工 学 博 士 学 位 论 文

基 于 统 计 语 言 模 型 的 汉 语 词 法 分 析 研 究

赵 岩

哈尔滨工业大学

2005 年 6 月

图书分类号：TP391.2
U.D.C.:681.37

工学博士学位论文

基 于 统 计 语 言 模 型 的 汉 语 词 法 分 析 研 究

博士研究生： 赵 岩
导　　师： 王晓龙　教 授
申 请 学 位： 工学博士
学 科、专 业： 计算机应用技术
答 辩 日 期： 2005 年 6 月
授予学位单位： 哈尔滨工业大学

Classified Index : TP 391.2
U.D.C : 681.37

A Dissertation for the Doctoral Degree in Engineering

Research on Chinese Morphological Analysis Based on Statistical Language Model

Candidate : Zhao Yan
Supervisor : Prof. Wang Xiaolong
Academic Degree Applied for : Doctor of Engineering
Speciality : Computer Application
Date of Defence : June, 2005
Degree-Conferring-Institution: Harbin institute of Technology

摘 要

词法分析是自然语言处理领域中最基础的处理步骤，尤其对汉语这种没有分割符的语言来说更是如此。本文研究的汉语词法分析主要包括自动分词、词性标注和词义相似度计算三个方面。词法分析是句法分析的先期处理步骤，其错误会沿处理链条扩散，并最终影响信息检索、机器翻译等面向最终用户的应用系统的质量；同时，词法分析所用的技术也可以直接应用到音字转换和语音识别等应用系统中，所以对它的研究具有极其重要的意义。

本文在统计语言模型方面主要探讨了N-gram模型、最大熵模型、支持向量机模型和矢量空间模型。重点研究了三个方面的内容：传统N-gram模型的改进方法；利用触发对提高矢量空间模型的质量；在最大熵模型中加入转换触发对特征。最后利用以上统计语言模型的研究成果对汉语词法分析进行了深入研究。主要内容包括四个方面：

第一、从两个方面改进了传统N-gram模型：首先，提出了改进的N-gram模型平滑算法；其次，抽取了能够承载长距离约束信息的触发对。数据稀疏会给N-gram模型带来零概率的问题，影响Vertibi解码算法的正常使用。本文利用词性信息解决了Katz平滑算法处理固定搭配时存在的问题；同时，利用词义词典进行了词语聚类的研究，并将聚类的结果应用到Uni-gram模型的平滑算法中。试验结果证明：以语言模型的交叉熵为量度，利用语言学知识可以提高现有平滑算法的性能。在触发对抽取方面，从理论上比较了互信息(Mutual Information, MI)和平均互信息(Average Mutual Informaiton, AMI)两种量度的不同，利用AMI抽取了200万词触发对。同时，在词触发对的基础上，提出了转换触发对的概念，应用在本文后续的分词、词性标注、音字转换研究中，提高了最终结果的准确度。

第二、分词是汉语词法分析中最基本的步骤，所有的汉语自然语言处理都要基于分词的结果。困扰汉语分词的主要问题是歧义消解和名实体识别。针对歧义问题，首先提出了基于递归枚举算法(Recursive Enum Algorithme, REA)的K-best分词模型以便有效地进行歧义词的识别，同时利用最大熵模型构造一个二值分类器来消解分词歧义。试验结果证明：K-best分词模型可以有效识别分词的歧义字段。通过局部特征配合转换触发对信息，分词歧义消解的正确率达到了92%。针对名实体中的人名识别，对中文姓(名)用字进行了必要的分类并以此为基础建立了人名识别多源知识表。与统计方法配合使用后，在提高识别准确率

的同时，可以保证识别召回率不受影响。最后利用有限自动机理论对时间、数字等因子词进行了识别。

第三、词性标注可以看成是噪声信道的解码问题。传统的HMM模型有两个缺点：首先它用联合概率解决一个条件概率问题，而且它不能包含长距离词法特征。针对以上问题，本文分别利用支持向量机模型和最大熵模型对复杂兼类词标注进行了研究，试验结果证明两种模型都可以有效降低兼类词标注的错误。在此基础上，利用最大熵模型对基于句子的词性标注进行了研究，重点研究了长距离聚类转换触发对“$w_A \rightarrow w_B / t_B$”特征的加入以及用于系列分类的Beam Search搜索算法。最后，利用与词性标注相同的技术对音字转换做了初步的探讨，主要试验了简单和复杂两种特征模板。试验结果证明：与HMM模型相比，融合了聚类转换触发对的最大熵语言模型使词性标注的错误减少了34%，音字转换的错误率在包含10万句的训练语料上减少了4%。

第四、词义是词法分析中的核心问题，本文重点利用矢量空间模型对词义相似度计算进行了研究。首先，提出了词分辨力量度，并优先选择分辨力大的词作为矢量空间的坐标轴词；其次，为克服传统矢量空间模型中“词袋”效应带来的噪声问题，建立了基于触发对的包含语言结构信息的矢量空间模型，并利用这个模型进行了词语聚类的研究。试验结果证明：以7组同义词词义相似度分布方差为评价量度，新的模型质量提高了32%，词语聚类的结果满足实际应用的需要。

关键词：统计语言模型；汉语词法分析；词义相似度计算；音字转换；触发对

Abstract

Morpheme analysis is the basic step in nature language processing, especially for Chinese that doesn't have separator. In this dissertation, Chinese morpheme chiefly analysis consists of word segmentation, Part-of-Speech (POS) tagging and word similarity calculation. As preceding processing of syntax analysis, error of morpheme analysis will cascade through the chain and impact the application such as information retrieve and machine translation. Simultaneously, the technology of morpheme analysis can be put to use of Pinyin to character (PTC) conversion and phonetic recognition, so the work is of great significance.

N-gram model, Maximum Entropy Model (MEM), Support Vector Machine (SVM) and Vector Space Model (VSM) are studied mainly. The research focus on smoothing algorithm for N-gram, improving the VSM with trigger pairs and adding conversion trigger pairs to MEM. Lastly, the above research achievements are utilized in Chinese morpheme analysis. The dissertation concerns the following aspects.

1) Betterment of traditional N-gram Model includes two themes. Smoothing algorithm is improved, and trigger pairs, which contain long-distance constrain information, are extracted. Data sparsness will cause the problem of zero probability, and influence the decoding algorithm Vertibi. Firstly, the problem, which Katz smoothing faced when deal with collocation, is resolved with the help of POS information. Secondly, the word clustering is made based on semantic dictionary, and the result of clustering is used in Uni-gram smoothing. Evaluated by perplexity of language model, the experiment shows that knowledge of language can improve the smoothing. In the research of trigger pairs, Average Mutual Information (AMI) is compared with Mutual Information (MI) theoretically, and 200 millions trigger pairs are extracted by AMI. On the basis of word trigger pair, the conversion trigger pairs, which are beneficial to increase the accuracy of POS tagging and word segmentation, are introduced.

2) Chinese word segmentation is the basic step in Chinese morpheme analysis. All Chinese nature language processing must come from result of it. Problems of Chinese segmentation are overlapping and cover disambiguity and name entity

recognition. For ambiguity problem, in order to recognize ambiguous phrase effectively, K-best word segmentation model based on REA is put forward. Then, a Boolean classifier is built based on MEM. Experiment shows that system can identify the most ambiguous phrases and local and trigger pairs feature can deal with 92% ambiguities. For name entity task, this paper classifies the surname and last name characters and then builds the multi-source table. With the help of it, the statistical system can increase the accuracy and keep the recall. In the end, the finite automation is applied to recognize the time and number word.

3) POS tagging can be regarded as the decoding problem of noisy source channel. There are two shortcomings in HMM. Firstly, it presents the joint probability to deal with the condition problem. Secondly, it cann't involve long-distance lexical feature. To address these issues, this dissertation applies MEM and SVM to complex POS tagging respectively. The result shows that the two models can decrease the error of complex POS tagging. Based on MEM, the research centres on fusion of the clustering conversion trigger pair "$w_A \rightarrow w_B / t_B$" and Beam-Search algortihem in sentence POS tagging. In the end, the same tecnnology of POS tagging is tested in the CTP conversion, the simple and complex feature template being tried respectively. The result shows that error of POS reduces by 34%, and accuracy of PTC conversion increases 4% on a small size training corpus which contains 100 thousand sentences.

4) Semantic is the key problem of morpheme analysis. This paper calculates the word similarity based on VSM. Firstly, discrimination measure of word is proposed and words of high discrimination ability are selected as axis of vector space, Secondly, in order to resolve the noise problem which is caused by "word bag" in traditional VSM, a new VSM, which includes struct information based on trigger pairs, is built and applied to word clustering. The experiment shows that quality of new VSM inceases by 32%, evaluated on variance of seven pairs of synonyms similarity.

Key Word: Statistical Language Model, Chinese Morpheme Analysis, Word Similarity Calculation, Pinyin-to-Character Conversion, Trigger Pairs

目 录

Contents

第1章 绪论

1.1 研究的目的和意义

自然语言是音、义结合，包含词汇信息和语法规则的复杂体系，不同于计算机所用的任何程序语言，自然语言是人类特有的交流、思想的工具，本身充满着众多的歧义现象。随着人类社会科技文明日新月异地发展，自然语言所承载的信息越来越大，人们也越来越希望计算机能够对自然语言承载的信息进行有效地转换、传输、存贮、分析和应用。计算语言学(Computational Linguistics)就是在这一背景下产生的一门交叉学科。作为计算语言学的一个核心研究部分，语言模型的主要研究内容是对自然语言内部规律的挖掘、描述以及使用。

语言建模的方法主要可以分成两个方向：第一个是用语言学家掌握的语言学知识和领域知识来建立模型的理性主义(Rationalism)方向；第二个是通过自动的方法从大规模真实语料库中获得实证知识和统计规律来建立模型的经验主义(Empiricism)方向。这两个方向的主要区别如表1-1所示：

表1-1 两种语言建模方法的对比[1]

Table 1-1 Comparison between two language modelings

经验主义方向	理性主义方向
非受限真实文本	受限的规范文本
概率化参数方法	基于规则的句法-语义分析
对语言现象的发现和归纳	对语言现象的直觉和反思
颗粒度细	颗粒度粗
上下文相关	多半上下文无关
注重形式	注重意义
以定量化评测驱动研究	例不十，法不立；例外不十，法不破
从文本整体把握	注重文本局部，放大并细化
覆盖面广	覆盖面窄

理性主义的建模方法由于主要采用语言学家的语感和直觉，知识的颗粒度粗且覆盖面窄。由于受到规则完备性(Completeness)和一致性(Consistency)的限制，这种方法只能面向受限的良构语言。随着计算机技术的飞速发展和网络文本的海量增长，人们对大规模、病构、真实文本进行实时处理的要求越来越迫切。这就要求采取以经验主义为主，以理性主义为辅的研究路线。本文的研究在整体上也遵循这一路线。

基于经验主义的统计语言模型最先在语音识别和音字转换领域获得较大的成功，这也给自然语言处理的其它应用带来了巨大的希望。但是经过近20年的发展，一些涉及到语言理解的应用如机器翻译、文摘、问答等系统依然不能完全满足用户的需要。这就使得研究人员开始重新审视统计语言模型本身的问题：首先是数据稀疏的问题；同时不能很好地融合词义信息；如何从大量的、灵活的、不确定的、包含冗余信息的语言现象中获得直观的语言学知识也没有获得根本地解决。这说明统计语言模型的研究在深度和广度上都有较大的空间。在汉语词法分析领域，目前的研究主要体现三个特点：1）更细，对某一特定子问题研究较深入，如分词中的人名识别和词性标注中的未登录词标注等；2）更新，一些新的机器学习理论被引入到汉语词法分析中，如统计学习理论等；3）更全，试图通过构建统一的模型来避免分步的处理。以上研究虽然取得了一定的成果，但从实用化的角度上考察依不尽人意。基于以上背景，本文利用统计语言模型对汉语词法分析进行了相关的研究。

研究的主要目的是：在理论上探索面向汉语词法分析的统计语言模型的建模方法；在实践上构造一个能够完成分词、词性标注和处理部分词义问题的词法分析软件，从而为其它自然语言处理研究提供底层的功能支持和充足的语料库数据支持。研究的主要意义在于：通过对汉语词法分析的研究可以有效地解决词法分析中的各种歧义问题，这样才可以进一步解决词法分析后续的句法分析中的结构歧义问题。否则，将无法从自然语言处理(Nature Language Processing)上升到自然语言理解(Nature Language Understanding)。同时，由于生物信息和自然语言之间存在如表1-2所示的类比关系，统计语言模型的研究成果也可以应用到计算生物信息学领域[2, 3]。其基本的思想是：基因可以当成一种特殊的语言。在语言学中，一些词可以排列成有意义的句子；而在生物学中，氨基酸是有意义的词，蛋白质是由一定的氨基酸排列而成，从而分析蛋白质的结构和功能。

表1-2 自然语言与生物信息的类比关系

Table 1-2 Analogy between nature language and biological information

生物信息	语言信息
核苷酸	字母
氨基酸	词
外显子	短语
蛋白质	词意
蛋白质电路	句子
生物功能	语义学
基因表达	语言的产生

1.2 主要统计语言模型

统计语言模型的研究促进了自然语言处理技术的普及，并在语音识别[4]、机器翻译[5]、手写识别[6]、音字转换[7, 8]和信息检索[9]中得到广泛地应用。值得一提的是，虽然基于转换学习(Transformation-Based Learning)和基于粗糙集(Rough Set)理论的机器学习算法在词法分析领域中的词性标注[10, 11]、词义消岐[12]等领域应用较广，但本文并没有对其进行研究。原因在于它们对语言现象的描述基本属于 *If …Then…* 的形式，而不是以概率值的方式给出。这就带来了一个主要的问题，即它不能被当成一个组件用在一个更大的基于概率的整体框架中。当面临系列分类问题的时候，给出每一步不同分类的概率值然后进行总体寻优要比给出单个的分类结果要更可靠一些。

在统计语言模型中，自然语言被看作是一个随机过程，其中每一个语言单位：包括字、词、句子、段落和篇章等都被看作是有一定概率分布的随机变量。为计算一个自然语言句子 S 的概率值 $p(S)$，假定 S 由最小的结构单位词 $w_1, w_2, ..., w_n$ 组成，直接计算 $p(S)$ 将很困难，通常利用离散概率的乘法定律将 $p(S)$ 分解为条件概率的乘积，见式(1-1)：

$$p(S) = p(w_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} p(w_i \mid h_i) \tag{1-1}$$

其中，$h_i \overset{def}{=} \{w_1, w_2, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{n-1}, w_n\}$ 称为 w_i 的上下文。实际应用中，由于当前词 w_i 只和前面若干个词相关，同时由于统计语言模型特有的数据稀疏问题，所以通常只考虑一定范围内的上下文。

1.2.1 N-gram模型

N-gram模型其实质是N-1阶马尔可夫模型，N-gram模型利用马尔可夫过程中时间不变特性(Time Invariant)和水平有限特性(Limited Horizon)减少参数估计的维数，见式(1-2)：

$$p(w_i \mid h_i) = p(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) \tag{1-2}$$

模型中 n 的大小要考虑估计中有效性和描述能力的折衷。n 值越大，描述能力越强，但是估计的有效性越差。N-gram模型有两个主要的问题：首先，模型的参数空间随着 n 值呈指数性增长，由于存储空间有限，从而极大限制了 n 值的

大小，所以目前常用的N-gram模型是Bi-gram和Tri-gram模型。过小的n值使得模型不能包含长距离的词法信息，而这种信息在语言现象中是广泛存在的；其次，即使Tri-gram模型可以解决存储空间的问题，但由于自然语言遵循最小力气定律(Zip'f Law)[13]，使得大量的语言现象不能出现在训练语料中，从而带来数据稀疏的问题。数据稀疏问题不仅使我们不能准确地预测某些小概率语言现象，更为严重的是，训练语料中未见的事件所带来的零概率问题会使整个N-gram模型不能通过动态规划算法来进行全局路径寻优。

众多的统计语言模型研究都致力于解决传统N-gram模型上的这两个问题。为了能包含长距离的约束信息，Matrin提出了一种基于skip的变长度N-gram模型[14]，其主要的思想就是“精确的上下文很少有机会重现，但是近似的上下文更容易重现”。例如一个Five-gram模型可以分成两个子集，分别为$p(w_i \mid w_{i-4}, w_{i-3}, w_{i-1})$和$p(w_i \mid w_{i-4}, w_{i-2}, w_{i-1})$。这种方法的主要缺点是分离的模型分割了训练数据，Goodman通过试验证明其效果并不十分理想[15]。与此类似，Ron和Niesler也分别提出了各自的变长度的N-gram模型[16, 17]。Siu利用一种特殊的存储树结构来减少存储空间以便增加n的值[18]，树的根节点对应于Uni-gram模型，第一层节点对应于Bi-gram模型，第二层节点对应于Tri-gram模型，依此类推，其父节点为在估计该词语的条件概率时的历史节点。同时Siu引入了树节点的剪切与合并算法，目的是将模型因为剪切而导致的性能损失限定在一个合理的范围内，这样在给定的存储空间中就可以尽可能地包含长距离的信息。但是即使引入了以上的算法，在面对大规模语料与实际应用时，n的值仍然只能限定在较短的上下文范围内。Zhou提出了一种基于触发对(Trigger-Pair)的统计语言模型[19]。这一模型将相互之间具有长距离约束关系的一对词构成一个触发对，然后将其与N-gram模型通过插值方式进行集成，在一定程度上解决了传统N-gram模型的长距离约束问题。但基于插值的方法对两种信息进行融合并不符合自然语言本身的规律。与此类似，Lau在最大熵框架下通过融合触发对信息解决长距离约束问题[20]。Ney等人提出了一种基于词语关联(Word Association)的语言模型[21]，该模型通过将一个词语与其上下文中的词语之间的长距离约束用词语关联对来进行表示，并将与同一个词语相关的多个关联词对的条件概率分布进行融合来确定一个词语的条件概率。

为解决N-gram模型中的零概率问题，研究人员提出了各种不同的针对N-gram模型的平滑算法。其主要的思想是“削富济贫”，即从已见的事件中折扣出一小部分概率，然后将折扣出的概率采用回退或插值的方式分给没有见过的

事件。平滑算法主要有：Good提出G-T平滑算法[22]，由于缺乏利用低阶模型对高阶模型进行回退或插值的思想，所以它一般不单独使用，多用于其它平滑算法中进行概率的折扣。Jelinek和Mercer提出了一种基于线性折扣的J-M平滑算法[23]，采用Baum-Welch算法计算相关的折扣系数，然后采用插值的方法与低阶的分布模型进行融合。Katz提出的平滑算法首先采用Good-Turing方法进行概率的折扣，然后采用回退的方法与低阶的分布模型进行融合[24]。W-B平滑算法由Bell和Witten提出，可以看成是J-M平滑算法的一个特例，所不同的是插值系数的计算根据高阶参数后面跟随的不同词的个数而计算[25]。绝对(Absolute)平滑算法由Ney和Kneser在1994年提出，通过减去固定的值进行概率的折扣，无论是大概率还是小概率事件都折扣出相同的值，所以称其为绝对折扣平滑算法[21]。在1995年Kneser和Ney对绝对平滑算法做了相应的改进，提出了K-N平滑算法[26]。

