合肥工业大学

硕士学位论文

基于规则的命名实体识别研究

姓名：周昆

申请学位级别：硕士

专业：计算机软件与理论

指导教师：胡学钢

20100401

基于规则的命名实体识别研究

摘　要

中文分词是自然语言处理的第一步。在实际应用中，分词受到诸多因素的制约，未登录词的切分就是影响分词正确率的重要因素之一。未登录词主要的形式包括人名，地名，机构名等命名实体。因此，将命名实体的识别融合到中文分词的过程中，对提高中文分词的准确率起着重要作用。另外，命名实体识别的研究对于信息抽取、信息检索、机器翻译、文本分类等应用系统的实现具有重要的理论意义和实践价值。

本文的主要研究内容如下：

(1) 提出了融合命名实体识别的中文分词模型，在分词的过程中同时进行命名实体的识别，减少了因为命名实体等未登录词的识别错误而引起的中文词法切分错误，从而提高了分词的准确率。

(2) 基于本体构建中文人名知识库的层次分类体系，将中文人名领域的知识分成若干个层次，低层次的领域知识是高层次的基础，高层次的领域知识是低层次的概括和总结，有效提高了人名知识库的可维护性。

(3) 构建命名实体识别的规则库，采用规则匹配的方法识别命名实体。识别系统具有自学习的能力，在识别命名实体的同时可以分析识别结果生成新的规则反馈给规则库，具有较好的命名实体识别的效果。

关键词：中文信息处理；命名实体识别；中文分词；本体

Research on Named Entity Recognition Based on Rules

Abstract

Chinese word segmentation is the first step in natural language processing. In practice, Chinese word segmentation subject to many constraints, unknown word is one of the important factors impact the accuracy. Unknown words mainly contains person's name, place name, organization name and other named entity. Therefore, integrate named entity recognition into the process of Chinese word segmentation plays an important role in improving the accuracy of Chinese word segmentation. In addition, the named entity recognition research has important theoretical significance and practical value for information extraction, information retrieval, machine translation, text classification applications, the realization.

Contributions of the dissertation are as follows:

(1) The model of integrating named entity recognition into the process of Chinese word segmentation is proposed, Recognize Named Entity in the process of Chinese word segmentation. This method can reduced the Chinese lexical segmentation error caused by he unknown words such as named entity and enhance the accuracy of Chinese word segmentation.

(2) Chinese name classification system is Constructed based on Ontology. Through this method, Chinese name knowledge will be divided into several levels. Low-level domain knowledge is the basis of high-level and high-level domain knowledge is the generalization and summary of low-level. This approach greatly improves the maintainability of the Chiese names knowledge.

(3) The rule system for named entity recognition is built. Then use rule matching method to recognize named entity. This recognition system has an ability of self-learning. Recognize named entity at the same time analyze the results to produce new rules and add them to rule system. Experiments show that through this metod we can get good results of named entity recognition.

Keywords: Chinese information processing; Named Entity Recognition; Chinese participle; Ontology

插 图 清 单

表 格 清 单

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标志和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 合肥工业大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签字：周昆　　签字日期：2010 年 4 月 30 日

学位论文版权使用授权书

本学位论文作者完全了解 合肥工业大学 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅或借阅。本人授权 合肥工业大学 可以将学位论文的全部或部分论文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、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文者签名：周昆　　导师签名：

签字日期：2010 年 4 月 30 日　　签字日期：2010 年 4 月 30 日

学位论文作者毕业后去向：

工作单位：　　电话：2901386

通讯地址：　　邮编：230009

致　谢

首先要感谢我的导师胡学钢教授，从课程的学习、论文的选题到文章的撰写、修改直至定稿，胡老师都给予了精心的指导。胡学钢老师治学严谨，学识渊博，在导师身边修学三载，耳濡目染，使我不仅接受了严格的学术锻炼，掌握了一定的研究方法，而且还明白了许多待人接物与为人处世的道理。没有导师付出的辛勤劳动，本论文是不可能完成的。在此，我对时时鼓励和悉心指导我完成学业的胡学钢教授表示真诚的谢意。

同时，我要深深地感谢[美]佛蒙特大学(University of Vermont)计算机系主任、合肥工业大学计算机与信息学院长江学者讲座教授吴信东教授对我的教导。吴老师诲人不倦、严格要求，对我的研究工作提出了许多中肯、宝贵的意见，令我受益匪浅。

感谢计算机与信息学院人工智能与数据挖掘实验室的吴共庆、张晶、张玉红等老师，感谢他们在积累资料和开展课题方面的启发，老师给予的许多珍贵的建议和帮助使我受益良多。感谢研究室的每个成员，与他们在一起学习生活是难忘而愉快的。

感谢合肥工业大学计算机与信息学院的各位老师和院系领导对我的帮助与支持。

作者：周昆

2010 年 3 月

第一章 绪论

命名实体的识别在自然语言的处理过程中扮演着不可缺少的角色，而且命名实体的识别在中文文本处理工作中是一项很有挑战性的问题。本章首先从中文信息处理的的研究背景和意义入手，引出了对命名实体识别的研究背景和研究意义出发，介绍了课题研究的来源，本文的主要工作，并给出了文章的详细结构安排。

1.1 研究背景及意义

中文信息处理涵盖了很多层面的信息加工处理任务，如字、词、短语、句子、篇章等任务，由于词比字更能精确的表达句子的含义，中文信息处理的研究已经从“字处理”的阶段进入到了“词处理”的阶段，这就是汉语的词法分析[1]。在汉语词法分析中，首先遇到的问题是切分问题，因而，正确分词成为处理中文文本的必要条件。

目前，未登录词的识别和歧义切分字段的处理是影响中文分词精度的两大主要因素。其中，命名实体识别就属于未登录词识别的范畴，而且是未登录词中数量最多、识别难度最大、对词法分析影响最大的一部分。

在实际的分词过程中，命名实体常因其内部成词或者是上下文成词而被切分成单个的字或者是词，影响了分词的效果，也严重地影响句子中语法和语义信息的获取。可以看出，命名实体识别技术是信息处理的关键技术之一，其性能决定了信息处理的发展水平，解决不好将会成为文本信息处理技术的瓶颈。

命名实体的识别就是要从文章中识别出实体的名字，命名实体的识别在自然语言的处理过程中扮演着不可缺少的角色。

命名实体的识别在中文文本处理工作中是一项很有挑战性的问题，这个挑战性远远高于对英文文本的处理。原因有两个方面，第一个原因是中文的分词。在英文文本中，句子是由一连串的词语组成的，而且这些词之间都有空格作为分隔，然后，中文就不是如此，在中文文本中，句子是由一连串的汉字组成的，没有简单明了的分隔标志。因此，识别出中文文本中每一个词语的边界就要比英文文本难得多。另外一个原因是词法。在英文文章中，每一个名字往往都是一个大写开头的词或者是一串大写开头的词以及一些小写的标志词，如连词、介词等。但是汉语却没有大写的模式。

命名实体识别的主要应用有信息抽取、信息检索、机器翻译、问答系统等等。鉴于命名实体识别的实用意义，2006 年国际计算语言学协会下属的中文处理专业委员会 SIGHAN，在其组织的比赛(bakeoff)[2]中除了分词测评之外，还增加了中文命名实体识别的项目。

本文的课题来源有：中国科学院自动化研究所“情报与安全信息学（ISI）创新团队国际合作伙伴计划”的子课题“HTML 新闻网页过滤与总结系统”和国家“九七三”重点基础研究发展计划项目，普适个性化信息处理基础理论与方法研究。

1.2 本文的工作

本文采用基于规则匹配的方法识别中文命名实体。首先构建了融合命名实体识别的中文分词模型，于分词的过程中识别源文本中的命名实体，识别采用基于规则匹配的方法，同时分析识别结果生成新的规则反馈给规则库。该方法能够对命名实体知识库进行有效的组织，同时具有一定的自学习的能力，可以获得比较好的命名实体识别效果。

具体来说，本文做了以下几点工作：

(1) 构建了融合中文命名实体识别的中文分词模型，将中文命名实体的识别过程融合到中文分词的过程中，在对输入文本进行分割处理后，产生字符串序列，这些序列作为中文分词和命名实体识别的基本单元，基于这些字符序列识别中文命名实体，与此同时展开分词。

(2) 基于本体构建中文人名知识库的组织模型，把人名的构成分成若干了类别，每一个类别分成若干个层次，这种把中文人名知识分类管理，分层组织的模式，使得低层次的知识为高层次的知识提供依据，高层次的知识概化低层次的知识，从而大大提高知识库的可维护性。

(3) 针对命名实体中出现频率较高的人名地名，构建命名实体识别的规则库。包括用于人名识别的规则，用于地名识别的规则。

(4) 采用基于规则匹配的方法识别命名实体，输入为中文的字符串序列，首先对源文本进行候选命名实体的提取，这一步的依据是已获取的统计信息，其中，人名的识别时根据已经构建的中文人名层次分类体系，然后根据规则匹配修正识别出的候选命名实体。

(5) 分析识别结果，产生新的规则反馈给规则库，以不断完善规则，从而提高命名实体识别的查全率和查准率。

针对以上工作，本文设计了一系列的实验，在最后一个章节有实验的详细设计和实验结果的分析。

1.3 本文的结构

本文首先从中文信息处理的的研究背景和意义入手，引出了对命名实体识别和中文分词相关知识的介绍。主要讨论了命名实体识别的一些方法和应用，以及中文分词的相关方法。大多数的汉语文章处理系统是依赖于词典来识别出文章中

的每个词语，但是词典中没有的词语变成为分词中值得特别注意的问题。而命名实体又是这些未知词语当中最为重要的一类，同时，命名实体识别的过程中，如果输入的是已经被分割的词语或者是字，那么一些被错误分隔开的命名实体就不能被识别出来，因此，在分词之前就对命名实体进别，然后将命名实体识别的结果和分词的结果一起输出无疑会提高分词的准确率。由于中文人名用字规律蕴含丰富的信息，为了提高中文人名知识库的可维护性，本文采用基于本体的模型组织中文人名知识库，对知识库分层次分类别的管理，从而提高了知识库的可维护性。在上述工作的基础上，实现了基于规则匹配的命名实体识别。该方法具有反馈自学习的能力。实验结果证明，随着命名实体识别模板的不断完善，识别的准确率也在不断提高

本文的内容安排如下：

第一章介绍了本文研究基于的主要背景，并列出了本文的主要工作。

第二章介绍研究中文命名实体识别的准备知识，包括命名实体识别的研究背景，命名实体识别的国内外的研究状况，并且详细介绍了命名实体识别的几种模型，包括基于分类的语言模型，隐马尔科夫模型，最大熵模型，AdaBoost 分类模型，支持向量机模型和风险最小化分类器模型。

第三章介绍了研究中文分词的准备知识，包括中文分词的背景，中文分词的研究现状，介绍了基于词典的中文分词方法，基于统计的中文分词方法和基于理解的中文分词方法，介绍了中文分词的主要方法，包括动态匹配算法，最大匹配算法，边界片段发现方法和 N 元分词模型，并介绍了融合命名实体识别的中文分词模型。

第四章主要介绍基于规则匹配的命名实体识别。包括基于本体的中文人名层次分类体系的构建，命名实体识别的规则库的构建，命名实体的识别以及反馈规则的产生，并给出了实验结果。

第五章进行总结和展望。

第二章 命名实体识别的相关研究

由于汉语书写的句子具有特殊性，通常包含一连串的字符，这些字符之间没有明显的分隔符，因此要对中文的文章进行处理，首先就要把文章分隔到可以理解的单位，那就是词语。但是分词的一个值得注意的问题就是未知词语的识别，而命名实体又是未知词语中出现量最大的一类，因而识别出中文命名实体很重要。本章主要介绍了命名实体识别所需的各种相关知识。首先，介绍了命名实体识别的研究意义；其次，介绍了命名实体研究的国内外现状；最后，介绍了几种命名实体识别的模型，包括基于分类的语言模型，隐马尔科夫模型，最大熵模型，AdaBoost 分类模型，支持向量机模型和风险最小化分类器模型。

2.1 命名实体识别的研究意义

汉语书写的句子通常包含一连串的字符，这些字符之间并没有明显的分隔符，因此要对中文书写的文章进行处理，首要的任务就是分词。大多数的汉语文章处理系统是依赖于词典来识别出文章中的每个词语，但是值得注意的是，这种方法是不可能完全识别文章当中的所有词语的，因为所使用的词典并不能包含所有的词语，因而词典中没有的词语变成为分词中值得特别注意的问题。

