，样本(x_i, y_i)总可以满足约束条件。但另一方面，和项$\sum_{i=1}^{l} \xi_i$与样本的分类错误相关并且体现了经验风险，必须限制它的大小。因此，我们得到“软化”后的最大间隔分类器的优化问题：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i$$

$$s.t.\ y_i(w.x_i+b) \geq 1-\xi_i,\ \xi_i \geq 0,\ i=1,...,l \tag{3.10}$$

其中，实常数$C > 0$称为“罚参数”，它在分类器的复杂度和经验风险之间进行权衡。采用类似公式(3.2)至(3.6)的推导过程，可以得到公式(3.10)的对偶优化问题。因而不再详细给出对偶问题的具体形式。

(2) 非线性的情况

解决线性不可分类问题的另外一个途径是用“超曲面”代替“超平面”，并寻找一个能够正确分类所有观测样本的“最大间隔超曲面”。但是，“最大间隔超曲面”是难以描述和直接求解的。通过引入由输入空间χ到某个高维空间H（一般是 Hilbert 空间）的非线性映射$\Phi(.)$：$\chi \rightarrow \eta$，能够把χ中的寻找非线性的“最大间隔超曲面”问题转化为在高维空间η中求解线性的“最大间隔超平面”的问题，从而更容易给出具体的模型进行求解。

其间，需要避免在η中进行高维的内积运算$(\Phi(x_i).\Phi(x_j))$。如果存在输入空间中定义的某个“核函数” $K(.,.)$且满足$K(x_i, x_j) = (\Phi(x_i).\Phi(x_j))$，就可以通过直接计算$K(x_i, x_j)$

的值而避免η中的内积运算，并且不需要知道映射函数$\Phi(.)$的显式形式。关于核函数的讨论参见第 3.1.3 节。

因此，综合前面两种处理线性不可分类问题的思想，我们得到更常用的“非线性软间隔支持向量机”，简称“支持向量机(SVM)”。它的原始优化问题(P)和对偶优化问题(D)分别如下：

原始优化问题(P)：

$$\begin{aligned}&\min_{w\in\eta,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i\\&s.t.\ y_i(w.\Phi(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0,\ i=1,...,l\end{aligned}\tag{3.11}$$

对偶优化问题(D)：

$$\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i.x_j)+\sum_{i=1}^{l}\alpha_j$$

$$s.t.\begin{cases}\sum_{i=1}^{l}y_i\alpha_i=0\\0\leq\alpha_i\leq C,\ i=1,...,l\end{cases}\tag{3.12}$$

求解对偶问题的最优解α^*后，支持向量机的决策函数为：

$$f(x)=\mathrm{sgn}(w^*.\Phi(x)+b^*)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^*y_iK(x,x_i)+b^*)\tag{3.13}$$

同样，根据 KKT 条件，优化公式(3.12)取最优解α^*时应该满足如下的条件：

$$\begin{cases}\alpha_i^*[y_i(w.\Phi(x_i)+b^*)-1+\xi_i^*]=0, i=1,2,...,l\\(C-\alpha_i^*)\xi_i^*=0,\end{cases}\tag{3.14}$$

结合公式(3.12)和公式(3.14)的约束条件，可以推导出如下重要结论：

若$\alpha_i^*=0$，则有$\xi_i^*=0$，且对应的样本x_i一定不是支持向量；

若$0<\alpha_i^*<C$，则有$\xi_i^*=0$和$y_i(w^*.\Phi(x_i)+b^*)=1$，且对应的样本称为“非边界支持向量”；

若$\alpha_i^*=C$，则有$\xi_i^*=0$和$y_i(w^*.\Phi(x_i)+b^*)<1$，且对应的样本称为“边界支持向量”。

可见，最优解α_i^*的“稀疏”性质同样满足，支持向量机的决策函数完全由$\alpha_i^*\neq 0$的支持向量决定。

核函数定义了由低维映射到高维的方式。在下一节中，会介绍几种常用的核函数。

3.1.3 核函数

核函数$K(\boldsymbol{x},\boldsymbol{x}_i)$实际上相当于就是$\boldsymbol{x}$和$\boldsymbol{x}_i$的相似度。对更一般的情况，需要这样的函数$K$对任意两个样本向量$\boldsymbol{x}$和$\boldsymbol{x}_i$，它的返回值$K(\boldsymbol{x},\boldsymbol{x}_i)$就是描述两者的相似度的一个数值，这样的一个函数就是所谓的核函数。在决策函数中，只涉及训练样本之间的内积运算，可以用原空间中的函数实现的，甚至没用必要知道变换的形式。根据泛函的有关理论，只要一种核函数满足 Mercer 条件，它就对应某一变换空间的内积。因此，在最优分类面中采用适当的内积函数就可以实现某一非线性变换后的线性分类。核函数存在性定理表明：给定一个训练样本集，就一定存在一个相应的函数，训练样本通过该函数映射到高维特征空间的线性可分的。

常见的核函数有 4 类[38]：

(1) 线性内积 Kernel 函数：$K(\boldsymbol{x},\boldsymbol{x}_i)=\boldsymbol{x}\cdot\boldsymbol{x}_i$

(2) 多项式 Kernel 函数：$K(\boldsymbol{x},\boldsymbol{x}_i)=[(\boldsymbol{x}.\boldsymbol{x}_i)+1]^d$，d 是自然数

(3) 径像基 Kernel 函数：$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left\{-\frac{|\boldsymbol{x}-\boldsymbol{x}_i|^2}{\sigma^2}\right\},\sigma>0$

(4) Sigmoid Kernel 函数：$\tanh(a(\boldsymbol{x}.\boldsymbol{x}_i)+t),a$，$t$是常数，tanh 是 Sigmoid 函数

3.1.4 多类支持向量机

SVM 本身是解决两类分类问题的，对于多类(k类)划分问题可将其转化为两类划分问题加以处理，目前主要有两种方法[39,40]：(1) pairwise 方法：在任意两个类别之间构造一个二值分类器，从而生成 $k(k\text{-}1)/2$ 个二值分类器，每个分类器训练两种不同类别的数据，在分类中使用投票策略：对于一个未知样本每个分类器都有一个选票，其结果是具有选票最多的类别。(2) one vs. others 方法：构造 k 个分类器，第 i 个分类器的训练数据是第 i 类的数据作为正例，其它类的数据作为负例，为每个类构造一个分类器，第 i 个分类器在第 i 类和其他类之间构造一个超平面，在多个两类分类器中具有最大输出的类别即是测试数据所属的类别。

以上两种多类划分的问题存在着一个共同的问题，它们都是将多类划分问题机械的划分为若干二类划分问题进行分类，之后应用投票策略将这些二类划分的结果合并起来成为多类划分的结果。在这个过程中势必会产生一些误差，从而影响最终的分类结果。

Crammer 和 Singer[41]提出了一种基于 Max-Margin 思想的直接对多类划分问题进行分类的方法。假设给定观察样本$S=\{(x_1,y_1),...,(x_l,y_l)\}$，$y\in\{1,2,...,k\}$，我们通过最大化正确的标记$t(\boldsymbol{x})$与其它候选标记间的差距$\gamma$，从而标记的可信度，如公式(3.15)所示：

$$wf(x,t(x)) - wf(x,y) \geq \gamma \tag{3.15}$$

其中，$f(x,y)$ 是特征函数，w 是特征函数的权重，它们的定义方式与第 2 章中 CRFs 模型中相同。

Taskar 将此间隔(margin)扩展，提出间隔(margin)应记录下正确标记序列 $t(x)$ 与其它标记序列 y 之间的不同原子标记的个数，如公式(3.16)所示：

$$\begin{aligned} \Delta t_x(y) &= \sum_{i=1}^{l} \Delta t_x(y_i) \\ \Delta t_x(y_i) &\equiv I(y_i \neq (t(x))_i). \end{aligned} \tag{3.16}$$

这样我们将此多类划分问题转化为以下的优化问题：

$$\begin{aligned} &\min_{w} \frac{1}{2}\|w\|^2 \\ &s.t.\ wf(x,t(x)) - wf(x,y) \geq \Delta t_x(y) \quad \forall x, y \end{aligned} \tag{3.17}$$

这样，我们得到线性约束的凸二次规划问题(3.17)：

同样，类似于第 3.1.3 节所述的二类 SVM 中的线性不可分的情况，若没有符合条件的 w，我们可以引入松弛因子 $\xi_i \geq 0$，可以得到“软化”的新约束条件：

$$s.t.\ wf(x,t(x)) - wf(x,y) \geq \Delta t_x(y) - \xi_i \quad \forall x, y$$

显然，当 ξ_i 充分大时，样本 (x_i, y_i) 总可以满足约束条件，但另一方面，和项 $\sum_{i=1}^{l} \xi_i$ 与样本的分类错误相关并且体现了经验风险，必须限制它的大小，我们引入惩罚因子 C，其中 $C>0$ 是一个常数，它控制对错分样本的惩罚程度，在分类器的复杂度和经验风险之间进行权衡，C 越大表示对错误的惩罚越重。因此，我们得到软化的分类器的优化问题：

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{x} \xi_x \\ &s.t.\ wf(x,t(x)) - wf(x,y) \geq \Delta t_x(y) - \xi_x \quad \forall x, y \end{aligned} \tag{3.18}$$

转化为相应的对偶问题为：

$$\begin{aligned} &\max\ -\frac{1}{2}\left\|\sum_{x,y} \alpha_x(y)\Delta f_x(y)\right\|^2 + \sum_{x,y} \alpha_x(y)\Delta t_x(y)\ ; \\ &s.t.\ \ \alpha_x(y) \geq 0,\ \ \forall x, y, \\ &\qquad \sum_{y} \alpha_x(y) = C,\ \ \forall x \end{aligned} \tag{3.19}$$

其中 $\Delta f_x(y) = f(x,t(x)) - f(x,y)$.

引入核函数得：

$$\max\ -\frac{1}{2}\sum_{x,y}\sum_{x',y'}\alpha_x(y)\alpha_{x'}(y')K(\Delta f_x(y),\Delta f_{x'}(y'))+\sum_{x,y}\alpha_x(y)\Delta t_x(y);$$

$$s.t.\ \alpha_x(y)\geq 0,\ \forall x,y,$$

$$\sum_y \alpha_x(y)=C,\ \forall x, \tag{3.20}$$

3.2 Max-Margin Markov Networks

Max-Margin Markov Networks(M^3Net) 模型于 2003 年，被 Ben Taskar 提出，M^3Net 的思想实际上是上文提到的 Crammer and Singer 的多类 SVM 思想在图结构的扩展。

通过 SVM 进行序列标注任务时，我们只关注于每个独立样本，而忽略了样本间的联系，而我们知道在实际任务中，相邻的节点间会有联系，这些联系对我们正确求解问题是很有帮助的。如第 2 章中我们介绍的判别无向图模型(CRFs)，它们将模型建立在无向图上，可以充分考虑这些结构化信息。我们同样可以将 Max-Margin 原理应用与无向图模型上，这样就将第 3.1.4 节中提出的多类 SVM 模型进行了扩展，便得到了 Max-Margin Markov Networks。

如前所述，SVM 模型将每个节点割裂开考虑，忽略了它们之间的联系，而 Max-Margin Markov Networks 模型中通过定义转移特征函数 $f(\boldsymbol{x},y_i,y_j)$=0 or 1，可以考虑到节点 i 与节点 j 之间的联系，具体来说，就是在处理节点 i 时，会同时考虑节点 j 的标记(也可考虑更多节点，与定义的无向图结构有关)。

为了更好的评价公式(3.20)中 $\alpha_x(y)$，我们定义 marginal dual variables：

$$\mu_x(y_i,y_j)=\sum_{y\sim[y_i,y_j]}\alpha_x(y),\ \forall x,y_i,y_j \tag{3.21}$$

$$\mu_x(y_i)=\sum_{y\sim[y_i]}\alpha_x(y),\ \forall x,y_i \tag{3.22}$$

那么公式(3.19)可以转换为公式(3.23)：

$$\max\ -\frac{1}{2}\left\|\sum_x\sum_{y_i,y_j}\mu_x(y_i,y_j)\Delta f_x(y_i,y_j)\right\|^2+\sum_x\sum_{i,y_i}\mu_x(y_i)\Delta t_x(y_i);$$

$$s.t.\ \sum_{y_i}\mu_x(y_i,y_j)=\mu_x(y_j),\ \forall y_j,\ \forall(i,j)\in E,\forall x,$$

$$\mu_x(y_i,y_j)\geq 0,\ \forall x,y_i,y_j,\ \forall(i,j)\in E,$$

$$\sum_{y_i}\mu_x(y_i)=C,\ \forall x,i,y_i. \tag{3.23}$$

同样的，将其转换为对偶形式并引入核函数得：

$$\max\ -\frac{1}{2}\sum_{x,x}\sum_{y_i,y_j}\sum_{y_r,y_s}\mu_x(y_i,y_j)\mu_{x'}(y_r,y_s)K(f_x(y_i,y_j),f_{x'}(y_r,y_s))\ +\sum_{x}\sum_{i,y_i}\mu_x(y_i)\Delta t_x(y_i);$$

$$s.t.\ \sum_{y_i}\mu_x(y_i,y_j)=\mu_x(y_j),\quad \forall y_j,\ \forall(i,j)\in E,\ \forall x,$$

$$\mu_x(y_i,y_j)\geq 0,\quad \forall x,y_i,y_j,\quad \forall(i,j)\in E,$$

$$\sum_{y_i}\mu_x(y_i)=C,\quad \forall x,i,y_i. \tag{3.24}$$

3.3 序列最小优化(SMO)

从公式(3.24)可以看出 M^3Net 的学习算法实际是针对凸二次规划问题进行求解的问题。求解二次规划问题有很多种方法，本文介绍一种较为常用的序列最小优化(SMO)方法。

1998 年，John C. Platt 提出 SMO[42](Sequential Minimal Optimization)算法。将工作样本集的规模减到最小两个样本。之所以需要两个样本是因为等式线性约束的存在使得同时至少有两个 Lagrange 乘子发生变化。由于只有两个变量，而且应用等式约束可以将其中一个用另一个表示出来， 所以迭代过程中每一步的子问题的最优解可以直接用解析的方法求出来。这样，算法避开了复杂的数值求解优化问题的过程；此外，Platt 还设计了一个两层嵌套循环分别选择进入工作样本集的样本， 这种启发式策略加快了算法的收敛速度。SMO 算法的主要优点在于: 两个变量的联合最优化问题可以通过解析求解，因而不需要迭代地求解二次规划问题，不需要专门的优化软件包。与通常的分解算法比较，尽管它可能需要更过的迭代步，但是由于每步之需要很少的计算量，该算法常表现出整体的快速收敛性质。但是子问题的规模和迭代的次数是一对矛盾，SMO 实际上是将求解子问题的耗费转嫁到迭代上，然后在迭代上寻求快速算法。此外，在挑选工作集时，SMO 算法所采用的策略是针对 Lagrange 乘子的改进而不是针对目标函数的，因而还有改进的空间。