本文将在第2章对以上两个问题作更深入地探讨。为解决汉语词法分析中的长距离约束问题，在触发对的基础上，提出了基于二元触发思想的转换触发对的概念。针对零概率问题，提出了融合词义的Uni-gram平滑算法，并解决了传统Katz平滑算法在处理某些固定搭配时无法折扣概率的问题。

作为一种较成熟的统计语言模型，N-gram模型已经成功地应用在自然语言处理的各个领域中，几乎所有的商业化自然语言处理应用产品都使用了某种形式的N-gram模型。N-gram模型可以认为是状态序列可见的马尔可夫模型，而隐马尔可夫模型(HMM)是一种特殊的马尔可夫模型[27, 28]。它与一般的马尔可夫模型的最大区别在于它的状态序列是不可见的，但是它可以输出一个可见的观察序列。在词性标注工作中，经常使用HMM为词性序列建立语言模型，在语音识别中，利用HMM为声学信号建立声学模型。

1.2.2 最大熵模型

最大熵理论最先由Jaynes于上世纪50年代提出[29]。它是一种具有朴素哲学思想的统计学习方法，其基本思想是将每个信息源视为一组约束条件。最大熵建模的目的在于：在满足所有约束条件的一组概率分布中，找寻其中熵最大的分布，即最均匀的分布。用于找寻这个分布的迭代算法是一个逐步适应的过程，新的约束条件可以被随时地添加到模型当中。在现存的约束条件下，最大熵模型肯定存在唯一的解。DellaPietra将最大熵理论第一次用于统计语言建模[30]。与N-gram模型的计算方式不同，最大熵模型用指数的形式来计算条件概率 $p(w_i \mid h_i)$，见式(1-3)：

$$p(w_i \mid h_i) = \frac{1}{Z(h_i)} \cdot \exp[\sum_j \lambda_j \cdot f_j(h_i, w_i)] \tag{1-3}$$

其中，$Z(h_i)$为归一化常量。f_j代表上下文中的特征，通常为二值布尔函数。每一个特征f_j对应一个参数λ_j，代表这个特征的权重，λ_j可通过迭代缩放(Generalized Iterative Scaling, GIS)算法自动计算[31]，并保证唯一收敛。

最大熵模型在语言建模领域有较突出的优点：首先，它是一种统计语言模型，可以在模型中加入灵活的、彼此不需要独立性假设的、颗粒度很细的、基于词的各种上下文特征信息。它对上下文环境中的信息有非常强的融合能力，同时对这些信息源没有任何约束和前提假设条件；其次，最大熵模型具有通用性。对任何事件空间的任何子集的概率估计都可以使用最大熵模型。任何现存的语言模型的知识都可以被添加到最大熵模型中，传统的N-gram模型、2间距的N-gram模型和长距离触发对模型都可以被编码成特征函数[32]。由于具有以上的优点，随着计算机运算能力的发展，经过Berger，Rosenfeld等人在自然语言处理中对其进一步的研究[33]，最大熵模型在自然语言处理的歧义消解[34]、文本分类[35]、名实体识别[36]和词性标注[37]等领域中获得了广泛的应用，并取得了比较理想的结果。

最大熵模型也有明显的缺点。虽然GIS算法保证收敛，但却无法确定算法迭代次数的理论上限，而且算法的训练时间复杂度非常高；同时，最大熵模型对特征的质量要求很高。Pietra利用KL距离描述了一个从给定候选集中选择特征的自动迭代方法[38]。Rosenfeld描述了一个特征抽取人机交互过程[39]。

1.2.3 支持向量机模型

支持向量机(Support Vector Machine，SVM)是一种用于分类问题的有指导机器学习算法[40, 41]。给定一个L维的矢量空间，存在一个如公式(1-4)定义的超平面，它能够把训练数据$\{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathrm{R}^L, y_i \in \{\pm 1\}\}$分成两个类，从图1-1的左半部分可以看出，存在有很多这样的超平面。SVM的任务在于发现一个最优的超平面，使得这个超平面和最近分类点的距离最大，如图1-1右半部分所示：图中灰色点代表的数据即为支持向量。

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0, \qquad \boldsymbol{w} \in \mathrm{R}^L, b \in \mathrm{R} \tag{1-4}$$

一旦找到这样的超平面，分类决策函数如公式(1-5)定义：

$$y_i = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \tag{1-5}$$

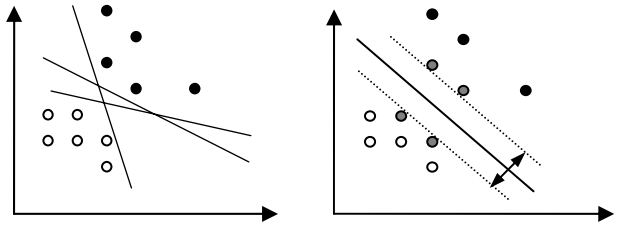

图1-1 支持向量机图示

Figure 1-1 Illustration of Support Vector Machine

针对一个线性不可分问题，特征向量 $\boldsymbol{x}$ 可以用一个非线性的函数 $\Phi(\boldsymbol{x})$ 映射到一个线性可分的更高维的空间中，这种映射一般具有较高的时间复杂度。在支持向量积模型中，从公式(1-4)和公式(1-5)可以看出涉及到的运算只有内积运算。所以我们并不需要真地将数据点映射到高维空间中去，而是通过核函数来模拟数据点在高维空间中的内积运算，见式(1-6)：

$$\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{1-6}$$

常用的核函数为多项式函数，见式(1-7)：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d \tag{1-7}$$

SVM模型可以利用统计学习理论(Computational learning theory)中的概念进行理论的分析，图1-1右部分中的最优超平面将使得测试错误的期望值最小。另外，通过对核函数的应用，简单线性分类算法也可以处理一些非线性问题，同时所有必要的计算都在原始的输入空间中完成，避免了在高维空间中进行运算。基于以上优点，SVM模型在自然语言处理中的文本分类[42]、组块分析[43]、手写识别[44]、词性和词义标注[45, 46]等领域得到了广泛的应用，而且也取得了较理想的结果。

在自然语言处理中利用SVM模型时，通常将二值SVM分类器转换为多值分类器[47]，同时，将SVM的输出转化为概率值也有助于处理序列加标问题[48]。

1.2.4 矢量空间模型

矢量空间模型从严格的角度来说并不属于统计语言模型，本文利用它作为

词义的量化模型来使用，所以也放入统计语言模型这一节进行论述。矢量空间模型最先应用在信息检索领域，具体应用时是将文档和查询都表示为高维词空间中的矢量，文档间的相似度利用矢量间的夹角余弦计算。这种表示方法基于这样一种现象：如果两个文档在一定程度上相同，那么它们会包含一定数量的相同的词。Schútze将矢量空间模型引入到词法分析的词义领域[49]。与信息检索领域中所用的矢量空间模型相同，它也是一个高维、离散和基于词的空间。空间中的每一维是从词典中根据特定量度选出来的一个词，通常称为坐标轴词。所不同的是，在词法分析领域，一个词的特定上下文表示为空间中的一个矢量。这个模型的优点是它可以利用更大范围的上下文，同时不需要一般统计语言模型的参数有指导学习，自动化程度较高。

利用矢量空间模型处理词义问题时，首先通过词语上下文建立所对应的上下文矢量，这些矢量代表词语在不同上下文中的词义信息。然后对矢量进行必要的聚类和归约，以便将词义知识在矢量空间中表示与量化。借鉴这种模型，主要的研究有词义量化模型[50, 51]、有导词义消歧[52]。

它的缺点是训练需要大量的语料；同时，如果将上下文简单的看成一个“词袋”，将会丢失语言中蕴含的结构化信息从而引入大量的噪声。

纵观以上这4种统计语言模型，基本的变化趋势是所利用的上下文范围越来越大，但同时描述能力却越来越弱。N-gram模型通常只考虑前面一个或两个词，对这两个词的约束能力通过计算条件概率给出，具有非常精确的描述和刻画能力，但是基本上不能包含任何长距离的上文信息和任何下文的信息；最大熵模型其本质为一个分类模型，在最大熵整体框架下给出每一种分类结果的概率值。它可以考虑更多的上下文，但是必须经过必要的特征选择。支持向量机模型中的超平面只依赖于支持矢量，所以它包含自动的特征选择过程。同时，它对训练语料的规模要求不高。矢量空间模型可以充分利用很大范围内的上下文信息，信息被均匀地分布到矢量的各个分量中，刻画能力最弱，包含信息最多，通常用来处理蕴含在句子或段落范围内的词义问题，但它需要大规模的训练语料。

表1-3 四种统计语言模型的比较

Table 1-3 Comparison among four statistical language models

	上下文范围	稀疏问题	顺序信息	适用领域	训练语料规模
N-gram模型	局部	严重	有	词	中规模
最大熵模型	单句	一般	有或无	词或词义	中规模
支持向量机模型	单句	一般	有或无	词或词义	中小规模
矢量空间模型	单句或段落	不严重	无	词义	大规模

本文主要以上面这4个统计语言模型为基础展开研究。首先，在第2章研究了N-gram模型的改进方法，第3章利用最大熵模型解决了分词的歧义问题，第4章利用支持向量机和最大熵模型进行了汉语词性标注的研究，第5章利用矢量空间模型进行了词语聚类的研究。

1.3 汉语词法分析

1.3.1 汉语词法分析的研究内容

完整的自然语言处理过程包括词法分析和句法分析两个层次。几乎所有的应用都需要首先进行词法分析，然后将词法分析的结果输入到句法分析过程中以便得到语言内部的结构信息。

汉语词法分析的研究主要集中在分词、词性和词义三个层次上。与英文相比，分词是汉语特有的问题。由于汉语的书写习惯，汉语词之间没有显式的分隔标志，所以基于汉语的应用都必须首先经过分词处理。汉语分词就是把没有分割标记的汉字串转换到符合语言意义的切分词串。目前困扰分词的主要问题就是歧义消解和名实体的识别。词性标注是自然语言处理中的一个基本问题，其任务就是根据一个词在某个句子中的特定上下文，为这个词标注正确的词性。其实质是研究词语所表现的语法功能的聚合关系。它要解决的主要问题是词性歧义(词性兼类)和未登录词词性的确定。词义的研究集中在两个方面，对单义词需要找到它的同义词；对多义词，类似于词性标注的研究，需要在特定上下文中为它标注正确的词义。

1.3.2 汉语词法分析的研究现状

汉语分词主要可以分为3种方法：简单模式匹配方法、基于规则的方法和基于统计的方法。简单模式匹配方法也被称为正向最大匹配和逆向最大匹配方法[53]。其主要思想就是根据词典从输入的文本中匹配长度 L 最大的词（通常 $L \le 10$ ）。同时使用正向和逆向最大匹配方法可以有效地发现交叉歧义，但是不能发现组合歧义，而且这两种方法也不能有效地消解歧义。在基于规则的方法中，Chen利用一系列启发规则消解分词歧义[54]；Hockenmaier和Palmer分别利用基于转换学习的方法来处理分词问题[55, 56]；Wu利用50个Word-Formation规则进行名实体识别[57]。在统计的方法中，Sproat将汉语分词当成一个随机过程，用

WFST模型进行了相关的研究。同时首先提出了一套全面的分词评测方法，分别对歧义、人名和组词分析三个部分进行评测[58]；在此基础上，Gao使用改进的信道模型通过建立七个不同的名实体词类进一步解决了汉语分词中名实体识别的问题[59]；Xue使用最大熵马尔可夫模型(Maximum Entorpy Markov Model)基于汉语单字进行了汉语分词方面的研究，其突出的特点是这种方法不需要词典[60]；Chiang使用鲁棒的自适应学习算法，在错误结果上通过惩罚函数调整当前模型的参数以便进一步提高分词算法的精度[61]；Teahan使用文本压缩算法进行了分词方面的研究[62]。在交叉歧义和组合歧义消解的研究中：Luo提出了基于词义的组合歧义消解算法[63]；孙茂松通过建立分词歧义库，通过对伪歧义记录固定的切分方式来对歧义进行消解[64]；李沐讨论了贝叶斯分类器在处理交叉歧义时的相关问题[65]。

词性标注系统主要可以分为基于统计和基于规则两种。规则模型主要有基于转换学习的模型[10]。统计模型主要有隐马尔克夫模型[66]和统计决策树模型(SDT)[67]。Ratnaparkhi提出了基于最大熵模型的词性标注模型，在华尔街语料上取得了96%的正确率[37]。目前英语的词性标注技术基本上已经成熟，许多研究者提出了各自的词性标注算法，标注也达到了较高的准确率。

词义问题通常包含两个方面：单义词的词义问题通常可以归结为聚类问题，即找到它的同义词；多义词的词义问题通常可以归结为分类问题，即它在特定上下文中的正确词义。聚类需要的词义相似度计算通常借助于可计算的词义词典[68]或上下文的分布特征[69]。应用较广泛的汉语词义词典是知网(HowNet)[70]。与词义相似度计算不同，词义消歧更类似于词性标注，但与词性标注不同的是：词义的自动消歧研究还处在探索阶段。早期的研究通常借助于人工智能的方法，这些方法通常只用在特定、受限的领域中。近期的研究主要可以分为基于知识的方法和基于语料库的方法。基于知识的方法也要通过可计算的词义词典来对词义进行自动地消歧。考虑到构建知识词典的巨大人工代价和词典构造者认识上的不确定的主观因素，基于语料库的方法显示出了更大的灵活性、经济性和可移植性。

基于语料库的方法一般是通过某种机器学习的方法依靠训练语料库构建一个分类器来对词义进行自动消歧。在词义自动消歧的研究历史中，众多的机器学习方法被一一尝试。在有指导统计学习方面，主要的工作有Bayesian方法[71]，在有指导符号学习方面，主要的尝试有决策列表方法(Decision Lists)[72]、决策树方法(Decision Tree)[73]、转换学习的方法等[74]。所有这些方法的一个基本的共同

点在于：它们都在经过人工标注的训练语料上从歧义词的上下文中获得必要的特征信息，然后应用这些特征信息对新的歧义词实例进行分类。所不同的是，在获取信息方面，所有的方法都作了不同的假设。在Bayesian方法中，对上下文中获得的所有信息要求独立性的假设，即假设上下文中的所有词的出现彼此独立，这显然不符合自然语言的规律。决策列表方法中，通过对数似然量度来对所有的特征信息进行排序，然后只用一个具有最好分类特性的特征信息来进行分类。它虽然避免了Bayesian方法中的独立性假设，但是却忽略了其它的对正确分类有帮助的信息。决策树方法似乎可以综合上述两种方法，但是，在处理基于词的自然语言问题时，它有很严重的数据稀疏问题，使得它的性能在很大程度上依赖于剪枝技术和为了增加泛化能力的聚类技术。转换学习的方法作为一种独立的分类器比较理想，这是因为它对一次分类的每种可能不会给出对应的概率值，这种非统计的本质限制了它在系列分类问题中的应用。

以上介绍的方法都是有指导学习模型。有指导学习一般面临知识获取的瓶颈问题。这是因为有指导学习都需要一个足够大的训练语料，而且要保证训练语料的基本正确，否则其中的错误将会使训练出的模型面临“垃圾输入，垃圾输出”(Garbage In, Garbage Out)的问题。通常我们不能完全通过自动的方法加工训练语料，这就使得加工过程需要巨大的人工代价。无指导的参数学习可以有效地克服这一问题，它是一个具有隐含变量的参数训练过程，所依赖的训练集可以是不完全数据(Incomplete Data)，因而不需要事先进行人工加工。

在词义消歧领域基于无指导学习的方法主要有EM方法[75]、Bootstrapping方法[76]、矢量空间模型[50]。EM算法是一个由交替进行的“期望(E过程)”和“极大似然估计(M过程)”两部分组成的迭代过程[77, 78]：对于给定的不完全数据和当前的参数值，“E过程”从条件期望中相应地构造完全数据的似然函数值，“M过程”则利用参数的充分统计量，重新估计概率模型的参数，使得训练数据的对数似然估计最大，见式(1-8)：

$$\phi_{n+1} = \arg\max_{\phi} Q(\phi \mid \phi_n) \tag{1-8}$$

EM算法的每一次迭代过程必定单调地增加训练数据的对数似然值，于是迭代过程渐进地收敛于一个局部最优值。EM算法通常与Bayesian和HMM联合使用，主要作用是自动地估计这些模型中的参数。EM算法已形成许多变型，如隐马尔可夫模型中的Forward-Backward算法或Baum-Welch算法[79]和PCFG中的Inside-Outside算法[80]。Bootstrapping方法是一种统计量估计中的重采样技术，在

词义无指导消歧中得以应用的一个最主要的原因在于自然语言的一个特性，那就是，词义针对于一个固定搭配是稳定的。首先，根据某些固定搭配训练一个基于统计的分类器，然后用这个分类器去标注所有数据，选出高可靠性的一小部分加入训练集后，继续训练这个分类器，如此反复，直到标记完所有输入集合。这种方法的优点在于它充分利用了语言的特性而使得方法本身非常的简洁。缺点是需要给出正确的种子搭配，同时还需要一个基于统计的分类器配合使用。矢量空间模型首先是从信息检索领域借鉴得来。在这种方法中，歧义词的上下文被表示为矢量空间中的一个矢量，然后通过自动聚类来对词义进行无指导消歧。

虽然无指导学习算法减少了人工参与的费用，但是在实际应用中单纯的无指导学习算法并没有取得很好的结果[81]。实验证明：采用有指导学习一般优于无指导学习。

目前汉语词法分析研究的主要困难在于系统的评测。在分词领域规范的不统一主要集中在分词词表的多样性上。1992年国家标准局颁布了作为国家标准的《信息处理用现代汉语分词规范》。规范使用“结合紧密，使用频繁”为成词的基本原则，在具体操作上留下了较大的空间。由于不同的系统使用不同的词表，给跨系统评测带来了困难。同样的，词性标注集和词义分类体系也有多个标准，面临同样的问题。在词义问题上，目前常用的方法是找到若干个歧义词构成测试集，然后根据某种词义分类体系对其进行手工标注，进而评测方法的精度。这种评测方法不仅使系统不能跨测试集评测，而且人工标注的正确率也不能得到根本的保证。

1.3.3 词法分析后续处理步骤——句法分析的研究

作为自然语言处理的两个步骤，词法分析和句法分析是彼此密不可分的，对句法分析的研究有助于更好地定制词法分析的输出。句法分析包含理论和分析算法两个部分，理论是分析算法的基础，分析算法是理论得以实际应用的工具。