这些词典中不包括的词，形象的称之为未登录词。在所有未登录词当中，命名实体是最为重要的一类。命名实体组成的一片文章的主要内容。因此，命名实体的识别成为许多自然语言处理应用的主要任务，像信息抽取，问题回答系统还有机器翻译等等。

命名实体识别属于信息抽取的范畴，MUC 会议定义了信息抽取的五个子任务，包括命名实体识别、指代消解、模板元素抽取、模板关系抽取和场景抽取。命名实体识别是信息抽取任务的第一步，是其它任务的基础。所谓命名实体识别是要抽取出文本当中的特殊信息，如人名、地名、组织机构名等实体，还有时间表达式如日期、时间等以及数字表达式如货币值、百分数等[3]。

一篇文章当中，实体是最基本的信息元素，它们往往指示了文章的主要内容，因此，识别出文章当中的命名实体是对文章进行理解的重要前提，同时命名实体识别的质量也会直接影响到一系列的后续工作，因此，命名实体识别已经越来越成为自然语言处理中的关键技术，命名实体的识别成为多种自然语言处理技术的基础，如信息抽取、信息检索、机器翻译和问答系统等[4]。

中文命名实体的识别也是汉语自动分词的难点之一，分词是深层次中文信息处理的基础，现有的中文自动分词方法中，基于词典的分词方法占有主导地位，中文分词的主要困难不在于词典中词条的匹配，而是在于切分歧义消解和未登录

词语的识别[5,6]。汉语通过派生，复合，缩写等形式产生了很强的造词功能，在词典中不存在的词成为未登录词[7]。未登录词是信息处理中的难点和热点，对中文词法切分发挥着重要作用。中文词法切分的大部分错误是由未登录词的识别错误造成的[8]，未登录词的主要形式包括人名，地名，组织机构名等命名实体，以及时间词，数量词和普通的语法派生词等[9]。

2.2 国内外研究现状

对于英文命名实体识别的研究起步较早，英文命名实体的识别已经达到了较高的水平。英文命名实体的识别主要采用基于统计模型的方法，常常采用的模型有隐马尔科夫模型、最大熵模型和支持向量机等[10-14]，这些统计模型能够很好地利用英文命名实体的词频、词缀等统计信息，对首字母大写的词串进行分析，并结合一定的上下文信息、句法信息、语义信息确定该词串是否是命名实体。

与英文相比，中文命名实体识别还存在一些困难，如缺乏明显的特征标志，分词影响命名实体的识别，不同种类的命名实体间存在歧义问题，使得中文命名实体识别的准确率和召回率还达不到英文命名实体识别的水平。

最近几年的研究中，人们从大范围的关于命名实体识别系统的评估结果中了解到，效果最好的命名实体识别的系统是一种介于统计的分类方法，这种方法典型的特点是将线性的分类算法和大量的经过精心设计的语言学特征结合在一起。在这里，使用相关的线性的简单的分类算法是很重要的，这是因为，对于一些更加复杂的学习模型来说，这些模型很难有效的利用那些经典的语言特征。众所周知，成功的集成这些有用的语言特征是一个好的命名实体识别系统所必须具备的要素之一。总的来说，要构建一个高质量的命名实体识别系统，关键的地方就是如何去编排有效的语言信息，从而使得线性的分类算法可以有效地利用这些语言信息。这种模型的设计是由任务决定的，并且对取得好的性能至关重要。

可以想象，中文命名实体的识别要难于英文命名实体的识别。因为，中文作为一种语言，它更具有灵活的语言特征。首先，没有词语的精确的定义，词语之间也没有间隔作为标志，也没有外部标志比如大写去帮助系统识别中文命名实体，因此，中文命名实体的识别具有更多困难。其次，中文命名实体中的汉字常常也会用于其它普通词语的构成，有些更会被作为单字词用于句子中，因此，仅仅依靠汉字用字情况去识别中文命名实体也是比较困难的。

由于中文的复杂性，希望通过一系列关于中文命名实体的规则去有效的识别出命名实体是不太可能的。手工写过则不仅很消耗时间，并且投入很大，需要专门的语言学家修订规则。修订的规则还具有难以扩展的特点，一遇到新的语言新的特征就需要重新增加新的规则，规则多了之后不同规则之间还可能引发识别冲突。随之产生的是使用统计技术识别中文命名实体。

为了识别出中文命名实体，内在的和文脉上的信息都是必须的，而且语义信

息的使用将能更好的增加系统的表现。但是在实际的应用中，想要获得这些短语的语义花费也是巨大的。研究者希望可以识别出文章当中的中文命名实体，并使得识别方法有效简单且花费较合理。

规则与统计匹配相结合的方法是当前的主流方法，这种识别方法首先采用统计方法识别出文本中的命名实体，称之为候选实体，然后利用规则机制对候选命名实体进行校正过滤。

虽然规则和统计相结合的方法可以兼备基于规则的方法和基于统计的方法这两者的优点，但是这种方法在实际应用中仍然存在一些不足：由于命名实体的识别在分词之后进行，一些内部成词或者上下文成词的命名实体大都被切分为单字或单词碎片，从而使得部分命名实体很难召回。由于规则来源于事先观察到的语言特征，因而缺乏有效的学习机制，不具备接收识别结果的反馈信息的能力。

对于命名实体的识别可以使用自相似来检测是命名实体或者不是命名实体。这是一种相应的库，使用这个库可以检测这种关系或者是非关系的集中性。如果自相似的值是取 0 到 1 之间的一个正小数，可以这样理解，这表明这种关系或者非关系的集中程度是非常紧密的。相反的，集中程度就是松散的。

2.3 命名实体识别的模型介绍

2.3.1 基于分类的语言模型

基于分类的语言模型应用到中文命名实体的识别，可以取得较好的效果。每一种命名实体被定义成语言模型中的一大类，如人名、地名、机构名字。其他的剩余的词语也被定义相应的类别。假如选择人名、地名、机构名三类中文命名实体，那么结果是，一共有 N+3 个大类在这个基于分类的语言模型中。其中 N 表示的是词汇的数量。基于分类的语言模型可以被描述成以下内容。

给出一串汉字串，W=w1,...,wn，中文命名实体识别的任务就是从这串字符串中寻找出最佳子串 X=x1,...,xm(m≤n)，所谓最佳子串是要使得概率 P(X|W)达到最大值。可以用下面的公式表示以上的含义：

X=arg max P(X|W)=arg max P(X)*P(W|X)，称之为基于分类的语言模型。

基于分类的语言模型包括两个分支模型，上下文模型 P(X)和实体模型 P(W|X)。上下文模型隐含着这样的含义，从已有的上下文产生命名实体的可能性有多少。P(X)是一个前验概率，它可以通过计算得出：P(X)=∏P(xi|xi-1,xi-2) (i=1,2,...,m)。

我们采用最大似然估计法去计算 P(xi|xi-1,xi-2)，使用训练集计算如下：

P(xi|xi-1,xi-2)=count(xi,xi-1,xi-2)/count(xi-1,xi-2)。

实体的模型可以参数化如下：

P(W|X)=P(w1,...,wn|x1,...,xm)

=P([w1,...,wcj-Start]...[wcj-End,...,wn]|x1,...,xm)

= ∏P([wcj-Start,…,wcj-End]|xj) (j=1,2,…,m)。

2.3.2 隐马尔科夫模型

隐马尔科夫模型(Hidden Markov Model，HMM)是一种统计模型，是用来描述随机过程的方法。在隐马尔科夫模型中，状态转移过程是不能被观察的，能被观察到的事件是状态的随机函数。这也就是说，隐马尔科夫模型指的是马尔科夫模型的内部状态外界不可见，外界只能看到各个时刻的输出值。

隐马尔科夫模型可以完整表示为一个五元组(ΩX，ΩO，A，B，Π)，以二元隐马尔科夫模型为例，各种参数的定义为：

(1) ΩX 表示状态的有限集合 ΩX = { q1,...,qn)；

(2) ΩO 表示观察值的有限集合 ΩO ={vl,...,vm)；

(3) A 表示转移概率矩阵 A= {aij}, aij=P{Xt+1=qj|Xt=qi}；

(4) B 表示输出概率矩阵 B= {bik},bik= P{Ot=vk|Xt=qi}；

(5) Π表示初始状态分布概率Π={Πi},Πi=P{Xi=qi}；

基于隐马尔科夫模型的训练过程是把隐马尔科夫模型应用到实际当中的三个问题之中的一个，那就是如何训练使得隐马尔科夫模型的参数 A，B，Π去适应之前观察到的观察序列值。也就是转移概率矩阵[aij]和输出概率矩阵[bik]的求解，求解的过程可以用经典的 Viterbi 算法实现。

2.3.3 最大熵模型

最大熵模型是一种用于概率估计的模型。给定背景知识或者是上下文信息，最大熵模型可以发现一个可能的分配，使得其满足所有约束，而基于的事实是使得熵达到最大。最大熵模型可见简单的描述如下，模型化所有已知的，猜想未知的部分。

根据最大熵模型，基于已知的信息去推断未知信息的时候，选择符合已知条件的，熵最大的那个模型，这个模型是唯一准确的表达。

对应到汉语文章的处理过程，一篇文章包括很多个词语，词语 y 的产生过程中，可能会受到其上下文信息 x 的影响。大量的训练数据集一般都会包含有 x 和 y 的共现信息，但是想要对所有可能出现的(x，y)对，完全精确地确定 p(y|x)一般是不能做到的。那么，最大熵模型为这个问题提供了解决的办法，就是基于 x 和 y 的统计特征，可靠地估计出一个概率模型 P(y|x)。

给出一组记录集合 D={d1,d2,…,dn}，类集合 C={c1,c2}，使用每个成员的类信息熵去离散化边界信息，对于给定的 n 个指示函数，希望模型 p 能属于 P 的子集 C 中，这个子集 C 可以定义为：C={p∈P|Ep(fi) = Ep'(fi')for(i∈{1K n})}，最大熵模型的思想就是在属于子集 C 的模型中选择那个最均匀的分布，用数学的形式表示就是：p*=arg max(H(p))(p∈C)。

2.3.4 AdaBoost 分类模型

AdaBoost 是 Boosting 家族的一个代表算法[15]。算法的原则是转换或者说推进大量的弱学习算法为强学习算法。这个算法已经被证明，如果弱学习算法的结果好于随机猜想而且数量趋于无限大，那么最终的强学习算法的错误率将接近于零。

基本的 AdaBoost 算法解决的是二元分类问题，但是在现实世界中，分类问题多数是多类别的，有很多扩展方法可以使得 AdaBoost 解决多类别问题，AdaBoost.M1 就是对其最简单易懂的概化[16]。它的描述如下：

给定(x1,y1), (x2,y2)...,(xn,yn)，其中 xi 表示训练样本集合，yi 是类标记，yi ∈Y={1,...,k}，k 表示类别的数目。如果 p 是正确的，那么 I(p)=1，否则 I(p)=0。初始化 D(i)=1/n，i 从 1 循环到 T(循环次数)，不断地训练弱分类器，训练时使用可分配的 Dt，εt=Dt(i)I(ht(xi)!=yi)，如果出现 εt>1/2，则可以转到输入阶段，同时设置 T=t-1。否则，则要更新 Dt+1=Dt(i)exp(-αiIht(xi)=yi))/Z，其中，αi=ln((1-εt)/εt)。循环结束后，可以转到输出阶段：

H(x)=arg max f(x,y)=arg max(∑αiI ht(x)=y)。

AdaBoost 是很容易编码实现的，它可以使弱分类器转化为强分类器，而且不需要设置任何参量，只需要弱分类器是确定的即可。但是 AdaBoost 需要大量的训练数据，并且需要较长的训练时间才能达到好的效果。

2.3.5 支持向量机模型

支持向量机(Support Vector Machine)是一种功能强大的机器学习方法[17]，这种模型已经被应用的自然语言处理的很多任务，如词性标注、未知词语猜测，并且达到了很好的效果。众所周知，支持向量机在高维特征空间和高水平的特征冗余的广义性有显著的性能。