4 条件随机场(CRFs)命名实体识别的研究

本文深入研究了当前中文命名实体识别的各种实现方法，由于条件随机域(CRFs)已表现出很多优于已有机器学习方法的性能，因此本章采用基于条件随机域进行中文命名实体自动识别。在第 4.3 节，采用单纯的 CRFs 模型进行命名实体识别；在第 4.4 节和第 4.5 节分别介绍了两种基于 CRFs 的混合模型，应用于命名实体识别任务。

4.1 BIO 分类标记

命名实体识别任务可以抽象为序列标注问题，所以需采用恰当的标记来表示序列的标注结果。1999 年 Tjong Kim Sang 等人提出了四种短语组块的表示方法，分别是 IOB1，IOB2，IOE1 和 IOE2。Uchimoto 于 2000 年提出了 Start/End 模型来表示短语组块，后来在此模型的基础上增加了三个标志：PRE，POST 和 MID 形成了 S/E+模型。
下面是这六种模型的简单描述：

IOB1：对于短语组块 X，如果两个短语组块并列出现，那么第二个实体的第一个字符标记成‘B-X’其余字符标记成‘I-X’。

IOB2：对于短语组块 X，第一个字符被标记成成‘B-X’，其余字符标记成‘I-X’。

IOE1：对于短语组块 X，如果有两个短语组块并列出现，则第一个实体的最后一个字符标记成‘E-X’，其它字符记为‘I-X’。

IOE2：对于短语组块 X，X 的最后一个字符标记成‘E-X’其它字符标记成‘I-X’。

S/E(Start/End)：对于短语组块 X，如果 X 是单个字符，则标记成‘S-X’，如果 X 有两个或两个以上字符组成，则第一个字符标记成‘B-X’，最后一个字符标记成‘E-X’，上述情况之外的所有字符标记为‘I-X’。

S/E+：此模型对短语组块内部的标记和 S/E 模型一样，但是在短语组块的上下文标记上略有差别。对于短语组块 X，如果 X 前面的字符不属于短语组块字符，则此字符标记成‘PRE-X’，同理，如果 X 后面的字符不属于短语组块字符，则次字符标记成。

‘POST-X’，如果两个实体中间存在一个非短语组块字符，则次字符标记成‘MID-X’。

本文采用 IBO2 的组块(Chunk)表达方法来标识命名实体，即将每个字分为三类：B-命名实体首字、I-命名实体中部、O-命名实体外部，这里一个组块(*BI* 或 *B*)视为一个命名实体。对训练文本中的每个字进行 IBO2 标注，即 $y_i \in \{B, I, O\}$，这样，用 CRFs 识别中文文本中的命名实体就是对文本中的每个字进行 *B*，*I*，*O* 标记。

4.2 命名实体特征的抽取

基于 CRFs 的中文命名实体识别，关键在于抽取命名实体的合适特性。通过对中文命名实体的特点进行分析定义命名实体的特征。

由于命名实体识别是对自动分词结果进行的，分词错误可能会影响命名实体的正确识别，如：

李/长/顺利/用/这次/机会/。

兴/城市/原/种/场/种子公司/还/拖欠/。

为了解决分词错误导致命名实体的错误识别，这里，必须将每一个词分解为一个一个的字，按字抽取特性，最后对每一个字进行分类识别。

下面分别说明人名识别和地名识别所选取的特征。

4.2.1 人名特征的抽取

根据中文文本中人名的特点，抽取了以下特征(表 4.1 给出了人名识别所选取的特性类型及相应值)。

(1)“单字”特性指该字本身

由于中文姓名的姓氏用字相对比较集中，名字用字分布较姓氏虽然要分散，但相对整个汉字集而言依然相对集中。针对这一特点，“单字”可以作为人名的一个特性。

表 4.1 人名特征的类型及相应值

Tab. 4.1 Person name features and their values

特性类型	值
单字(Character)	字本身
单字词性(POS)	n-B, v-I, p-S, …
是否在姓氏表中(PSur)	Y or N
人名用字的概率(PN)	Y or N
是否为人名的前一个字(BeforeP)	Y or N
是否为人名的后一个字(BehindP)	Y or N

(2)“基于字的词性”

此特征为该字所属词的词性加上其位置属性，标注方法如表 4.2 所示。例如：若一个词包含三个字，第一、二、三个字的词性标注分别为：词性-B、词性-I、词性-E，单字词的词性标注为：词性-S。其中“词性”为该词(多字词或单字词)的词性，这里采用北大词性标注规范。

表 4.2 基于字的词性标注方法

Tab. 4.2 POS tags in a word

词性标注	字类型
词性—S	单字词
词性—B	多字词首字
词性—I	多字词(至少三字词)中间字
词性—E	多字词尾字

(3) 判断当前字是否在姓氏表中

《姓氏人名用字分析统计》对人口普查中抽出的2.5万人名(去掉重复后总计174900个)进行了统计：在中国使用的有 737 个姓氏，姓氏虽多，但使用集中在少数大姓上，姓氏使用出现次数在 10 次以上的有 379 个，约占频率为 99.085%；剩下的为出现次数在 10 次以下的有 350 个，只占单姓姓氏频度总数的 0.643%，其中仅出现一次的有 143 个，共占频度总数的 0.144%。因为姓氏在中文姓名中出现的频度很高，所以“是否在姓氏表中”是中文姓名的一个很重要特性，如果该字为中国姓氏，则该特性的值为 Y，否则为 N。

(4) 单字用作姓名的概率

从 98 年 1 月《人民日报》中抽取全部人名得到一张姓名表，然后计算表中每个汉字用作人名的概率，若此概率大于某个阈值，则其属性值为 Y，否则为 N。用字概率计算公式如下：

$$\text{姓名用字概率} = \frac{\text{汉字作为姓名出现的次数}}{\text{汉字出现的总次数}} \times 100\%$$

(5) 单字用作人名之前一个字的概率

人名之前(后)的一个字可以在一定程度上帮助系统识别人名，因此我们从训练语料中抽取每个汉字用作人名前一个字的概率，若此概率大于某个阈值，则其属性值为 Y，否则为 N。用字概率计算公式如下：

$$\text{姓名用字概率} = \frac{\text{汉字作为姓名前字出现的次数}}{\text{汉字出现的总次数}} \times 100\%$$

(6) 单字用作人名之后一个字的概率

从训练语料中抽取每个汉字用作人名后一个字的概率，若此概率大于某个阈值，则其属性值为 Y，否则为 N。用字概率计算公式如下[43]：

$$\text{姓名用字概率} = \frac{\text{汉字作为姓名后字出现的次数}}{\text{汉字出现的总次数}} \times 100\%$$

特征模板是系统抽取特征遵循的规则，一个科学的特征模板可以大大提高系统的性能。本文定义的人名识别的特征模板如下：

- Character(n) (n=-2,-1,0,+1,+2)
- PSur(n) (n=-2,-1,0,+1,+2)
- POS(n) (n=-2,-1,0,+1,+2)
- PN(n) (n=-2,-1,0,+1,+2)
- BeforeP(n) (n=-2,-1,0,+1,+2)
- BehindP(n) (n=-2,-1,0,+1,+2)
- PSur(n)PN(n+1) (n=-1,0,+1)
- PSur(n)PN(n+1) PN(n+2) (n=-1,0,+1)
- BeforeP(n-1)PSur(n)PN(n+1) (n=-1,0,+1)
- BeforeP(n-1) PSur(n)PN(n+1) PN(n+2)(n=-1,0,+1)

4.2.2 地名特征的抽取

根据中文地名的特点，抽取了以下特征，表 4.3 给出了地名识别所选取的特性类型及相应值)：

表 4.3 地名特征的类型及相应值

Tab. 4.3 Location name features and their values

特性类型	值
单字(Character)	字本身
单字词性(POS)	n-B, v-I, p-S, …
是否在地名特征表中(LC)	Y or N
是否为地名的前一个字(BeforeL)	Y or N
是否为地名的后一个字(BehindL)	Y or N

(1)“单字”特征指该字本身

地名用字可以充分反映地名的特点，因此将“单字本身”作为地名的特征。

(2)“基于字的词性”

与人名的此特征相同，标注方法如表 4.2 所示 。

(3) 是否在特征词表中

因为地名结尾经常有地名特征词出现，所以“是否在特征词表中”是地名的一个很重要特征，如果该字为地名特征词，如：“省”、“市”等，则该特征的值为 Y，否则为 N。

(4) 单字用作地名之前一个字的概率

地名之前(后)一个字可以在一定程度上帮助系统识别地名，因此我们从训练语料中抽取每个汉字用作地名前一个字的概率，若此概率大于某个阈值，则其属性值为 Y，否则为 N。

(5) 单字用作地名之后一个字的概率

我们从训练语料中抽取每个汉字用作地名后一个字的概率，若此概率大于某个阈值，则其属性值为 Y，否则为 N。

地名识别的特征模板如下定义：

- POS(n) (n=-2,-1,0,+1,+2)
- Character(n) (n=-2,-1,0,+1,+2)
- LC(n) (n=-2,-1,0,+1,+2)
- BeforeL(n) (n=-2,-1,0,+1,+2)
- BehindL(n) (n=-2,-1,0,+1,+2)
- LC(n)BehindL(n+1) (n=-1,0,+1)

4.3 基于 CRFs 的中文命名实体识别

基于 CRFs 的人名识别和地名识别的过程相同，下面以人名识别为例来说明。具体的识别步骤如下：

(1) 对训练语料及测试语料进行自动分词和词性标注(基于字的标注)，建立训练集和测试集。

① 对人工标注好的训练语料重新标注：去掉人工标注结果还原到原始文本，并记录人名标注位置，然后用 ICTCLAS 系统[44](ICTCLAS 是中科院开发的基于层叠马尔可夫模型的分词和词性标注系统)进行自动分词和标注系统，并进行基于字的词性标注(词性-S，B，I，E)，再根据记录人名的位置对语料中的每个字进行 IBO2 自动标注；用相同方法对测试语料同样进行自动分词和基于字的词性标注。

② 建立训练集和测试集。

(2) 基于 CRFs 模型对训练集进行学习。学习的过程主要分为生成特征函数和训练得到每个特征函数的权重两部分，以下分别对这两部分进行说明。

① 生成特征函数：

下面通过一个实例具体说明 CRFs 系统是如何生成特征函数的。

假设训练集中的一个观察序列为 $\boldsymbol{x}$：“主席胡锦涛”，它相应的状态序列为 $\boldsymbol{y}$：“*OOBII*”，我们对此 $\boldsymbol{x}$，$\boldsymbol{y}$ 序列进行训练(假设以当前的用字作为特征)，系统从起始位置(i = 1)遍历序列，遍历到第三个样本时，以 i = 3(x_i = “胡”，y_i = “*B*”)为例，系

统会产生若干转移特征函数$t_i(y_{i-1}, y_i, \boldsymbol{x}, i)$和若干状态特征函数$s_j(y_i, \boldsymbol{x}, i)$。其中$y_{i-1}$，$y_i$分别是前一个位置的标记和当前位置的标记，由于$y_{i-1}$ y_i共有 9 种不同的组合值(“*BB*”，“*BI*”，“*BO*”，“*IB*”，“*II*”，“*IO*”，“*OB*”，“*OI*” 和“*OO*”)，同样y_i共有 3 种不同的候选标记(“*B*”，“*I*”，“*O*”)，所以当系统遍历到“胡”(i = 3)时，会产生 9 种转移特征函数和 3 种状态特征函数，它们分别是：

$$t_1(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}B\text{","}B\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_2(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}B\text{","}I\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_3(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}B\text{","}O\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_4(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}I\text{","}B\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_5(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}I\text{","}I\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_6(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}I\text{","}O\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_7(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}O\text{","}B\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_8(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}O\text{","}I\text{"} \\ 0, & \text{其它} \end{cases}$$

$$t_9(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_{i-1}, y_i \text{ 是 "}O\text{","}O\text{"} \\ 0, & \text{其它} \end{cases}$$

$$s_1(y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡", } y_i \text{ 是 "}B\text{"} \\ 0, & \text{其它} \end{cases}$$

$$s_2(y_i, \boldsymbol{x}, i) = \begin{cases} 1 & , \quad \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡"}, y_i \text{ 是 "}I\text{"} \\ 0 & , \quad \text{其它} \end{cases}$$

$$s_3(y_i, \boldsymbol{x}, i) = \begin{cases} 1 & , \quad \boldsymbol{x} \text{ 在 } i \text{ 位置是 "胡"}, y_i \text{ 是 "}O\text{"} \\ 0 & , \quad \text{其它} \end{cases}$$

CRFs 模型中产生的特征函数都是二值的，每个二值特征函数相当于一个条件，如果该特征函数的条件被满足，那么此特征函数的值为 1，否则特征函数的值为 0。由于上例中，当 x_3 = "胡" 时，y_{i-1} = "O"，而 y_i = "B"，可见此时转移特征函数 $t_7(y_{i-1}, y_i, \boldsymbol{x}, i)$ 和状态特征函数 $s_1(y_i, \boldsymbol{x}, i)$ 的条件被满足，所以只有转移特征函数 $t_7(y_{i-1}, y_i, \boldsymbol{x}, i)$ 和状态特征函数 $s_1(y_i, \boldsymbol{x}, i)$ 的值为 1，其余特征函数的值都为 0。

② 训练得到每个特征函数的权重

第一步生成了若干特征函数，每个特征函数对于最终的标注都会有些贡献，然而它们的贡献不是完全相同的，因而我们需要通过训练得到每个特征函数的权重，如第 2.2.3 节所述，本文使用 L-BFGS 算法实现对目标函数的优化求解。L-BFGS 是一种充分利用以前的梯度和修改值来近似曲率值的一阶方法，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 训练只要求提供似然函数的一阶导数。

具体训练算法如下：

Step 1. 初始化。将语料库划分成 K 个训练单位 $\{\boldsymbol{x}^k \mid k = 1...K\}$。$\boldsymbol{x}^k$ 通常是一个句子或段落。令特征权重向量 λ、特征权重梯度向量 $\Delta\lambda$、目标函数 L_A 初值为 0。

Step 2. 计算特征梯度向量 $\Delta\lambda$。

① 若 $k < K$，从语料库中取出一个训练单位 $\boldsymbol{x}^k$，$k = k+1$；否则转 Step 3。

② 构建训练单位 $\boldsymbol{x}^k$ 的全切分图，并用 λ 计算矩阵 $\boldsymbol{B}$，$\boldsymbol{E}$ 以及 $\boldsymbol{a}$ 和 $\boldsymbol{\beta}$。修正 $\Delta\lambda$。

③ 用公式计算目标函数值 L_A，转 Step 2。

Step 3. 计算特征权重 λ。

将梯度向量 $\Delta\lambda$ 和目标函数值 L_A 代入 L-BFGS 算法器中，得到修正后的 λ。如果满足 L-BFGS 算法的停止条件或超过最大迭代次数，转 Step 4；否则，转 Step 2。