随着语言学家对语言的认识不断地深入，语言学家和计算机工作者共同合作，创建了很多句法理论以及基于这些理论的分析算法。其中比较有代表性的有基于对语言知识认知的功能主义句法如：乔姆斯基转换生成句法(TG)、特尼埃尔的依存句法(DG)、菲尔格的格句法、韩礼德代表的系统功能句法、Langacker倡导的认知句法。另外就是基于对语言知识表达的形式主义句法如：短语结构

句法(PSG)、扩充转移网络(ATN)、支配约束理论(GB)、功能合一句法(FUG)、词汇功能句法(LFG)、中心词驱动的短语结构句法(HPSG)、广义短语结构句法(GPSG)、范畴句法(CG)、链接句法(LG)和树邻接句法(TAG)等。如果以句法所基于的基础上来看，又可以分为两类：一类是基于句法范畴的，一类是基于词的，如图1-2所示：

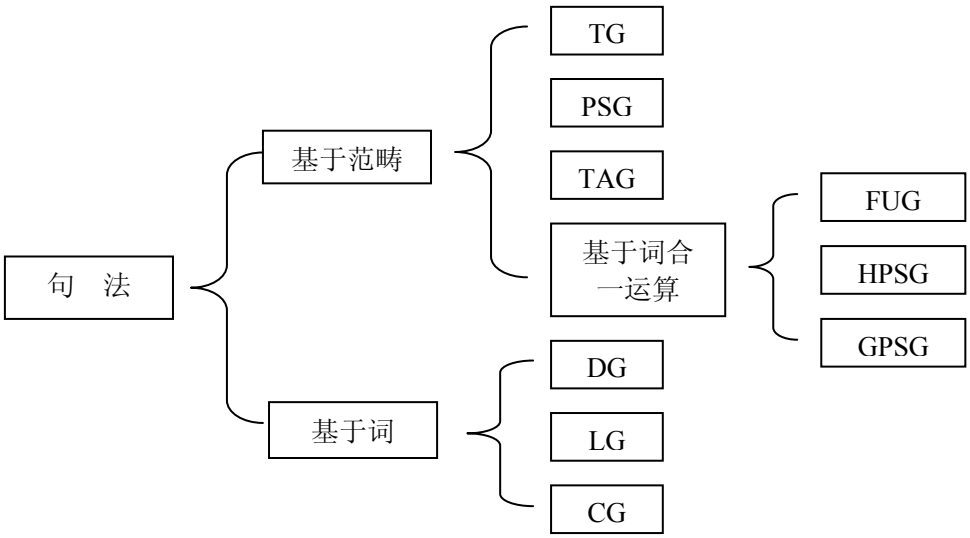

图1-2 句法分类体系

Figure 1-2 Classification of grammer

在以上基本句法理论的支持下，结合语料库语言学的发展，派生出了多种概率型句法。以短语结构句法(PSG)为基础，提出了概率上下文无关句法(Probabilistic Context-Free Grammar, PCFG)。对PCFG的研究主要集中在两个方面：1、PCFG推导及参数学习：Lari和Young首次采用Inside-Outside算法自动估计PCFG参数[80]；在汉语领域，王挺和周强先后采用Inside-Outside算法研究了汉语的PCFG自动推导[82, 83]。2、分析算法：Fujisaki采用VB和CYK相结合的算法实现了一个基于CNF形式的PCFG的句法分析系统[84]；Stolcke将Earley算法与PCFG结合，给出了概率版本的Earley算法[85]。

PCFG目前已经形成较完整的体系，具有形式简洁、参数空间小和分析效率高等优点。但是由于它没有引入词义的成分，消歧能力十分有限。如“牛吃草”和“草吃牛”在句法结构上都是一样的，短语结构句法没有办法加以区别。要解决这个问题，必须要引入更精确的词义定义和约束。如定义“吃”的主语必须是一个动物等。而要引入这些约束必须对词这个句法范畴进行更多的定义，

这样就引入了复杂特征集，以及基于复杂特征集的合一运算。Briscoe将合一句法同概率广义LR表(Probabilistic Generalized LR,PGLR)算法结合，以增强PCFG的上下文描述能力[86]。在Briscoe工作的基础上，朱胜火等人也对基于GLR的句法分析算法进行了有益的研究[87]。以依存句法为基础，Simmons首次提出上下文依存句法(Context-Dependent Grammar, CDG)[88]，并基于CDG针对英语受限子集(Newswire stories)实现了一个英语句法获取和分析系统。借鉴CDG思想，周明实现了一个汉语依存句法交互标注工具[89]。以树邻接句法(TAG)为基础，Schabes提出一种概率树邻接句法(Probabilistic Tree-Adjoining Grammar，PTAG)[90]，这种句法在上下文无关句法中的标准替换规则基础上增添了一种附加规则，以提高规则的上下文敏感性。

通过以上的介绍，可以发现句法分析不仅需要词义的支持，更需要“知识”。相比词法分析来说，句法分析是一个更为困难的任务。试图将词法分析中的问题留给句法分析去解决，类似于用一个更难的问题去解决一个难的问题。所以本文并没有采用这样的研究路线，而是争取在词法分析领域内部去解决相关的问题。这也是通过研究句法分析理论得到的结论。

1.4 用于词法分析的数据资源建设

词法分析的研究需要基于一定的数据资源：它们主要包括词典和语料库两个部分，下面分别给予介绍。

1.4.1 词法词典的建设

词典是词法分析的基础，也是词法分析的重要知识来源。没有词典，任何词法分析的研究都无从谈起。目前，对自然语言处理的应用从保存、检索上升到问答、文摘、直至包含语言理解和生成的机器翻译。为满足上述要求，词典已经不能只停留在一个单独的词表上，目前词典通常都包含丰富的词义和句法信息。

如果要构建包含词义的词典，需要了解什么是词义。词义通常可定义为在一定的语言环境中所阐明的内容[91]。想要描述一种词语所表述的意义，常见的方法有三种：第一种就是同义词分类的方法，这种方法出现的时间最早，所基于的概念也最简单，一个词的词义完全可以用属于同一个集合中的其它词来表示；第二种是基于成分词汇语义学(Componential Lexical Semantics)的表示方法。

这种方法也称为义原分析法，就是把一个词的意义分析为更小的概念原子的组合，这类原子也可以称为义原。不过，定义一套概念原子却非易事；第三种是基于关系词汇语义学(Relational Lexical Semantics)的表示方法。随着这方面研究的增多，越来越多的认知心理学家和计算语言学家意识到：除了利用义原分析法定义词义，还可以利用“网”的形式来描述词义。词汇所表示的概念相互之间存在着联系，彼此构成了一个知识网络，因此这类词义词典在整体上就是一个词义网络或者是知识网络。

第一类的词义词典代表为《现代汉语辞海》[92]和《同义词词林》[93]。《同义词词林》提供了一个汉语词义的分类体系，其分类体系分为大、中、小3级，共分12个大类、94个中类、1428个小类。其中每个大类以大写英文字母表示，中类以小写字母表示，小类则以阿拉伯数字表示。例如：词“个人”在《同义词词林》中的编号为Aa01。A说明它属于“人”大类，小写字母a代表它是“泛称”中类，数字01代表它是“人民”小类。它的不足之处是很多常用词没有被收录。在英语中，较有代表性的同义词词典为Roget的《International Thesaurus》，目前已被广泛用于机器翻译[94]、信息检索[95]及文章内容分析[96]等领域。

第二类的词义词典以董振东建立的《知网》(HowNet)为代表。它是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[70]。首先通过对约六千个汉字进行考察和分析来抽取一千多个义原。义原是知网中最基本的，不能再分割的意义最小单位，将其作为解释知识词典的基本要素，其它的词条全都由这些义原来定义。例如《知网》中医生的定义为DEF=human|人,#occupation|职位,*cure|医治,medical|医。其中，“人、职位、医治和医”，就是用来定义词义的四个原子义原，分别表示医生是人，与职位相关，是医治的施事者。

第三类的词义词典代表为WordNet[97]，该词典集成了许多在词义消岐中常用到的词汇特征。另外，WordNet还提供了表示词义关系的连接，包括词义间的下义关系、反义关系等。它的名词和动词都是分层级组织词语之间词义关系的，在名词中，有上下位关系(Hyper-hyponymy)和整体--部分(Meronymy)关系等；在动词中有下位(Troponymy)关系和继承(Entailment)关系等；其中动词中的继承关系类似名词中的整体--部分关系。

国内较著名的句法词典是由北大俞士汶教授等主编的《现代汉语语法信息词典》[98]，该词典共收录了五万多词条，每个词条都根据其句法功能给出了其词性信息，词性依据朱德熙[99]先生提出的词组本位句法体系作为设置各项句法

范畴的理论基础。并结合汉语的特点给出了每个词的拼音、同形等丰富的词法信息；不仅如此，词典更给出了详细的句法信息。值得一提的是国内学者黄曾阳积多年研究心得，提出面向整个自然语言理解的理论框架——概念层次网络理论(HNC)[100]，对传统的基于句法知识的语言表述及处理模式提出了挑战，代之以词义表达为基础来对汉语进行理解。目前也取得了较多的研究成果。

1.4.2 语料库的建设

语料库是作为信息载体的大量的语言资料的集合。语料库有以下几种分类标准，以语料库设计结构分类：可分为均衡结构语料库和无结构的随机开放式语料库；以语料库的来源分类：可分为单语种语料库和多语种语料库；以语料的时效分类：可分为共时语料库与历时语料库；以语料库的处理方式分类：可分为未经加工的生语料和经过加工的熟语料[101]；对生语料进行加工的过程就是添加相应的“显性”解释性语言学信息(词、词性、句法成分和义项等)的过程，通常称为语料库标注或加工[102]。从标注过的语料库中我们可以获取语言学信息的统计分布知识，用于语言学分析或建立语言模型。

国外的语料库建设开始于60年代，至今已发展了三代，60年代，美国Brown大学建立了世界上第一个标准语料库——Brown语料库。该语料库规模为100万词次的美国英语。70年代开发了与Brown语料库类似的基于英语的LOB语料库。这两个语料库可以称为第一代语料库的代表。80年代初由Collins出版社资助的COBUILD语料库达到了2000万词次，覆盖英语和美式英语。进入90年代，随着各种加工技术和统计方法日益成熟，语料库发展也进入了第三代。美国Pennsylvania大学对百万词次的英语语料进行了全面的词性和句法标注，建立了大规模的Penn树库[103]。成为计算语言学领域的一个重要资源。

由于汉语书写的特殊性，汉语的语料库加工还包括分词。目前大多数的汉语语料库加工集中在分词和词性标注上，句法树库正处于一个迅速发展时期。而其它层次的加工，如语义和言语等还处于起步阶段。清华大学按照系统性原则收集了5000万汉字的原始语料库；北京大学也建立了1998年人民日报带有词性和名实体标注的汉语新闻类语料库；哈尔滨工业大学机器翻译和信息检索教研室也都在各自的网站上公布了部分汉语句法树库。

语料库的建设与加工是构造统计语言模型必不可少的基础性工作。无论是什么应用系统，都需要一定的语料库自动加工技术，这是非常明显的。目前，语料库建设的重要性已经被越来越多的研究人员所认识。

1.5 本文主要工作

1.5.1 本文研究内容

本文在统计语言模型方面首先利用词法信息对N-gram模型的平滑方法进行了改进；其次，利用AMI提取了描述长距离约束关系的触发对；然后，建立了基于触发对的矢量空间模型；最后，建立了融合转换触发对的最大熵模型。利用以上统计语言模型领域中的研究成果对汉语词法分析领域的分词、词性标注和词义三个子领域以及音字转换进行了研究，全文组织如下：

第1章为绪论部分。首先讨论了本文研究的主要目的和意义，对本文应用的统计语言模型进行了综述和比较；其次介绍了汉语词法分析领域主要的研究内容、现状和面临的问题。然后给出了用于词法分析的数据资源——词典和语料库方面的研究进展。

第2章主要介绍了传统N-gram模型的改进方法。针对数据稀疏问题，首先讨论了各种主要的N-gram模型平滑技术；其次改进了Katz平滑算法；然后利用HowNet词典提出了一种新的Uni-gram平滑方法；最后介绍了用于解决长距离约束问题的词触发对和转换触发对的抽取。

第3章主要介绍了汉语分词方面的研究。首先介绍了基于REA算法的K-best分词模型;其次介绍了如何对特定的句子计算对应的K值以便有效地从伪歧义字段中识别出真歧义字段；然后利用最大熵模型进行了分词歧义消解的研究；针对人名识别问题，建立了人名识别多源知识表以配合统计方法使用；针对英语、时间、数字等因子词，采用了有限自动机理论对其进行了识别。

第4章主要介绍了词性标注方面的研究。首先介绍了传统HMM模型的问题；其次引入最大熵模型和支持向量机模型对复杂兼类词的标注进行了研究；然后引入最大熵模型对基于句子的词性标注进行了研究，重点探讨了长距离特征的加入以及用于序列分类的Beam Search搜索算法；最后利用词性标注的研究成果对音字转换进行了探索性的研究。

第5章主要讨论了利用矢量空间模型计算词义相似度。首先利用词分辨力筛选出矢量空间模型中的坐标轴词；然后利用词触发对库建立基于触发对的矢量空间模型；最后在此模型上进行了词义相似度的计算和词语聚类的研究。

1.5.2 主要创新点

本文以汉语词法分析和音字转换为测试平台，对统计语言模型的平滑、解码算法、长距离约束三个方面提出了相应的改进方法，主要创新点如下：

一、以往的N-gram模型平滑算法主要基于统计学知识，而没有针对语言这个特定的待平滑对象进行研究。本文以此为切入点，针对传统Katz平滑算法在处理某些固定搭配时存在无法折扣出概率的问题，利用词性信息提出了新的Katz平滑折扣系数，在模型交叉熵量度上取得了比Abs平滑和W-B平滑更好的结果。在此基础上，利用HowNet词典提供的词义相似度计算功能改进了Uni-gram平滑算法。试验结果证明：利用词法信息可以改进传统平滑算法的性能。

二、针对分词中的歧义问题，提出了基于REA算法的K-best分词模型以及相应的K值计算方法，同时利用最大熵模型对分词歧义进行了消解。针对人名识别中的竞争和歧义问题，综合统计与语言学知识，建立了人名识别多源知识表。基于以上研究成果开发的分词系统在2003年全国863评测中，在分词的交叉歧义和组合歧义消解指标上取得最好成绩。通过在统计方法中使用人名识别多源知识表，可以保证在不降低召回率的情况下，提高识别的准确率。

三、为解决HMM在词性标注中不能包含长距离约束信息的问题，本文首先利用平均互信息抽取出转换触发对“$w_A \rightarrow w_B / t_B$”，然后利用同义词建立聚类触发对以解决转换触发对的稀疏问题，最后在最大熵框架下与局部特征进行了融合，有效地解决了语言中的长距离约束问题。试验结果证明：在1998年第6个月人民日报语料上进行测试，融合了转换触发对的最大熵标注模型比传统的HMM相比，词性标注的错误减少了1/3。

四、对矢量空间模型研究的工作包括：提出了词分辨力量度并选择高分辨力的词作为矢量空间的坐标轴词。针对矢量空间模型中“词袋”假设带来的噪声，利用触发对模拟一个依存句法分析器，使得新的矢量空间模型包含了语言中的结构信息。试验结果证明：与传统的矢量空间模型相比，新模型的质量提高了32%。基于新矢量空间模型的词语聚类研究也取得了比较理想的结果。

第2章 N-gram模型改进方法研究

2.1 引言

不可否认的是，在自然语言处理中的词法分析领域，N-gram模型是目前应用最广的统计语言模型。它有数据结构简单、易于实现、训练和运行速度快、内存占用小等优点。与基于规则的语言模型相比，N-gram在大量的商用系统中都取得了非常好的结果。但是，N-gram模型也面临着两个主要的问题，首先，它不能包含长距离的约束信息，同时，也面临数据稀疏带来的挑战，本章将对这两个问题作进一步的研究。

由于自然语言遵循Zip'f定律，所以不能简单地通过增加训练语料规模来解决数据稀疏带来的零概率问题，通常研究人员利用平滑算法来解决这个问题。本章首先系统回顾并总结了平滑算法领域前人研究的成果，分别实现了图灵(Good-Turning, G-T)、绝对(Absolute, Abs)、W-B和Katz四种平滑算法。平滑算法可以看成是由概率的折扣与重新分配两个部分组成，折扣和分配都需要一定的知识来作为基础。已有平滑算法的研究重点基本集中在统计知识的发掘和使用上，目前还很少应用语言学领域的知识，尤其是词法信息进行模型平滑的研究。本章以此为突破口进行了相应的探索。首先利用词性信息对Katz平滑算法进行了改进。在此基础上，利用HowNet2004知识词典提出了一个融合词义信息的Uni-gram模型平滑算法。这种算法可以被当成一个基本的组件用在其它各种平滑算法中，改进其它高阶N-gram模型平滑算法的性能。

为解决N-gram模型不能包含长距离约束信息的问题，本章利用平均互信息量度抽取了200万词触发对。与互信息相比，平均互信息可以过滤掉由特定搭配引起的一部分噪声。在此基础上，针对词法分析中特定的分词歧义、词性标注歧义等问题，提出了形如“$w_A \rightarrow w_B / t_B$”和“$y_A \rightarrow y_B / c_B$”转换触发对的概念。在后续的词法分析研究中通过加入转换触发对有效地提高了最终结果的精度，我们将在第3章第4章中详细论述。

本章的主要研究内容如下：第2.2节系统地总结了已有的平滑算法研究工作，并介绍了两种改进的平滑算法；第2.3节讨论了触发对的抽取过程；第2.4节给出了试验结果；最后是本章的小结。

2.2 改进N-gram模型平滑算法

2.2.1 已有平滑算法综述

2.2.1.1 最大似然估计和G-T平滑

任何统计语言模型都涉及到模型参数的学习问题，在语言建模过程中主要有两种参数学习算法，即有指导参数学习和无指导参数学习。有指导参数学习在训练样本中提供类别标志和分类代价，学习算法对给定的问题寻找与样本最佳拟合的分布或能降低总体代价的方向。在自然语言处理领域中，N-gram模型通常用最大似然估计(Maximum Likelihood Estimation, MLE)对模型的参数进行训练。如果有足够规模的已进行相应语言层次消歧或标注的熟语料，那么很容易利用最大似然估计通过从训练语料中观察到的相对频度(Relative Frequency)准确地估计相对应的统计语言模型的参数。这种训练方法可以使模型最佳地拟合训练语料。下面给出最大似然估计在N-gram模型中的公式化描述，$c(w_{i-n+1}^{i})$ 代表 n 元词串 w_{i-n+1}^{i} 在训练语料中出现的次数。基于最大似然估计的上下文条件概率计算见式(2-1)：

$$p_{ML}(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i})} \tag{2-1}$$

通常，语言模型的训练文本 T 的规模及其分布存在着一定的局限性和片面性，许多合理的语言搭配现象没有出现在 T 中。当 $c(w_{i-n+1}^{i}) = 0$ 的时候，该词串对应的上下文条件概率 $p_{ML}(w_i \mid w_{i-n+1}^{i-1}) = 0$，从而导致该词串所在的语句 S 的出现概率 $p(S) = 0$。为了解决这个问题，研究人员提出了G-T平滑算法。首先，引入基于 n 元词串 w_{i-n+1}^{i} 相关定义，其中 $r = c(w_{i-n+1}^{i})$，n_r 定义见式(2-2)：

$$n_r = \sum_{w_{i-n+1}^{i}} \delta(c(w_{i-n+1}^{i}), r) \tag{2-2}$$

函数 δ 的定义见式(2-3)：

$$\delta(a,b) = \begin{cases} 1 & \text{当} a = b \\ 0 & \text{其它} \end{cases} \tag{2-3}$$