支持向量机的原则可以被简要描述如下：开始于一组输入—输出对(x1,y1), (x2,y2)...,(xn,yn)，其中，X∈Pn ，yn∈{-1,1}。它们可以被一个超平面分割开，t.x+b=0。可以把这个超平面这样描述：yi [(t.xi+b)]-1≥0。在这些超平面中，最合适的分割超平面是如此选择的：它距离最近点的距离是最大的。图 2.1 显示了一个简单的二维样本。

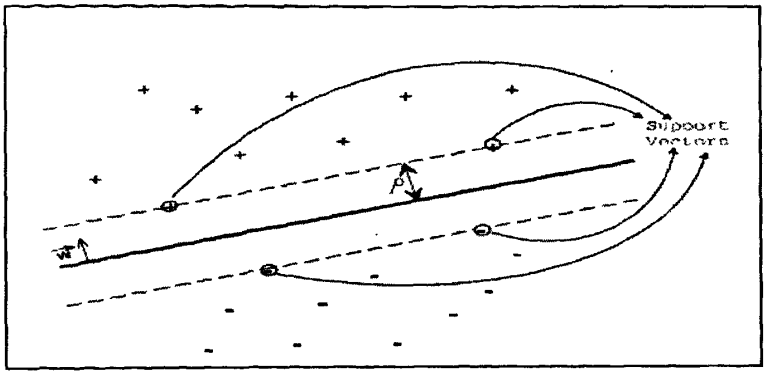

图 2.1 最佳分割超平面

基于支持向量机的学习模型的目标是，找出那些最佳的分割超面，把这些训练数据分成 N 个维度，分割的标准正如上面所描述的那样。然后，用这个标准去把真实数据分割成为同样的维度即可。

支持向量机的决定函数可以根据相近的核心函数得到：

$f(x)=sign(\sum \alpha iyiK(xi,yi)+b)(i,j=1,\ldots,n)$

其中，n 是样本的数量，K(xi,yi)是核心函数，b 是修正斜线。

2.3.6 风险最小化分类器模型

这种方法是把命名实体识别的任务看成是序列化的分类问题。使用 wi (i=0,1,...,n)去表示已经被标记过的文本序列，这个文本序列是作为系统的输入数据，那么每一个标记 wi 都应当被分配给一个类别，给它一个类标签 ti。对于命名实体的识别，所有类标签形成一个序列{ti}，这个序列记录着命名实体的详细信息，如 BIO 就是这种那个命名实体的信息。

在具体的命名实体识别系统，风险最小化分类器的工作方式是这样的，每一个标记 wi 连同它的类标记 ti 通过计算条件概率可以预先得知，对于每一个可能的类标记值 c 条件概率 P(ti=c/xi)，而 xi 代表的是标记 wi 的特征向量。

假设 P(ti=c/xi)=P(ti=c/wi,{tj}j≤i)，特征向量 xi 依赖于经过精确预测的类标记{tj}j≤i，但是，这个依赖性具有局部的代表性。在风险最小化算法中，上述的条件概率模型可以用参数形式表示如下：

P(ti=c/xi,ti-l,...,ti-1)=T(wcT xi+bc)，其中，T(y)=min(1,max(0,y))。

T(y)是 y 在[0,1]区间的切断点。wc 是一个线性权向量，bc 是一个常量。参数 wc 和 bc 可以通过训练数据得到估计。

2.4 小结

汉语书写的句子通常包含一连串的字符，这些字符之间并没有明显的分隔

符，因此要对中文书写的文章进行处理，首要的任务就是分词。大多数的汉语文章处理系统是依赖于词典来识别出文章中的每个词语，但是值得注意的是，这种方法是不可能完全识别文章当中的所有词语的，因为所使用的词典并不能包含所有的词语，因而词典中没有的词语变成为分词中值得特别注意的问题。命名实体就是这种词典中没有的词语。一篇文章当中，实体是最基本的信息元素，它们往往指示了文章的主要内容，因此，识别出文章当中的命名实体是对文章进行理解的重要前提，同时命名实体识别的质量也会直接影响到一系列的后续工作，因此，命名实体识别已经越来越成为自然语言处理中的关键技术。

第三章 融合命名实体识别的中文分词模型

在书写汉语的时候，一个词语在一句话中没有被明显的标记出来，因此，为了能理解汉语，一个最重要的任务就是必须把这些连续的中文汉字给分隔开来。本章主要介绍了中文分词的各种相关知识。首先，介绍了中文分词研究的背景；其次，介绍了中文分词的研究现状，介绍了基于词典的中文分词方法，基于统计的中文分词方法和基于理解的中文分词方法；接着，介绍了中文分词的几种方法，包括动态匹配算法，最大匹配算法，边界片段发现方法和 N 元分词模型；最后，介绍了中文分词存在的难点，进而引出了融合命名实体识别的中文分词模型。

3.1 中文分词的研究背景

中文分词之所以能够成为中文信息处理中的一个重要环节，是由汉语本身的特点所决定的。英文是以词为单位的，词和词之间是有空格隔开，而汉语与英语等其他语言不同，汉语是以字而不是词作为语言的基本构造单位的，即句子中所有的字连起来才能描述一个意思。汉语这种特有的书写特点，使得中文信息处理必须经过分词这样一层的基本处理阶段，才能够进入上层的句法和语义阶段的处理。如果不经过分词的处理，那么在句法和语义的分词阶段计算机要直接面对一系列的独立汉字组成的汉字串，这样就丢弃了汉语当中能够作为相对独立的成分并具有相对独立意义的词中的重要信息，从而加大了上层的复杂度，有时甚至根本无法继续进行或者形成完全错误的分析结果。所以，中文分词仍然是现阶段计算机中文信息处理的基础和不可逾越的阶段。对分词技术深入透彻的研究和对分词精确度的进一步提高，是对现代中文信息处理卓有成效的贡献[18,19]。

分词这个问题看起来似乎很简单，容易，但是分词并不是一个琐碎的价值不高的事情，事实上，分词是很具有积极意义的，它对计算机语言的研究具有主要作用，也越来越被中文语言处理协会所重视。为了能够完成分词的任务，各种各样的技术发展开来。

对于分词问题自身来说，有两个重要的问题吸引研究者的注意，其中一个问题是背景模糊，另外一个就是未知词语(未登录词)。背景模糊的问题只要是由汉字符可以出现在不同的词语中，可以出现在词语的不同位置。至于未知词的问题被认为在系统的词典中不存在的词语。总的来说，想要识别出这些词语还是较为困难的，因为在系统缺少和它们相关的知识。

中文分词是中文信息处理最重要的预处理过程，被广泛地应用于机器翻译，自动分类，信息检索以及搜索引擎研究等多个方面。

(1) 机器翻译

所谓机器翻译，顾名思义，就是利用计算机自动地实现不同语言之间的转换。机器翻译一般要经过分析和生成两个步骤。最关键的还是对源语词语的分析，当汉语作为源语时，分词就是源语分析工作的基础，因此，中文分词是机器翻译不可缺少的一个环节。

(2) 自动分类

自动分类系统是信息处理的重要研究方向之一。文本分类涉及到文本特征集的构建，而对汉语来讲，文本的特征集主要由一些具有较大权值的词语构成，因此自动分词是文本分类面临的首要问题，分词的效果将直接关系着文本分类的结果。

(3) 搜索引擎

中文搜索引擎的关键点在于抽取出重要的信息，而这其中的难点就是中文分词。随着因特网在我国的发展和普及，中文搜索引擎研究有了重大突破，在短期内就涌现出了许多重要的中文搜索引擎，并得到了广泛应用。但是，中文搜索引擎研究开发仍然存在大量的问题，如在信息组织、检索速度、检准率和检全率等方面还有较大的发展空间。今后，中文搜索引擎的研究方向应是将中文自动分词、信息检索、自然语言理解和人工智能等与搜索引擎研究相结合。

3.2 中文分词的研究现状

中文自动分词技术，主要从上个世纪八十年代开始研究的，经过二十多年来的发展，问分词技术已经在中文信息处理的很多领域得到了应用。从简单的基于词典的分词方法，到基于规则的分词方法，再到基于统计的分词方法，分词的研究方法变得越来越多也越来越有特色[20-22]。

早期的分词系统主要以基于词典的分词方法为主。例如北京航空航天大学计算机系于 1983 年设计实现的 CDWS 分词系统，山西大学计算机系研制的 ABWS 自动分词系统等。由于基于词典的分词方法，精度不高，不能达到实际应用中对于分词精度的要求，研究的方向开始转向考虑能否利用句法、语法、语义等语法和语义知识来对文本进行切分。即构造一个知识库来处理分词中遇到的一些歧义的问题。由于自然语言的模糊性和复杂性，这种研究方法很难包括所有的语言知识，因此在后来一段时间这个研究方向受到了一定的限制。到了上世纪九十年代，随着大规模语料库的出现，开始利用统计的理论进行分词的处理，取得了较好的效果。例如，哈尔滨工业大学设计实现的基于统计的分词系统。目前中文分词技术的研究反向大都集中在规则的方法和统计的方法两类，这两种方法各有优点，也都存在一定的不足之处。因此，有很多自动分词系统将规则和统计的方法结合在一起进行分词，取得了不错的分词效果。例如北京大学计算语言所的 ICTCLAS 分词系统和海量科技的分词系统。

与此同时，也提出了许多算法，具有代表性的算法有：最大匹配算法，包括

向前匹配，向后匹配和前后匹配，最大似然方法，特征词典的方法，邻接矩阵的方法，人工神经网络的方法，关联回路的方法等等。

文献[23]人提出了一种词频统计中文分词技术，她的文章中详细地介绍了一个基于词频统计的中文分词系统的设计和实现。基于词频的系统主要选用了三种统计原理分别进行了统计：互信息、N 元统计模型和 t—测试。论文还对这三种原理的处理结果进行了比较，分析各种统计原理的统计特点，以及各自所适合应用的地方。这几人又提出了基于支持向量机的词频统计中文分词研究，详细介绍支持向量基在基于词频统计的中文分词系统中的应用。使用这个方法的好吃是可以是词典中不重复地存储了每次处理中得到的汉语，以及这些词语出现的频率。选用了互信息原理进行统计。

为了给出关于中文分词在普通测试集合上结果的全面的比较，三届国际中文分词 Bakeoffs 分别在 2003 年，2005 年和 2006 年举行。在所有被提出的方法中，基于字符标签的方法替代的传统的基于词汇的方法，取得较好的成绩。

自从 2003 年国际中文分词评测活动 Bakeoff 开展以来，基于字标注的统计学习方法引起了广泛关注。

在 Bakeoff2003 中各种分词技术的优劣尚难分伯仲，但是既然未登录词对分词精度的影响比分词歧义大至少 5 倍以上，能获得最高 OOV 召回率的分词方法自然会得到人们的青睐。

Bakeoffs 的成功不仅在于它给出了一些公开的分词的标准，而且在于它提出了一种基于文集的分词标准代表，这种标准取代了传统的已知词典和手工分隔方法。因此，中文分词越来越倾向于基于文集标准的机器学习的处理。

文献[24]提出了一种基于层叠隐马尔科模型的中文文章分词方法。在这篇文章中提出了一种将中文分词方法和词性标注以及切分排歧还有未知词识别集成在一起的，一个完整的理论框架的模型方法。在分词方面，采取的是基于隐马尔科夫模型，在这层隐马尔科夫模型中，未知词语的处理方式和词典中的普通词一样。实验结果显示，层叠隐马尔科夫模型的各个层面对汉语词法分析都发挥了积极的作用。文献[24]还介绍了基于层叠隐马模型的汉语词发分词系统 ICTCLAS 的实现。

然而，中文自动分词仍然是限制中文信心处理发展的一个主要的问题，模糊不清的背景也仍然是影响分词系统正确性的主要事实。模糊背景的种类主要包括粗糙的交叠的模糊背景和混合的模糊背景。

很多已经存在的系统常常训练处一系列的不同的模型去处理分词中出现的问题。这种处理导致了训练模型的协调性和可扩展性都很差，因此需要的是少重复性多层次性的处理。

基于统计的方法首先要为语言处理问题建立统计模型，并且使用训练语料库来训练产生统计模型中的参数，然后把参数应用到分词的处理过程中。近几年来，

基于统计的分词方法占了主要的地位。

基于统计的分词算法的基本思想是：对输入字串进行全切分，找到的所有可能切分结果，对每种切分结果利用能够反映语言特征的统计数据计算它的出现概率，从结果中选取概率最大的一种。概率的计算方法依赖于所建立的语言模型。基于统计的分词方法所需要的参数可以通过对大规模语料库进行训练获得，这种方法也得到越来越广泛的使用。