Step 4. 输出所有特征及相应的 λ。

(3) CRFs 模型进行测试

在测试过程中，系统同样遍历测试集，并遵循与训练过程相同的原则生成若干特征函数，根据 CRFs 公式计算得到每个节点各个候选标记的概率，之后通过 Viterbi 算法解码得到最优标记的序列。

基于 CRFs 模型进行命名实体识别的具体流程如图 4.1 所示：

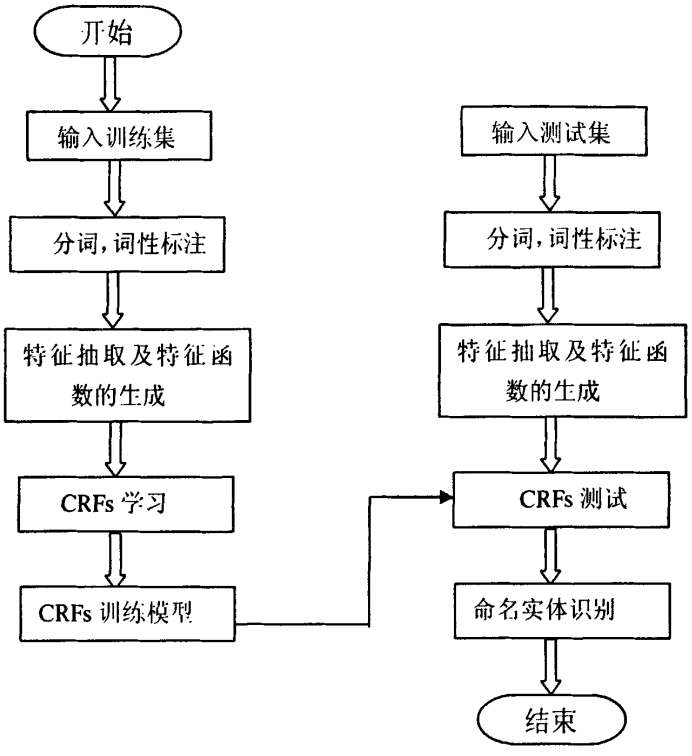

图 4.1 CRFs 命名实体识别系统流程图

Fig. 4.1 Recognition process of Chinese NER based on CRFs

4.4 基于 CRFs 与边界模板的人名识别

为了优化 CRFs 的识别效果，本文对 CRFs 识别错误的情况进行分析发现，CRFs 系统给出的错误标记大部分拥有较低的边缘概率，如果采取一种较好的方法对边缘概率较低的样本进行重新标记，便会进一步改善中文命名实体识别的效果。边界模板方法被用作人名识别并取得较高的准确率[45]，因此，本文提出了一种基于 CRFs 与边界模板的混合模型进行人名识别，具体的说，如果 CRFs 模型中样本的边缘概率较低，我们使用边界模板方法代替 CRFs 模型对其进行标记，否则仍使用 CRFs 模型进行标记。实验表明，该混合模型综合了 CRFs 和边界模板两种方法的优势，识别效果好于单纯的 CRFs 方法。

4.4.1 边界模板

如第 1.1.2 节所述，人名的用字较为随意，人名又包括简称、外文译名，这给人名的识别带来了不小的困难。然而，不管人名如何变化，人名的前词和后词是存在一些规律的。我们正是通过人名前词和后词的组合来建立边界模板的。

(1) 边界词语

如果汉语句子中含有单词序列W_1pW_2，其中 p 为人名，则W_1称为人名左边界，W_2为人名右边界。W_1，W_2均可为空，W_1为空时记W_1 = BOS，表示句子的开始；W_2为空时记W_2 = EOS，表示句子的结束。

(2) 边界模板

我们称四元组$q = < W_1, f_{lb}(W_1), W_2, f_{rb}(W_2) >$为人名边界模板，其中$W_1$，$W_2$分别为人名的左、右边界词语，$f_{lb}(W_1)$为在训练语料库中作为人名左边界的频度，$f_{rb}(W_2)$为在训练语料库中$W_2$作为人名右边界的频度，定义边界模板的频度$f(q)$为：

$$f(q) = \begin{cases} f_{lb}(W_1) + f_{rb}(W_2), f_{lb}(W_1) f_{rb}(W_2) \neq 0 \\ 0, otherwise \end{cases} \tag{4.1}$$

在人名识别过程中，每个候选人名(左右两侧词语分别为W_1，W_2)对应一个边界模板$q = < W_1, f_{lb}(W_1), W_2, f_{rb}(W_2) >$，如果$f(q) = 0$则$f_{lb}(W_1)$和$f_{rb}(W_2)$至少有一个为 0，即该模板的左右边界词至少有一个在训练语料中没有出现，此时我们要淘汰该候选人名。这一严格的标准确保了算法定位人名的准确性，它对召回率的影响(数据稀疏引起)可通过篇章范围内的扩散操作加以弥补。

(3) 扩散操作

在识别人名时，如果某位置上的字符串被识别为人名，则将文章中出现的所有该字符串都被认为是人名，我们把这一动作定义为扩散操作。其中，此扩散操作只针对长度大于 1 个字符的人名。由于只含一个字符的姓名或人名(如“高”、“和”)在文章中以非人名的身份出现的频率较高，如果被扩散会导致大量人名识别错误。

4.4.2 基于边界模板的人名识别模型

本模型的基本思想是：首先构造候选人名的边界模板，并计算候选人名边界模板的频度，如果此频度为 0，则认为此候选词不是人名，否则如果此频度值大于阈值，则输出 Y，否则输出 N。最后，进行扩展操作，得到最终判别结果。

具体方法如下：

(1) 首先通过训练语料统计每个词作为人名前词和后词的概率$f_{lb}(W_1)$和$f_{rb}(W_2)$。

(2) 对于每个候选人名(左右两侧词语分别为W_1，W_2)构建一个边界模板$q=<w_1, f_{lb}(w_1), w_2, f_{rb}(w_2)>$。

(3) 计算边界模板的频度$f(q)$

(4) if $f(q)=0$

输出 N

else if $f(q)>\varepsilon$

输出 Y

else 输出 N

(5) 进行扩散操作(第 4.4.1 节所述)

(6) 得到最终标记结果

4.4.3 基于 CRFs 与边界模板的人名识别方法

本文对 CRFs 模型标记的结果进行分析发现，大多数的错误标记拥有较低的边缘概率，边缘概率的定义如公式(2.26)所示，因此，CRFs 模型中每个标记的边缘概率可以看成 CRFs 模型对其给出标记的信心，当边缘概率较高时，CRFs 很可能给出的标记是正确的，否则 CRFs 很可能给出了错误标记，下面举例说明边缘概率的含义。

例子 1：测试集中有一个观察序列x：“胡锦涛”，它可能的状态序列有 27 种，分别为“*BBB*”，“*BBI*”，“*BBO*”，“*BIB*”，“*BII*”，“*BIO*”，“*BOB*”，“*BOI*”，“*BOO*”，“*IBB*”，“*IBI*”，“*IBO*”，“*IIB*”，“*III*”，“*IIO*”，“*IOB*”，“*IOI*”，“*IOO*”，“*OBB*”，“*OBI*”，“*OBO*”，“*OIB*”，“*OII*”，“*OIO*”，“*OOB*”，“*OOI*”，“*OOO*”。其中每一个候选的状态标记都会有一个概率值，显然它们的概率之和是 1。

由于 CRFs 模型求解的是最优序列，因此单个标记的边缘概率也是基于序列的概率而求解的。具体来说，第一个节点标为“*B*”的边缘概率就是所有第一个标记为“*B*”的状态序列的概率之和。27 种候选状态序列中共有 9 个序列的第一个标记为“*B*”，它们是“*BBB*”，“*BBI*”，“*BBO*”，“*BIB*”，“*BII*”，“*BIO*”，“*BOB*”，“*BOI*”，“*BOO*”。则第一个节点标记为“*B*”的边缘概率$P(y_1="B"|x)$，就是这 9 种标记的概率之和：

$$\begin{aligned}P(y_1="B"|x)=&P(y="BBB"|x)+P(y="BBI"|x)+P(y="BBO"|x)\\&+P(y="BIB"|x)+P(y="BII"|x)+P(y="BIO"|x)\\&+P(y="BOB"|x)+P(y="BOI"|x)+P(y="BOO"|x)\end{aligned}$$

经过 CRFs 模型求解，得到第一个节点标记为“*B*”的边缘概率为 0.995，即 $P(y_1="B"|x)=0.995$。由于该边缘概率的值较大，那么 CRFs 给出第一个节点是“*B*”状态很可能是正确的。

例子 2：测试集中有一个观察序列 x：“望子成龙”，它可能的状态序列有 81 种，同样每一个候选的状态标记都会有一个概率值，显然它们的概率之和也是 1。

同样，第三个节点标为“*B*”的边缘概率就是所有第三个标记为“*B*”的状态序列的概率之和。81 种候选状态序列中共有 27 个序列的第三个标记为“*B*”，则第三个节点标记为“*B*”的边缘概率，就是这 27 种标记的概率之和：

$$\begin{aligned}P(y_3="B"|x)=&P(y="BBBB"|x)+P(y="BBBI"|x)+P(y="BBBO"|x)\\&+P(y="BIBB"|x)+P(y="BIBI"|x)+P(y="BIBO"|x)\\&+P(y="BOBB"|x)+P(y="BOBI"|x)+P(y="BOBO"|x)\\&\cdots\qquad\qquad\cdots\qquad\qquad\cdots\\&+P(y="OOBB"|x)+P(y="OOBI"|x)+P(y="OOBO"|x)\end{aligned}$$

经过 CRFs 模型求解，得到第三个节点标记为“*B*”的边缘概率为 0.875，即 $P(y_3="B"|x)=0.875$，由于该边缘概率的值相对较低，那么 CRFs 给出第三个节点是“*B*”状态很可能是错误的。因此，本文引入边界模板的方法对那部分边缘概率较低的节点进行重新标注，以修改 CRFs 方法中可能存在的错误。那么这个区分两种方法的阈值究竟取多少较为合理，可以通过实验来确定。下面我们详细说明 CRFs 与边界模板混合模型的中文人名实体识别的方法。

具体算法如下：

(1) 对训练语料和测试语料进行自动分词和词性标注，建立训练集、测试集。

① 对人工标注好的训练语料用 ICTCLAS 系统进行自动分词，并进行基于字的词性标注(词性-S，B，I，E)，再根据记录人名的位置对语料中的每个字进行 IBO2 自动标注；用相同方法对测试语料进行自动分词和基于字的词性标注。

② 建立训练集和测试集。

(2) 采用 CRFs 算法对训练集进行训练，对测试集进行标记并计算每个样本的边缘概率。(其中(1) (2)与第 4.3 节方法相同)

(3) if 边缘概率< ε

用边界模板方法(第 4.4.2 节所述)对样本进行标记

else

保持 CRFs 的标记

(4) 得到最终标记结果

基于 CRFs 与边界模板的人名识别的流程图，如图 4.2 所示：

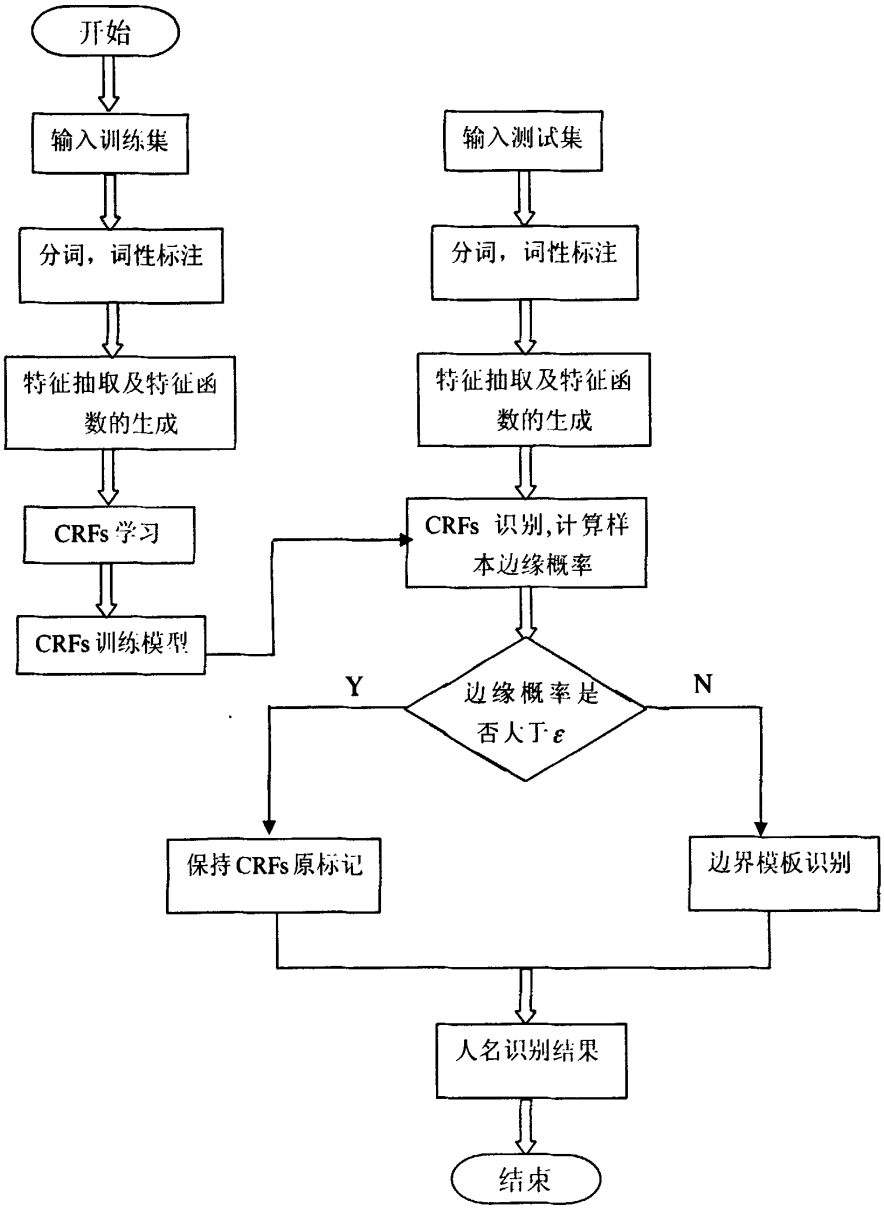

图 4.2　基于 CRFs 与边界模板的人名识别系统流程图

Fig. 4.2　Recognition process of person NER based on CRFs and boundary templates

4.5 基于 CRFs 与概率统计的命名实体识别

如第 4.4 节所述，CRFs 系统给出结果中的错误标记大部分拥有较低的边缘概率，在本节中，介绍一种概率统计的方法对这部分样本进行修正，期望进一步改善中文命名实体识别的效果。概率统计方法是中文命名实体研究中较为成熟的技术，已经得到了广泛的应用，因此本文提出一种基于 CRFs 和概率统计组合算法的中文命名实体识别模型。具体的说，对于 CRFs 标记中边缘概率较低的测试样本采用概率统计方法代替 CRFs 方法进行识别；对于 CRFs 方法中边缘概率较高的测试样本仍使用 CRFs 标记。实验表明，该混合模型综合了 CRFs 和概率统计两种方法的优势，结果好于单纯的 CRFs 方法。