基于以上定义，对于N-gram模型中出现 r 次的 n 元词串 w_{i-n+1}^{i}，基于G-T平滑算法的上下文条件概率计算见式(2-4)：

$$p_{GT}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\frac{n_{r+1}}{n_r} \cdot (r+1)}{\sum_{w_i} c(w_{i-n+1}^{i})} \tag{2-4}$$

在式(2-4)中由于n_r不能为零，所以Gale和Sampson提出一种用于n_r的改进平滑算法[104]，统计语言模型中通常使用这种方法进行G-T平滑。

2.2.1.2 J-M平滑算法

J-M平滑是一种基于线性插值的平滑方法。该数据平滑方法主要利用低阶N-gram模型与高阶N-gram模型进行线性插值。定义见式(2-5)：

$$p_{JM}(w_i \mid w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} \cdot p_{ML}(w_i \mid w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}}) \cdot p_{JM}(w_i \mid w_{i-n+2}^{i-1}) \tag{2-5}$$

N-gram模型可以递归地定义为由最大似然估计原则得到的n阶p_{ML}和$(n-1)$阶p_{JM}的线性插值。为了结束以上的递归定义：令0-gram模型为一个均匀分布模型，见式(2-6)：其中V代表词表中所有词构成的集合。

$$p_{0-gram}(w_i) = \frac{1}{|V|} \tag{2-6}$$

一般利用Baum-Welch算法计算插值系数$\lambda_{w_{i-n+1}^{i-1}}$[27]。其基本思想为：使用经过数据平滑的模型概率参数，计算一个测试集T的对数似然概率$\log p(T)$，当$\log p(T)$为极大值时，对应的$\lambda_{w_{i-n+1}^{i-1}}$为最优值。因此可以求解$\log p(T)$对应于每个$\lambda_{w_{i-n+1}^{i-1}}$的偏导数，令$\frac{\partial \log p(T)}{\partial \lambda_{w_{i-n+1}^{i-1}}} = 0$。通过对该方程求解，可以得到$\lambda_{w_{i-n+1}^{i-1}}$的迭代计算公式，见式(2-7)：

$$\lambda'_{w_{i-n+1}^{i-1}} = \frac{1}{c(w_{i-n+1}^{i-1})} \sum_{w_i} c(w_{i-n+1}^{i}) \frac{\lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i \mid w_{i-n+1}^{i-1})}{\lambda_{w_{i-n+1}^{i-1}} P_{ML}(w_i \mid w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}}) p_{JM}(w_i \mid w_{i-n+2}^{i-1})} \tag{2-7}$$

由于需要计算的$\lambda_{w_{i-n+1}^{i-1}}$参数众多，Jelinek和Mercer建议对$\lambda_{w_{i-n+1}^{i-1}}$参数空间进行桶式分类，Baul进一步提出根据$c(w_{i-n+1}^{i-1})$的值对$\lambda_{w_{i-n+1}^{i-1}}$进行分类，属于同

一类的 $\lambda_{w_{i-n+1}^{i-1}}$ 设置为相同的值[105]。

2.2.1.3 W-B平滑算法

W-B平滑算法是J-M线性插值平滑算法的一个特例。该平滑算法与J-M平滑算法的不同之处在于插值系数 $\lambda_{w_{i-n+1}^{i-1}}$ 的计算方式。J-M平滑算法采用Baum-Welch重估计算法训练 $\lambda_{w_{i-n+1}^{i-1}}$，而Witten-Bell平滑算法采用如下的公式计算 $\lambda_{w_{i-n+1}^{i-1}}$，定义见式(2-8)：

$$1-\lambda_{w_{i-n+1}^{i-1}}=\frac{N_{1+}(w_{i-n+1}^{i-1}\bullet)}{N_{1+}(w_{i-n+1}^{i-1}\bullet)+\sum_{w_i}c(w_{i-n+1}^{i})} \tag{2-8}$$

其中，符号 $N_{1+}(w_{i-n+1}^{i-1}\bullet)$ 的定义见式(2-9)：

$$N_{1+}(w_{i-n+1}^{i-1}\bullet)=|\{w_i \mid c(w_{i-n+1}^{i-1},w_i)>0\}| \tag{2-9}$$

位置符号“$\bullet$”代表在训练语料库中出现在词串 w_{i-n+1}^{i-1} 之后的任意一个词。$N_{1+}()$ 表示括号里的处于位置“$\bullet$”且出现次数大于0的词的个数。将 $\lambda_{w_{i-n+1}^{i-1}}$ 的值带入公式(2-5)，可以得到W-B平滑算法的公式，见式(2-10)：

$$p_{wb}(w_i \mid w_{i-n+1}^{i-1})=\frac{c(w_{i-n+1}^{i})+N_{1+}(w_{i-n+1}^{i-1}\bullet)\cdot p_{wb}(w_i \mid w_{i-n+2}^{i-1})}{\sum_{w_i}c(w_{i-n+1}^{i})+N_{1+}(w_{i-n+1}^{i-1}\bullet)} \tag{2-10}$$

2.2.1.4 Katz平滑算法

Katz提出了一种采用回退方式进行数据平滑(Backing-off Smoothing)的算法。该数据平滑算法的主要思想为：当一个 n 元词串 w_{i-n+1}^{i} 的出现次数 $c(w_{i-n+1}^{i})$ 大于一个阈值时，$p_{ML}(w_i \mid w_{i-n+1}^{i-1})$ 是 w_{i-n+1}^{i} 可靠的概率估计。而当 $c(w_{i-n+1}^{i})$ 小于一个阈值，采用G-T平滑算法，将其部分概率折扣给未出现的 n 元词串 w_{i-n+1}^{i}。当 $c(w_{i-n+1}^{i})=0$ 时，模型回退到低阶模型，给未出现的 n 元词串按着 $p_{Katz}(w_i \mid w_{i-n+2}^{i-1})$ 比例来分配被折扣出来的概率。综合上述思想，Katz平滑算法见式(2-11)：

$$p_{Katz}(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} p_{ML}(w_i \mid w_{i-n+1}^{i-1}) & if\,(c(w_{i-n+1}^{i}) \geq k) \\ d_r \cdot p_{ML}(w_i \mid w_{i-n+1}^{i-1}) & if\,(1 \leq c(w_{i-n+1}^{i}) < k) \\ \alpha(w_{i-n+2}^{i-1}) \cdot p_{Katz}(w_i \mid w_{i-n+2}^{i-1}) & if\,(c(w_{i-n+1}^{i}) = 0) \end{cases} \quad (2\text{-}11)$$

其中，$p_{ML}(w_i \mid w_{i-n+1}^{i-1})$ 为最大似然估计，阈值 k 为一个常量，Katz建议 $k = 5$。

参数 $\alpha(w_{i-n+2}^{i-1})$ 定义见式(2-12)：

$$\alpha(w_{i-n+2}^{i-1}) = \frac{1 - \sum_{w_i : c(w_{i-n+1}^{i}) > 0} p_{Katz}(w_i \mid w_{i-n+1}^{i-1})}{\sum_{w_i : c(w_{i-n+1}^{i}) = 0} p_{Katz}(w_i \mid w_{i-n+2}^{i-1})} \quad (2\text{-}12)$$

其作用是保证模型参数满足概率的归一化约束条件，即 $\sum_{w_i} p_{Katz}(w_i \mid w_{i-n+1}^{i-1}) = 1$。为了结束公式(2-11)的递归定义，令0-gram模型为均匀分布模型。

根据G-T平滑算法，被折扣给所有未出现的 n 元词串 w_{i-n+1}^{i} 的次数之和等于 n_1。见式(2-13)：

$$\sum_{r=1}^{k} n_r (1 - d_r) r = n_1 \quad (2\text{-}13)$$

通过对式(2-13)的求解，可以求出 d_r 的值，见式(2-14)：

$$d_r = \frac{\frac{n_{r+1}(r+1)}{n_r r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \quad (2\text{-}14)$$

显然，与线性插值平滑算法相比，回退式数据平滑算法的参数较少，而且可以直接确定，无需通过某种迭代重估计算法反复训练，因此它的实现更为方便。

2.2.1.5 绝对(Abs)平滑算法

原始的绝对平滑算法定义见式(2-15)：

$$p_{abs}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c(w_{i-n+1}^{i}) - D, 0)}{\sum_{w_i} c(w_{i-n+1}^{i})} + (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{abs}(w_i \mid w_{i-n+2}^{i-1}) \quad (2\text{-}15)$$

Ney采用Leaving-One-Out方法自动优化折扣系数 D，见式(2-16)：可以看出 $0 < D < 1$。

$$D = \frac{n_1}{n_1 + 2n_2} \tag{2-16}$$

为了保证 $\sum_{w_i} p_{abs}(w_i \mid w_{i-n+1}^{i-1}) = 1$，令：

$$1 - \lambda_{w_{i-n+1}^{i-1}} = \frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1} \bullet) \tag{2-17}$$

这里 $N_{1+}(w_{i-n+1}^{i-1}\bullet)$ 的定义与公式(2-9)相同。从式(2-15)可以看出，绝对平滑与J-M平滑不同，J-M平滑通过乘以插值系数 $\lambda_{w_{i-n+1}^{i-1}}$ 来进行折扣，而绝对平滑的折扣过程是通过减去一个固定的常数 D 。在绝对平滑算法中，如果 $c(w_{i-n+1}^{i})$ 比较大，我们认为这种估计是可靠的。减去 D 后对它影响不大，等同于对其进行较小的折扣；如果 $c(w_{i-n+1}^{i})$ 很小，我们认为这种估计是不可靠的。减去 D 后对其影响很大，等同于对其进行较大的折扣。这与基于线性折扣的J-M平滑算法恰好相反。从这个角度上看，绝对平滑更加符合语言现象的规律。

2.2.1.6 K-N平滑算法

K-N平滑算法是绝对平滑算法的改进，其最大特点是将语言学的特点和知识应用到了平滑算法中。在自然语言中，存在一种固定的搭配现象，如汉语中的一个搭配“宏观/调控”，“调控”一词出现的概率 $p_{abs}(w_i)$ 很高，根据公式(2-15)，在Bi-gram模型中，针对任何的 w_{i-1}，条件概率 $p_{abs}(\text{调控} \mid w_{i-1})$ 都相对较大。这并不符合自然语言分布的现象和规律，因为“调控”一词基本上都跟随在“宏观”一词的后面。为此，Kneser提出了一种新的计算低阶分布的方法，以Uni-gram概率为例，见式(2-18)：

$$p_{KN}(w_i) = \frac{N_{1+}(\bullet w_i)}{N_{1+}(\bullet\bullet)} \tag{2-18}$$

其中，分子和分母的定义见式(2-19)和(2-20)：

$$N_{1+}(\bullet w_i) = |\{w_{i-1} : c(w_{i-1}, w_i) > 0\}| \tag{2-19}$$

$$N_{1+}(\bullet\bullet)=\sum_{w_i}N_{1+}(\bullet w_i) \tag{2-20}$$

公式(2-19)代表的是词w_i前面出现的所有词的个数，而不是频数。例如，如果“宏观/调控”出现10次，但“调控”一词只出现在“宏观”一词后面，那么$N_{1+}(\bullet\text{调控})=1$。这里可以看出K-N平滑所依据的是语言的分布信息而不是简单的频数统计。采用这种方法，一阶的概率分布更加符合自然语言的现象。这样就可以有效解决固定搭配带来的问题。K-N平滑算法可以给我们一些启示，那就是在对语言模型进行平滑的过程中，仅仅从数学的角度来调整概率的分布是不够的，还要综合考虑语言这个样本的一些特有的规律和约束。

将公式(2-18)推广到高阶模型中，得到式(2-21)：

$$p_{KN}(w_i\mid w_{i-n+2}^{i-1})=\frac{N_{1+}(\bullet w_{i-n+2}^{i})}{N_{1+}(\bullet w_{i-n+2}^{i-1}\bullet)} \tag{2-21}$$

其中，分子和分母的定义见式(2-22)和(2-23)：

$$N_{1+}(\bullet w_{i-n+2}^{i})=\mid\{w_{i-n+1}\mid c(w_{i-n+1}^{i})>0\}\mid \tag{2-22}$$

$$N_{1+}(\bullet w_{i-n+2}^{i-1}\bullet)=\mid\{(w_{i-n+1},w_i)\mid c(w_{i-n+1}^{i})>0\}\mid \tag{2-23}$$

综上所述，Kneser-Ney平滑公式被定义为：

$$\begin{aligned}p_{KN}(w_i\mid w_{i-n+1}^{i-1})=&\frac{\max\{c(w_{i-n+1}^{i})-D,0\}}{\sum_{w_i}c(w_{i-n+1}^{i})}\\&+\frac{D}{c(w_{i-n+1}^{i})}\cdot N_{1+}(w_{i-n+1}^{i-1}\bullet)\cdot p_{KN}(w_i\mid w_{i-n+2}^{i-1})\end{aligned} \tag{2-24}$$

其中，$p_{KN}(w_i\mid w_{i-n+2}^{i-1})$由公式(2-21)计算，而$D$值由公式(2-16)计算。$N_{1+}(w_{i-n+1}^{i-1}\bullet)$的定义与式(2-9)相同。

2.2.2 已有平滑算法的总结

Kneser以统一的公式概括了上面介绍过的各种平滑算法[26]，见式(2-25)：

$$p(w_i\mid w_{i-n+1}^{i-1})=\begin{cases}\alpha\cdot p(w_i\mid w_{i-n+1}^{i-1}) & if\ c(w_{i-n+1}^{i})>0\\ \gamma(w_{i-n+1}^{i-1})\cdot p(w_i\mid w_{i-n+2}^{i-1}) & if\ c(w_{i-n+1}^{i})=0\end{cases} \tag{2-25}$$

式(2-25)中，α为折扣变量$0<\alpha<1$，针对在训练语料中出现过的事件

$c(w_{i-n+1}^{i})$，通过α减少了一部分的概率值。值得注意的是式(2-25)所表示的只是一种折扣方式，并不代表通过乘法来对概率值进行折扣。针对那些没出现过的事件$c(w_{i-n+1}^{i})=0$。我们采用相对低阶的事件$c(w_{i-n+2}^{i})$来进行预测，直到回退到Uni-gram。如果Uni-gram也存在零概率的问题，就回退到0-gram分布，即均匀概率分布上。式中$\gamma(w_{i-n+1}^{i-1})$为归一化参数，用来保证所有事件的概率之和等于1。

Chen在不同的语料规模上进行了不同平滑算法之间的比较，同时也给出了各种平滑算法间的主要区别[106]，如表2-1所示：

表2-1各种平滑算法的比较

Table 2-1 Comparsion among smoothing algorithms

	分配		折扣		
	回退	插值	图灵	线性	绝对
W-B		√		√	
Abs	√				√
Katz	√		√		
J-M		√		√	
K-N	√				√

在表2-1中，最左列给出五种不同的平滑算法。减少可见事件概率值的过程称为折扣，常用的折扣方法有三种：分别为图灵折扣、线性折扣和绝对折扣。折扣出来的概率需要重新进行分配，常用的分配方式可以分为回退方式和插值方式两种。表2-1中给出的分配方法代表原始文献中所用的方法，但并不代表只能用这一种分配方式。例如，本文中所用的绝对平滑算法就是以插值的方式实现的。这两种分配方式的主要区别之处在于插值方式将折扣出来的概率值在整个分布上进行重新分配。而回退方式只是将折扣出的概率值在低阶的分布上进行重新分配。例如：以绝对平滑算法为例：如果用公式(2-26)来定义，就为插值分配方式：

$$p_{abs}(w_i \mid w_{i-n+1}^{i-1})=\begin{cases}\dfrac{\max(c(w_{i-n+1}^{i})-D,0)}{\sum_{w_i} c(w_{i-n+1}^{i})} & if\ c(w_{i-n+1}^{i})>0 \\ \gamma(w_{i-n+1}^{i-1})\cdot p_{abs}(w_i \mid w_{i-n+2}^{i-1}) & if\ c(w_{i-n+1}^{i})=0\end{cases} \tag{2-26}$$

其中，$\gamma(w_{i-n+1}^{i-1})$的计算见式(2-27)：

$$\gamma(w_{i-n+1}^{i-1}) = \frac{D \cdot N_{1+}(w_{i-n+1}^{i-1}\bullet)}{\sum_{w_i} c(w_{i-n+1}^{i}) \cdot \sum_{w_i:c(w_{i-n+1}^{i})=0} p_{abs}(w_i \mid w_{i-n+2}^{i-1})} \tag{2-27}$$

与公式(2-15)对比可以发现，基于回退的绝对平滑算法将折扣出来的概率只在$c(w_{i-n+1}^{i}) = 0$的事件上进行了分配。通常情况下，基于插值的分配方式在全部的事件上分配信息，所以效果上要略好于基于回退的方式。

2.2.3 基于词性信息改进Katz平滑算法

2.2.1.4节介绍的Katz平滑算法在处理某些具有固定搭配性质的语言现象的时候，存在无法进行概率折扣的问题。下面以Bi-gram模型为例进行说明，Tri-gram模型与此类似。传统的Katz平滑算法定义见式(2-11)：原始定义中规定当频度大于等于5时，不进行概率折扣。但在自然语言现象中，存在如下的现象，那就是$N_{1+}(w_{i-1}\bullet) = 1$，也就是说，词$w_{i-1}$后面只出现过一个词，且$c(w_{i-1}, w_i) \geq 5$，或者是$N_{1+}(w_{i-1}\bullet) > 1$，但全部的$c(w_{i-1}, w_i) \geq 5$，部分实例见表2-2：

表2-2 部分固定搭配

Table 2-2 Partial Collocations

w_i	w_i	$c(v$	w_{i-}	w_i	
老	地	11	短	雷	
麻	理	9	高	公	
不	的	29	教	组	

从表2-2中可以看出，这类现象有些来源于固定词组，如：“老少边穷/地区”，这是因为“老少边穷”后面基本上只出现“地区”一词；有些则来源于训练语料规模偏小而带来的数据稀疏问题，如：“高等级/公路”，这是因为“高等级”一词也可能出现在“铁路”一词前面。无论是什么原因，这类语言现象都会给传统的Katz平滑算法带来影响。由于$c(w_{i-1}, w_i) \geq 5$，不能从中折扣出概率而分配