3.2.1 基于词典的中文分词方法

最大匹配法最早由苏联学者提出。这个方法提出的理论依据是，汉语的词汇通常可以作为构成词的基本资料初级资料构成新的词语[25]。文献[26]首次将这个方法应用到大规模的中文自动分词的系统中。在最大匹配算法中，根据扫描句子的方向，可以分成正向最大匹配算法和逆向最大匹配算法两种。根据梁等人的实验结果，改方法在词典构建完备并且没有任何其他知识的情况下，最大匹配算法发生的错误分割率为 1 次 / 169 字～1 次 / 245 字，并且具有简单、快速的优点。Guo J.更是对最大匹配算法的工作原理作了严格的形式解释。

(l) 正向最大匹配法

其基本思想为：设 D 为词典，max_length 表示 D 中的最大词长，str 为待切分的字串。正向最大匹配法是每次从 str 中取长度为 max_length 的子串与 D 中的词进行匹配。如果能够成功地匹配当前子串，则认为该字符串是一个词语，指针后移 max_length 个汉字后继续匹配，否则子串逐次减一进行匹配。该方法设计思想简单，易于机器实现，时间复杂度也比较低。但也有一些不足之处：最大词长 max_length 难以确定，词库难以容纳所有词汇，每次都从长到短对子字串进行匹配，否认了词中含词这一语言现象，出错率高。因此，正向最大匹配一般不单独使用，而是作为一种基本的机械切分方法同其他方法配合使用。

(2) 逆向最大匹配法

逆向最大匹配的方法其实和正向最大匹配的方法有很多相同之处，唯一不同的地方是，逆向最大法的扫描方向不同，这种方法是按照从右至左的方向依次取出子汉字串和词典进行匹配。统计结果表明，逆向最大匹配法在切分的准确率上比正向最大匹配法有很大提高。

(3) 双向匹配法

双向最大匹配法即对同一个字符串分别按照正向最大匹配和逆向最大匹配的方法进行切分处理，如果能够得到相同的切分结果，则认为切分成功，否则要做进一步的分析处理。Sun M.S.和 Benjamin K.T 注意到：汉语书写的文章中 90% 左右的句子，正向最大匹配方法和逆向最大匹配方法的切分完全重合且正确，9% 左右的句子正向最大匹配方法和逆向最大匹配方法的切分不同，但其中必有一个是正确的。

双向匹配法使得算法的复杂度有所提高。为了使词典更能支持正向和逆向两

种顺序的匹配和搜索，词典的结构要比一般的词典结构要复杂一些。

(4) 最少切分法

也成最点路径切分法，该方法通过查找词典，找出字符串中存在的所有词，构造一个有向无环图。采用层进式最短路径法来得到最后的切分结果。但是由于大多数汉字均可构成单字词，所以按最少切分法匹配的分词结果往往因分得太细而不合要求。

(5) 最佳匹配方法

对分词词典进行必要的处理，按词的出现频率的大小排列词条，高频率的词排在前，低频率的词排在后，从而缩短查询分词词典的时间，加快分词的速度，使分词达到最佳的效果。这种分词方法对于分词的算法没有什么改进，只是改进了分词词典的排列顺序，它虽然降低了分词的时间复杂度，却没有提高分词的正确率，而且这种方法的空间复杂度也稍有增加。

基于词典的分词方法有很多优点，易于实现是这种分词方法最值得指出的一个优点；当然也不能忽视基于词典方法的缺点，基于词典分词法的缺点是分词的速度相对较慢，而且还会存在模糊的交集问题以及存在岐义切分的问题，没有经过证明有效的公认的词典，词典中词语的定义没有一个可参照的统一的标准，使用不同的词典还有可能会产生的切分差异。

对于基于标记的分词方法，这种方法通过实际证明，标记的设置对于分词精度的提高没有帮助。而且，基于标记的分词方法还有可能引起本来不会产生的分词的错误。

基于词典的分词方法加上对岐义切分的修正是对基于词典的分词方法的一种改进。这种改进方法增加了规则，利用新增加的规则对切分产生的岐义进行校正，希望可以提高分词的准确率。经过实际的应用证明，这种相结合的方法改良是很有效的。

3.2.2 基于统计的中文分词方法

根据常识，汉语中词语的组成还是有一定规律的。词语中字与字共同出现给研究者们一些启示：即相邻的几个字共同出现的次数越多，那么它们构成一个词语的可能性就越大，因此可以这样理解：字与字共同出现的频率能够很明显的告知它们构成词语的可信度。

基于以上的分析，做出以下定义：字的共现信息。这个共现信息其实简单来说，就是表现了汉字之间的关系。当几个字之间的关系紧密到一定程度的时候，可以认为这几个汉字有可能构成了一个词，这个紧密的程度习惯上使用阈值来衡量。对于这种方法，首先要做的是对文章中的字与字出现的次数进行统计，而且使用这种方法的一个好处是不需要分词的词典，因此，我们也可以称之为无词典的分词方法。

当然，不可否认这种方法也还是存在一定缺点的，经常会分割出一些共同出现频率高，但是其实并不是词语的字符串。使用该方法在某些情况下，时空开销也很大。

一个基于统计的计算语言模型以概率分布的形式描述了任意语句 W 属于某种语言集合的可能性。假定词是一个句子的最小结构单位，并假设一个语句 W 由词($x_1,x_2,x_3,\ldots,x_n$)组成，可定义语句 W 的 N-Gram 模型：

$P(W)=P(x_1)P(x_2|x_1)P(x_3|x_1x_2),\ldots,P(x_n|x_1x_2,\ldots,x_{n-1})=\prod_{i=1,\ldots,n}P(x_i|x_1,\ldots,x_{i-1})$

对于概率 $P(x_i|x_1,\ldots,x_{i-1})$可以采用最大相似度估计的方法，用以下公式估算：

$P(x_i|x_{i-1})=count(x_{i-1},x_i)/\sum count(x_{i-1},x_i)$

其中 $count(x_{i-1},x_i)$为词对 x_{i-1},x_i 在训练语料库中出现的次数，用于估算基于统计的计算语言模型中的概率分布的训练语料库的文本称为训练数据，根据训练数据估算 $P(x_i|x_{i-1})$的过程称为训练。

基于统计的分词方法在应用到实际过程中时，为了提高分词的准确率通常都会为之配备一个基础性的词典，那么再使用统计的方法去识别一些未知词，就可以发挥出两者相结合的优点，提高词语切分的正确性。

基于统计的分词方法经常采用的经典的模型有以下几种：隐马尔科夫模型、AdaBoost 模型、支持向量机模型和最大熵模型等[27-31]。上述的这些统计模型一个重要的实现就是，利用词语之间的概率信息作为分割的的依据。

3.2.3 基于理解的中文分词方法

基于理解的分词方法的基本思想是，让计算机模拟人的大脑对于汉语句子的理解方式，像人脑那样的机制工作，从而实现对词语的切分[28]。

基于理解的分词方法，是在分词的时候把中文的句法分析和语义分析涵盖进去，利用这些信息来处理概念模糊和未知词语的现象。人工智能是对信息进行智能化处理的一种模式，主要有两种处理方式：一种是基于心理学的符号处理方法。模拟人脑的功能，构造推理网络，经过符号转换，从而可以进行解释性处理。一种是基于生理学的模拟方法。神经网络旨在模拟人脑神经系统机构的运作机制来实现一定的功能。应用到分词方法上，产生了专家系统分词法和神经网络分词法。人工智能分词技术的关键是如何在分词过程中引入有用的词法、句法及语用等各种语义知识进行有条件的切分。

专家系统分词法是把知识推理应用到分词的过程中，搭建推理网络。把汉语的词法信息、句法信息以及语义信息单独作为维护的结构体。分词的时候，构建词法语义树，这个树的节点就是已经分割的词语还有未被识别出的词语。那么，分词的过程其实就是生成词法语义树的过程，当然还有利用实现构建的规则修正模糊的部分。

神经网络的分词方法的关键在于构建分词知识库，建立推理规则。分词的时

候也要利用词法语法规则，包括词语内部的规则还有词语外部的规则。分词的时候要构建一个动态的分词网络，不断调整权值进行匹配，最后实现分割结果的输出。

3.3 中文分词方法介绍

3.3.1 动态匹配算法

动态匹配算法的思想不是从左到右做出边界决定，而是在一整个句子里寻找最合适的切分。为了完成这个任务，下一个最长的词语的标准已经换成最长的平均词语的长度。这实际上是一种简单的找寻词语的方法，使得一个句子中词语的个数最少。

动态匹配的算法也是基于词典的，每一个词典中的词语都有相同的评分，为了找出整个句子中最合适的切分，可以使用动态规划。假设，最长的词包括四个汉字，那么可以定义分割的分数，$\delta(C(1,\ldots,t))=\min\{\delta(C(1,\ldots,t-1))+\theta(C(t)), \delta(C(1,\ldots,t-2))+\theta(C(t-1,t)), \delta(C(1,\ldots,t-3))+\theta(C(t-2,t)), \delta(C(1,\ldots,t-4))+\theta(C(t-3,t))\}$。其中，如果 C(i,...j)在词典中，则 $\theta(C(i,\ldots j))=1$；否则 $\theta(C(i,\ldots j))$为负无穷大。C(1,...,t)是汉字序列($c_1,\ldots,c_t$)，$\delta(C(1,\ldots,t))$是 t 时刻的计算分数值，实际上是计算到第 t 个汉字时的最小词语数目。

动态匹配算法没有把词语出现的频率考虑进去，一种简单的冠以动态匹配算法的拓展是把词语修正值 $\theta(C(i,j))$换成负的对词语概率取 log 值。

3.3.2 最大匹配算法

根据噪音信道模型，可以这样理解，通过信道的传送，一个词语串失去了词语的标志，这是因为噪音的干扰。接着自动分词是这样的过程：指出一个词串，这个词串具有这样的特征，在所有已知的汉语字符串中具有最大的概率[32]。公式表达 W=arg max P(W|Z)。通过贝叶斯公式转换为：W=arg max P(W|Z)=ard max (P(W)P(Z|W)/P(Z))。其中，P(Z)是中文汉字符串的概率。对于所有候选的词语串它都是相同的，所有不必对它做过多的关注。P(Z|W)是一个条件概率，是从词语串到汉字符串的条件概率。假如词语串已知的情况下，则相应的汉字符串的概率是 1，也就是说从词语串到汉字符串的转换的数量只有一种。于是，唯一需要考虑的是 P(W)，称之为词语串概率。因此，公式可以进一步简化为：W=arg max P(W)。

这就意味着，具有最大概率的词语串是最合适的词语串。但是，词语串的概率或许可以使用 N 元模型区计算出。如果我们使用二元模型，那么 $P(W)=\Pi P(wi|wi-1)$ (i=1,...,n ,w0 是普通的首词语串)。

但是，使用二元模型去计算概率进而分词，概率的转换矩阵从一个词语到另外一个词语的规模是相当巨大的。这是因为，基本的词表包含了大量的词汇，导

致转移矩阵的规模也相当大。因此，只能使用一元模型区解决词语串的概率问题：$P(W)=\prod P(wi)$ (i=1,...,n)。

每一个词语的概率可以通过它出现的频率而被估计出来，因此，一个词语就必须具有两个属性，并且词语频率要至少出现在词语表中。

基本思想是：首先，对于输入的汉字符串，根据词语表可以找出所有的可能的词汇，接着，找出所有可能的分词路径，并且计算出每条路径的概率，那么概率值最大的那条路径就是最佳路径，也就是可以输出的路径。

因为每个词语的概率都是一个很小的小于 1 的正数，如果汗字符串很长，那么结果可能是每一个词语串的概率值约等于 0，所以机器就不能成功地把这些词语分割出来。这个可以这样解决：计算出每个词语的概率的对数和，这样乘法可以转换成加法，但是计算的结果是一个负数，取结果的相反数然后它会变成一个正数，这个正数我们称它为“FEE”。很明显，对于一个词语串，概率越高 FEE 的值越低。$FEE(W)=\Sigma$ -log Pi(W) (i=1,...,n; -log Pi(W)是第 i 个词语的 FEE)。