其中，概率统计方法的基本思想是：首先计算候选命名实体的构词可信度和基于词性的二元接续可信度，然后将得到的构词可信度和二元接续可信度的值代入命名实体评价函数，便可得到该候选命名实体的概率估值。最后比较此概率值和给定的阈值，如果此概率值大于阈值输出 Y，否则输出 N。下面我们将具体给出命名实体识别的概率统计模型。

4.5.1 人名识别的概率统计模型

(1) 所用的资源

为了计算人名的构词可信度和接续可信度，首先要构造人名表，本文对大连理工大学近几年在校生中 50000 人的名字进行统计后建立了姓氏字表(LastName)和名字用字字表(FirstName)。然后从 1998 年《人民日报》上抽取 60 万字的语料，作为基本语料库，统计语料库中的人名的前词的词性和后词的词性，建立人名前后词的词性频度表(PersonPOS)。

(2) 构词可信度

我们将中国人名(PN)定义为：$PN=LF_1F_2$，其中 $L \in LastName$ 是待评价人名中的姓氏，$F_i \in FirstName\ (i=1,2)$ 是待评价人名的第 i 个名字用字。

首先分别对姓 L 及名字用字 F_i 计算其可信度，然后再计算人名 PN 的可信度，公式定义如下：

① 对于任意的 $L \in LastName$，定义其姓氏用字可信度 $P_l(L)$ 如下：

$$P_l(L) = \frac{P_{l0}(L)}{\sum_{y \in LastName} P_{l0}(y)} \tag{4.2}$$

其中，$P_{l0}(L) = \log_2^{N(L)+2}$，$N(L)$ 是汉字串 L(可以是单姓，也可以是复姓)作为姓氏在姓氏表中出现的次数。

② 对于任意的 $F_i \in FirstName$，定义其名字用字可信度 $P_f(F_i)$如下：

$$P_f(F_i) = \frac{P_{f0}(F_i)}{\sum_{y \in FirstName} P_{f0}(y)} \tag{4.3}$$

其中，$P_{f0}(F_i) = \log_2^{N(F_i)+2}$，$N(F_i)$是汉字串 F_i 作为名字的一部分在名字用字表中出现的次数。

③ 人名 PN 的构词可信度定义为：

$$\begin{aligned} LP(PN) &= P_l(L) \times P_f(F_1) && if(PN = LF_1) \\ LP(PN) &= P_l(L) \times C_b \times (P_f(F_1) + P_f(F_2)) && if(PN = LF_1F_2) \end{aligned} \tag{4.4}$$

其中，C_b 是单双名调节系数，我们取 C_b 为 0.844[21]。

(3) 基于词性的二元接续可信度

中国人名在真实文本中有其典型的上下文词性分布特点，例如，“对张帅说”，人名“张帅”的前词的词性是 p-S，后词的词性是 v-S。因此把人名的前词的词性和后词的词性抽取出来，得到人名前后词的词性频率表，并在此基础上计算基于词性的二元接续可信度。

基于词性的二元接续可信度定义为：

$$CP(PN) = \frac{personPOS(\langle lpos, rpos \rangle)}{TotalPOS} \tag{4.5}$$

其中，$lpos$ 是人名的前词的词性，$rpos$ 是的后词的词性，$personPOS(\langle lpos, rpos \rangle)$是表示在词性频度表中，人名 PN 其前字词性为 $lpos$，后字词性为 $rpos$ 时出现的次数。$TotalPOS$ 表示词性频度表中所有人名的前后词的词性信息的总数。

(4) 评价函数

中国人名的评价函数 $TotalFrequency\ (PN)$定义为：

$$TotalFrequency(PN) = \alpha LP(PN) + (1-\alpha) CP(PN) \tag{4.6}$$

其中，$LP(PN)$和 $CP(PN)$分别是公式(4.4)和公式(4.5)中定义的构词可信度和基于词性的二元接续可信度。α 是平衡系数[46]，用于平衡评价函数中构词可信度和基于词性的二元接续可信度的作用。当 α=0.5 时，保证两者对人名评价函数的影响力相同。通过调节 α 的大小，可以调节构词可信度和接续可信度的权重。实验证明当 α=0.4 时，此概率统计模型对人名识别的效果最好，因此本文使用 α=0.4 进行人名识别实验。

4.5.2 地名识别的概率统计模型

(1) 所用的资源

① 地名用字知识库(PlaceName)。收集《中国地名录》中出现的所有地名的用字信息，包括地名首字、地名中字和地名尾字的频度信息。

② 地名上下文词性频度表(PlacePOS)。从 1998 年《人民日报》上抽取 150 万字的语料，作为基本语料库，统计语料库中的地名的前词词性和后词词性，建立地名前后词的词性频度表(PlacePOS)。

(2) 地名构词可信度

我们将中文地名(*LN*)定义为：$LN = F_0F^+S$，其中 F_0 为地名首字，F^+ 为地名中部，$F^+ = F_1...F_n,(i=1,...,n)$，$S$ 为地名尾字。

① 地名首字可信度 $P_h(F_0)$ 定义为：

$$P_h(F_0) = \frac{P_{h0}(F_0)}{P'_{h0}(F_0)}, \tag{4.7}$$

其中：$P_{h0}(F_0) = \log_2(C(F_0)+2)$，$C(F_0)$ 是汉字 F_0 作为地名首字在中国地名库中出现的次数。$P'_{h0}(F_0) = \log_2(C'(F_0)+2)$，$C'(F_0)$ 是汉字 F_0 在中国地名库中出现的总次数。

② 地名中部可信度 $P_f(F^+)$ 定义为：

$$P_f(F^+) = \sum_{i=1}^{n} \frac{P_f(F_i)}{P'_f(F_i)} \tag{4.8}$$

其中：$P_f(F_i) = \log_2(C(F_i)+2)$，$C(F_i)$ 是汉字只作为地名中部在中国地名库中出现的次数。$P'_f(F_i) = \log_2(C'(F_i)+2)$，$C'(F_i)$ 是汉字月在中国地名库中出现的总次数。

③ 地名尾字可信度 $P_l(S)$ 定义为：

$$P_l(S) = \frac{P_l(S)}{P'_l(S)}, \tag{4.9}$$

$P_l(S) = \log_2(C'(S)+2)$，$C(S)$ 是汉字 S 作为地名尾字在中国地名库中出现的次数。

$P'_l(S) = \log_2(C'(S)+2)$，$C'(S)$ 是汉字 S 在中国地名库中出现的总次数。

④ 地名构词可信度：

$$LN = (P_h(F_0)+P_f(F^+)+P_l(S))/len(LN), \tag{4.10}$$

其中：$Len(LN)$ 为地名 LN 的长度。

(3) 基于词性的二元接续可信度

中文地名在文本中出现时都有自己的特点，它们会对上下文的词性分布造成影响。如：“在重庆市举行”，地名“重庆市”前词的词性为 p，后词的词性为 v。把地名的前词词性和后词词性抽取出来，得到地名前后词的词性频度表(PlacePOS)，并在此基础上计算基于词性的二元接续可信度。

基于词性的二元接续可信度$CP(LN)$定义为：

$$CP(LN)=\frac{placePOS(\langle lpos,rpos\rangle)}{TotalPOS} \tag{4.11}$$

其中，$lpos$是地名的前词词性，$rpos$是的后词词性，$placePOS(\langle lpos,rpos\rangle)$表示在词性频度表中，$LN$作为中文地名时前词词性为$lpos$，后词词性为$rpos$时出现的次数。$TotalPOS$表示词性频度表中所有地名的前后词词性信息的总数。

(4) 评价函数

中文地名的评价函数$TotalFrequency(LN)$定义为：

$$TotalFrequency(LN)=\alpha LP(LN)+(1-\alpha)CP(LN) \tag{4.12}$$

其中，$LP(LN)$为公式(4.10)定义的地名构词可信度；$CP(LN)$为公式(4.11)基于词性的二元接续可信度；α为平衡地名构词可信度与基于词性的二元接续可信度的可比性系数，通过调节α的大小，可以调节构词可信度和接续可信度的权重，实验证明当α=0.2时，此概率统计模型对人名识别的效果最好，因此本文使用α=0.2进行人名识别实验[43]。

4.5.3 基于 CRFs 与概率统计的命名实体识别方法

CRFs 是一种优秀的机器学习模型，具有较强的学习能力，但它在边缘概率较低的标记上表现较差，因此我们引入概率统计模型代替 CRFs 模型对这些原子进行标记，它综合了 CRFs 模型与概率统计模型的优点，此模型可以用于分词、命名实体识别等序列标注任务。

CRFs 和概率统计命名实体识别算法的基本思想是：当 CRFs 方法标记的边缘概率大于阈值ε，用 CRFs 标记；否则采用概率统计模型代替 CRFs 模型进行标记。算法如下：

(1) 对训练语料和测试语料进行自动分词和词性标注。

(2) 采用 CRFs 算法对训练集进行训练，对测试集进行标记并计算每个样本的边缘概率。

(其中(1)(2)与第 4.3 节方法相同)

(3) if 边缘概率< ε

用概率统计方法(第 4.5.1 节和第 4.5.2 节所示)对样本标记

else 保持原来 CRFs 的标记

(4) 得到最终标记结果

CRFs 与概率统计的命名实体识别系统的流程图，如图 4.3 所示：

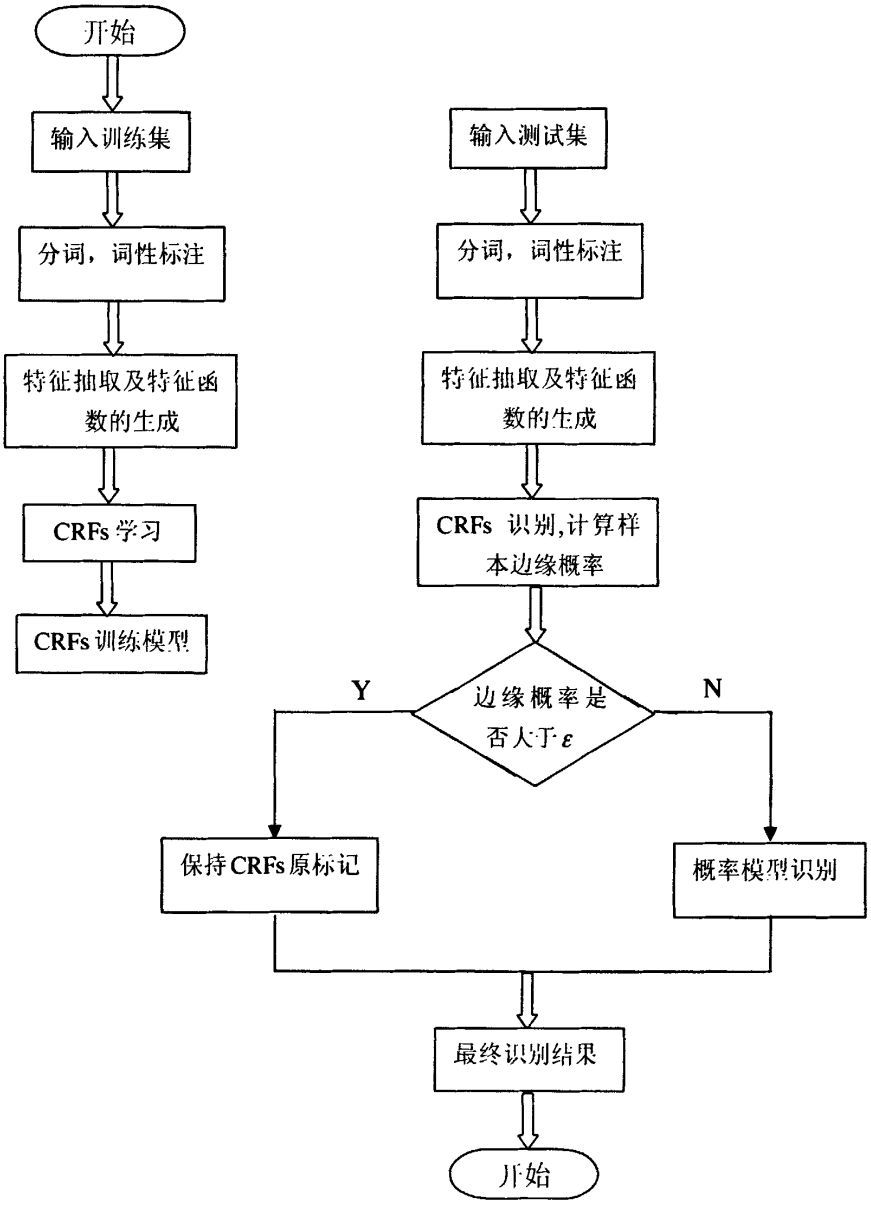

图 4.3 基于 CRFs 与概率统计的中文命名实体识别系统流程图

Fig. 4.3 Recognition process of Chinese NER based on CRFs and statistical method

5 基于 Max-Margin Markov Networks 的地名识别

在这一章中，同样将地名识别任务看成一个序列标注问题，从而采用基于 Max-Margin Markov Networks 的方法进行标注。选取与第 4 章中基于 CRFs 模型中相同的特征与特征模板。

Max-Margin Markov Networks 是一种机器学习模型，它的主要原理是将 Max-Margin 的思想应用于图结构中。Max-Margin 思想是人工智能领域较为常用的思想，就是通过最大化正确标记与其它标记之间的差距来增加给出标记的信心程度。我们熟悉的 SVM 模型就是来源于 Max-Margin 思想，通过最大化两类超平面的距离来增加分类的信心。而 Max-Margin Markov Networks 模型不像 SVM 模型将每个节点割裂开，分别求解每个节点的最优标记，它是构造一个马尔可夫链，考虑到相邻节点的联系，从而求解最优的序列。

在命名实体识别任务中，节点之间的联系是很重要的。如第 4 章中的例子，假设一个观察序列为 $\boldsymbol{x}$：“胡锦涛”，我们知道它的正确的状态序列为“*BII*”，Max-Margin 的思想就是通过最大化正确标记“*BII*”与其它标记之间的差距来增加标记正确的可能性。另外，我们知道相邻节点的标记是有联系的，如序列中第一个节点的标记是“*B*”，那么第二个节点的标记很可能是“*I*”，又如第二个节点的标记是“*I*”，第一个节点不可能是“*O*”。为了考虑到节点之间的联系，Max-Margin Markov Networks 模型实际上将 Max-Margin 思想建立在马尔可夫网络上，从而求解最优的序列。