给未见事件。这就出现了 $p_{Katz}(w_i \neq \text{地区} \mid \text{老少边穷}) = 0$ 的情况，从而进一步地影响对应的Tri-gram模型参数的平滑。

传统Katz平滑算法采用的是图灵折扣，其折扣系数 d_r 的计算来源于图灵平滑的思想，但是这种计算方法却不能处理具有固定搭配性质的语言现象。为有效地解决这个问题，需要从 $c(w_{i-1}, w_i) \geq 5$ 的事件中折扣出部分概率，为此我们需要重新定义Katz平滑算法的折扣系数 d_r。根据对已有平滑算法的研究，本文认为新的折扣系数要满足三方面的要求：首先，$c(w_{i-n+1}^{i})$ 越大，应该折扣出较小的值，这是因为频度越高，代表它是越可靠的估计；同时，如果 $N_{1+}(w_{i-1}\bullet)$ 越大，代表它有可能跟随更多的词，所以应该折扣出更大的值；最后，从表2-2中可以看出，针对Bi-gram统计，如果词 w_i 是个名词词性，如：“短时/雷雨”，那么很有可能存在同义的其它的名词词性的词，如“短时/冰雹”、“短时/大风”等。这种不是固定搭配性质的二元统计应该多折扣出一些概率。综合以上三种要求，借鉴W-B平滑算法，本文定义了一个新的折扣系数 d_r，见式(2-28)：

$$d_r = \begin{cases} = \dfrac{\sum_{w_i} c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i}) + N_{1+}(w_{i-n+1}^{i-1})} & if\,(w_i\text{的词性不是名词}) \\ = \dfrac{\sum_{w_i} c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i}) + c(w_i)/N_{1+}(w_{i-n+1}^{i-1})} & if\,(w_i\text{的词性是名词}) \end{cases} \tag{2-28}$$

从式(2-28)可以看出，如果 w_i 是名词词性，那么折扣出来的概率值见式(2-29)：

$$\frac{c(w_i)/N_{1+}(w_{i-n+1}^{i-1})}{\sum_{w_i} c(w_{i-n+1}^{i}) + c(w_i)/N_{1+}(w_{i-n+1}^{i-1})} \tag{2-29}$$

上面已经论述过，对于名词词性的 w_i，应该多折扣一些。这里我们用 $c(w_i)$ 来模拟预测应该出现的次数，通过这种方法来提高折扣出来的概率值。否则，折扣出来的概率值见式(2-30)：

$$\frac{N_{1+}(w_{i-n+1}^{i-1})}{\sum_{w_i} c(w_{i-n+1}^{i}) + N_{1+}(w_{i-n+1}^{i-1})} \tag{2-30}$$

下面对式(2-28)进行分析，我们已经提到，$c(w_{i-n+1}^{i})$ 越大，应该折扣出较小的值，在式(2-29)和式(2-30)中，$c(w_{i-n+1}^{i})$ 在分母的位置，符合我们的要求；同时，$N_{1+}(w_{i-1}\bullet)$ 越大，应该折扣出较大的值，式中 $N_{1+}(w_{i-1}\bullet)$ 出现在分子的位置上，

也符合我们的要求；最后，如果词 w_i 是个名词词性，我们在分子的位置用 $c(w_i)/N_{1+}(w_{i-n+1}^{i-1})$ 来代替 $N_{1+}(w_{i-n+1}^{i-1})$，进一步提高了折扣出来的概率；可以看出，新的折扣系数符合我们的期望。综上，改进的Katz平滑算法见式(2-31)：

$$p_{Katz}(w_i \mid w_{i-n+1}^{i-1}) = \begin{cases} d_r \cdot p_{ML}(w_i \mid w_{i-n+1}^{i-1}) & if\,(\forall w_i, c(w_{i-n+1}^{i}) \geq 5) \\ \alpha(w_{i-n+2}^{i-1}) \cdot p_{Katz}(w_i \mid w_{i-n+2}^{i-1}) & if\,(c(w_{i-n+1}^{i}) = 0) \end{cases} \tag{2-31}$$

式(2-11)和式(2-31)一起，构成了最后的改进Katz平滑算法。

2.2.4 基于词义相似度的Uni-gram平滑算法

2.2.4.1 传统的Uni-gram平滑算法

如果基于研究的目的，完全可以从训练语料中抽取所有的词以建立词典。这样就不存在Uni-gram模型的稀疏问题了。但是在实际处理大规模真实语料的时候，词典通常是固定的而且通常包含超过10万的词条。这样就使得Uni-gram模型也面临数据稀疏带来的零概率问题。在以上介绍的所有平滑算法中，无论是基于插值还是回退的分配方案，都递归终止于Uni-gram模型，所以Uni-gram模型的零概率问题会反映到高阶N-gram模型中。为解决这个问题，需要对Uni-gram模型进行平滑。一个传统的平滑方法是构造0-gram模型，它是基于词典中所有词的一个均匀分布，用符号 D 代表词典中所有词条的数量。平滑后的Uni-gram模型参数计算见式(2-32)：

$$p(w_i) = \frac{c(w_i) + \lambda}{\sum_{w_i} c(w_i) + D \cdot \lambda} \tag{2-32}$$

当 $\lambda = 1$ 时，就是加一平滑(Add-One Smoothing)，如果词典中词条数量 D 较大的时候，折扣出来的值偏大，所以目前广泛使用的 $\lambda = 0.5$。式(2-32)可以写成式(2-33)定义的插值的形式。

$$p(w_i) = \mu \cdot \frac{c(w_i)}{\sum_{w_i} c(w_i)} + (1-\mu) \cdot \frac{1}{D} \tag{2-33}$$

其中，μ 的定义见式(2-34)：

$$\mu = \frac{\sum_{w_i} c(w_i)}{(\sum_{w_i} c(w_i) + D \cdot \lambda)} \tag{2-34}$$

2.2.4.2 利用HowNet词典计算词义相似度

与均匀分布的0-gram模型进行插值的Uni-gram模型平滑算法并不能反映自然语言的本质现象，本文提出了一种基于词义相似度的Uni-gram模型平滑算法，主要的思想是利用同义词集合对 $c(w_i)=0$ 的词 w_i 进行平滑。建立同义词集合所需的词义相似度计算利用知网(HowNet)词典完成。以HowNet2004版为例，它共包含78074个词，由于一个词可以有不同的义项，所以共有92595个义项，由于词典支持添加自己定义的词汇，所以以上数字大致为一个参考值。在2004版中不仅包含简单的词表，同时提供了一系列的接口函数对整个词典包含的知识进行检索和计算。接口主要分为四类：词典查询类、义原分类查询类、相关度计算类和词义相似度计算类。本文主要利用词典查询类和词义相似度计算类接口。在词典查询类接口中包括函数HowNet_Search_Relation。词义相似度计算类接口中包括函数HowNet_Get_Concept_Similarity(这里省略了函数对应的输入和输出参数，细节可查询HowNet2004帮助文件)。前一个函数返回与目标词词义相似度为1的所有词，后一个函数计算给定两个词的词义相似度。

我们用符号 S_{HowNet} 代表HowNet词典中包含的所有词构成的集合，用符号 S_{Dict} 代表词典中包含的所有词构成的集合，符号 S_{Uni} 代表Uni-gram模型中出现的词构成的集合。集合 $(S_{Dict}-S_{Uni})$ 代表 $c(w_i)=0$ 的那些词，即需要平滑的词。为了能够利用HowNet词典计算集合 $(S_{Dict}-S_{Uni})$ 中的词和集合 S_{Uni} 中的词之间的词义相似度，需要将集合 $(S_{Dict}-S_{Uni})$ 与 S_{HowNet} 取交集。为了论述方面，定义符号 S_0 及 $S_{>0}$。

$$S_0=(S_{Dict}-S_{Uni})\bigcap S_{HowNet} \tag{2-35}$$

$$S_{>0}=S_{Uni}\bigcap S_{HowNet} \tag{2-36}$$

集合 S_0 中词共计20192个，集合 $S_{>0}$ 中词共计43003个。为了能够正确地对集合 S_0 中每一个词进行概率的平滑，需要计算 S_0 中每一个词和集合 $S_{>0}$ 中每一个词的词义相似度。然后按照相似度列表得到相似度最大的若干词。计算词义相似度的函数为HowNet_Get_Concept_Similarity。这个函数的时间复杂度较高，在一台标准配置的P4机器上，将集合 S_0 中的一个词与集合 $S_{>0}$ 中的每一个词利用函数HowNet_Get_Concept_Similarity计算一遍平均需要3分钟，这样，如果循环完集合 S_0，就需要大约6万分钟。

为了解决这个问题，本文并没有利用词义相似度的计算函数，而是利用了

函数HowNet_Search_Relation。这个函数可以按照词义关系和事件关系分别进行查找，在词义关系中可用同义(Synonym)、反义(Antonym)、对义(Converse)和同类(SynClass)等四种关系进行查找。当查找同义词的时候，这个函数只是返回词义相似度为1的那些词。从这个角度上说，HowNet_Search_Relation在计算同义关系时是HowNet_Get_Concept_Similarity的一个特例，但是计算的速度非常快。

计算结果见表2-3：针对集合S_0中的第i个词w_i，可以在集合$S_{>0}$中找到一个子集S_i，S_i中包含$S_{>0}$中与w_i同义的所有词。表2-3中第一行为集合S_0中的6个词，每一个词对应两列，左边一列列举出集合S_i中的5个词，右面的一列显示的是这5个词在训练语料中出现的频度$c(w_i)$。

表2-3 部分同义词列表

Table 2-3 Partial synonyms list

啊哈		阿爹		粗笨		战船		劝架		情素	
哎	7	爸	19	笨重	7	兵舰	1	调处	11	情操	75
哎呀	1	爸爸	79	别扭	4	舰	9	调和	9	情调	19
哎哟	2	慈父	4	拙笨	1	舰船	21	调解	158	情感	205
唉	9	爹	4	拙劣	3	军舰	41	调停	21	情绪	197
兮	13	父	52	笨拙	9	战舰	5	讲情	1	情愫	10

对于S_0中的所有未见词，利用与之同义的$S_{>0}$中词的频度，可以模拟未见词的频度来达到对其平滑的目的。模拟的值可以分别为集合S_i中频度的最大值$Max(S_i)$、最小值$Min(S_i)$和平均值$Averge(S_i)$，例如，在词“啊哈”形成的集合中，最大值$Max(S_i)=13$最小值$Min(S_i)=1$和平均值$Averge(S_i)=6.4$。见式(2-37)：

$$c(w_i)=\begin{cases} Max(S_i) \\ Averge(S_i) \\ Min(S_i) \end{cases} \tag{2-37}$$

与以往的平滑算法不同，我们并没有直接对可见的概率进行减法或线性折扣，而是通过对未见词赋以一个频度来增加分母的值$\sum_{w_i} c(w_i)$来达到概率折扣的目的。在具体的试验中发现，这三种模拟方法都存在将过多的概率折扣出来的现象。平滑算法的一个原则认为，可见的事件一定是可靠的，对可靠的事件不应该减少过多的概率。上面三种折扣方式都折扣出较大的值，所以本文并没有采取以上三种方式，而是直接取一个最小的值1，见式(2-38)：

$$c(w_i)=\begin{cases}1 & \text{if } |S_i|>0\\ 0 & \text{if } |S_i|=0\end{cases} \tag{2-38}$$

对于集合$S_{Dict}-((S_{Dict}-S_{Uni})\bigcap S_{HowNet})$中剩余的$c(w_i)=0$的词，依然采用第2.2.4.1介绍的方法进行平滑。

2.3 长距离触发对的抽取

2.3.1 利用平均互信息抽取词触发对

在自然语言中，许多词对之间由于存在较强的词义关联度(Semantic Association)从而经常以固定搭配关系或上下文同现关系出现在同一个上下文语言环境中。如名词与名词之间的“计算机/软件”、动词与名词之间的“穿/衣服”以及某些固定搭配“越来/越”等。这些高度关联的词对通常跨越很长的上下文距离，承载着长距离的语言依存关系。本章中把它们称作词触发对，如定义2-1所示：

定义2-1：如果一个词单元A与另外一个词单元B明显相关，那么可以说A触发B，$A\rightarrow B$称为触发对，其中A称为触发词单元，B称为目标词单元。

直接计算词义关联度比较困难。一般通过计算词对的距离或频度来模拟计算词义关联度。Smadja通过两个词单元在真实语料中距离分布的方差来获取固定搭配[107]，低方差意味着两个词单元的出现带有一定的规律性，也就是具有一定的关联度，这种方法在计算固定搭配时比较有效，但是在本章中并不适用，原因在于本章中要抽取的触发对在距离的分布上具有很大的随机性。通过统计词对频度来计算词义关联度的方法主要代表为互信息(Mutual Information, MI)，见式(2-39)：

$$MI(A,B)=\frac{p(A,B)}{p(A)\times p(B)} \tag{2-39}$$

两个随机变量的互信息可以解释为知道一个随机变量的取值后对另一个随机变量的不确定性减少的量度，或者一个随机变量包含的另一个随机变量的信息量。互信息是非负、对称的量度，可以用来衡量两个随机变量的依赖程度(或者独立性)。当两个随机变量独立时，它们的互信息刚好为0，如果互信息值很小，可以认为A，B不相关；互信息的取值越大，应该表明两个随机变量的相

关程度越高。但是从式(2-39)中可以看出，互信息的值依赖于分母中两个单个词单元的频度；如果一个词对由两个非常少见的词单元组成，整个公式的分母就会很小，这样词对的互信息就很大，这显然不符合我们对触发对性质的期望。从另外一个角度上说，如果有两个词对，一个词对由两个低频词单元构成，另外一个词对由两个高频词单元组成，虽然第一个搭配互信息大，但我们更倾向于后者，因为后者在处理新语料的时候能提供更多的信息。

一个更好的抽取触发对的量度为平均互信息[32]，见式(2-40)：

$$\begin{aligned} AMI(A,B) = {} & p(A,B)\log\frac{p(B\mid A)}{p(B)} + p(A,\overline{B})\log\frac{p(\overline{B}\mid A)}{p(\overline{B})} \\ & + p(\overline{A},B)\log\frac{p(B\mid \overline{A})}{p(B)} + p(\overline{A},\overline{B})\log\frac{p(\overline{B}\mid \overline{A})}{p(\overline{B})} \end{aligned} \tag{2-40}$$

可以从以下两个方面来解释公式(2-40)：首先，式(2-40)中的第一项蕴含了互信息的公式 $\log(MI(A,B))$，互信息给出了由于知道 B 而使 A 的不确定性减少的量度，但是同时平均互信息更给出了由于知道非 B $(\overline{B})$ 而使 A 的不确定性减少的量度。这样，如果 A，B 本身频度比较少，式(2-40)中第一项比较大，但第二和第三项就比较小。使得平均互信息的总体值偏小。另外一个方面是，平均互信息可以看成是KL距离(Kullback-Leibler Divergence)。KL距离是衡量在同一个事件空间中两个概率分布 $p(x)$ 和 $q(x)$ 差异的一个量度。如果使用数学期望来描述，可定义为公式(2-41)：

$$D(p\,\|\,q) = E_p(\log\frac{p(x)}{q(x)}) \tag{2-41}$$

平均互信息可以看成是两个概率分布 $p(A,B)$ 和 $p(A)\times p(B)$ 之间的KL距离，见式(2-42)：在每个分布上，A，B 分别取0，1两个离散值。比起 $AMI(A,B)$，$MI(A,B)$ 只是度量了两个分布上(1,1)一点上的距离，从这个意义上来说，总体的分布距离要更加可靠一些，所以 $AMI(A,B)$ 更全面、准确地反映了词对之间的关联程度。

$$AMI(A,B) = D(p(A,B)\,\|\,p(A)\times p(B)) \tag{2-42}$$

为计算式(2-40)，需要统计式(2-43)中定义的四个频度信息，其中 $C(A,B)$ 代表词单元 A 和 B 在一定窗口范围内共现的频度。在汉语语言模型的研究中证明：窗口的大小为25时已经可以给出足够的信息[108]。本文在抽取触发对的时候也采

取这种窗口尺寸，公式中的M取12，也就是说，考虑目标词单元B的前12个和后12个触发词A分别构成前向触发对和后向触发对的候选。

$$\begin{aligned} N(A:B) &= C(A,B) = \sum_{d=1}^{M} C(A,B) \\ N(\overline{A}:B) &= \sum_{w\in V} C(w,B) - N(A:B) \\ N(A:\overline{B}) &= \sum_{w\in V} C(A,w) - N(A:B) \\ N(\overline{A}:\overline{B}) &= \sum_{w_1,w_2\in V} C(w_1,w_2) - \sum_{w\in V} C(A,w) - N(\overline{A}:B) \end{aligned} \tag{2-43}$$

我们从1974-1984年10年的新闻类语料——人民日报语料(700M)来抽取词触发对。所有的语料通过笔者开发的INSUN-LEX词法分析系统进行分词(部分分词和词性标注结果见附录A)。利用公式(2-40)和(2-43)来计算$AMI(A,B)$，共建立了包含200万词触发对的库，部分结果如表2-4所示：

表2-4 部分触发对列表和对应的平均互信息的值

Table 2-4 Partial trigger pairs and AMI value

A	*B*	*AMI*	*A*	*B*	*AMI*
一边	一边	1.01E-05	应	邀请	1.44E-05
留下	印象	1.09E-05	提高	素质	1.54E-05
分	秒	1.15E-05	有的	有的	1.79E-05
面积	亩	1.17E-05	包括	在内	1.84E-05
经	批准	1.24E-05	只有	才能	2.12E-05
由	组成	1.24E-05	虽然	但	2.47E-05
取得	成果	1.37E-05	调动	积极性	2.71E-05
无论	还是	1.37E-05	越	越	3.96E-05

2.3.2 用于词法分析的转换触发对

在汉语词法分析领域，有些长距离的约束关系不能单纯依靠词触发对来描述，例如：在汉语分词领域中的一个例句：“只有/在/市场/中/企业/才/能/发展/壮大/。” 针对“才能”这个词来说，“只有”是一个有效的长距离触发对。当“只有”出现在长距离上文中的时候，“才能”更倾向于“Just can”的词义，也就是说，它更倾向于分成两个词“才/能”。而在“施展/才能/的/广阔/的/舞台/。”例句中，当“舞台”出现在长距离下文中的时候，“才能”更倾向于“Ability”的词义，也就是说，它更倾向于组合成一个词“才能”。在汉语词性标注领域。