3.3.3 边界片段发现方法

给定一个词典，边界点就是句子当中那些只是词语边界的位置，无论这个句子按照哪种方式去分割。边界片段是一个次序，这个次序在两个相邻接的边界点之间。边界片段在一个句子中可以使用下面的算法来识别，图 3.1 是边界片段识别算法流程图。

这个算法可以通过简单的例证来证明。首先，算法在所有字符中寻找最长的可以匹配的词语，寻找的依据是词典。接着，算法检查这个最长的词语是否覆盖其他的词语。那些相交迭的词语相互重合在一起，它们仍然分隔在同一组。通过相结合，边界片段就在识别结果中输出。

使用这个算法，一个句子被分解为一个边界片段序列。这个分隔的根据是跟踪背景。一方面，边界片段的定义是根据不同边界片段的字符是不相关的，在相邻的边界片段是不会形成词语的。另一方面，所有相近相关的字符都存在于一个边界片段中。这种分隔保证了任何一个隐含的，边界不明确的序列都不会被分隔到两个不同的边界片段中。

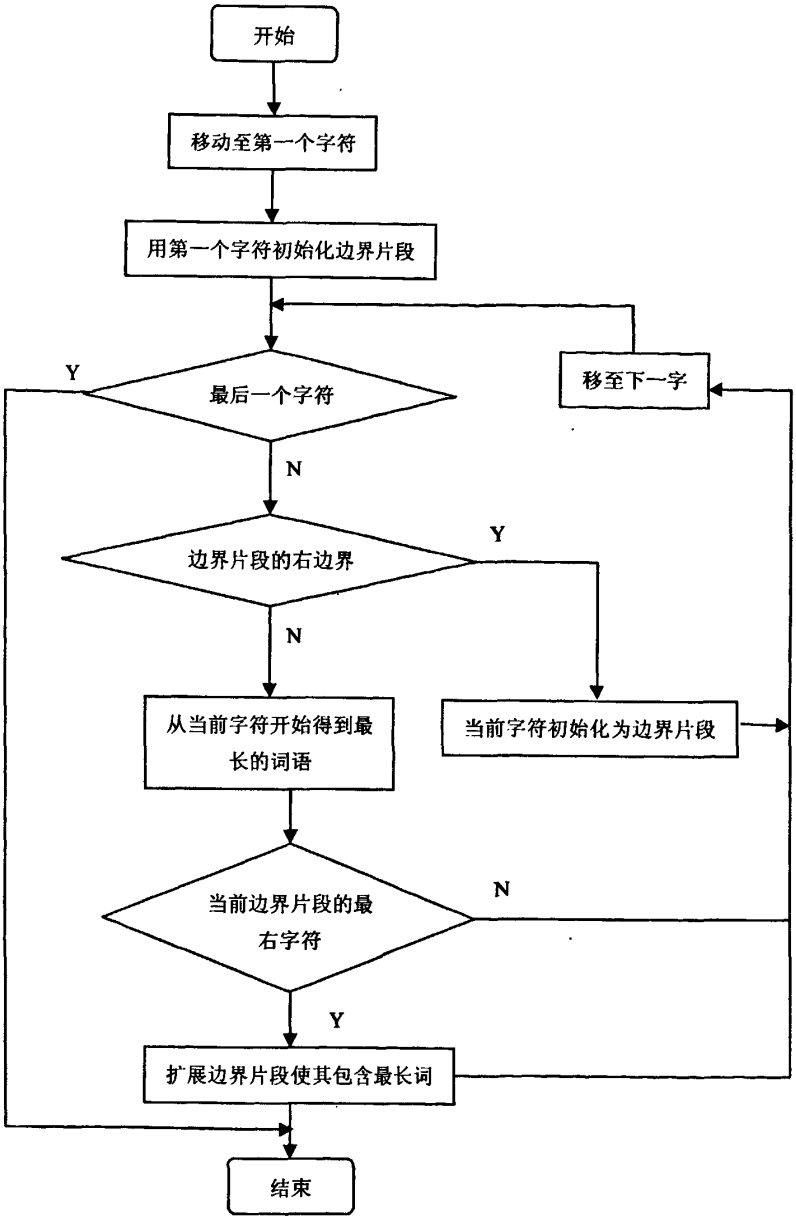

图 3.1 边界片段识别算法流程图

3.3.4 N 元分词模型

对所有可能的分割点进行分割，使用一元统计模型，企图发现最佳的词语序列，W_0=arg max P(W)=arg max ΠP(wi) (i=1,...,n)，很显然，这个效果要好得过直接匹配。

对所有可能的分割点进行分割，使用二元统计模型，最佳词语序列的格式演变成：W_0=arg max P(W)=arg max ΠP(wi|wi-1) (i=1,...,n)，那么值得期待的是，二元模型将会比一元模型的分割效果更好。

对所有可能的分割点进行分割，使用二元统计模型，同时相结合使用一元模型对二元模型进行平滑过滤：W_0=arg max P(W)=arg max (∏P(wi|wi-1) *(1-α)+ P(wi)* α) (i=1,...,n)， α 的取值根据实验结果得出。二元模型可能会带来的稀疏问题，通过上述方法计算可以很自然的解决这个问题。

对所有可能的分割点进行分割，使用二元统计模型，时相结合使用二元模型对二元模型进行平滑过滤：W_0=arg max P(W)=arg max (∏P(wi|wi-1) *(1-α)+ P(xi|xi-1) P(wi|ti)* α) (i=1,...,n)， α 的取值根据实验结果得出。值得注意，平滑采用的是不寻常的方法。它的工作方式是一种复杂的方式，相类似于基于二元模型的部分语言标记。

3.4 融合命名实体识别的中文分词模型

3.4.1 中文分词技术的难点

汉语的书写与英文是有区别的，英文是以词为单位的，词和词之间是有空格隔开，而汉语与英语等其他语言不同，汉语是以字而不是词作为语言的基本构造单位的，即句子中所有的字连起来才能描述一个意思。汉字和汉字，词语和词语，它们一个接着一个，换句话说，一个词语在一句话中没有被明显的标记出来。

分词这个问题看起来似乎很简单，很容易操作，其实则不然。分词并没有像想象中那么简单易操作，存在一些技术难点，而且分词也不是一个琐碎的价值含量不高的事情，事实上，分词是很具有积极意义的。对于分词问题自身来说，有两个重要的问题吸引研究者的注意，其中一个问题是背景模糊，而另外一个问题就是未知词语的识别，这些未知词语被形象地称之为未登录词。

中文自动分词问题中歧义字段和未登录词的切分是影响分词正确率的主要因素，是自动分词的两大难题。

背景模糊的问题只要是由汉字符可以出现在不同的词语中，可以出现在词语的不同位置。至于未知词的问题被认为在系统的词典中不存在的词语，命名实体就是属于未知词语的范畴，而且所占的比例比较大。总的来说，想要识别出这些词语还是较为困难的，因为在系统缺少和它们相关的知识。

因为各种中文处理系统都需要使用词频等信息，如果中文自动分词中对未登录词的识别不对，统计到的信息就会有很大误差。在实际操作中，超过60%的分词错误来源于新词，所造成的错误大大高于歧义引发的错误。因此在信息检索中，分词系统中的未登录词的识别十分重要。目前，未登录词识别的准确率已经成为评价一个分词系统好坏的重要标志之一。

3.4.2 融合命名实体识别的中文分词模型介绍

将命名实体的识别融合到中文分词的过程中，对提高中文分词的准确率起着重要作用。

中文分词模型如图一所示，首先对输入的汉语书写文章进行预处理，对文章进行适当的切分，得到下一步的出入文档，这是一系列的中文短句的集合；接着再对这些中文短句进一步的切分，这一步主要是将中文短句变成不可再分的字符串序列，这些字符串序列就可以作为中文分词和命名实体识别的基本单元；下一步，基于这些字符序列识别命名实体，与此也是展开分词；最后输出分词结果。

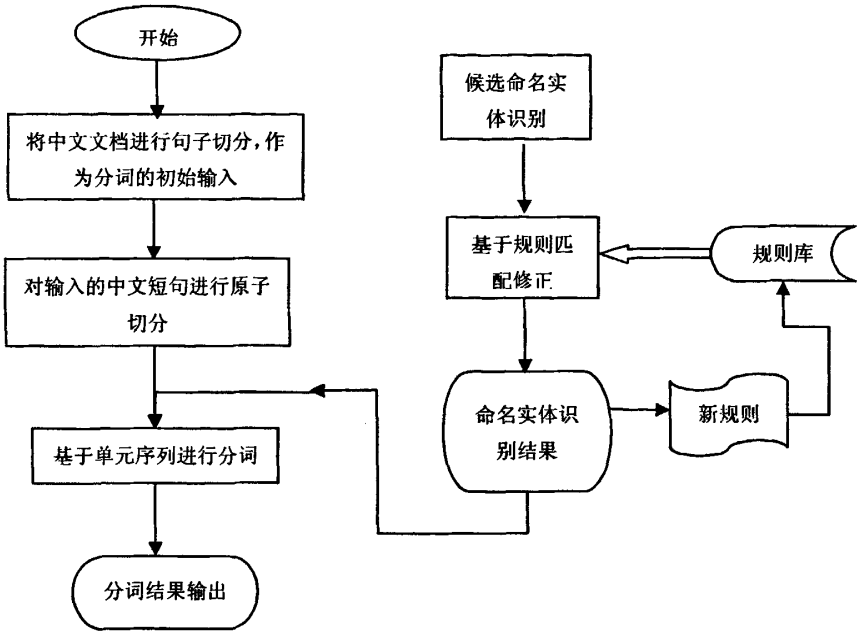

图 3.2 中文分词模型

3.5 小结

汉语和英文不同，英文是以词为单位的，并且英语文章中的词和词之间是有空格隔开的，而汉语与英语等其他语言不同，汉语是以字而不是词作为语言的基本构造单位的，即句子中所有的字连起来才能描述一个意思。

汉语这种特有的书写特点，使得中文信息处理必须经过分词这样一层的基本处理阶段，才能够进入上层的句法和语义阶段的处理。如果不经过分词的处理，那么在句法和语义的分词阶段计算机要直接面对一系列的独立汉字组成的汉字串，这样就丢弃了汉语当中能够作为相对独立的成分并具有相对独立意义的词中的重要信息，从而加大了上层的复杂度，有时甚至根本无法继续进行或者形成完全错误的分析结果。

对于分词问题自身来说，有两个重要的问题吸引研究者的注意，其中一个问题是背景模糊，另外一个就是未知词语(未登录词)。命名实体是未登录词的主要组成部分，在分词之前尽可能多地识别出文章当中的命名实体，对提高分词的查全率和查准率具有重要的意义。

第四章 基于规则匹配的命名实体识别

本章主要介绍了基于规则匹配的命名实体识别过程。首先，基于本体构建中文人名层次分类体系；其次，构造中文命名实体识别所需要的规则库，包括用于人名识别的规则库和用于地名识别的规则库；第三，介绍基于规则匹配的中文命名实体识别；最后给出实验结果。

4.1 基于本体的中文人名层次分类体系构建

本体论主要被应用于捕获一些感兴趣的领域的知识。本文是基于本体构建中文人名知识库的组织模型，把人名的构成分成若干了类别，每一个类别分成若干个层次，这种把中文人名知识分类管理，分层组织的模式，使得低层次的知识为高层次的知识提供依据，高层次的知识概化低层次的知识，从而大大提高知识库的可维护性[33-35]。

(1) 姓氏用字分析：

姓氏的个数有限，但是确是人名字结构中最重要的因素，本文整理得到的姓氏用字个数 737 个，其中单姓 729 个，复姓 8 个。

(2) 名字用字分析：

名字的用字范围相对于姓氏来说更为广泛，分布也更加的分散，并且值得注意的是在名字当中经常出现一些生僻字。

(3) 中文人名的构成特点分析：

中国人名字的构成比较有规律，常见的中国人的名字都是姓加上名字。

根据统计信息显示，父亲母亲的姓氏均包含在姓名内的形式也很常见，在此，我们拓展姓名的另外三种形式：单姓+单姓；单姓+单姓+单名；单姓+单名+单姓。

鉴于上述分析，本文采用 Mike Ushold 和 Micheal Gruninger 的骨架法对本体建模，构建中文人名知识库的层次分类体系的步骤如下[36-42]：