5.1 Max-Margin Markov Networks 模型的构建

本文根据不同阶数的马尔可夫网络[47]来构建模型。首先，我们定义了一阶 Max-Margin Markov Networks，它是定义在一阶的马尔可夫网络上，只考虑当前节点和前一个节点的关系；另外还定义了二阶 Max-Margin Markov Networks，它是定义在二阶的马尔可夫网络上，考虑当前节点和前两个节点的关系。如图 5.1 和图 5.2 所示：

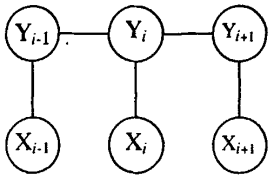

图 5.1 一阶马尔可夫链
Fig. 5.1 Structures of one order Markov Networks

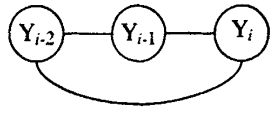

图 5.2 二阶马尔可夫链
Fig. 5.2 Structures of two order Markov Networks

如第 3.1.3 节所述，核函数规定了将样本从低维空间映射到高维空间的方式，从而使样本在高维空间线性可分。

本文选取最简单的线性核函数 $K(\boldsymbol{x},\boldsymbol{x}_i)=<\boldsymbol{x}\cdot\boldsymbol{x}_i>$ 来构建 Max-Margin Markov Networks 模型。

5.2 基于 Max-Margin Markov Networks 的地名识别方法

本章将地名识别任务看出一个序列标注问题，从而，通过基于机器学习模型(Max-Margin Markov Networks)的方法来解决。本节选取与第 4 章相同的特征与特征模板进行实验。

和 CRFs 模型一样，一阶的 Max-Margin Markov Networks 同样定义两种特征函数，分别为转移特征函数 $t_j(y_{i-1},y_i,\boldsymbol{x},i)$ 和状态特征函数 $s_j(y_i,\boldsymbol{x},i)$，其中转移特征函数 $t_j(y_{i-1},y_i,\boldsymbol{x},i)$ 正是为了考虑节点之间的联系而定义的。二阶的 Max-Margin Markov Networks 定义了转移特征函数 $t_j(y_{i-2},y_{i-1},y_i,\boldsymbol{x},i)$，通过定义此类特征函数可以考虑当前节点和前两个节点的关系。M^3Net 模型中所定义的特征函数也是二值的，定义方式和 CRFs 模型相同，见第 4.3 节，这里不在举例说明。

基于 Max-Margin Markov Networks 的命名实体识别方法的具体步骤如下：

(1) 对训练语料及测试语料进行自动分词和词性细标注(基于字的标注)

① 对人工标注好的训练语料重新标注：去掉人工标注结果还原到原始文本，并记录人名标注位置，然后用 ICTCLAS 系统(ICTCLAS 是中科院开发的基于层叠马尔可夫模型的分词和词性标注系统)进行自动分词和标注系统，并进行基于字的词性标注(词性-S，B，I，E)，再根据记录人名的位置对语料中的每个字进行 IBO2 自动标注，用相同方法对测试语料同样进行自动分词和基于字的词性标注。

② 建立训练集和测试集。

(2) 训练部分

① 生成特征函数

系统系统会产生若干转移特征函数 $t_j(y_{i-1},y_i,\boldsymbol{x},i)$ 和若干状态特征函数 $s_j(y_i,\boldsymbol{x},i)$，生成过程于 CRFs 模型中相同。

② 构建二次规划问题

如第 3.2 节所述的原理，构造二次规划问题。

③ 训练得到每个特征函数的权重

利用 SMO 算法训练得到每个特征函数的权重，如第 3.3 所述。

(3) M^3Net 模型进行测试

在测试过程中，系统同样遍历测试集遵循与训练过程同样原则生成若干特征函数，计算得到每个节点各个候选标记的概率，之后通过 Viterbi 算法解码得到最优标记的序列。

基于 Max-Margin Markov Networks 的中文地名识别流程图，如图 5.3 所示。

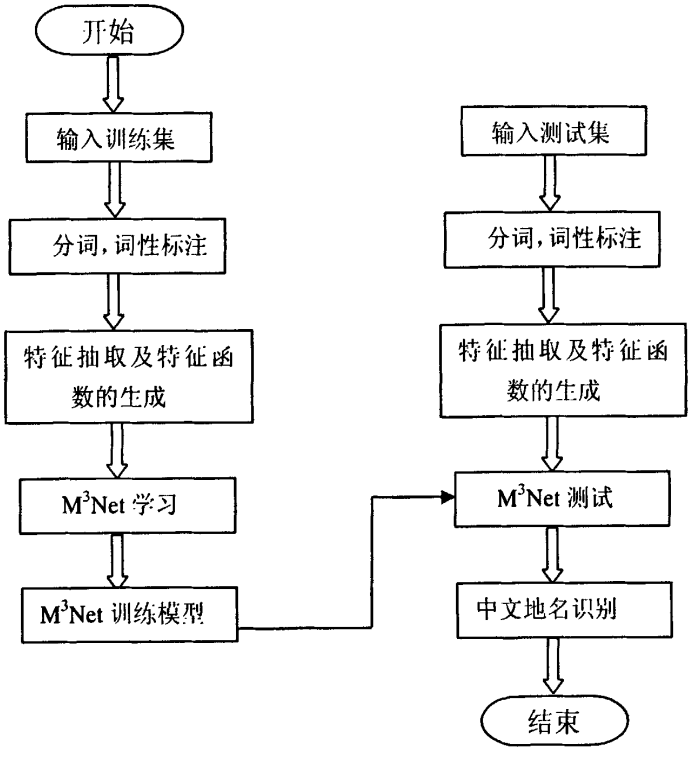

图 5.3 M³Net 地名识别系统流程图

Fig. 5.3 Recognition process of location NER based on M³Net

6 基于概率特征函数的 CRFs 模型

CRFs 模型是目前最优秀的机器学习方法之一，模型中定义的特征函数全部都是 0、1 二值的。在本章中，我们将这些二值特征函数融入概率信息，生成概率特征函数，然后基于这种概率特征函数构造条件随机场。此改进模型可以将学习到的有用的概率信息加入特征函数中，进而提高机器学习能力。在这一章中，我们通过命名实体识别任务验证此模型的学习能力。

6.1 概率特征函数的表示

如第 2 章所述，CRFs 模型中定义的特征函数都是 0、1 二值的，如公式(6.1)和公式(6.2)所示，分别为汉字“郭”和“于”作为人名姓氏的特征函数。显然，“郭”和“于”作为人名姓氏的概率不同。当“郭”出现时，其被用作人名姓氏的概率较大，而“于”出现时被用作人名姓氏的概率较小。CRFs 定义的特征函数(6.1)、(6.2)无法反映这种概率的差异，而这种概率的信息对于命名实体的识别是有帮助的。

$$s_1(y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x}\text{在 } i\text{位置为“郭”，} y_i = \text{"}B\text{"} \\ 0, & \text{其它} \end{cases} \tag{6.1}$$

$$s_2(y_i, \boldsymbol{x}, i) = \begin{cases} 1, & \boldsymbol{x}\text{在 } i\text{位置为“于”，} y_i = \text{"}B\text{"} \\ 0, & \text{其它} \end{cases} \tag{6.2}$$

本章提出一种基于概率的特征函数，即在定义特征函数时加入概率信息。例如：二值特征函数(6.1)、(6.2)对应的概率特征函数为：

$$s\text{-}p_1(y_i, \boldsymbol{x}, i) = \begin{cases} 0.985, & \boldsymbol{x}\text{在 } i\text{位置为“郭”，} y_i = \text{"}B\text{"} \\ 0, & \text{其它} \end{cases} \tag{6.3}$$

$$s\text{-}p_2(y_i, \boldsymbol{x}, i) = \begin{cases} 0.018, & \boldsymbol{x}\text{在 } i\text{位置为“于”，} y_i = \text{"}B\text{"} \\ 0, & \text{其它} \end{cases} \tag{6.4}$$

其中，公式(6.3)中的概率 0.985 表示当前字为“郭”时，当前的状态为“*B*”的概率。同理，公式(6.4)中的概率 0.018 表示当前字为“于”时，当前的状态为“*B*”的概率。

如第 2 章所述，CRFs 模型中定义了两种特征函数，分别是状态特征函数和转移特征函数。同样，转移特征函数的概率形式为：

$$t\text{-}p_7(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} 0.54 \ , & \boldsymbol{x} \text{ 在位置 } i \text{为 "郭"}, y_{i-1}, y_i \text{ is "}O\text{","}B\text{"} \\ 0 \ , & \text{其它} \end{cases} \tag{6.5}$$

其中，公式(6.5)中的概率 0.54 表示当前字为“郭”时，前一个标记与当前标记的组合为“*OB*”的概率。

6.2 概率特征函数的定义

前一节中，介绍了概率特征函数的基本形式。这一节将介绍概率特征函数的定义方法。首先，以状态特征函数(6.3)为例说明，特征函数的概率值的计算方法如下：

$$s\text{-}p_1(y_i, \boldsymbol{x}, i) = \begin{cases} \dfrac{\log_2(2 + num(x_i = \text{"郭"},\ y_i = \text{"}B\text{"}))}{\sum_{y \in BIO} \log_2(2 + num(x_i = \text{"郭"},\ y_i = y))} \ , & \boldsymbol{x} \text{在 } i \text{位置为“郭”，} y_i = \text{"}B\text{"} \\ 0 \ , & \text{其它} \end{cases} \tag{6.6}$$

其中，$num(x_i = \text{"郭"},\ y_i = \text{"}B\text{"})$表示训练语料中当前字为“郭”，当前标记为“*B*”的个数。另外，通过对数运算进行平滑操作。

同样，汉字“郭”的另外两种概率特征函数定义如下：

$$s\text{-}p_2(y_i, \boldsymbol{x}, i) = \begin{cases} \dfrac{\log_2(2 + num(x_i = \text{"郭"},\ y_i = \text{"}I\text{"}))}{\sum_{y \in BIO} \log_2(2 + num(x_i = \text{"郭"},\ y_i = y))} \ , & \boldsymbol{x} \text{在 } i \text{位置为“郭”，} y_i = \text{"}I\text{"} \\ 0 \ , & \text{其它} \end{cases} \tag{6.7}$$

$$s\text{-}p_3(y_i, \boldsymbol{x}, i) = \begin{cases} \dfrac{\log_2(2 + num(x_i = \text{"郭"},\ y_i = \text{"}O\text{"}))}{\sum_{y \in BIO} \log_2(2 + num(x_i = \text{"郭"},\ y_i = y))} \ , & \boldsymbol{x} \text{在 } i \text{位置为“郭”，} y_i = \text{"}O\text{"} \\ 0 \ , & \text{其它} \end{cases} \tag{6.8}$$

转移特征函数也可以按照同样的原理求解概率，以特征函数(6.5)为例：

$$t\text{-}p_7(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} \dfrac{\log_2(2 + num(x_i = \text{"郭"},\ y_{i-1} y_i = \text{"}OB\text{"}))}{\sum_{y'y} \log_2(2 + num(x_i = \text{"郭"},\ y_{i-1} y_i = y'y))} \ , & x_i \text{为 "郭"}, y_{i-1}, y_i \text{ 是 "}O\text{","}B\text{"} \\ 0 \ , & \text{其它} \end{cases}$$

其中，$num(x_i$="郭"，$y_{i-1}y_i$="OB")表示当前字为“郭”时，前一个的标记与当前标记的组合为“OB”的个数。

6.3 基于概率特征函数的 CRFs 的构建

在这一章中，介绍如何基于第 6.2 节中定义的概率特征函数构造条件随机场模型。

如第 2 章所述，在给定观测序列 $\boldsymbol{x}$ 的情况下，CRFs 等定义了标记序列 $\boldsymbol{y}$ 的概率是势函数(Potential Function)乘积的一个归一化形式，其中每个因子形式如式：

$$\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_k \mu_k S_k(y_i, \boldsymbol{x}, i)) \tag{6.9}$$

这里，我们用概率特征函数 $t\text{-}p_j(y_{i-1}, y_i, \boldsymbol{x}, i)$ 和 $S\text{-}p_k(y_i, \boldsymbol{x}, i)$ 来代替 CRFs 模型中的二值特征函数，则势函数的因子被定义为：

$$\exp(\sum_j \lambda_j t\text{-}p_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_k \mu_k S\text{-}p_k(y_i, \boldsymbol{x}, i)) \tag{6.10}$$

因此概率条件随机场对于一个给定观测序列 $\boldsymbol{x}$，其对应的标记序列 $\boldsymbol{y}$ 的概率为公式(6.10)所示：

$$P(\boldsymbol{y} \mid \boldsymbol{x}, \lambda) = \frac{1}{Z(\boldsymbol{x})} \exp(\sum_j \lambda_j F\text{-}p_j(\boldsymbol{y}, \boldsymbol{x})) \tag{6.11}$$

概率条件随机场的参数估计方法与 CRFs 模型相同，如第 2.2.3 节所述。标记过程同样使用 Viterbi 等动态优化方法，求出最优解 $\boldsymbol{y}^*$。

6.4 命名实体识别的实验

在本节中，应用基于概率特征函数的 CRFs 模型进行命名实体识别实验。通过与传统 CRFs 实验结果的对比，来验证改进模型的机器学习能力。具体步骤如下：

(1) 对训练语料及测试语料进行自动分词和词性标注(基于字的标注)，建立训练集和测试集。此步骤与第 4.3 节相同。

(2) 基于改进 CRFs 模型对训练集进行学习。学习的过程主要分为生成特征函数和训练得到每个特征函数的权重两部分，以下分别对这两部分进行说明。

① 生成特征函数

遍历训练集，计算每种特征函数的概率，生成概率特征函数(如第 6.2 节所述)。

② 训练得到每个特征函数的权重

第一步生成了若干特征函数，每个特征函数对于最终的标注都会有些贡献，然而它们的贡献不都相同，因而我们通过训练得到每个特征函数的权重。

(3) 改进 CRFs 模型进行测试

在测试过程中，系统同样遍历测试集，遵循与训练过程相同的原则生成若干特征函数，计算得到每个节点各个候选标记的概率，之后通过 Viterbi 算法解码得到最优标记的序列。

7 实验结果与分析

本文采用精确率(Precision)，召回率(Recall)和 F 值对识别结果进行评价，其中：

识别精确率(Precision)：$P = \frac{\text{正确命名实体数}}{\text{召回命名实体数}} \times 100\%$

识别召回率(Recall)：$R = \frac{\text{正确命名实体数}}{\text{文本命名实体总数}} \times 100\%$

$F\text{值} = \frac{2 \times \text{召回率} \times \text{精确率}}{\text{召回率} + \text{精确率}} \times 100\%$