针对词“为”，在例句“植物学/*n* 上/*f* 应/*v* 称/*v* 它/*r* 为/*v* 复叶/*n* ，/*w*”中是动词词性*v*，在例句“主/*n* 为/*p* 客/*Ng* 服务/*v* ，/*w*”中是介词词性*p*。在第一个例句中，长距离前向词特征“称”可以帮助我们把“为”正确标注为*v*。同理，第二个例句中的词“服务”也是词性*p*的一个明显长距离后向词特征。在音字转换领域，对例句“一/*yi* 枝/*zhi* 美/*mei* 丽/*li* 的/*de* 鲜/*xian* 花/*hua*”，拼音“*zhi*”对应着100多个汉字，使得这个转换任务相对较难。但在句中存在一个长距离的拼音特征“*hua*”可以帮助我们将拼音“*zhi*”转化为正确的汉字“枝”。如果仅仅通过局部的上下文，如“一/*yi*”、“美/*mei*”和“丽/*li*”等，是不能给出正确的转换结果的。这是因为同样的局部上下文，却可以有不同的音字转换结果，例如下面的例句“一/只/美/丽/的/小/猫”。拼音“*zhi*”应该转化为汉字“只”。从这个例子可以看出，长距离约束不仅仅是一个有益的补充，有时更是获得正确结果的必要条件。

在分词问题中，针对“才能”这个词，假设它有两个词性，“才能/0”代表切分，“才能/1”代表不切分，那么可以将其看成为一个与词性标注类似的问题。当利用平均互信息来提取词触发对的时候。平均互信息利用的是频度的信息，并没有依赖于抽象的词义，所以，平均互信息不仅可以提取词触发对，还可以用于两个词单元。以此为基础，本文提出了形如“$w_A \to w_B / t_B$”的前向和后向转换触发对用来描述长距离触发词与目标词对应词性的约束关系。其中，w_A为触发词，w_B为目标词，词性标记t_B为目标词w_B对应的正确词性。这种触发对也可以被看成是一种二元触发对，即w_A和w_B联合触发出正确的词性标记t_B。在上面的例句中，可以根据平均互信息量度抽取出较为明显的前向转换触发对“只有→才能/0”、“称→为/*v*”和后向转换触发对“舞台→才能/1”、“服务→为/*p*”。针对音字转换问题，提出了形如“$y_A \to y_B / c_B$”的前向和后向转换触发对用来描述长距离触发拼音与目标拼音对应的字的约束关系。其中，y_A为触发拼音，y_B为目标拼音，标记c_B为目标拼音对应的汉字。如上例中的后向转换触发对“*hua*→*zhi*/枝”。为论述统一，本章中将上面的“$w_A \to w_B / t_B$”和“$y_A \to y_B / c_B$”两种长距离约束关系统称为转换触发对。

用于抽取转换触发对的文本窗口与抽取词触发对的窗口尺寸一致，也为25个词。抽取分词转换触发对所用的语料与抽取词触发对的一致，也是1974-1984年10年的新闻类语料。与抽取词触发对不同的是，分词转换触发对的目标单元“w_B / t_B”是较固定的，通常为高频的交叉歧义和组合歧义字段，部分结果如表2-5所示：

表2-5 分词转换触发对和对应的平均互信息的值
Table 2-5 Word segmentation conversion trigger pairs and AMI

正向触发对			反向触发对		
A	*B*	*AMI*	*A*	*B*	*AMI*
单位	个人	2.7E-05	个人	贷款	2.4E-05
只有	才/能	1.6E-05	一/起	事件	1.2E-05
在	一起	1.1E-05	一/起	案件	6.0E-06
高	人才	7.0E-06	从/小学	中学	2.0E-06
查处	一/起	2.0E-06	才能	舞台	1.0E-06
只有	人/才	1.0 E-06	人才	科技	1.0E-06

本文从北大计算语言研究所加工的带有词性标志的1998年前5个月人民日报语料库[1]中抽取用于词性标注的转换触发对。转换触发对的抽取包含以下两步：首先，从语料库中找到兼类词性同时出现50次以上的复杂兼类词，例如词“为”，动词词性*v*出现了11284。介词词性*p*出现了12973。共得到2000余个满足条件的复杂兼类词。然后以这些词和对应的兼类词性当成目标词单元，利用2.3.1节中介绍的方法计算前向和后向转换触发对的平均互信息。最后选取AMI>1.0E06的共2.6万个转换触发对，部分结果如表2-6所示。例如：针对单词“中”，如果后面有“关系”一词，则它的词性更倾向于简称词性“中/*j*”。

表2-6 词性标注转换触发对和对应的平均互信息的值
Table 2-6 POS tagging conversion trigger pairs and AMI

前向转换触发对			后向转换触发对		
w_A	w_B/t_B	*AMI*	w_A	w_B/t_B	*AMI*
、	等/*u*	6.75E-4	中	在/*p*	3.56E-4
年	来/*f*	3.55E-4	上	在/*p*	2.8E-4
在	上/*f*	3.09E-4	来	年/*q*	2.32E-4
记者	报道/*v*	2.79E-4	为	以/*p*	2.15E-4
图片	张/*q*	2.66E-4	干部	领导/*n*	1.48E-4
以	为/*v*	2.37E-4	关系	中/*j*	1.01E-4
从	到/*p*	9.8E-6	国	中/*j*	8.3E-5
称	为/*v*	3.0E-6	服务	为/*p*	6.6E-5

抽取用于音字转换的转换触发对也来源于北大计算语言研究所提供的1998

1 本语料由北大计算语言研究所提供，在此表示衷心感谢!

年前5个月人民日报语料库，针对这些语料编写了字音转换的程序对其进行拼音标注。字音转换程序具有一定的识别多音字的能力，部分转换触发对结果如表2-7所示：

表2-7 音字转换触发对和对应的平均互信息的值
Table 2-7 Pinyin-to-character conversion trigger pairs and AMI

前向转换触发对			后向转换触发对		
y_A	y_B / c_B	*AMI*	y_A	y_B / c_B	*AMI*
、	*deng*/等	3.91E-4	*yue*	*hua*/华	2.06E-4
zai	*shang*/上	1.38E-4	*ri*	*hua*/华	1.92E-4
zai	*zhong*/中	1.30E-4	*de*	*de*/的	1.86E-4
xue	*xue*/学	7.7E-5	*zhang*	*fu*/附	1.23E-4
zuo	*gong*/贡	5.0E-5	*shang*	*zai*/在	1.22E-4
jing	*xun*/讯	4.8E-5	*li*	*deng*/邓	4.5E-5
shi	*de*/的	4.3E-5	*wei*	*yi*/以	3.8 E-5
、	*he*/和	3.9E-5	*bao*	*ri*/日	3.5 E-5
yue	*ri*/日	3.8E-5	*le*	*de*/地	3.4 E-5

针对标点符号，我们认为其对应的拼音也是存在的，就用标点符号自身来表示。之所以如此考虑是因为标点符号对正确的音字转换也有一定的提示作用。如转换触发对“、→deng/等”，如果前面出现顿号，则拼音“deng”更倾向于汉字“等”。

2.4 试验结果

2.4.1 改进Katz平滑算法试验结果

平滑算法的性能一般根据它在测试文本上的交叉熵(Cross Entropy)或迷惑度进行度量[109]。给定语言模型 M，同时给定一个测试文本 S，S 由 l 句子序列 $(s_1, s_2, \ldots, s_l)$ 组成，包含的总词数为 N_S。基于以上定义，交叉熵 E 的计算方法见式(2-44)：

$$E = -\frac{1}{N_S}\sum_{i=1}^{l}\log_2 p(s_i) \tag{2-44}$$

在交叉熵的基础上，可以计算出语言模型的迷惑度，见式(2-45)：

$$P = 2^E \tag{2-45}$$

可以看出，交叉熵和迷惑度这两个量度是一致的。作为数据平滑算法的另一种性能评价标准，迷惑度 P 代表在总体平均的情况下，当前状态的下一状态将面临 P 个等概率的选择。显然，语言模型的交叉熵或迷惑度值越小，说明该统计语言模型的性能越好。

试验所用词典包含122664个词，所用的语料来源于1998年前半年人民日报，前5个月为训练语料，测试采用开放测试，测试语料为第6个月语料，包含13万个句子和120万个词。在2.2.2节已经简单介绍了平滑算法的基本分类方式，本章在分配方式上全部采用插值方式，在折扣方式上分别以3种平滑算法为代表，分别为绝对(Abs)平滑算法(绝对折扣方式)、W-B平滑算法(线性折扣方式)和Katz平滑方式(G-T折扣方式)。试验的最后结果见表2-8：

表2-8不同平滑算法的试验结果比较

Table 2-8 Experiment result of different smoothing algorithms

分类		语言模型交叉熵
Bi-gram	Abs平滑	9.63141
	W-B平滑	9.63373
	改进Katz平滑	9.59621
Tri-gram	Abs平滑	9.23823
	W-B平滑	9.33768
	改进Katz平滑	9.17259

从表2-8给出的结果中可以看出，改进的Katz平滑算法无论在Bi-gram模型上还是Tri-gram模型上，其交叉熵量度都要略低于Abs平滑算法和W-B平滑算法。试验结果证明了利用词性信息改进Katz平滑算法是有效的。

2.4.2 改进Uni-gram模型平滑算法试验结果

这一部分采用与上面试验相同的词典，以交叉熵量度进行评测。试验所用的训练语料也来源于1998年前5个月人民日报。测试语料从1974-1984十年的人民日报随机抽取6部分语料进行开放测试。每一部分语料平均包含30万个句子，采用最大正向匹配方法进行分词后，每一部分平均包含300万个词。首先试验了两种不同的Uni-gram模型平滑算法，其中算法1采用与均匀分布进行插值的平滑方法，算法2采用了第2.2.4节介绍的基于词义信息的平滑方法。以此为基础，同时给出了Bi-gram和Tri-gram模型在不同测试语料上的结果，高阶N-gram模型采用Abs平滑算法递归终止于采用不同平滑算法的Uni-gram模型上。

表2-9 两种Uni-gram分布的比较
Table 2-9 Comparsion between two Uni-gram distributions

	Uni-gram		Bi-gram		Tri-gram	
	算法1	算法2	算法1	算法2	算法1	算法2
语料1	11.4798	11.4588	11.247	11.2161	11.205	11.1746
语料2	11.4409	11.4218	11.2476	11.2188	11.2018	11.174
语料3	11.4272	11.4081	11.2187	11.19	11.1686	11.1406
语料4	11.5006	11.4808	11.3662	11.337	11.3353	11.3071
语料5	11.4754	11.456	11.2981	11.2692	11.2559	11.228
语料6	11.4803	11.4611	11.3188	11.2901	11.2808	11.2531

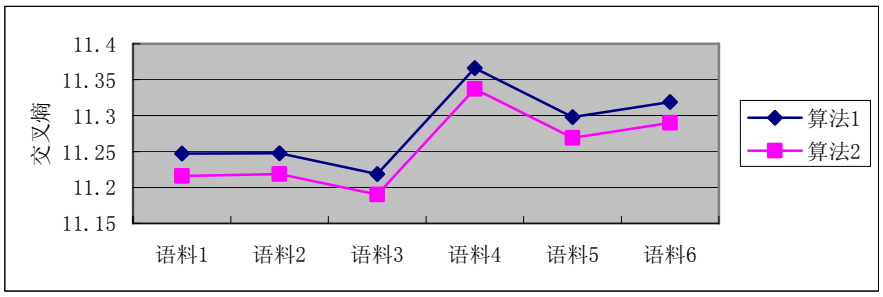

图2-1 Bi-gram模型的交叉熵结果
Figure 2-1 Result of cross entropy of Bi-gram model

从试验结果中可以看出，基于词义的Uni-gram模型平滑算法在不同语料上均减少了模型的交叉熵。图2-1给出了表2-9中Bi-gram模型的图形化表示，可以看出作为一个基本组件，利用词义信息平滑后的Uni-gram模型也可以改善其它平滑算法在高阶N-gram模型上的性能。

与此同时，我们发现交叉熵降低的幅度较少，这主要由于Uni-gram模型在高阶平滑中所占的比例较少。如果能在高阶平滑中引入词义信息，会有效地降低模型的交叉熵，但这就要求更加精确的词义描述，同时还需要准确的词法搭配信息进行约束，如汉语“打/毛衣”是可以的，但如果仅仅依赖于词义将“打/上衣”平滑为较高的概率就是错误的。由此可见，高阶平滑需要更加精准的知识，这也是为什么本文只对Uni-gram模型进行了研究的原因。

2.5 本章小结

本章致力于解决传统N-gram模型的平滑和长距离约束问题。与英语不同，汉语是一种轻结构，重意义的语言，汉语中较自由的语序反映了汉语的这一特点，这样就给汉语的平滑带来了一定的挑战。同时也带来了一定的机遇。本章

首先对已有的平滑算法做了系统地回顾和总结，并分别实现了绝对平滑、W-B平滑和Katz平滑三种平滑算法。针对Katz平滑在处理汉语固定搭配时存在的问题，利用词性信息对概率进行了折扣。在语言模型的交叉熵量度上获得了一定程度的下降。试验的成功使我们相信加入语言学知识会提高平滑算法的质量。本章研究的切入点是利用HowNet词典提供的词义知识，将其应用到语言模型的平滑问题中。具体的方法是利用HowNet提供的词义相似度计算功能重新计算Uni-gram分布以代替传统的与平均分配进行插值的平滑方法。试验证明：新的平滑算法在试验中使交叉熵量度获得了小幅度的下降。另外，以KL距离的角度说明了互信息和平均互信息之间的区别，并利用平均互信息建立了词触发对库。针对汉语词法分析中的问题，提出了形如“$w_A \to w_B / t_B$”和“$y_A \to y_B / c_B$”的转换触发对的概念。从试验的结果上看这种转换触发对具有较明显的关联关系，承载了一定的长距离约束信息。

第3章 基于REA算法的K-best汉语分词模型研究

3.1 引言

自然语言处理的第一个步骤是把输入的文本转化为一系列独立的词，因为词是基本的意义承载单位。分词过程对英文这种以空格作为分隔符的语言来说非常简单，但对汉语来说比较困难，这是因为汉语输入的是一串连续的汉字字符，没有明显的分割标志。虽然汉语分词并不直接面向应用，Gao却指出分词的结果对统计语言模型的质量有非常大的影响[110]。同时，早期的分词错误会通过级联的方式传递到面向最终用户的应用系统中，这在需要语言理解的应用系统如机器翻译中表现得更为明显。由此可见，汉语分词是一个特殊的、对自然语言理解有重要作用的基础性研究。

汉语分词的研究可以算是一个古老的话题，至今已有20多年的研究历史。这个听起来相对简单的任务从实用的角度上看还不能满足人们对自然语言深度理解的要求。目前问题主要集中在以下三个方面：1、分词的规范不同。在Wu的文章中列举了4种不同的汉语分词标准[57]，在SIGHAN国际汉语分词评测中也针对这4种标准给出了四个不同的评测集。不同的分词规范必须要求有不同的基本词表，而基本词表的不同给系统的横向评测带来了一定的困难。例如：在第一届SIGHAN国际汉语分词评测大会上，各个分系统在不同的评测集上取得的名次也很不一致[111]；2、组合歧义和交叉歧义的消解。歧义主要分为真歧义和伪歧义两种，其中伪歧义一般可以通过统计方法来进行识别和消解，但是真歧义往往需要更加精确的词义和语法信息才能有效地进行消解；3、名实体的识别。主要可以分为专有名词、词形派生词和因子词。专有名词主要包括人名、地名和机构名等；词形派生词主要包括重叠词；因子词主要包括时间、日期和数字词等。

针对汉语分词存在的以上问题，面向大规模实际应用，本章建立了基于递归枚举算法的K-best汉语分词模型。首先根据输入的句子建立对应的词网格结构，将分词问题转换为图路径寻优问题。针对汉语分词所建立的词网格的特点，采用递归枚举算法在词网格中进行K-shortest路径寻优，K-shortest路径对应的K-best分词结果可以有效地对句子中的真交叉歧义和组合歧义进行定位。然后，

利用最大熵模型进行了歧义消解的研究。在名实体识别领域，针对中文人名识别问题建立了人名识别多源知识表，并利用多源知识表配合统计方法进行了人名识别的研究；最后利用有限自动机理论进行了因子词识别的研究。整个研究内容如图3-1所示：

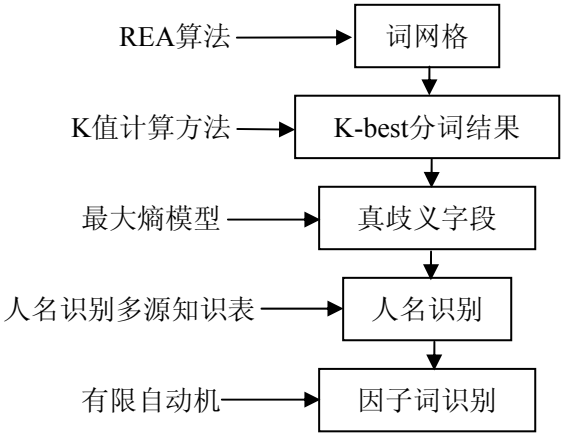

图3-1 分词部分研究内容

Figure 3-1 Research contents of Chinese word segmentation

本章的主要研究内容如下：在3.2节主要介绍了基于递归枚举算法的K-best汉语分词模型；3.3节主要讨论了基于最大熵模型的分词歧义消解的研究。3.4节建立了人名识别多源知识表；3.5节基于有限自动机理论对因子词进行了识别；3.6节给出了最后的试验结果和分析；最后为本章的小结。

3.2 基于K-best分词模型的歧义词发现

3.2.1 词网格的建立

作为汉语分词基本数据结构的词网格最早来源于语音识别，其实质为一个有向无环图(Directed Acyclic Graph)。图中每一个节点对应一个从输入的汉字字符串中抽取出来的词，这些词或者来源于词典，或者来源于为了识别某些名实体而制定的一些规则。针对例句“市场中国有企业才能发展”，图3-2中给出了与其对应的词网格，一共有17个节点。为描述简洁，这里省略了开始和终止节点。

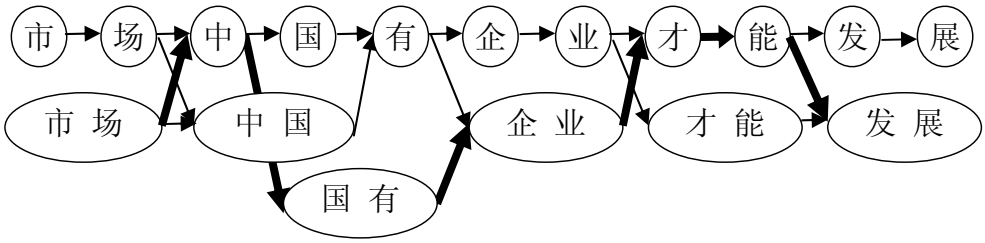

图3-2 例句的词网格

Figure 3-2 Word lattice of an example sentence

在图3-2中，可以发现交叉歧义字段“中国有”以及组合歧义字段“才能”的不同切分方式都出现在词网格中不同的路径上，所以词网格本身是一个自包含(Self-Contained)的网络。从始点到终点的每一条路径分别对应一个最终的分词结果。为了发现用粗箭头线连接的正确的路径，需要计算路径所包含的词串的概率 $p(W)$，见式(3-1)：

$$p(W) = p(w_1, \cdots, w_{i-1}) \cdot \prod_{i=n}^{N} p(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) \tag{3-1}$$