①首先确定中文人名知识库的层次分类。体系框架的第一层显然是，中文人名。

②获取中文人名的特征概念。

③提取中文人名的特征属性，用特征属性表示特征概念。

④将中文人名的特征属性和特征概念有机的联系起来构建中文人名知识库层次分类体系。

⑤不断完善中文人名知识库的层次分类体系。

中文人名知识库的层次分类体系的部分实例如下图所示：

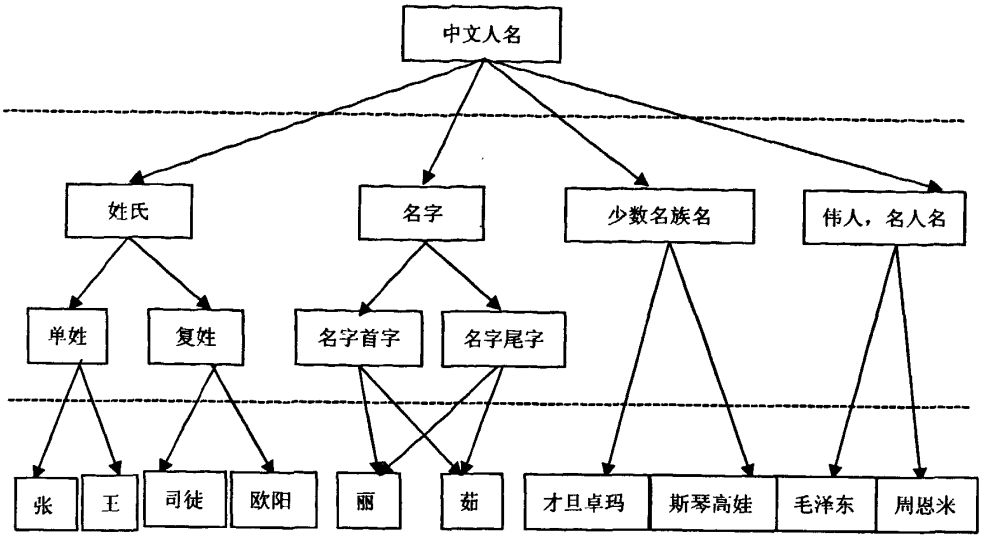

图 4.1 中文人名知识库层次分类体系

本体作为建模工具，在语义的层次上描述知识，可以实现知识的表达、知识的共享和知识的重用。本文采用分类和分层的方法对中文人名知识进行组织，将领域知识分成不同类别不同层次来处理，低层次的领域知识是高层次的基础，高层次的领域知识是低层次的概括和总结，这种方法可以大大提高知识库的可维护性。

4.2 命名实体识别的规则库的构建

4.2.1 用于人名识别的规则

(1) 中文人名的构成特点(Com_Feature)

首先，考虑常规的中文人名字的构成规律，中国传统的起名字的方式都是父亲的姓氏或者是母亲的姓氏，后面跟一个字或者是两个字作为名字。那么可以总结其中的规律为，一般由两部分构成，即姓氏和名字。姓氏和名字又分别由一个或两个字构成，当然我们所说的都是通常的情况。为了方便规则的组织，本文的规则中使用 X 代表姓氏用字，M 代表名字用字。那么，常见的姓名组合可分为四种形式：

规则一：X_0+Y_0，如：周华，李荣；

规则二：$X_0+Y_0Y_1$，如：张福生，刘晓莉；

规则三：$X_0X_1+Y_0$，如：欧阳维，司马光；

规则四：$X_0X_1+Y_0Y_1$，如：皇甫韶华，司徒婉儿。

根据统计信息显示，父亲母亲的姓氏均包含在姓名内的形式也很常见，在此，我们拓展姓名的另外三种形式：

规则五：X_0+X_1，如：王梁，张洪；

规则六：$X_0+X_0+Y_0$，如：李肖晓，朱李丽；

规则七：$X_0+Y_0+X_0$，如：李茹程，方丽朱；

根据识别结果，发现新的中文姓名的构成特征，可做扩展。

(2) 构词能力(Word_Feature)

根据是否具有构词能力，名字用字可以分成三类：不能构成词语，可与上下文构成词语，可单独构成词语。

规则八：不能构成词语(Not)

有些姓氏名字用字，不能和上下文的字符构成词语，也不可以内部单独组成词语，这类名字可以轻松的识别出姓名的边界标志。

如“张筱”的“筱”字，“李逵”的“逵”字。

规则九：可与上下文构成词语(Context)

名字可以和上下文的字符组成词语的情况可以分成一下两种：姓氏与上文的字符或者是字符串组成词；名字和下文的字符或者是字符串组成词语。

如：正确的理解：李白 天天喝酒。

名字与下边界成词：李 白天 天喝酒。

正确的理解：改进 程芳 的方法。

姓氏与上边界成词：改 进程 芳的方法。

规则十：可单独构成词语(Inside)

王明想看日出。

这个例子中的“想”这个字既可以作为姓名的组成部分，也可以以单字词。从而导致上述句子有两种理解方式：

王明 想看日出。

王明想 看日出。

(3) 具有指示意义的语义信息(Denote_ Feature)

在汉语书写的文章中，姓名的出现无法估计，看起来好像就有随意性，但其实它们是有一定的规律可遵循的，在姓名出现的周围，会有具有指示作用的信息，它们能起到暗示的作用。上下文指示信息按其出现的位置可分为 L 指示语义信息和 R 指示语义信息。

规则十一：L 指示语义信息(L_Denote)

包括前称谓词、前指示动词和标点符号。

规则十二：R 指示语义信息(R_Denote)

包括后称谓词、后指示动词和标点符号。

(4) 特殊的姓氏(Especial_ Feature)

有些特殊的姓氏处在特定的上下文环境中不充当姓氏，这种情况我们要单独考虑。

规则十三：有些姓氏前面跟数词的时候，不识别为姓氏；

规则十四：特殊标记字“于”，前面跟特定的词时只作为普通用字。可以把这些词构建成一个知识表。

(5) 是否包含父母双姓(Double_ Feature)

规则十五：姓名当中包含父母双姓。

(6) “和”作为连接词出现在两个姓名的中间(He_ Feature)

语料中常出现这样的句子，XX 和 YYY……，通常情况下，“和”这个字会被误识别作为第一个姓名的尾字，要对此进行修正。

规则十六：“和”字作为连接词出现在两个姓名之间。

(7) 识别结果反馈的信息(Feedback_ Feature)

根据识别结果，一篇文章中的姓名可以分成四类情况：人名被正确识别，人名未被识别，非人名被误识别，非人名没被识别。围绕这四种情况，可以获取相关规则作为每次识别的反馈信息存入规则库。

规则十七：人名且被正确识别。

规则十八：人名未被正确识别。

规则十九：非人名被误识别。

规则二十：非人名没被识别。

4.2.2 用于地名识别的规则

规则一 (Statistical_ Feature)：居民聚落名称；

规则二 (Statistical_ Feature)：全国各级行政区域名称；

规则三 (Statistical_ Feature)：自然地理实体名称；

规则四 (Statistical_ Feature)：名胜古迹，遗址等的名称。

规则五 (Embed_ Feature)：嵌入在其他实体当中的地名。

(值得注意的是：嵌入或者是限定其他实体的地名不能识别为地名)

规则六 (Indicate_ Feature)：能够表明地名特征。

规则七 (Denote_ Feature)：具有地名指示作用。

规则八 (Especial_ Feature)：特定的别名。

(值得注意的是：非特定的地点或者非地点性质的不能作为地名标记)

规则九 (Significant_ Feature)：具有特殊指定意义的建筑物名称。

规则库的命名组织形式见下表：

表 4.1 人名规则库

R01：X_0+Y_0	R11：L 指示语义信息
R02：X_0+Y_0Y_1	R12：R 指示语义信息
R03：X_0X_1+Y_0	R13，R14：特殊姓氏
R04：X_0X_1+Y_0Y_1	R15：姓名当中包含父母双姓
R05：X_0+X_1	R16：“和”字作为连接词出现两个姓名之间
R06：X_0+X_0+Y_0	R17：人名且被正确识别
R07：X_0+Y_0+X_0	R18：人名未被正确识别
R08：不能构成词语	R19：非人名被误识别
R09：可与上下文构成词语	R20：非人名没被识别
R10：可单独构成词语	

表 4.2 规则库的命名组织形式

用于人名识别的规则	用于地名识别的规则
R01-R07 Com_Feature	R1-R4 Statistical_ Feature
R08-R10 Word_Feature	R5 Embed_ Feature
R11,R12 Denote_ Feature	R6 Indicate_ Feature
R13,R14 Especial_ Feature	R7 Denote_ Feature
R15 Double_ Feature	R8 Especial_ Feature
R16 He_ Feature	R9 Significant_ Feature
R17-R20 Feedback_ Feature	

4.3 命名实体的识别

4.3.1 基于规则匹配的命名实体识别

命名实体识别的基本处理过程如下：输入数据为中文的字符串序列，首先对输入的源文本当中的命名实体进行提取，提取出来的命名实体先称作候选命名实体，这一步预处理的依据是已获取的统计信息，其中，考虑到人名和地名特征的差异性，人名的识别是根据已经构建的中文人名层次分类体系；然后根据规则匹配修正候选命名实体，这是一个双向的过程，在利用已有规则对候选命名实体进行修正并产生识别结果的同时,分析识别结果，根据四种情况：是命名实体被正确识别，是命名实体没有被识别出来，不是命名实体被认为是命名实体以及不是命名实体没有被错误认为是，产生新的规则，这些主动生成的规则将被吸收进规则库。

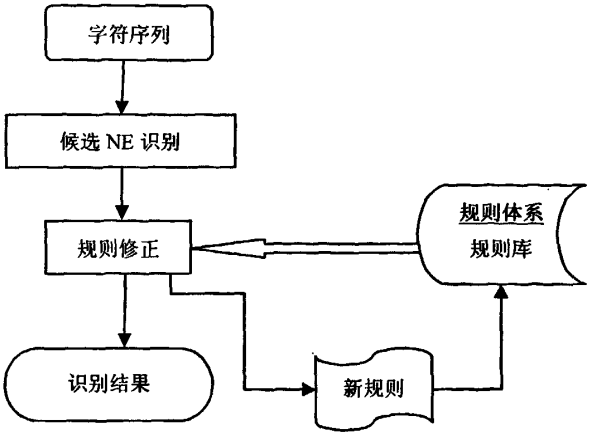

图 4.2 命名实体识别模型

鉴于人名与地名的区别性，在规则匹配识别命名实体之前，先对文章中的候选人名字进行分析处理，具体步骤如下：

(1) 读取输入文本的一个原子序列 P=Si+1Si+2...Si+j。

(2) 匹配人名层次分类体系中的专有名特征属性，若 P 中存在的专有人名，则依次添加专有人名至候选人名链，跳转至第三步；若 P 中没有匹配到专有名，则直接跳转至第三步。

(3) 扫描 P，匹配人名层次分类体系中常用名的姓氏特征属性，判断是否存潜在姓氏，如匹配到潜在姓氏则存进潜在姓氏链（注意：字符串 P 中可能包含不止一个潜在姓氏），同时跳转至第四步；否则读入后续原子序列 P= Si+j+1Si+j+2...Si+j+j，跳转至第二步。

(4) 根据潜在姓氏逐一匹配人名层次分类体系中常用名的名字特征属性，确定候选姓名的右边界，若成功匹配到人名的右边界，则添加该人名至候选人名链，同时，读入后续原子序列 P= Si+j+1Si+j+2...Si+j+j，跳转至第二步；否则读入后续原子序列 P= Si+j+1Si+j+2...Si+j+j，跳转至第二步。

输入：文本 A，单姓库 ***single_list***，复姓库 ***double_list***，名字库 ***name_list***，特殊名字库 ***special_list***；
输出：候选姓名列表 ***Candidate***；

```
while(!A.end)
{   Search A 中所有包含在 special_list 中的字符串 sub_str;
    Candidate.add(sub_str);
}
From A.begin to A.end
{  Get_familyname:
   读取 A 中两个字 A.i A.i+1;
   If (A.i A.i+1∈double_list )
```

```
{   family_name.add(A.i A.i+1);
      i=i+2;
      goto Next_step;
 }
 else if(A.i ∈single _list)
 {  family_name.Add(A.i);
      i=i+1;
      goto Next_step;
 }
 else
 {   i++;
      goto Get_familyname;
 }
Next_step:
  if (singel name)
    扫描姓氏名字用字知识库确定姓名的右边界，并返回 Candidate;
  else if(double name)
    扫描名字用字知识库确定姓名的右边界，并返回 Candidate;
  else goto Get_familyname;
}
```