7.1 基于 CRFs 模型的命名实体识别

本节以 CRFs 作为机器学习的方法，分别采用单纯 CRFs 模型和两种基于 CRFs 的混合模型进行实验。实验采用的训练集与测试集是 Bakeoff2007 NER 任务的 MSRA 语料。训练语料约有 120 万字，其中包含人名 9028 个，地名 18522 个，测试语料约有 22 万字其中包含人名 1864 个，地名 3658 个。

7.1.1 单纯采用 CRFs 模型

采用第 4.3 节中介绍的单纯使用 CRFs 模型的方法进行中文人名和地名识别的实验，实验结果如表 7.1 所示。

表 7.1 单纯使用 CRFs 的命名实体识别结果

Tab. 7.1 Results of NER based on sole CRFs

	Precision	Recall	F-measure
人名识别	95.74%	89.16%	92.33%
地名识别	95.32%	85.13%	89.94%

7.1.2 基于 CRFs 与边界模板的人名识别

从表 7.1 的结果可以看出，单纯使用 CRFs 标记时，召回率偏低。正如第 4.4 节中所分析，单纯使用 CRFs 标记时，边缘概率较低的标记很可能是错误的，因而我们引入边界模板方法对这部分样本重新标记，即边缘概率大于某阈值时认为 CRFs 给出的标记是正确的，如果边缘概率小于某阈值时，认为标记是有问题的，则通过边界模板方法重新标记，阈值 ε 是多大时，系统才能得到最好的结果，我们通过实验来确定。表 7.2 是 CRFs 与边界模板人名识别的结果。

表 7.2 基于 CRFs 与边界模板的人名识别结果

Tab. 7.2 Results of person NER based on CRFs and boundary templates

	Precision	Recall	F-measure
ε =0.85	96.14%	90.83%	93.41%
ε =0.90	96.14%	90.88%	93.44%
ε =0.92	*96.15%*	*91.15%*	*93.58%*
ε =0.93	95.94%	91.20%	93.51%
ε =0.95	95.94%	91.31%	93.57%
ε =0.98	95.46%	91.36%	93.37%

我们观察到当ε从 0.85 到 0.98，基于混合模型识别的 F 值均好于单纯的 CRFs 模型，当ε=0.92 系统得到最好的 F 值，召回率和 F 值分别提高了 1.09%和 1.25%。

7.1.3 基于 CRFs 与概率统计的命名实体识别

在本节中，采用概率统计的方法对边缘概率较低的概率进行重新标记。如第 4.5 节所述，概率统计方法通过计算候选命名实体的构词可信度、词性接续可信度来评价候选词是否是命名实体。本混合方法充分综合了 CRFs 方法与概率统计方法的优点。表 7.3 和表 7.4 分别给出了基于 CRFs 与概率统计混合模型的人名和地名识别实验结果。

表 7.3 基于 CRFs 与概率统计的人名识别结果

Tab. 7.3 Results of person NER based on CRFs and statistical method

	Precision	Recall	F-measure
ε =0.80	95.65%	90.83%	93.18%
ε =0.85	95.60%	90.99%	93.24%
ε =0.88	95.53%	91.63%	93.54%
ε =0.89	95.52%	91.58%	93.51%
ε =0.90	95.42%	91.63%	93.49%
ε =0.91	95.47%	91.68%	93.54%
ε =0.92	*95.28%*	*92.01%*	*93.61%*
ε =0.93	95.07%	92.11%	93.57%
ε =0.95	94.50%	92.11%	93.29%

从表 7.3 观察到当ε从 0.80 到 0.95，基于混合模型识别的 F 值均好于单纯的 CRFs 模型，其中，召回率与 F 值都有很大程度的提高，当ε=0.92 系统得到最好的 F 值，此时，召回率和 F 值分别提高了 2.85%和 1.28%。

在基于混合模型的地名识别算法中，同样应用概率统计的方法对 CRFs 模型中边缘概率较低的标记进行修正，具体算法如第 4.5 节所述，实验结果如表 7.4 所示。

表 7.4 基于 CRFs 与概率统计的地名识别结果

Tab. 7.4 Results of location NER based on CRFs and statistical method

	Precision	Recall	F-measure
ε =0.75	94.58%	88.33%	91.35%
ε =0.80	94.45%	88.39%	91.31%
ε =0.85	94.31%	88.55%	91.33%
ε =0.90	94.23%	89.26%	91.67%
ε =0.91	94.23%	89.34%	91.72%
ε =0.92	*94.21%*	*89.42%*	*91.75%*
ε =0.93	94.18%	89.31%	91.68%
ε =0.95	94.02%	89.39%	91.65%

同样，我们观察到当ε从 0.75 到 0.95 的变化过程中，基于混合模型识别的 F 值均好于单纯的 CRFs 模型，当ε=0.92 时，系统得到最好的 F 值，召回率和 F 值分别提高了 4.29%和 1.81%。

7.1.4 几种基于 CRFs 方法的比较

这一节将本文第 4 章所提出的几种基于 CRFs 模型的实验结果进行比较。表 7.5 给出了基于单纯 CRFs 模型、基于 CRFs 与概率统计的混合模型、基于 CRFs 与边界模板混合模型的人名识别的结果的比较。其中，基于 CRFs 与概率统计的混合模型取得了最好的结果。

从表 7.5 中，可以看出基于 CRFs 的方法召回率较低，这是因为 CRFs 模型的泛化能力不够，对训练集中没有出现的情况，较难召回。而混合模型中分别通过引入概率模型与边界模板模型召回了一些正确的情况，大大提高了识别的召回率与 F 值。又因为概率模型同时考虑了候选人名构词的信息和词性接续的信息，而边界模板方法只考虑了候选人名前后词对的信息，因此基于 CRFs 与概率统计的混合模型效果好于基于 CRFs 与边界模板的混合模型。

表 7.6 给出了基于 CRFs 模型、CRFs 与概率模型结合的混合模型的地名识别结果的比较。混合模型中通过引入概率方法对 CRFs 识别结果进行优化，大大提高了识别的召回率和 F 值。

表 7.5 人名识别方法的比较

Tab. 7.5 Results of person NER based on different methods for comparison

	Precision	Recall	F-measure
基于 CRFs	95.74%	89.16%	92.33%
CRFs 与边界模板	96.15%	91.15%	93.58%
CRFs 与概率模型	*95.28%*	*92.01%*	*93.61%*

表 7.6 地名识别方法的比较

Tab. 7.6 Results of location NER based on different methods for comparison

	Precision	Recall	F-measure
基于 CRFs	95.32%	85.13%	89.94%
CRFs 与概率模型	*94.21%*	*89.42%*	*91.75%*

7.2 基于 Max-Margin Markov Networks 模型的地名识别

在本节中，基于 Max-Margin Markov Networks 模型，进行地名识别的实验。采用与第 7.1 节相同的语料，选取与 CRFs 模型中相同的特征与特征模型，建立地名识别系统，进行实验，具体方法如第 5 章所述。下面是基于 Max-Margin Markov Networks 的地名识别的结果。

(1) 基于不同惩罚因子 C

惩罚因子 C 表明对错分因子的惩罚程度。此实验中，选取不同的惩罚因子 C 进行实验，实验结果如表 7.7 所示。

表 7.7 基于 M^3Net 的地名识别

Tab. 7.7 Results of location NER based on M^3Net

	Precision	Recall	F-measure
$C = 1$	*93.51%*	*87.81%*	*90.57%*
$C = 5$	93.37%	87.75%	90.27%
$C = 10$	93.45%	87.75%	90.51%

我们可以看到当 $C = 1$ 时，实验的结果最好。

(2) 基于不同阶数的 M^3Net

我们可以选取不同的阶数的马尔可夫链来构造 M^3Net 模型(如第 5 章所述)，此实验中选取一阶与二阶的 M^3Net 进行地名识别的实验。实验结果如表 7.8 所示。

表 7.8 基于不同阶数的 M³Net 的地名识别结果比较

Tab. 7.8 Results of location NER based on different order M³Net

	Precision	Recall	F-measure
1 阶 M³Net	93.51%	87.81%	90.57%
2 阶 M³Net	*93.75%*	*88.35%*	*90.96%*

从表 7.8 中，观察到基于 2 阶的 M³Net 模型获得更好的识别效果。这是由于在 2 阶模型中，标记当前样本时，模型会同时考虑前两个样本的信息。由于融入了更加丰富的结构特征，2 阶的 M³Net 模型的识别效果好于 1 阶的模型。

(3) 三种机器学习方法的比较

表 7.9 给出了采用多类 SVM，一阶的 CRFs 和一阶的 Max-Margin Markov Networks 三种机器学习模型进行地名识别任务的实验结果，实验均选取相同的特征和特征模板。可以看出 1 阶 M³Net 与 1 阶 CRFs 模型的效果要明显好于 SVM，这是因为无向图模型具有可以充分考虑节点之间的关联信息的优点。另外 1 阶 M³Net 的效果要好于 1 阶 CRFs。

表 7.9 基于三种机器学习模型的地名识别结果比较

Tab. 7.9 Results of location NER based on three models for comparison

	Precision	Recall	F-measure
SVM(pairwise)	92.50%	80.89%	86.31%
1 阶 CRFs	95.32%	85.13%	89.94%
1 阶 M³Net	*93.51%*	*87.81%*	*90.57%*

7.3 基于概率特征函数的 CRFs 的命名实体识别

基于概率特征函数的 CRFs 是传统 CRFs 的一种改进(如第 6 章所述)。本节通过命名实体任务对此改进模型的学习能力进行评测。采用北大标注的语料，其中，训练集中包含人名 7499 个，地名 4494 个，测试集中包含人名 1979 个，地名 1195 个。在相同的语料，相同的特征、特征模板的情况下，分别采用传统 CRFs 模型与基于概率特征函数的 CRFs 进行实验，比较二者的结果。如表 7.10 和表 7.11 所示，基于概率特征函数的 CRFs 的实验效果都要好于传统 CRFs 的实验效果。这是因为在基于概率特征函数的 CRFs 模型中，通过定义概率特征函数，使模型学习到有用的概率信息，以提高模型的机器学习能力。

表 7.10 基于概率 CRFs 的人名识别结果

Tab. 7.10 Results of person NER based on probability CRFs

	Precision	Recall	F-measure
CRFs	94.47%	86.30%	90.32%
概率 CRFs	*95.45%*	*86.50%*	*90.75%*

表 7.11 基于概率 CRFs 的地名识别结果

Tab. 7.11 Results of location NER based on probability CRFs

	Precision	Recall	F-measure
CRFs	94.26%	90.11%	92.14%
概率 CRFs	*94.41%*	*91.52%*	*92.81%*

7.4 与其它文献的比较

本节将基于 CRFs 与概率统计的命名实体识别结果与文献[15]、[48]进行比较。

文献[15]采用基于 SVM 与概率统计的混合模型进行命名实体识别实验，这种方法认为，SVM 分类超平面附近的样本有可能存在着分类错误，因此对于这部分样本，引入概率统计模型(如第 4.5.1 节和第 4.5.2 节所述)进行分类，其它样本仍采用 SVM 分类。我们应用文献[15]的方法，在与本文相同的语料上进行实验。从表 7.13 可以看出，本文提出的基于 CRFs 与概率统计的混合模型的实验结果要好于文献[15]。

本文较文献[15]的优点在于：本文选取 CRFs 模型作为机器学习的方法，CRFs 是一种无向图模型，可以更好的融入信息，并且考虑到相关节点的关联，它更适合于序列标注任务。而文献[15]中所用的 SVM 模型将每个样本割裂开来，无法考虑样本间的关联；另外，SVM 模型更适用于二值分类问题，对于多分类问题它采用投票的方法，这其中会产生一些误差，而影响标注结果。

文献[48]采用了 CRFs 模型结合后期处理的方法进行实验，实验语料与本文使用的完全相同。从表 7.13 中可以看出，本文提出的基于 CRFs 与概率统计的混合模型的实验效果要好于文献[48]。

本文较文献[48]的优点在于：本文通过边缘概率科学的定位出 CRFs 模型中出现错误的可能性较大的标记，通过引入一种概率统计的方法对这些原子进行重新标记，进而提高了识别效果。而文献[48]只通过规则对 CRFs 标记进行后期处理，对整个的识别效果提高不大。

表 7.12 与文献[15]、[48]的比较

Tab. 7.12 Results of our method and reference[15],[48] for comparison

		Precision	Recall	F-measure
人名识别	文献[15]	94.83%	89.68%	92.18%
	文献[48]	90.83%	92.16%	91.49%
	CRFs与概率统计	***95.28%***	***92.01%***	***93.61%***
地名识别	文献[15]	93.36%	88.22%	90.71%
	文献[48]	89.89%	91.66%	90.77%
	CRFs与概率统计	***94.21%***	***89.42%***	***91.75%***

7.5 实验结果分析

第 7.1 节的实验结果表明：以 CRFs 作为机器学习的方法，完成命名实体识别任务，可以取得较好的效果。但由于 CRFs 模型的泛化能力不够，导致召回率较低。为了优化识别结果，通过边缘概率定位出 CRFs 模型中错误机会较大的标注，并分别通过边界模板、概率统计的方法进行修正。这两种修正方法大大提高了系统的识别效果，基于 CRFs 与边界模板的人名识别的召回率与 F 值比单纯 CRFs 方法分别提高了 1.09%和 1.25%；另外，基于 CRFs 与概率统计的人名和地名识别的 F 值比单纯 CRFs 方法分别提高了 1.28%和 1.81%。

第 7.2 节的实验结果表明：采用 Max-Margin Markov Networks 模型进行命名实体识别的效果好于 CRFs 模型。这是由于 Max-Margin Markov Networks 模型综合了 SVM 与无向图模型的优点，具有更好的机器学习能力。

第 7.3 节采用基于概率特征函数的 CRFs 模型进行命名实体识别实验，由于此概率 CRFs 模型可以将学习到的有用的概率信息加入特征函数中，进而提高机器学习能力。所以，它在命名实体识别中的表现都要好于传统的 CRFs 模型。

结　论

本文将命名实体识别问题看成一个序列标注任务，通过基于机器学习的方法进行识别。主要的工作有以下两点：

(1) 采用 CRFs 模型进行命名实体识别任务，并提出 CRFs 与边界模板混合算法、CRFs 与概率统计混合算法，对 CRFs 识别结果进行优化。实验证明这两种混合算法的识别效果明显好于 CRFs 模型。

① 通过对基于 CRFs 人名识别结果分析发现，CRFs 系统识别中产生错误的标记大都具有较低的边缘概率，因此我们引入边界模板方法对边缘概率较低的样本进行重新标记，边界模板方法充分考虑候选人名的前词与后词的组合信息，更好的评价人名。CRFs 与边界模板的混合方法可以在一定程度上修正 CRFs 模型中的错误标记，从而大大提高系统识别的 F 值与召回率。实验表明：基于 CRFs 与边界模板的人名识别的召回率与 F 值比单纯 CRFs 方法分别提高了 1.09%和 1.25%。