词节点之间的路径权重可分别在Bi-gram或Tri-gram模型中计算。例如，词节点“中”和词节点“国有”之间的路径权重在Bi-gram模型中用条件概率 $p(\text{国有} \mid \text{中})$ 计算。零概率的问题通过第2章介绍的改进Katz平滑算法来解决。为了方便计算和防止计算结果下界溢出，通常采用对数运算将乘法运算转化成加法运算，这样公式(3-1)又可以写成公式(3-2)。

$$p(W) = \log(p(w_1, \cdots, w_{i-1})) + \sum_{i=n}^{N} \log(p(w_i \mid w_{i-n+1}, \cdots, w_{i-1})) \tag{3-2}$$

为有效地进行路径寻优以便获得一个最好的分词结果，通常采用Viterbi算法。Viterbi算法是一种动态规划算法，通过纪录中间计算结果避免重新计算而减少整个算法的时间复杂度。但Viterbi算法只能输出一个分词结果，可以认为它是一个One-best分词方法，这对消解分词的伪歧义有较大的作用，但是却不能有效地消解分词中的真歧义。通常真歧义可以认为是一个词义问题，需要更长的上下文和更复杂的特征，而这些特征没有办法加入到Bi-gram或Tri-gram模型中。本章采用另外一个解决方法，首先输出K-best分词结果，通过K-best分词结果准确地识别出真歧义字段，然后送到下一个特定的消解步骤中去处理，具体的消解方法在第3.3节中给予详细地介绍。

3.2.2 递归枚举算法

K-best分词等同于寻找词网格中K-shortest路径问题。目前图论中有较多的K-shortest路径寻优算法，为了在理论上比较这些算法在汉语分词问题上的性能，首先介绍词网格本身的一些特点，然后简要介绍递归枚举算法的主要步骤。

从图3-1中可以看出，词网格中不包含带有负数权重的路径，也不包含环路。同时还有以下的特点：1、用符号$\overline{d}$代表词网格中每个节点的平均入度。在汉语中，词长通常小于10，就算有大于10的词，也可以分解成长度小于10的几个分离的词而不影响对其正确词义的理解，所以可以认为词网格中$\overline{d} \le 10$。2、汉语里每一个单字都可以认为是一个词。这样，词网格中就包含很多单字形成的节点。这就使得最后的最优路径通常只经过图中个数较少的节点。词网格的以上特性使得REA算法更加适合于在其中进行K-shortest路径寻优。

REA算法首先由Jimenez提出[112]。定义符号如下：符号$G=(V,E)$代表一个有向无环图，式中$V=\{v_1,v_2\cdots v_{|V|}\}$代表节点的集合，$E\subseteq V\times V$代表边的集合。$l(u,v)$表示边$(u,v)$的权重，如果不存在边$(u,v)$，则$l(u,v)=+\infty$。给定节点$v$，它的所有后继节点集合为$\Gamma(v)=\{u\,|\,(v,u)\in E\}$。所有前继节点集合为$\Gamma^{-1}(v)=\{u\,|\,(u,v)\in E\}$。节点$u$和$v$之间的一条路径可以定义为一系列的节点$\pi=\pi_1\cdot\pi_2\cdots\pi_{|\pi|}\in V^+$，其中，$\pi_1=u,\pi_{|\pi|}=v$。路径的权重为$L(\pi)=\sum_{1\le i<|\pi|}l(\pi_i,\pi_{i+1})$，以节点$s$开始并且终止于节点$v$的所有路径可以表示为$C(v)$，并且可以迭代的表示为：$C(v)=\{\pi\cdot v:\pi\in C(u),u\in\Gamma^{-1}(v)\}$。$C(v)$中第$K$个最短的路径定义见式(3-3)：

$$\pi^k(v)=\underset{\pi\in C(v)-C^{k-1}(v)}{\arg\min}\ L(\pi) \tag{3-3}$$

权重定义见式(3-4)：

$$L^k(v)=L(\pi^k(v))=\min_{\pi\in C(v)-C^{k-1}(v)}L(\pi) \tag{3-4}$$

$C^k(v)=\{\pi^1(v),\pi^2(v),\cdots\pi^k(v)\}$代表$C(v)$中$K$个最短的路径。.

基于以上符号，REA算法的主要过程如算法3-1描述：

算法3-1 递归枚举算法：

A1. 对所有的$v\in V$，计算$\pi^1(v)$

A2. 对$k=2,\ldots,K$ 计算 $NextPath(t,k)$ 对$k\ge 2$，当$\pi^1(v),\ldots,\pi^{k-1}(v)$已经

计算完毕后，递归函数 *NextPath*(v,k) 用如下的过程计算 $\pi^k(v)$：

B1. 如果 $k=2$，那么初始化候选路径集合

$C[v] \leftarrow \{\pi^1(u)\cdot v : u \in \Gamma^{-1}(v)\}$

B2. 使 $\pi^{k-1}(v)$ 成为 $\pi^j(u)\cdot v$

(a) 如果 $\pi^{j+1}(u)$ 还没有被计算, 那么调用 *NextPath(u,j+1)*

(b) $C[v] \leftarrow (C[v]-\{\pi^j(u)\cdot v\}) \cup \{\pi^{j+1}(u)\cdot v\}$

B3. 使 $\pi^k(v)$ 成为 $C[v]$ 的最短路径

下面，将REA算法与其他K-shortest路径搜索算法做一个理论上的比较。在所有的K-shortest路径搜索算法中，一个较早提出的算法是Dreyfus算法[113]，其本质上是一个贪婪算法。它在图中计算从 s 到每个节点的最短距离，即使我们只对图中开始节点和终止节点间的最短路径感兴趣，所以它并不适合本章中的汉语分词问题。在REA算法中，候选路径 $C[v]$ 通常用线性数据结构的堆(heap)来实现，所以算法3-1中B2的第二个步骤(b)的时间复杂度为 $O(\log|\Gamma^{-1}(v)|)$，B1步骤为初始化过程，其时间复杂度为 $O(|\Gamma^{-1}(v)|)$，就象以前提到的那样，词网格的 $\bar{d} \le 10$，也就是说 $\Gamma^{-1}(v) \le 10$，这样在REA算法中堆的初始化代价较小。所以REA算法比Matins算法[114]和Eppstein算法[115]要快。在 $\bar{d} \le 10$ 的情况下，该结论已经在相关试验中得到了验证[116]。另外一个在分词领域应用较广的算法是A*算法[117]。但是在A*算法中，当计算完 $\pi^1(t),\cdots,\pi^{k-1}(t)$ 以后，计算 $\pi^k(t)$ 的时间复杂度为 $O(\min(|V|,|\pi^{k-1}(t)|)\cdot(d_{out}+\log(K))$，而REA算法的时间复杂度小于A*算法，为 $O(\min(|V|,|\pi^{k-1}(t)|)\cdot\log(\min(d_{in},K))$。根据REA算法的描述，函数 *NextPath*(t,k) 产生的递归调用次数至多为 $\min(|V|,|\pi^{k-1}(t)|)$。从这个公式可以看出，REA算法非常适合于那些最短路径只包含图中小部分节点 $|\pi^{k-1}(t)|$ 的图路径寻优问题。而这正是上面提到过的词网格的第二个特点。不仅如此，REA算法只依赖于简单的数据结构，并且较容易实现。综上所述，在汉语分词领域，REA算法是较合适的算法之一。

3.2.3 计算*K*值

如果在一个句子中没有任何交叉和组合歧义，我们不需要计算K-best个分词结果，只需要一个最好的分词结果就可以了。计算 K 的主要目的在于对特定的句子能够有效地区分伪歧义字段和真歧义字段，以便将真歧义字段送入下一个

特定的消歧阶段中。计算 K 的主要困难在于歧义字段的数量与句子的长度虽然有一定的关系，却没有明显固定的线性关系。

为了正确地计算 K 值。首先，建立一个包含组合和交叉歧义字段的歧义词库。部分歧义字段如表3-1所示(目前收录的字段还比较有限，需要不断地扩充)。当进行分词时首先对整个句子进行扫描，计算收录在歧义词库中歧义字段的数量，用数量 N_{cvr} 来代表组合歧义字段数量， N_{olp} 代表交叉歧义字段数量。 K 的初始值计算见式(3-5)：

$$K = 2^{N_{cvr}+N_{olp}} \tag{3-5}$$

表3-1 部分真交叉歧义和组合歧义字段

Table 3-1 Partial overlap and cover ambiguious phrases

真交叉歧义字段	中国有	我国有	项目的	和平等	出现在	和平时
真组合歧义字段	有的	以上	大和	中等	不同	所有

在开放的自然语言中，通常没有办法收集到所有的真歧义字段。本章提出了一个利用统计知识逐步试探求解 K 值的算法。主要通过 $(K+1)-best$ 分词结果和 $K-best$ 分词结果之间的差值来作为试探终止的条件。为了详细说明该算法的原理，分别以真歧义字段“中国有”和伪歧义字段“在建设”为例，在Bi-gram模型下计算歧义字段在“市场中国有企业”和“文化在建设中”两个上下文中不同切分的概率值。Bi-gram模型采用1998年人民日报前5个月语料进行训练，并采用改进Katz平滑算法进行平滑，计算的结果见表3-2：

表3-2 真伪歧义字段的计算结果

Table 3-2 Result of true and pseudo ambiguious phrases

真交叉歧义两种切分			伪交叉歧义两种切分		
$A \mid B$	$p(B \mid A)$	$-(\log(p(B \mid A)$	$A \mid B$	$p(B \mid A)$	$-(\log(p(B \mid A)$
市场|中国	0.00024	8.299	文化|在建	1.01E-06	13.803
中国|有	0.0043	5.444	在建|设	1.35E-05	11.209
有|企业	0.0005	7.459	设|中	0.000935	6.974
SUM	**0.0051**	**21.202**	**SUM**	**0.00095**	**31.986**
市场|中	0.00953	4.653	文化|在	0.006798	4.991
中|国有	4.4E-05	10.031	在|建设	0.001399	6.571
国有|企业	0.517	0.658	建设|中	0.031455	3.459
SUM	**0.527**	**15.342**	**SUM**	**0.039652**	**15.021**

为了计算方便，我们将概率值 $p(B|A)$ 取负对数。从表3-2的结果可以看出，伪歧义字段通常只有一种切分结果，所以它的错误切分对应的SUM值偏大，真歧义字段一般都有两种切分结果，所以两种切分的SUM值相差不多。这里用符号 $L^K(S)$ 代表第 $K-best$ 个分词结果的数值，β 代表第 $(K+1)-best$ 分词结果和 $K-best$ 分词结果的差值。如果 β 小于一定的门槛值 Ω，说明 $(K+1)-best$ 分词结果可能包含一个伪歧义。通常 Ω 与训练语料的大小和平滑算法有关，在上例中，$\Omega=12$。综上，计算 K 的具体算法如算法3-2描述：

算法3-2 K 值计算算法：

1) 初始化 K 值 $\begin{cases} K=2^{N_{cvr}+N_{olp}}, if((N_{cvr}+N_{olp})\geq 1) \\ K=1, \quad otherwise \end{cases}$

2) 计算 $L^K(S)$,$L^{K+1}(S)$ 和 $\beta=L^{K+1}(S)-K^K(S)$

3) 如果($\beta<\Omega$) 那么 $K=K+1$;转到步骤2

否则 转到 步骤4;

4) 结束

3.3 基于最大熵模型的分词歧义消解

计算机编译系统可以很快地把程序语言转换为对应的机内二进制码。这是因为程序语言本身没有歧义。但对于自然语言，目前还没有一种能够大规模实用的机器翻译系统，其主要原因在于自然语言中存在各种各样的歧义现象。以分词中的交叉和组合歧义为例进行说明。在网上以“市场中国有企业才能发展”为例句分别试验了两个机器翻译系统，得到以下两个结果，分别为“Market China have business just can develop”和“There is enterprise's ability development in China on the market”。这两个结果都是错的，第一个分词的结果为“市场/中国/有/企业/才/能/发展”，没有正确的切分交叉歧义字段“中国有”，第二个分词的结果为“市场/中国/有/企业/才能/发展”。不仅没有正确识别交叉歧义字段“中国有”，对组合歧义字段“才能”也没有正确切分。最后的结果就是，早期的分词歧义消解错误造成最后的机器翻译系统对整个句子的翻译错误。

本章采用有指导统计学习中的最大熵模型来处理汉语分词中的组合歧义和交叉歧义问题。最大熵在自然语言处理词法分析领域通常定义在 $H\times T$ 上，H 代表所有上下文中特征的集合，T 代表所有可能的标记集合。给定一个特定上下

文h_i，特定标记t_i的条件概率见式(3-6)：

$$p(t_i \mid h_i) = \frac{p(h_i, t_i)}{\sum_{t' \in T} p(h_i, t_i')} \tag{3-6}$$

$$p(h_i, t_i) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i, t_i)} \tag{3-7}$$

在式(3-7)中，$f_j(h_i, t_i)$为一个二值特征函数，代表着第j个特征是否出现在h_i这个上下文中。给定h_i,当$f_j(h_i, t_i) = 1$的时候，代表着这个特征包含有一定的可以预测特定标记t_i的信息。每一个特征对应着一个权重α_j，α_j的值在模型训练过程中通过GIS[31]算法自动获得。GIS算法具体描述见算法3-3：

算法3-3 GIS算法

设定α_j的初值，$j = 1,2,\cdots,m$

for $j = 0$ to m do

保持α_j，寻找满足第j个约束的α_j的最优值α_j^*，

$Ef_j = \sum_x p(x) f_j(x)$ 且 $p(x) = \pi \prod_{j=1}^{l} (\alpha_j)^{f_j(x)}$

$\alpha_j^* = \alpha_j [\frac{\widetilde{E}f_j}{Ef_j}]^{\frac{1}{C}}$，$C$为最大特征数

令$\alpha_j = \alpha_j^*$

end for

GIS算法可以保证其收敛性，但为了减少训练时间过程，通常我们设定一个迭代步数，本章中将步数限制为100。当训练达到特定步数后，基本上可以保证模型精度的需要。

在本章的研究中，分词的歧义可以被认为是一个二值分类问题，所以T包含两个标记，0代表分开，1代表不分开。H的范围为前后各两个词 $h_i \overset{def}{=} \{w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2},\}$。对应的特征模板如表3-3示：其中$X$和$Z$的值在训练的过程中通过训练语料实例化。

表3-3 特征模板
Table 3-3 feature template

$w_{i-2} = X \ \& t_i = Z$
$w_{i-1} = X \ \& t_i = Z$
$w_{i+1} = X \ \& t_i = Z$
$w_{i+2} = X \ \& t_i = Z$

在上面的例句中，针对组合歧义字段“才能”，以前一个词为特征，特征函数定义见式(3-8)：

$$f_j(h_i, t_i) = \begin{cases} 1 & \text{if } w_{i-1} = \text{企业} \ \& \ t_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3-8}$$

在ME模型中可以加入了长距离的上下文特征信息，这种特征可以通过2.3.2节介绍的转换触发对获得，如表2-5所示：可以将转换触发对信息作为特征加入到最大熵模型中去，例如，触发对“只有→才能/0”的特征函数见式(3-9)：

$$f_j(h_i, t_i) = \begin{cases} 1 & \text{if } (w_{\text{target}} = \text{"才能"} \& w_{\text{trigger}} = \text{"只有"} \text{且出现在} h \text{中} \& \ t_i = 0) \\ 0 & \text{otherwise} \end{cases} \tag{3-9}$$

值得一提的是，本章并没有加入 w_i 的特征，这种特征其实就是歧义分布的先验概率，也就是说，在训练语料中，歧义分布的自然状态。本章在试验中构造了一个平衡的测试和训练语料，这样可以重点考察模型本身只利用上下文特征时的消歧能力。当利用这种模型对真实语料进行处理时，可以通过加入 w_i 到 h_i 中而获得先验概率的知识，这样就会获得更好的歧义消解结果。

3.4 基于多源知识表的人名识别研究

名实体的识别可以看成是信息抽取的过程，一个高精度、高召回率的名实体识别系统不仅对汉语分词非常重要，同时对提高机器翻译、问答和文摘等应用系统的性能也具有十分重要的意义。本节重点对中文人名识别进行了研究。

根据国际消息理解大会(MUC)的定义，人名识别是名实体识别中的一个子专题[118]。同时，在汉语名实体识别的研究领域中，人名识别也有其独特的特点：首先，中文人名是个完全开放的集合，不可能用穷举的方法登录到词典中；其次，在真实语料中，中文人名又占有比较大的比重；最后，中文人名识别有比较明显的驱动因素和左边界，那就是中文姓氏。这就给中文人名识别提供了直

接的线索和一定的便利。但是，由于中文不同于英文，中文姓名不具有英语语言中的形态特征(如大小写)作为识别姓名的依据；同时，中文姓名的结构复杂，出现形式多样；最后，中文人名中的姓(名)用字也包含相当一部分常用字，不仅可以自身成词，而且还能与其相邻的字构成词，这样就造成了人名识别的竞争现象[119]，这些现象都增加了中文人名识别的难度。

目前常用的方法是利用统计语言模型配合语料库来进行人名的识别[120]，常用的统计语言模型主要有隐马尔可夫模型[121]、最大熵模型[36]以及支持向量机模型[122]等。这些统计语言模型无一例外地面临着数据稀疏带来的挑战，所以没有办法从小规模和不确定的语料中获得具有一定深度和广度的语言学知识。至今为止，关于中文人名识别所用知识的深入研究相对较少。孙茂松探讨了有关姓(名)用字的概率分布规律和相关的句法知识[123]；李建华初步探讨了姓(名)用字的使用概率和出现概率的问题[124]。众所周知，受限于特定的传统和文化，中文姓(名)用字集中在大约2000个汉字上，基本上可以认为是封闭的。同时，中文人名具有较强的规则性，这些特点决定了基于规则的方法有较大的用武之地。本节以此为出发点，以统计方法为手段，对姓(名)用字进行了统计分析，并结合词性和词义知识，构造一个用于中文人名识别的多源知识表以便提高中文人名识别的精度，更好地解决竞争问题。