经过基于统计的初步筛选后，绝大多数人名，地名都能够被识别出来，称之为侯选命名实体，只要知识库构建的充分，就会尽可能多地识别出候选命名实体，但是此时被识别出来的侯选实体中，有很多被错误识别的实体，因此我们需要建立规则来进行进一步的筛选，以提高识别的准确率，这个建立规则进行筛选的过程实际上是一个分类的过程。

决策树模型是机器学习里应用最广泛的分类模型之一，该模型是一个决策机制，算法的核心是贪心算法，它以自顶向下递归的各个击破方式构造决策树，通过在决策树的内部结点进行属性值的比较，并根据不同的属性值判断从该结点向下的分支，在决策树的叶子结点得到结论。所以，从根到叶结点的一条路径就对应着一条合取规则，整棵决策树就对应着一组析取表达式规则[43]。

如何正确构造决策树成为下边需要解决的问题。我们知道，决策树是在训练数据的基础上训练产生的；如何挑选训练对象，如何对训练对象进行描述从而构造训练数据，成为问题解决的关键。

本文采用 ID3 算法[43]来完成分类任务，完成分类首先要明确训练样本以及样本的条件属性和决策属性。本文中训练样本从第一个步骤获得，即是经过统计提取出的候选命名实体，而决策属性也是显而易见的，“是”或者“不是”命名实体，那么问题的关键聚焦在条件属性的选择上。

具体实现采取的方法是：从含有命名实体边界标注的语料库中抽取所有命名实体对应的规则序列，同时还需要有命名实体对应的上下文特征序列，通过训练

利用以上获取的知识，保留正确率高的符合特征规则的模板，而这每一个模板都对应一棵决策树，被保留下来的决策树规则模板便可投入到命名实体识别过程中。

4.3.2 反馈规则的产生

根据识别的正确与否，一篇文章中的命名实体可以分成四种情况：

(1) 字符序列 P 是 NE 且被识别；

(2) 字符序列 P 是 NE 未被识别；

(3) 字符序列 P 不是 NE 被误识别；

(4) 字符序列 P 不是 NE 且没被误识别。

围绕这四种情况，分析字符序列 P 的自身构成信息和上下文信息，作为识别反馈信息吸纳进规则库，使得规则库不断丰富。

以下是分析识别记过，为规则库增加了新的难以发现的规则：

例句 1 政府有何安全保护措施？

这句话中“何安全”被错误识别为人名。需要引入语义关联，通过上下文分析，当“安全”的上下文出现与其有语义关联的词如“措施”、“保护”等，“安全”并非作为人名出现。

例句 2 黎巴嫩游击队今天黎明用轻武器攻击了以军阵地。

这句话中有两个词“黎巴嫩”，“黎明”被错误识别为人名。其中，“黎巴嫩”作为国家的名字，应当收录进专有名词表，而“黎明”的上下文出现与其有语义联系的词“今天”，那么这个词表示时间的概率大于作为一个人名的概率。

例句 3 胡杨是唯一天然成林的树种。

这句话中“胡杨”被错误识别为人名。“胡杨”作为树木的名称，应被收录进专有名次表，并且当上下文出现与树木相关的语义时，其作为树木名称的概率高于作为人名的概率。

例句 4 加之深圳信息有限公司是一家从事系统集成和提供网络技术服务的高科技企业。

这句话中“高科技企业”被错误识别为组织机构名称。“高科技企业”因为带有“企业”特征后缀词而被错误识别，应当向前追溯，看其前端词语的词性，以确定是否是修饰性后缀。

例句 5 高深集团是广州的一家公司。

“一家公司”被错误识别为组织机构名称。“一家公司”因其特征后缀“公司”而被错误识别，应当向前追溯，考虑其前一个词语的词性，以确定是否属于修饰性后缀。

4.4 实验

为了评价实验结果，给出两个评价指标：

召回率 = 系统识别出的正确姓名个数 / 语料中所含的姓名总数；

准确率 = 系统识别出的正确姓名个数 / 系统识别出的姓名个数；

实验采用 ID3 算法训练命名实体识别的模板，我们从 2009 年 1 月的人民日报中随机抽取了 500 篇文章作为训练语料，其中包含 2419 个中文姓名。系统识别出中文姓名 2587 个，其中正确的有 2302 个，中文姓名识别的召回率 95.16%，精确率 89.29%。

图 4.3 规则库简单图示

输入的文章摘自人民日报一则新闻报道：

“记者潘帝都报道。部长下矿贺新年。你们的高水平生物工程成果却在研究室里睡大觉。最佳邮票评选近日于蓉城揭晓。韩国老板使王海清醒了。张兴民，严世军，周振林报道。”

“游人们兴高采烈的游玩了长城，故宫博物馆。下一站天堂杭州，西藏布达拉宫。”

“加之深圳信息有限公司是一家从事系统集成和提供网络技术服务的高新

企业。”

具体实验结果演示见下图：

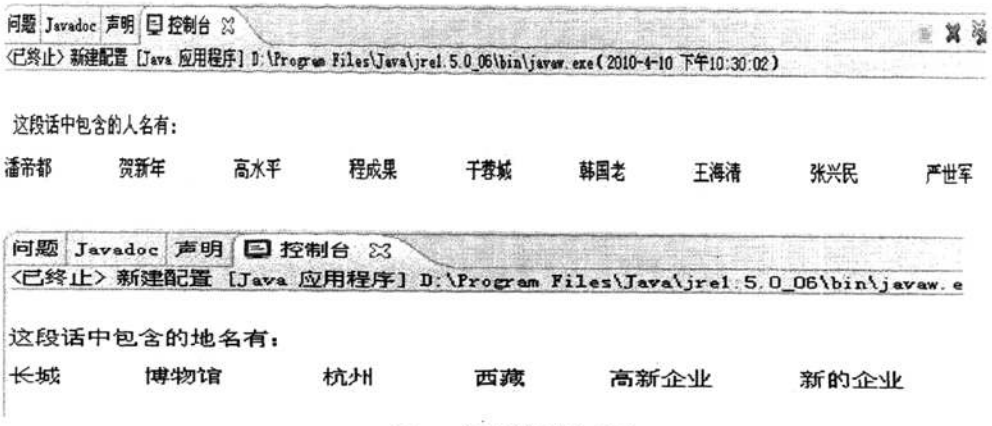

图 4.4 运行结果演示图

在训练分类模板时，采用经典的 10 折分层交叉验证的方法[44]，根据十次测试结果的好坏对每次所得到的规则进行加权合并。

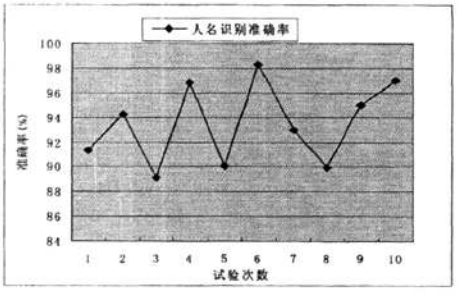

图 4.5 人名识别准确率

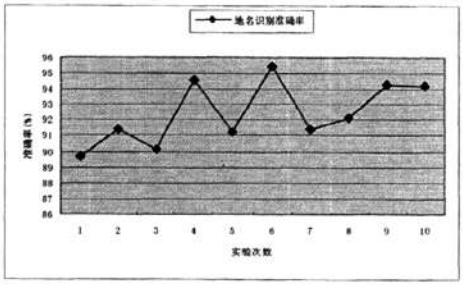

图 4.6 地名识别准确率

表 4.3 人名识别结果

指标 / 序号	包含人名个数	识别人名个数	正确识别个数	准确率(%)	召回率(%)
1	2419	2532	2209	87.21	91.33
2	2419	2582	2279	88.25	94.22
3	2419	2574	2156	83.76	89.13
4	2419	2947	2341	79.45	96.77
5	2419	2489	2179	87.55	90.08
6	2419	2603	2377	91.32	98.26
7	2419	2567	2249	87.61	92.97
8	2419	2688	2175	80.92	89.91
9	2419	2555	2298	89.94	95.00
10	2419	2703	2346	86.79	96.98
加权合并	2419	2587	2302	89.29	95.16

表 4.4 地名识别结果

指标 / 序号	包含地名个数	识别地名个数	正确识别个数	准确率(%)	召回率(%)
1	897	942	804	85.34	89.65
2	897	956	820	85.78	91.42
3	897	910	808	88.76	90.18
4	897	1042	848	81.34	94.56
5	897	905	817	90.3	91.22
6	897	954	856	89.66	95.43
7	897	925	820	88.61	91.44
8	897	1005	826	82.13	92.12
9	897	963	845	87.76	94.3
10	897	969	845	87.2	94.21
加权合并	897	929	839	90.25	93.6

将规则反馈融合到对命名实体模板的训练过程中，从 2008 年 1 月的人民日报中随机抽取了 800 篇文章作为训练和测试语料。把语料分为 10 个互不相交的子集 A1 ，A2 ，...，A10 ，每个子集大小基本相同，每次实验针对其中一个子集进行，实验结束后分析识别结果，结合这四种情况分析：字符序列 P 是 NE 且被识别；字符序列 P 是 NE 未被识别；字符序列 P 不是 NE 被误识别；字符序列 P 不是 NE 且没被误识别。围绕这四种情况，分析字符序列 P 的自身构成信息和上下文信息，作为识别反馈信息吸纳进规则库生成新的规则反馈给规则库，可以发现随着试验次数的增加，正确率有了明显的提高。

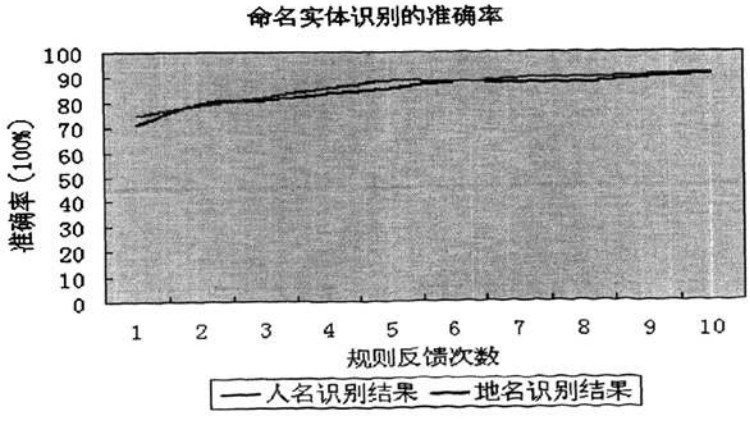

图 4.7 命名实体识别的准确率对比

表 4.5 融合规则反馈的人名识别结果

指标 / 序号	包含人名个数	识别人名个数	正确识别个数	准确率(%)	召回率(%)
1	302	389	276	70.92	91.33
2	241	285	227	79.45	94.22
3	369	406	323	80.92	89.13
4	255	295	247	83.76	96.77
5	343	360	309	85.79	90.08
6	298	332	293	88.25	98.26
7	316	335	294	87.61	92.97
8	288	296	259	87.55	89.91
9	368	389	350	89.94	95.00
10	360	382	349	91.32	96.98

表 4.6 融合规则反馈的地名识别结果

指标 / 序号	包含地名个数	识别地名个数	正确识别个数	准确率(%)	召回率(%)
1	102	122	91	74.66	88.94
2	141	161	127	78.83	90.23
3	169	185	151	81.79	89.55
4	95	105	90	85.7	94.77
5	76	78	69	88.9	91.08
6	99	108	95	88.25	95.76
7	124	127	114	89.67	91.63
8	188	188	169	89.8	90.22
9	132	131	119	90.65	91.22
10	106	103	95	91.67	90.77

4.5 小结

本文首先采用分类和分层的方法对中文人名知识进行组织，将领域知识分成不同类别不同层次来处理，低层次的领域知识是高层次的基础，高层次的领域知识是低层次的概括和总结，这种方法可以大大提高知识库的可维护性。