② 为了更好的修正 CRFs 模型中的错误，我们又尝试引入概率统计方法对 CRFs 模型中边缘概率较低的样本进行重新标记。概率统计方法通过计算候选词的构词可信度和前后词词性的接续可信度，来评价候选命名实体，此方法可以充分考虑候选词的构词和前后词词性的二元特征。通过 CRFs 与概率统计的混合方法可以在一定程度上修正 CRFs 模型中的错误标记，从而大大提高系统识别的 F 值与召回率。实验表明：基于 CRFs 与概率统计的人名和地名识别的 F 值比单纯 CRFs 方法分别提高了 1.28%和 1.81%。

(2) 本文提出一种基于概率特征函数的 CRFs 模型；另外，介绍了一种较新的机器学习模型 Max-Margin Markov Networks(M^3Net)模型。通过命名实体识别实验证明，这两种模型比 CRFs 模型具有更强的机器学习能力。

① 提出了一种基于概率特征函数的 CRFs 模型，它将 CRFs 模型中定义的 0、1 二值特征函数融入概率信息，生成概率特征函数，然后基于这种概率特征函数构造条件随机场。此模型可以将学习到的有用的概率信息加入特征函数中，进而提高机器学习能力。在命名实体识别实验中，基于概率特征函数的 CRFs 识别效果好于传统的 CRFs。另外，此模型可以应用于其它任务，概率特征函数可以根据不同任务的特点重新定义，以达到更好的效果。

② Max-Margin Markov Networks 模型综合了 SVM 模型和无向图模型的优点，可以通过使用核函数处理高维向量，又可以充分考虑相关节点间联系，具有较好的机器学习能力。本文通过中文地名识别任务进行实验，实验证明，基于 Max-Margin Markov

Networks 的地名识别的 F 值好于 CRFs 方法。此模型可以应用于自然语言处理的其它任务。

通过对以上工作进行总结发现，CRFs 模型是目前最好的机器学习模型之一，在基于 CRFs 进行识别任务时，可以通过边缘概率定位出 CRFs 中较可能出错的标记，并引入其它优秀的方法对这些样本重新标记。本文提出了两种混合模型，其中 CRFs 与概率统计的混合模型取得了最好的效果，召回率与 F 值明显好于单纯的 CRFs 模型。

另外，基于概率特征函数的 CRFs 模型和 Max-Margin Markov Networks 模型都比 CRFs 模型具有更好的学习能力。它们也可以与其它方法结合进行命名实体识别任务，会取得更好的识别效果。

为了进一步提高中文命名实体识别的精度，可以对机器学习的模型进行改进和优化，如一些 CRFs 的优化方法，Dynamic CRFs 和 Semi-Markov CRFs 等；同样，可以对 Max-Margin Markov Networks 进行一些改进，将 Max-Margin 思想用于较为复杂的图模型上。另外，我们可以针对实验结果中的某些错误加入一些规则进行修正，从而取得更好的识别效果。

参 考 文 献

[1] 赵铁军. 机器翻译原理. 哈尔滨: 哈尔滨工业大学出版社, 2001.

[2] 刘开瑛. 中文文本自动分词和标注. 北京: 商务印书馆, 2000.

[3] Tan H Y, Zheng J H, Liu K Y. Research on method of automatic recogniton of Chinese place name based on transformation. Journal of Software, 2001, 12(11): 1608-1613.

[4] 吕雅娟, 赵铁军, 杨沐昀等. 基于分解与动态规划策略的汉语未登录词识别. 中文信息学报, 2001, 15(1): 28-33.

[5] 罗智勇, 宋柔. 现代汉语自动分词中专名的一体化, 快速识别方法. 国际中文电脑学术会议, 新加坡, 2001: 323-328.

[6] 郑家恒, 刘开瑛. 自动分词系统中姓氏人名的处理策略讨论. 见: 陈力为编. 计算语言研究与应用. 北京: 北京语言学院出版社, 1993.

[7] 张华平, 刘群. 基于角色标注的中国人名自动识别研究. 计算机学报, 2004, 27(1): 85-91.

[8] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger. Proceedings of 40th Annual Meeting of the ACL, Philadelphia, 2002: 473-480.

[9] Borthwick A. A maximum entropy approach to named entity recognition: [PhD Dissertation]. New York: New York University, 1999.

[10] 秦文, 苑春法. 基于决策树的汉语未登录词识别. 中文信息学报, 2004, 18(1): 14-19.

[11] Collins M. Ranking algorithms for named entity extraction: boosting and the voted perceptron. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002: 489-496.

[12] Carreras X, Marquez L, Padro L. Named entity extraction using AdaBoost. The 6th Conference on Natural Language Learning, Taipei, 2002: 167-170.

[13] Takeuchi K, Collier N. Use of support vector machines in extended named entity recognition. The 6th Conference on Natural Language Learning, Taipei, 2002: 119-125.

[14] Goh C L, Asahara M, Matsumoto Y. Chinese unknown word identification using character-based tagging and chunking. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, 2003: 197-200.

[15] Li L Sh, Mao T T, Huang D G et al. Hybrid models for Chinese named entity recognition. Proceeding of SIGHAN5 Workshop in COLING-ACL, Sydney, 2006: 72-78.

[16] 宋柔, 朱宏. 基于语料库和规则库的人名识别法. 见: 陈力为编. 计算语言研究与应用. 北京: 北京语言学院出版社, 1993: 150-154.

[17] 黄德根, 马玉霞, 杨元生. 基于互信息的中文姓名识别方法. 大连理工大学学报, 2004, 44(5): 744-748.

[18] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the 7th Conference on Natural Language Learning, Edmonton, 2003: 188-191.

[19] Chen W L, Zhang Y J, Isahara H. Chinese named entity recognition with conditional random fields. Proceedings of 5th SIGHAN Workshop on Chinese Language Processing, Sydney, 2006: 118-121.

[20] Lu P, Yang Y P, Gao Y B et al. Hierarchical conditional random fields(HCRF) for Chinese named entity tagging. The Third International Conference on Natural Computation, Haikou, 2007: 24-28.

[21] 黄德根，杨元生，王省等. 基于统计方法的中文姓名识别. 中文信息学报，2001，15(2)：31-37.

[22] 黄德根，岳广玲，杨元生. 基于统计的中文地名识别. 中文信息学报，2003，17(2)：36-41.

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

[24] Fukuda T, Izumi M, Miura T. Word segmentation using domain knowledge based on conditional random fields. Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 2007: 436-439.

[25] Huang D G, Sun X, Jiao S D et al. HMM and CRF based hybrid model for Chinese lexical analysis. In Sixth SIGHAN Workshop on Chinese Language Processing, Sydney, 2008: 133-137.

[26] Okanohara D, Miyao Y, Tsuruoka Y et al. Improving the scalability of semi-markov conditional random fields for named entity recognition. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, 2006: 465-472.

[27] Sha F, Pereira F. Shallow parsing with conditional random fields. Proceedings of Human Language Technology/North American Chapter of the Association for Computational Linguistics Annual Meeting, Columbia University, 2003: 213-220.

[28] Tan Y M, Yao T S, Chen Q et al. Applying conditional random fields to Chinese shallow parsing. Proceedings of CICLing-2005, Mexico City, 2005: 167-176.

[29] Nocedal J, Wright S J. Numerical optimization. New York: Springer-Verlag, 1999.

[30] Taskar B, Guestrin C, Koller D. Max-Margin markov networks. Neural Information Processing Systems Conference, Vancouver, 2003.

[31] Taskar B. Learning structured prediction models: a large margin approach: [PhD Dissertation]. California: Stanford University, 2004.

[32] Tsochantaridis I, Joachims T, Hofmann T et al. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 2005, 6(9): 1453-1484.

[33] Crammer K, Singer Y. Ultraconservative online algorithms for multiclass problems. Journal of Machine Learning Research, 2003, 3(1): 951-991.

[34] Chechik G, Heitz G, Elidan G et al. Max-margin classification of data with absent features. Journal of Machine Learning Research, 2008, 9(1): 1-21.

[35] Boser B, Guyou L, Vapnik V. A training algorithm for optimal margin classfier. Processings of 5th Annual Workshop on Computational Learning Theory, 1992: 144-152.

[36] Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

[37] Vapnik V, Golovich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: Advances in Neural Information Processing System, Cambridge, MIT Press, 1997: 281-287.

[38] 袁亚湘, 孙文瑜 著. 最优化理论与方法. 北京: 科学出版社, 1997.

[39] Vapnik V N. Statistical learing theory. New York: John Wiley & Sons, 1998.

[40] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[41] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2001, 2(3): 265-292.

[42] Platt J C. Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 1999.

[43] 毛婷婷. 中文专有名词识别的研究: (硕士学位论文). 大连: 大连理工大学, 2006.

[44] Zhang H P, Liu Q, Cheng X Q et al. Chinese lexical analysis using hierarchical hidden markov model. Proceedings of the 2nd SIGHAN Workshop on Chinese Language, Sapporo, 2003: 63-70.

[45] 李中国, 刘颖. 边界模板和局部统计相结合的中国人名识别. 中文信息学报, 2006, 20(5): 44-50.

[46] 干俊伟, 黄德根. 汉语介词短语的自动识别. 中文信息学报, 2004, 19(4): 17-23.

[47] Sutton C, Rohanimanesh K, McCallum A. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. Journal of Machine Learning Research, 2007, 8(3): 693-723.

[48] Feng Y Y, Huang R H, Sun L. Two step Chinese named entity recognition based on conditional random fields models. Proceedings of the 6th SIGHAN Workshop on Chinese Language Proceeding, Sydney, 2008: 120-123.

攻读硕士学位期间发表学术论文情况

[1] Lishuang Li, Zhuoye Ding, Degen Huang. Recognizing Location Names from Chinese Texts Based on Max-Margin Markov Network. Proceeding of IEEE NLP-KE08. Beijing, China, 2008.(EI 检索) (正文第 5 章)

[2] Lishuang Li, Zhuoye Ding, Degen Huang, Huiwei Zhou. A hybrid model based on CRFs for Chinese Named Entity Recognition, In International Conference on Advanced Language Processing and Web Information Technology (Alpit 2008), Dalian, 2008: 127-132.(EI 检索) (正文第 4 章)

[3] Lishuang Li, Zhuoye Ding, Degen Huang. Recognize person names from Chinese texts based on clustering SVM, In Intelligent Systems and Control, Cambridge, USA, 2007. (正文第 3,4 章)

致　谢

本论文从选题到完成，自始至终得到了我的导师黄德根教授的悉心教育和精心指导，倾注了导师的心血。我的每一点进步，每一份成绩都与导师的谆谆教诲和支持息息相关。黄老师高尚的人品、精湛的学术水平、渊博的学识，敏锐的科研思维以及严谨的作风和敬业的精神都给我留下了深刻的印象。衷心感谢尊敬的导师悉心栽培，使我受益匪浅，终生难忘。

感谢实验室的李丽双副教授，一直都在指导和协助我完成整个论文，在此深深表示感谢！李老师对待学术的刻度钻研的精神以及一丝不苟的作风让我深深感到敬佩，多年来李老师一直在工作、学习、生活上给予我无微不至的关怀、指导和照顾。感谢李老师三年来对我的教育和帮助，使我收获良多。

本文还得到林鸿飞教授和唐达副教授的指导和评阅，林老师对学术精益求精的态度和对教育的热情，和唐老师耐心的态度、细致的作风使我受益匪浅，在此表示深深的感谢！

从刚刚进入自然语言处理实验室到整个课题的完成，实验室的全体成员都给予了我极大的帮助。感谢孙晓师兄，毛婷婷师姐在本课题研究中对我的帮助，他们总是不厌其烦地给我解答我所不知道的问题，他们是值得我学习的榜样。感谢的师弟师妹，对帮助过我的所有同学表示深深感谢！

中文命名实体识别的研究

作者： 丁卓冶
学位授予单位： 大连理工大学

相似文献(10条)

1.学位论文 吴宝琪 中文命名实体的识别方法研究及其实现 2007

命名实体识别是自然语言处理领域的研究焦点之一，在信息检索、信息抽取、机器翻译等领域都有着十分重要的应用。与英文命名实体识别取得的成绩相比，中文命名实体识别还有很多问题需要研究。本论文的主题就是对中文命名实体识别进行初步的研究。

最大熵模型在自然语言处理领域有着广泛的应用，并取得了很好的成绩。本文首先对如何将最大熵模型应用到中文命名实体识别领域进行了探讨。然后，结合本文原型系统，从最大熵模型的特征、特征选择、模型训练和测试文本标注几个方面，对最大熵模型在中文命名实体识别领域的具体实现进行了介绍。

由于最大熵模型是在训练样本上进行数据挖掘来获得信息的，会受到训练样本的局限，发现不了那些训练样本中未出现的规律。比如在文档中，有些重要词语会重复出现多次，而这些重复词语中有些正是命名实体。本文将这种词语重复出现的信息应用到中文命名实体识别领域，提出了一种将词语重现信息与最大熵模型相结合的复合的中文命名实体识别方法，并对这种复合识别方法的实现进行了详细介绍。

最后，本文在MET-2会议的数据集上，对文中提出的复合识别方法进行了测试。结果表明，这种复合的识别方法比单纯的最大熵模型在中文命名实体识别领域具有更好的性能。

2.学位论文 张友书 基于动态条件随机场的中文命名实体识别 2009

命名实体识别就是把文本中出现的命名实体包括人名、地名、组织机构名、日期、时间、和其他实体识别出来并加以归类。命名实体识别属于自然语言处理的基础研究领域，是信息抽取、信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础，对命名实体识别进行研究具有很大的实用意义。

本文主要研究以人名、地名和组织名识别为主的中文命名实体识别问题。具体说来，本文的主要内容如下：本文首先介绍了命名实体识别的定义及其特点，并简要介绍了国内外的中文命名实体识别研究情况，以及现有的主要命名实体识别方法。目前，中文命名实体识别主要采用统计机器学习的方法，而其中条件随机场的效果最好。所以紧接着详细介绍了链式和动态条件随机场的定义、模型表示、参数估计和训练方法等。进一步地，将链式和动态条件随机场模型应用于中文命名实体识别任务，提出了三种基于条件随机场的命名实体识别方法：使用链式条件随机场基于字的方法、使用链式条件随机场基于词的方法、基于动态条件随机场的方法。最后，通过实验对这三种方法进行了比较分析。

本文的主要贡献是：第一，首次尝试将动态条件随机场模型应用到中文命名实体识别中来。第二，利用动态条件随机场进行命名实体识别，将中文分词与命名实体识别过程融合在一起，使二者相互影响，能够同时改善中文分词和命名实体识别效果，改进了现有的命名实体识别技术。第三，基于动态条件随机场实现了一个命名实体识别系统。

3.期刊论文 冯元勇.孙乐.李文波.张大鲲.FENG Yuan-yong.SUN Le.LI Wen-bo.ZHANG Da-kun 基于单字提示特征的中文命名实体识别快速算法 -中文信息学报2008,22(1)

近年来条件随机场(CRF)模型在自然语言处理中的应用越来越广泛.标准的线性链(Linear-chain)模型一般采用L-BFGS参数估计方法,收敛速度慢.本文在分析模型复杂度的基础上提出了一种改进的快速CRF算法.该算法通过引入小规模单字特征降低特征的规模,并通过在推理过程中引入任务相关的人工知识压缩Viterbi和Baum-Welch格搜索空间,提高了训练的速度.在中文863命名实体识别评测语料和SIGHAN06语料集上进行的实验表明,该算法在不影响中文命名实体识别精度的同时,有效地降低了模型的训练代价.