然后，针对命名实体中出现频率较高的人名地名，构建命名实体识别的规则库，包括用于人名识别的规则库和用于地名识别的规则库。在经过基于统计的初步筛选后，采用基于规则匹配的方法识别命名实体。识别结束后分析识别结果，根据识别的正确与否，一篇文章中的命名实体可以分成四种情况： 字符序列 P 是 NE 且被识别；字符序列 P 是 NE 未被识别；字符序列 P 不是 NE 被误识别；字符序列 P 不是 NE 且没被误识别。分析结果产生新的规则反馈给规则库，以不断完善规则。实验证明该方法可以有效地提高命名实体识别的准确率和召回率。

第五章 总结与展望

在信息化技术迅猛发展的今天，自然语言的处理水平和处理量已经成为衡量一个国家是否步入信息社会的重要标准之一，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，这种技术的发展为人类与计算机之间的信息交流提供了更加便捷、高效的方法。中文命名实体识别是自然语言处理中一项基础性研究，如果计算机能够自动地识别出各种命名实体以及其类型，无疑将使信息的处理和加工更加得准确和高效。

在书写汉语的时候，汉字和汉字，词语和词语，它们一个接着一个，换句话说，一个词语在一句话中没有被明显的标记出来，在这个方面，汉语和英语完全不同，英语的句子中有空格可以把词语分开。因此，为了能理解汉语，一个最重要的任务就是必须把这些连续的中文汉字给分隔开来。

大多数的汉语文章处理系统是依赖于词典来识别出文章中的每个词语，但是值得注意的是，这种方法是不可能完全识别文章当中的所有词语的，因为所使用的词典并不能包含所有的词语，因而词典中没有的词语变成为分词中值得特别注意的问题。

这些词典中不包括的词，形象的称之为未登录词。在所有未登录词当中，命名实体是最为重要的一类。命名实体组成的一片文章的主要内容。因此，命名实体的识别成为许多自然语言处理应用的主要任务。

5.1 总结

命名实体的识别在中文文本处理工作中是一项很有挑战性的问题，这个挑战性远远高于对英文文本的处理。原因有两个方面，第一个原因是中文的分词。在英文文本中，句子是由一序列的词语组成的，但是这些词之间都有空格作为分隔，然后，中文就不是如此，它没有分隔标志。在中文文本中，句子是由一连串的汉字组成的，没有简单明了的分隔标志。因此，识别出中文文本中每一个词语的边界就要比英文文本难得多。另外一个原因是词法。在英文文章中，每一个名字往往都是一个大写开头的词或者是一串大写开头的词以及一些小写的标志词，如连词、介词等。但是汉语却没有大写的模式。

本文主要实现了基于规则匹配的命名实体识别研究，具体工作如下：

(1) 一篇文章当中，实体是最基本的信息元素，它们往往指示了文章的主要内容，因此，识别出文章当中的命名实体是对文章进行理解的重要前提，同时命名实体识别的质量也会直接影响到一系列的后续工作，因此，命名实体识别已经越来越成为自然语言处理中的关键技术。本文首先从中文信息处理的的研究背景

和意义入手，引出了对命名实体识别和中文分词相关知识的介绍。主要讨论了命名实体识别的一些方法和应用，以及中文分词的相关方法。

(2) 大多数的汉语文章处理系统是依赖于词典来识别出文章中的每个词语，但是值得注意的是，这种方法是不可能完全识别文章当中的所有词语的，因为所使用的词典并不能包含所有的词语，因而词典中没有的词语变成为分词中值得特别注意的问题。而命名实体又是这些未知词语当中最为重要的一类，同时，命名实体识别的过程中，如果输入的是已经被分割的词语或者是字，那么一些被错误分隔开的命名实体就不能被识别出来，因此，在分词之前就对命名实体进别，然后将命名实体识别的结果和分词的结果一起输出无疑会提高分词的准确率。中文分词模型首先对输入的汉语书写文章进行预处理，对文章进行适当的切分，得到下一步的出入文档，这是一系列的中文短句的集合；接着再对这些中文短句进一步的切分，这一步主要是将中文短句变成不可再分的字符串序列，这些字符串序列就可以作为中文分词和命名实体识别的基本单元；下一步，基于这些字符序列识别命名实体，与此也是展开分词；最后输出分词结果。

(3) 由于中文人名用字规律蕴含丰富的信息，为了提高中文人名知识库的可维护性，本文是基于本体构建中文人名知识库的组织模型，把人名的构成分成若干了类别，每一个类别分成若干个层次，这种把中文人名知识分类管理，分层组织的模式，使得低层次的知识为高层次的知识提供依据，高层次的知识概化低层次的知识，从而大大提高知识库的可维护性。

(4) 在上述工作的基础上，实现了基于规则匹配的命名实体识别。首先构建人名地名规则库，然后使用规则匹配的方法识别命名实体，同时分析识别结果根据识别的正确与否，一篇文章中的命名实体可以分成四种情况：字符序列 P 是 NE 且被识别；字符序列 P 是 NE 未被识别；字符序列 P 不是 NE 被误识别；字符序列 P 不是 NE 且没被误识别，分析结果可以形成反馈规则完善规则库。实验结果证明，随着命名实体识别模板的不断完善，识别的准确率也在不断提高。

5.2 展望

对于命名实体识别的研究工作，本文主要采用了基于规则匹配的识别方法，还存在以下一些需要进一步研究的任务：

(1) 本文命名实体识别的工作主要针对出现频率较高人名和地名进行，如何将识别工作继续扩展到其他实体的识别是下一步的工作。

(2) 本文采用的机器学习的模型是决策树模型，那么下一步可以考虑把其他统计模型应用到命名实体识别的过程当中，如 AdaBoost、支持向量机模型等等。

(3) 由于书写汉语的时候，汉字和汉字，词语和词语，它们一个接着一个，一个词语在一句话中没有被明显的标记出来，在这个方面，汉语和英语完全不同，汉语的语言的特性更为复杂，因此基于机器学习的中文命名实体的识别可能效果

不那么理想。如何更好的利用汉语的语言特性，从而更好的发挥机器学习方法的特点，使其在中文命名实体识别系统中的性能也能有所提高是一直在考虑的问题。

参考文献

[1] 黄昌宁, 夏莹. 语言信息处理专论. 北京: 清华大学出版社, 1996.

[2] 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.

[3] Sundheim BM. Namd entity task definition, version2.1. In: Proc. of the 6th Message Understanding Conf. 1995.319-332.

[4] Liu Qun, Zhang Huaping, Yu Hongkui, et al. Chinese lexical analysis using cascaded hidden Markov model. Journal of Computer Research and Development, 2004, 41(8): 1421-1429.

[5] 姜维, 王晓龙, 关毅, 赵健. 基于多知识源的中文词法分析系统[J]. 计算机学报, 2007,30(1).

[6] 罗智勇, 宋柔. 现代汉语通用分词系统中歧义切分的实用技术[J]. 计算机研究与发展, 2006,43(06):1122-1128.

[7] K.J.Chen, Ming-Hong Bai.Unknown word detection for Chinese by a corpus-based learning method. International Journal of Computational Linguistics and Chinese Language Processing,1998,3 (1):27～44.

[8] 崔世起, 刘群, 孟遥, 于浩, 西野文人. 基于大规模语料库的新词检测[J]. 计算机研究与发展, 2006年05期.

[9] H Luo, Z Ji. Inverse Name Frequency Model and Rules Based on Chinese Name Identifying. Natural Language Understanding and Machine Translation. Beijing: Tsinghua University Press, 2001. 123-128.

[10] Danil M.Bikel, Schwarta R and Weischedel R. An Algorithm that Leams what's in A Name[J]. Machine Leanring Jounral Special Issue on Natural Language Leanring, 1999,34:211-231.

[11] Guohong Fu, Kang-Kwong Luke. Chinese Named Entity Recognition using Lexicalized HMMs[J].ACM SIGKDD Explorations,2005,7(1):19-25.

[12] Andrew Borthwick. A Maximum Entropy Approach to Named Entity Recognition[D]. PhD thesis, New York University,1999.

[13] Uchimoto K, Murada M, Ma Q, Ozaku H and Isahara H. 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, 2000: 326-335.

[14] Hideki Isozaki, Hideto Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition[C]. Proceedings of the 17th Intenrational Conference on Computational Linguistics,Taipei,Taiwan,20 02:390-396.

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

[16] G Eibl, K.P Pfeiffer. How to Make AdaBoost.M1 Work for Weak Base Classifiers by Changing Only One Line of the Code. In Proceedings of the 13th European Conference on Machine Learning, 2002, 72-88.

[17] 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.

[18] 文庭孝, 邱均平, 侯经川. 汉语自动分词研究展望[J]. 现代图书情报技术, 2004, 112(7): 6-10.

[19] 邱均平, 文庭孝, 周黎明. 汉语自动分词与内容分析法研究[J]. 情报学报, 2005, 24(3): 309-317.

[20] 左军. 基于概念的中文分词模型研究[D]. 硕士论文, 南京邮电大学, 2007.

[21] 张金柱. 基于字位的中文分词方法研究与实现[D]. 硕士论文, 中国科学技术信息研究所, 2008.

[22] 张彬. 面向中文网络信息检索的自动分词系统设计与算法实现[D]. 硕士论文, 华东师范大学, 2007.

[23] 朱小娟, 陈特放. 基于SVM的词频统计中文分词研究. 微计算机信息, 2007, (30).

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

[25] 刘涌泉. 语言学现代化和计算机. 武汉: 武汉大学出版社, 1988.

[26] 梁南元. 书面汉语自动分词系统-CDWS. 北京航空航天大学, 1987.

[27] xiongyou LIANG, Youngsheng XUE. An Algorithm of Solving Interlink Overlapping Ambiguity and Combinatorial Ambiguity and Compound Ambiguity in Chinese Word Segmentation[J]. Journal of Computational Information Systems3, 2007(3), 1189-1200.

[28] Xue N. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processmg, 2003.

[29] 罗智勇, 宋柔. 现代汉语通用分词系统中歧义切分的实用技术[J]. 计

算机研究与发展, 2006,43(06):1122-1128.

[30] Wu A , Jiang Z. Statistically-enhanced new word identification in a rule-based Chinese system. Proceedings of the 2nd Chinese Language Processing Workshop, 46-51, 2000.

[31] Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. Subwmd-based tagging by Conditional Random Fields for Chinese Word Segmentation. P orceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics anmual meeting, 193-196, 2006.

[32] Pak-Kwong Wong, Chorkin Chan. Chinese Word Segmentation based on Maximum Matching and Word Binding Force. COLING96, 1996, 200-203.

[33] 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.

[34] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[35] Studer, Benjamins V R. FenselD Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering, 1998, 25(12): 161-190.

[36] 邓志鸿, 唐世渭, 张铭等. Ontology研究综述[J]. 北京大学学报, 2002, 38(5): 730-738.

[37] OntoKnowledge. http://www.ontoknowledge.org

[38] XOL. http://www.ai.sri.com/pkarp/xol.

[39] RDF. http://www.w3.org/rdf.

[40] Uschold M. Ontologies principles, methods and applications[J]. Knowledge Engineering Review, 1996, ll(2): 93-155.

[41] Gruninger M, Fox M. Methodology for the design and evaluation of ontologies[J]. Workshop on Basic Ontological Issues in Knowledge Sharing, 1995, 4: 253-280.

[42] Maedche A. Ontology learning for the Sematic Web[M]. Boston: Kluwer Academic, 2002.

[43] Tom M. Mitchell, Machine Learning[M]. The McGraw -Hill Companies, Inc., 1997.

[44] Quinlan J R Induction to decision trees[J]. Machine Learning, 1986,1:81-106.

附录一 攻读硕士学位期间主要科研工作及成果

参与科研项目

(1) 参加了中国科学院自动化研究所“情报与安全信息学（ISI）创新团队国际合作伙伴计划”子课题“HTML 新闻网页过滤与总结系统”。

(2) 参加了国家“九七三”重点基础研究发展计划项目（No. 2009CB326203），普适个性化信息处理基础理论与方法研究。

攻读硕士学位期间发表的学术论文

(1) 周昆，胡学钢，一种基于本体论和规则匹配的中文人名识别方法。微计算机信息(已录用)。

(2) 周昆，胡学钢，本体论和机器学习相结合的中文姓名识别研究。The 12th China Conference on Machine Learning(已录用)。