4.学位论文 赵琳瑛 基于隐马尔科夫模型的中文命名实体识别研究 2008

随着信息时代的到来和Internet的发展，用自然语言作为人机交互已是必然趋势，这对自然语言处理的深度和广度提出了越来越高的要求。自命名实体识别技术在1995年的MUC-6(Message Understanding Conference)会议上提出以来，越来越受到自然语言处理研究者的关注，并成为很多应用中的关键技术。

本文对命名实体识别的方法进行了研究，分析了基于规则的方法和基于统计的方法的优缺点。由于获取上下文信息的多少和数据平滑的程度是评价识别性能的两个重要参数，而以前的统计模型获取上下文信息有限，本文提出了一种基于三阶隐马尔科夫模型的命名实体识别方法，该方法使用语言知识进行约束，兼顾了准确率和召回率，取得了较好的识别效果。自动分词和词性标注直接影响命名实体的识别，本文采用了海量智能分词系统对文本进行分词和标注。在统计词频方面，本文使用了改进的K均值方法对参数进行估计，并采用线性差值法对参数结果进行平滑处理。在命名实体识别方面，本文采用改进的Viterbi算法对初始观察序列重新标注，并求出最佳的状态序列。本文识别的主要内容为实体词，即人名、地名和机构名。目前，中文命名实体识别实验仍处于初期阶段，还有不少工作有待进一步完善。今后的工作将进一步研究规则的制定和数据平滑技术，以期进一步提高命名实体的识别率。

5.学位论文 杨华 基于最大熵模型的中文命名实体识别方法研究 2008

命名实体识别是信息抽取的子任务，同时也是机器翻译、自动问答等多种自然语言处理技术的基础。由于受中文自身特点的限制，中文命名实体识别一直相当困难。为了促进其它中文自然语言处理技术和应用的发展，研究中文命名实体的识别技术是很有意义，也是非常重要的。

本文利用最大熵模型(Maximum Entropy，ME)进行中文命名实体识别。尝试了在不同特征模板集下，命名实体识别的性能，深入研究了最大熵模型在中文命名实体识别中的特点，发现最大熵模型不能自动组合特征，模型性能很大程度上依赖于特征模板。因此，设计合理的特征模板是基于最大熵模型中文命名实体识别的关键。

汉语中存在大量的隐含语义特征，可以帮助命名实体的识别，而最大熵模型的一个重要优点就是能融合不同粒度和不同层次的特征。针对这一特点，本文通过从语料库中抽取信息的方式，建立了大量的中文命名实体语义知识库。但是，由于语料库的规模有限，并且基于统计的方法普遍存在数据稀疏的问题，导致很多重要的知识不能被挖掘出来。为了解决这一问题，本文首次将语义扩展的思想应用在命名实体识别中，充分发挥了有限语言资源的作用，深度挖掘了有限资源的信息和知识，在不扩大语料库的前提下，挖掘出更丰富的知识，一定程度上缓解了数据稀疏问题。实验证明，相对于扩展前的知识库，利用扩展后的知识库，平均识别召回率提高了1.17%，F值提高了0.41%。特别是结构比较复杂的机构名识别准确率提高了0.24%，召回率提高了1.39%，F值提高了0.86%。

6.学位论文 周俊生 基于统计学习的中文信息抽取技术研究 2007

Web的发展使得电子文档数目巨大且迅猛增长，大量的信息存在于非结构化的自然语言文档中，为了能高效地利用存在于自然语言文本中的信息，信息抽取技术提供了一条有效的途径，利用它可以将非结构的文本转化为结构化的信息，以便于信息的后续处理(如：数据挖掘等)。信息抽取系统的实现涉及自然语言处理的一系列难点，是当前自然语言处理的一个研究热点。本文主要基于统计学习方法，围绕实现中文信息抽取过程的几个关键问题展开研究，主要工作包括：

1. 提出一种基于层叠条件随机场模型的中文命名实体识别算法。条件随机场是一种新的概率无向图模型，本论文在充分利用条件随机场模型优势的基础上，结合中文命名实体的特点，设计了一种层叠条件随机场模型用于中文命名实体的识别。在层叠条件随机场模型中，低层模型的识别结果将传递到高层模型，为高层条件随机场模型对复杂命名实体的识别提供决策支持。实验结果显示，该算法取得了很好的识别效果。

2. 提出一种基于大间隔方法的中文组块识别算法。首先给出了中文组块的定义，将中文组块识别问题转化为序列化标注问题；然后根据大间隔思想给出判别式的序列化标注函数的优化目标和训练算法，并针对中文组块识别问题，设计了一种改进的F1损失函数，使得F1损失值能依据每个句子的实际长度而相应缩放，实现间隔值的动态调整，从而能够引入更有效的约束不等式。通过在LDC的CTB4数据集上的实验数据显示，该算法优于当前的其它中文

组块分析算法。

3．提出一种有监督的关联聚类算法实现对中文实体提及的指代消解。首先将指代消解过程看成图的关联聚类问题，它从全局的角度实现对共指等价类的划分，而不是孤立地对每一对名词短语分别进行共指决策；然后给出了关联聚类的推导算法；最后设计了一种基于梯度下降的特征参数学习算法，实现从训练语料中自动学习各个特征的权值，从而使得训练出的特征参数能够较好的拟合关联聚类的目标。在ACE中文语料上的实验结果显示，该算法优于传统的“分类—聚类”指代消解学习算法。

4．针对当前中文指代标注训练语料非常缺乏的现状，提出一种无监督聚类算法实现对中文实体提及的指代消解。通过将指代消解问题转化为图划分问题，引入一个有效的模块函数作为目标函数实现对图的自动划分，依据该函数值来自动选择最优的聚类数目，并设计了基于贪心法的聚类算法。聚类过程避免了阈值选择问题，是一种有效可行的无监督指代消解算法。

5．提出一种基于新的合成核的中文实体关系抽取方法。论文首先设计了一种能够直接利用浅层语言特征的混合谱核来描述关系实例的上下文，并给出了基于广义后缀树的高效核函数值计算方法；然后再通过与实体核的组合生成合成核，该合成核既表示了两个关系实例出现的上下文之间的相似特征，又考虑了两个实体对之间的相似特征，核的计算不需要依赖于中文句法分析结果，且具有较低的计算复杂度。在ACE中文语料上的实验结果显示，基于这种新的合成核的中文关系抽取方法获得了较好的实验结果。

7.学位论文 乔永波 规则与统计相结合的中文命名实体识别 2007

自然语言处理作为人工智能的重要研究领域之一，是利用计算机进行语言知识的获取、表示以及应用的技术，为人与计算机之间的信息交流提供了更加高效、便捷的方法。由于汉语的书写习惯，词与词之间的边界标志是隐含的，对于大多数汉语处理系统来讲，首先要做的工作就是分词。而在实际应用中，分词仍然受到诸多因素的制约。其中，命名实体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的精度以及其后的词性标注和句法分析的精度。另外，命名实体识别的研究还有利于信息抽取、信息检索、机器翻译、文本分类等应用系统的实现。因此，研究命名实体的自动识别具有重要的理论意义和实践价值。

目前，国内外关于中文命名实体识别的研究仍然存在着识别的自动化程度不高，忽视了词法、句法及语义信息的作用等问题，并且大部分的研究只是针对人名的识别，而对于地名和机构名识别的研究还不够成熟。

针对上述不足，本文以中文人名、译名、地名和机构名的识别为研究重点，提出了一种规则与统计相结合的一体化解决方案，该方案采用了双层命名实体识别模型来识别包括嵌套地名和机构名在内的多种命名实体。该双层命名实体识别模型的实现思想是：首先，在分词之前建立第一层命名实体识别模型，该模型由名实体检索算法实现，该算法利用命名实体的特征词，如人名的姓氏、地名的后缀词来引发命名实体的识别，并根据词法规则信息和命名实体的用字统计信息来识别部分命名实体；然后，在分词之后所得到的N个合法分词序列的基础上，引入第二层命名实体识别模型——基于隐马尔科夫的统计模型，该模型可以识别人名、译名、地名和机构名，并利用第一层模型所识别出来的命名实体识别嵌套的地名和机构名。

本文重点讨论了如何在分词之前和分词之后分别设计和实现命名实体识别模型，并考虑将该双层模型结合到已建立的汉语句法分析系统的分词子系统中，既保证命名实体识别与并发检索一综合排歧分词子系统的兼容性，又能够较好地支持基于二元关系模型的汉语句法分析系统。

在双层识别模型中，第一层模型能够很好地支持第二层模型识别出复杂结构的命名实体，二者相辅相成，很好地解决了由分词导致的命名实体误识别和漏识别问题。并且，为了保证隐马尔科夫模型识别命名实体的时效性，还采用了一种基于动态规划思想的过滤解码算法。通过对系统的测试发现，该模型识别命名实体的准确率和召回率都达到了90％以上，能够较好地保证汉语句法分析系统正确分析包含命名实体的句子的结构。因此，本文所提出的双层命名实体识别模型具有一定的研究意义和实用价值。

8.会议论文 冯元勇.孙乐.张大鲲.李文波 基于单字提示特征的中文命名实体识别快速算法 2007

近年来条件随机场(CRF)模型在自然语言处理中的应用越来越广泛。标准的线性链(linear-chain)模型一般采用L-BFGS参数估计方法，收敛速度慢。本文在分析模型复杂度的基础上提出了一种改进的快速CRF算法。该算法通过引入小规模单字特征降低特征的规模，并通过在推理过程中引入任务相关的人工知识压缩Viterbi和Baum-Welch格搜索空间，提高了训练的速度。在中文863命名实体识别评测语料和SIGHAN06语料集上进行的实验表明，该算法在不影响中文命名实体识别精度的同时，有效地降低了模型的训练代价。

9.学位论文 何楠 基于统计机器学习的两阶段中文命名实体识别研究 2008

作为信息抽取的基本任务，也是重要任务之一，命名实体识别已经成为自然语言处理的研究热点之一。从1998年开始，由美国国防高级研究计划委员会资助的消息理解会议就把命名实体识别当作它的子任务之一，并明确定义命名实体包括1．实体(组织名、人名、地名)；2．时间表达式(日期、时间)；3.数字表达式(货币值、百分数)。之后的自动内容抽取评测更加拓宽的命名实体识别的范围，把实体的提及、实体之间的关系都列为考察内容。

从2003年开始，计算语言协会下属的中文特别兴趣小组发起了中文分词和命名实体识别竞赛中，到2007年已经举办四次。前两次只在中文分词任务上展开评测，后两次加入了中文命名实体识别评测。SIGHAN定义中文命名实体包括人名、地名、机构名和地理信息名等四种，命名实体识别就是在未分词的语料中识别这四种实体的过程。

本文以SIGHAN竞赛的命名实体定义和评测标准为依据，提出了一种基于统计机器学习的两阶段命名实体识别方法，把命名实体识别分为边界检测和类型识别两个阶段，针对两个阶段的特点选取不同的机器学习方法，在几乎不损失精度的情况下大大减小了训练所需的时间复杂度和空间复杂度，这对训练代价特别大的条件随机场模型有着尤其重要的意义。

两阶段中文命名实体识别的过程是：首先进行实体边界检测，边界检测可以转化为一个序列标注问题，因此选用可以融入丰富特征并无标记偏置问题的条件随机场模型；然后使用最大熵模型进行实体类型识别，因为它符合满足已知约束情况下不对未知事物做出任何推断的哲学原理，并且在许多自然语言处理任务上有出色表现。

在进行边界检测时：第一，对比了常见的六种标记集，实验结果显示了同时强调实体开头和结尾的BIOE标记集有最好的性能；第二，对比了不同特征模板窗口大小对边界检测效果的影响，实验证明窗口数过大或过小都不好，过小的窗口可能损失上下文信息，而过大的窗口又会造成特征量过大，使训练代价提高，且会造成数据稀疏。

在进行类型识别时将所用特征归为两类，与实体本身相关的本地特征和与上下文相关的全局特征。本地特征只包含实体本身用字信息，而全局特征包含实体所处上下文用字的信息。把特征分成这两类的目的是考察实体本身和上下文用字对实体类型的区分性。实验结果发现，仅仅使用本地特征就可以取得很好的效果。分析原因发现同一实体在不同上下文中呈现不同类别的混淆现象很少，因此只使用实体本身的信息就可以很好的区别不同的实体。

接着把一阶段与两阶段实体识别进行了对比，发现两阶段与一阶段的识别准确率(F值)非常接近，略低于SIGHAN的最好结果。但两阶段的时间复杂度和空间复杂度只是一阶段的20％左右。本文的实验中，一阶段中文命名实体的时间消耗在20个小时以上，特征数量将近1亿，内存消耗12G；而采用两阶段方法后特征数量降为1千6百万，训练耗时3.5小时，内存消耗3.2G。

最后给出两阶段优越性的理论依据，指出了有待深入研究的问题。

10.学位论文 向晓雯 基于条件随机场的中文命名实体识别 2006

命名实体识别属于自然语言处理的基础研究领域，是信息抽取、信息检索、机器翻译、组块分析、问答系统等多种自然语言处理技术的重要基础。因此，对命名实体识别的研究具有很大的实用意义。

本文针对现代汉语文本的特点，主要研究以人名、地名和组织名的识别为核心内容的中文命名实体识别问题，我们以一种较新型的统计模型—条件随机场为基本框架，设计并实现了一个中文命名实体识别系统。具体说来，本文的主要内容如下：

本文首先分析了命名实体识别的难点，人名、地名、组织名的相关语言学知识，并对现有的一些命名实体识别方法和中文命名实体识别系统进行了简要介绍。

接着，详细介绍了条件随机场的定义、模型结构、势函数、参数估计和训练方法、概率计算方法等。进一步地，将条件随机场模型应用于中文命名实体识别任务，提出了适合于各类中文命名实体的特征模板，并通过实验进行验证，确定了有效特征。

本文最后，实现了一个基于条件随机场的中文命名实体识别系统，系统采用了层叠结构，以模型训练模块和命名实体识别模块作为系统的核心组成部分，在低层条件随机场模型中进行人名、简单地名以及简单组织名的识别，低层的识别结果传递到高层模型，再进行复合地名与复合组织名的识别。实验结果表明，基于条件随机场的中文命名实体识别系统能够获得较为满意的效果，在对2004年863中文命名实体识别评测语料的开放测试中，系统识别的精确率、召回率和F值分别为82.50％、76.04％和79.14％。