3.4.1 姓(名)用字的统计规律

为了统计姓(名)用字相关的统计规律，建立了人名资源语料库。在人名资源语料库中共收集了186583人名，其中17万来源于《姓氏人名用字分析和统计》[125]。另外来自于网络上搜集到的人名。作为人名资源语料库，已经具有较大的代表性。人名资源语料库中共有姓用字761个，名用字1727个。同时，我们利用了一个普通语料库统计了姓(名)用字在真实语料中的分布规律，这个普通语料为北京大学计算语言研究所在网上免费提供的1998年1月人民日报切分、词性标注语料库。其中一共出现了6449个人名，姓用字406个，名用字1375个。针对这两种语料库，分别定义了使用概率PU和出现概率PA，符号的定义如下：人名资源语料库中姓(名)用字 X 的频度定义为 $C(X)$；人名资源语料库中姓(名)用字的总频度分别定义为 $SUMN_{姓}(X)$ 和 $SUMN_{名}(X)$；普通语料库中 X 以姓(名)用字出现的频度定义为 $D(X)$；普通语料库中姓(名)用字 X 出现的的总频度分别定义为 $SUMC_{姓}(X)$ 和 $SUMC_{名}(X)$。

使用概率 $PU(X)$ 和出现概率 $PA(X)$ 的定义见式(3-10)和(3-11)：

$$PU(X)=\frac{C(X)}{SUMN_{姓}(X)(SUMN_{名}(X))} \tag{3-10}$$

$$PA(X)=\frac{D(X)}{SUMC_{姓}(X)(SUMC_{名}(X))} \tag{3-11}$$

在人名资源语料库中，761个姓用字中使用概率前10位的姓用字见表3-4：从中可以看出，前10位的覆盖率达到了43%以上。

表3-4 使用概率前十位的姓用字

Table 3-4 Usage probability first ten surname characters

姓用字	王	李	陈	张	刘	杨	黄	吴	林	周
概率	0.071	0.071	0.069	0.061	0.045	0.026	0.024	0.021	0.019	0.018

在人名资源语料库中，1727个名用字共出现了342759次。使用概率前10位的名用字见表3-5：从中可以看出前10位的覆盖率达到了14.9%以上。

表3-5 使用概率前十位的名用字

Table 3-5 Usage probability first ten firstname characters

名用字	华	英	玉	秀	明	文	珍	芳	国	兰
概率	0.021	0.021	0.018	0.017	0.014	0.013	0.012	0.011	0.011	0.011

对于姓(名)用字的出现概率，根据大小顺序抽取3个组成，分别见表3-6和表3-7：

表3-6 语料库中姓用字的出现概率

Table 3-6 Appearing probability of surname characters in coprpus

姓	$D(X)$	$SUMC_{姓}(X)$	出现概率
邝	1	1	1．0
张	701	1223	0.573100
和	3	10588	0．000183

表3-7 语料库中名用字的出现概率

Table 3-7 Appearing probability of firstname characters in coprpus

名	$D(X)$	$SUMC_{名}(X)$	出现概率
淑	41	41	1
国	455	17901	0.0254
上	5	3781	0．001322

从表3-6和表3-7可以看出，单纯考虑使用概率是不够的，必须要把使用概率和出现概率结合在一起进行考虑才能得出符合汉语规律的人名概率分布。可以这样考虑这两种概率，例如‘张’的使用概率是0.061,出现概率是0.5731。也就是说，当给一个人命名时，需要两个步骤：首先以0.061的可能性从761个姓

用字中取出‘张’，然后再从‘张’的两个分类：姓用‘张’和非姓用‘张’中以0.5731的概率取出姓用‘张’。整个过程符合概率论中的乘法原则。所以在本文中将出现概率和使用概率相乘作为最后的人名识别概率。考虑到运算的方便性，最后采用取对数的方法进行计算，见式(3-12)：

$$p(X)=\log(PU(X)\times PA(X))=\log(PU(X))+\log(PA(X)) \tag{3-12}$$

从实际的结果上看，虽然‘邝’与‘和’字的使用概率差不多，但通过出现概率，可以使一些通用姓‘和’在最后的结果中排名靠后，使一些专用姓如‘邝’在最后的结果中排名提前，这样就更加符合识别中文人名的需要。

3.4.2 姓(名)用字分类的目标

从第3.4.1节可以看出，姓用字的出现概率中隐藏着一个有用的信息，那就是有些姓用字如果出现，那么它一定就是以姓出现的。如：“邝、佘、诸葛”等，一旦在语料中出现这些字，就可以断定它和后面的一个或者两个字构成一个人名。这样我们就可以把姓用字分为专用姓和通用姓两种，专用姓是指那些只用于姓的字，它们对于人名的识别有很强的提示作用。对于通用姓来说，按两个指标进行分类，首先是按频度区分：可以分为常用姓和非常用姓两种。前者如“王，张”等，后者如“和、米、曾”等，如图3-3所示：进行这种分类是由于对于常用姓和非常用姓两种情况，应该采用不同的阈值，对于常用姓可以减少阈值以提高召回率，对于非常用姓，可以增大阈值来提高精度。动态的阈值对于提高整个系统的性能是必要的。另外，人名识别系统中通常存在姓用字竞争的情况：如下例：“我/和/杨/秀/平/一起/来/的/。/”进行人名识别时，由于是从左向右进行扫描，当扫描到“和/杨/秀/”时，就会错误判定这是个人名。这种现象就是因为‘和’同‘杨’产生了姓竞争的情况。为有效处理这种情况，我们把通用姓按照词义分为竞争姓和非竞争姓两种。竞争性如“和、同、由”。

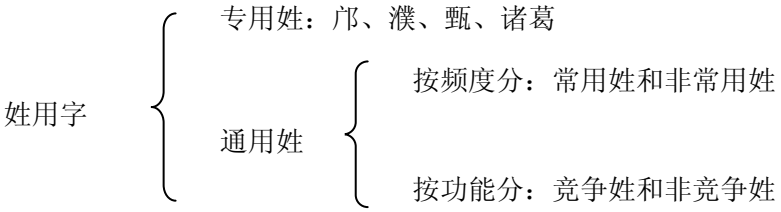

图3-3 姓用字分类
Figure 3-3 Classification of surname characters

对于名用字，也存在与姓用字类似的情况。如图3-4所示：如“喆、弼”等只会以名用字出现。对于通用名用字，不存在常用名用字和非常用名用字的分类。但是如果名用字出现在姓后的第三个位置，则会出现一种名用字竞争的情况，主要体现为单双名的竞争问题。如‘生’字，在下面两个例句中：“汪金生/来/了/吗”，“熊飞/生/了/一个男孩”，分别代表两种竞争的情况。而竞争情况是影响人名正确识别的主要问题。如果能够找到竞争用字，就可以通过词的二元统计、上下文词义消歧等技术来处理这种竞争字。但是这种处理过程一般是比较费时的，所以应该把这种带有竞争性质的名用字找到，然后就可以只对这些竞争字进行附加的处理。而对于非竞争字，又可以分成两种情况，一种情况是如果它们出现在姓后的第三个位置那么基本上可以认定它们是名尾字，如“岭，岩”等，而另外一些字如果出现，基本上就可以认为是非名尾字，如“从，向”等。这样对于非竞争字就可以不需要任何复杂的处理，做到既能保证正确率，又能保证算法的效率。

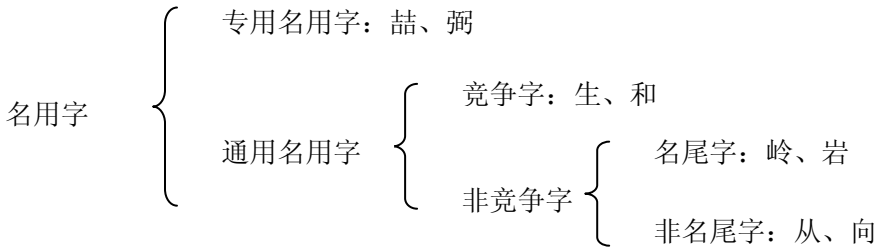

图3-4 名用字分类

Figure 3-4 Classification of fristname characters

3.4.3 姓(名)用字分类的具体方法

3.4.2节中提出的是姓(名)用字的最终分类目标，为达到这个目标还需要一些具体的加工步骤，对于姓用字和名用字的分类都需要四个步骤。

对于姓用字，第一步就是根据姓的出现概率和HowNet词典来把姓用字分为专用姓和通用姓两类，凡是出现概率为1的姓用字都被认为是专用姓。在HowNet中如果定义这个字的词义义项只是用于姓，那么也认为它是专用姓。如果词义义项即可用于姓，又可用于其它方面，如‘于’，其包含7个相应的词义义项，我们就认为它是通用姓。根据这两个指标，可以把姓用字分为专用和通用两类。其中专用姓共有429个。第二步我们主要处理剩余的通用姓，根据出现概率进行分割。凡是出现概率小于0.2的认为是非常用姓。一共找到111个非常用姓。这

样就完成了初步的分类。第三步是竞争姓和非竞争姓的分割，在通用姓中找出那些为连词词性和介词词性的词。如“同、和、由”，因为这些词性的词经常出现在人名前面而造成姓竞争。最后一步就是对初分类的结果进行人工校验。例如：由于语料库和人名库的数据稀疏问题，不可能包含全部的姓，对于那些不被语料库包含但是其语言知识比较明确的，也对其进行人工的分类。如，‘呼延’这个姓，虽然在人名资源语料库和普通语料库中都没有出现，但也应该作为专用姓。最后的分类结果部分实例如表3-8所示：

表3-8 姓用字的分类结果

Table 3-8 Classification result of surname characters

专用姓（429）	咎	俞	罗	戚	瞿	窦	霍	呼延	皇甫
常用姓（221）	曲	高	洪	曾	温	朱	汪	林	白
非常用姓（111）	年	门	于	和	向	爱	要	能	有
竞争姓（6）	和	同	由	要	有	占			
非竞争姓（326）	章	全	战	黄	盖	边	菜	演	钟

对于名用字，第一步就是把名用字出现概率为1的设置为专用名用字。一共有206个专用名用字，第二步要把通用名用字分成竞争字和非竞争字。首先统计语料库中人名后面最常出现有那些词性，也就是词性的Bi-gram统计，由于主要的任务就是找出相关的竞争字，所以主要考率人名/单字词这种统计情况。如“赵/nr　天衡/nr　背/v　着/u　书包/n”。被认为是nr/v共现一次。但是“李/nr 清林/nr　留下/v　500/m　元/q　钱/n”这种情况不算在内。统计出最后的结果如表3-9示：

表3-9 人名词性的二元同现统计结果

Table 3-9 Statistical result of POS name based on Bi-gram

名字后词性	v	p	u	d	c	Vg	M	n	R
总频数（次）	1357	1130	1074	537	406	370	192	94	28
总字数（个）	199	27	6	53	10	10	22	22	8

从表3-9可以看出，人名后动词出现的概率最多，介词第二，助词第三，副词第四。所以这些词性的单字最有可能成为竞争字。如‘为’字是动词，同时它还是名用字，这时，就可以认为‘为’字是个竞争字。这也符合语言的实际情况。如下面的例句：“李德/为/名誉/主编，蒋大为/的/歌声”。在这两句中，‘为’字就是一个竞争字。基于这种现象，可以把名用字首先按词性进行分类。虽然介词出现的总频数很大，但是出现的字数很少，基本集中在下面几个字上：在(533次)、对(138次)、从(58次)、向(53次)等，而动词和副词的字数较多，所以

重点在动词词性和副词词性的名用字中得到竞争字，其它的词性作为补充。这里有一点需要注意的是，人名后边接单字名词的可能性比较小。所以名词词性的单字一般不作为竞争字，同时，除动词和副词词性外的字基本上不可能成为竞争字。初步筛选出703个名用竞争字。第三步就是把一些非竞争字分成名尾字和非名尾字。分类的标准是大部分名词词性字都是名尾字，而一些动词、介词一般都认为是非名尾字。这里需要注意一点的是，分类的过程中词性只是参考指标，同时还要考虑人名的一般命名规律。如：‘在’、‘也’、等字一般很少出现在人名的第三个字。这样就可以把它们归结为非名尾字。但是‘和’字作为人名的第三个字却比较符合人名命名的一般习惯，那么就应该认为‘和’字也是竞争字。本步一共选出8个非名尾字。最后一步和处理姓用字一样，也是对初分类的结果进行人工校验。如‘家’字是一个名词，它可以经常出现在人名后面。这样‘家’字就也是一个竞争名用字。最后的分类结果部分实例如表3-10所示：

表3-10 名用字的分类结果

Table 3-10 Classification result of firstname characters

专用名用字（206）	笙	芷	婵	淑	歆	琪	钊	墉	煊
竞争字（703）	年	门	于	和	向	爱	要	能	有
名尾字（809）	豆	柳	月	柏	页	叔	西	春	北
非名尾字（8）	在	也	从	于	向	传	获	最	

结合使用概率和出现概率，以及上面提到的姓用字和名用字的分类，最终可以形成一个多源知识表。最后一个步骤就是建立一个人工标注的测试集，本文应用包含1979个人名的测试集，通过不断测试对多源知识表进行调整，使其分类趋向合理和稳定。

3.5 基于有限自动机理论的因子词识别

为了对中文因子词进行有效地识别，首先列举出常见的汉语因子词的主要特征和表现形式，如表3-11所示：以中文数字词“四万一千七百八十”为例说明识别的基本理论。识别这个数字词的时候不需要更长距离的上下文，只需要确定当前的状态为一个中文数字，同时保证下一个输入也为一个中文数字，直到下一个输入不为中文数字为止，就可以保证得到正确的识别，如图3-5所示：以上这个识别过程正好符合正规句法的特点。

表3-11 因子词分类、例子和特点

Table 3-11 Classification, examples and characteristic of factoids

类别	语言	例子	特点说明
数字	中文	四万一千七百八十	包含基本中文数字
		百分之三十	数字间包含“分之”
		十四点二六	数字间包含“点”
	英文	13000	包含基本英文数字
		3,999,345　1.34	数字间包含符号
		3/4　56%　1.34E-10	包含各种特殊符号
日期时间	中文	一九九年三月	数字间包含“年、月”
		三点二十分	数字间包含“点、分”
	英文	25/3/1999　25-03-1996	数字间包含符号
		3:20am　1:20pm	数字后面接特殊符号
其他	中文	三比二　第二百个	数字和特殊符号“比、第”
		三块四毛五	数字间包含“块、毛”
	英文	http://www.insun.hit.edu.cn	英文字母和符号构成网址
		IBM　NEC　BBC	英文简写

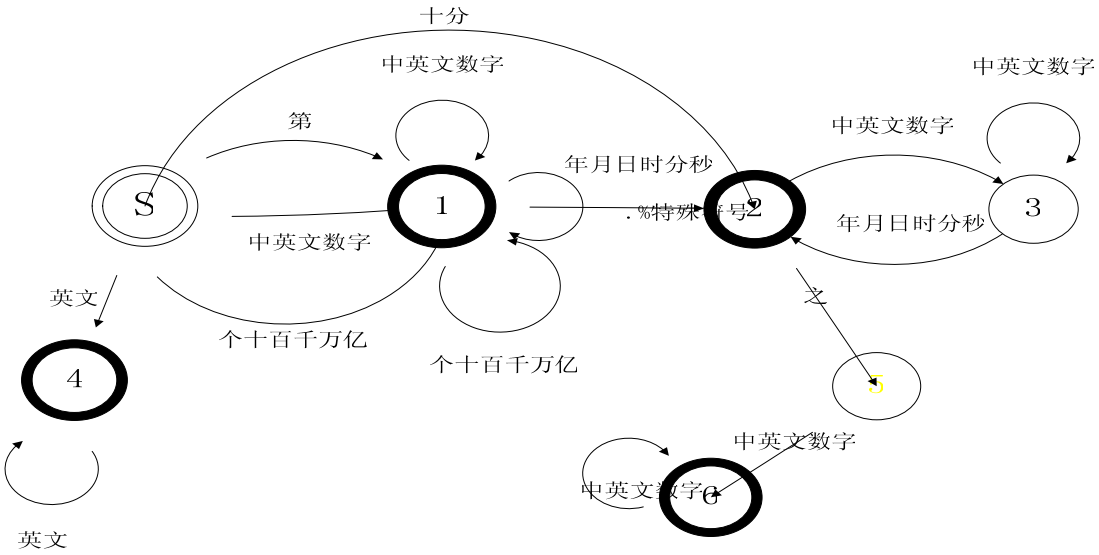

图3-5 状态转移图

Figure 3-5 Figure of states transformation

为了能对因子词进行自动的识别，需要写出识别所用的正规句法。正规句法包括 $A \to \alpha B$ 和 $A \to B$ 两种转换规则。其中符号 A 和 B 代表非终结符，符号 α 代表终结符。正规句法规则通常需要根据所识别的对象进行手工制定，制定这些句法规则的时候，需要注意一点的是必须使句法规则满足没有自嵌套性，否则就变成了上下文无关句法，有限自动机就不能正确地进行识别。为节省篇幅，

表3-12中只给出识别数字和日期因子词的部分句法规则。

表3-12 识别因子词的正规句法规则

Table 3-12 Regular grammer rules for recognition of factoids

基本转换规则	<digit> -> [0 .. 9]; <chinese_digit> -> 零|一|二|三|四|五|六|七|八|九|十|〇; <十百千万> -> 十|百|千|万|亿|百万|千万|百亿|千亿|万亿; <letter_lower> -> [a..z]; <letter_upper> -> [A..Z]; <letter> -> <letter_lower>|<letter_upper>;
数字转换规则	<cn_integer> -> {<chinese_digit><十百千万>*}{<chinese_digit>+}; <en_integer> -> {<digit>+}; <base_integer> -> <en_integer> | <cn_integer>; <integer> ::= <base_integer>; <real> ::= <integer>(.|．|•|点)<integer>; <real> ::= <integer>分之<integer>; <real> ::= <integer>分之<chinesedigit>; <real> ::= <real>(％|%|‰); <real> ::= <integer>(％|%|‰); <integer> ::= <integer><十百千万>; <real> ::= <real><十百千万>; <real> ::= 百分之(<real>|<integer>); <cn_en_integer> -> <integer> | <cn_integer>; <orderinteger> ::= 第<integer>; <integer> ::= 几(十*)<十百千万>; <integer> ::= 两(<十百千万>)(<base_integer>*); <integer> ::= <十百千万>;
日期转换规则	<日> -> <cn_en_integer>日; <月> -> <cn_en_integer>月; <年> -> <digit><integer>年; <年> -> <chinese_digit><cn_integer>年; <date> ::= <年><月><日>; <date> ::= <年><月>; <date> ::= <月><日>; <date> ::= <年>; <date> ::= <月>; <date> ::= <日>; <年代> -> <integer>年代; <年代> -> 上个世纪<integer>年代; <世纪> -> <integer>世纪; <date> ::= <年代>; <date> ::= <世纪>; <date> ::= <世纪><年代>; <date> ::= (公元前|公元|公元后)<integer>年;

3.6 试验结果

3.6.1 分词试验结果

3.6.1.1 基于Viterbi算法的One-best分词试验结果

本章以两个指标来衡量分词的最后结果，分别为精度 P 和召回率 R，见式(3-13)和(3-14)：以这两个指标为基础定义最后的 F 量度，见式(3-15)：

$$P = \frac{\text{正确分词的个数}}{\text{分词系统分出的总词个数}} \tag{3-13}$$

$$R = \frac{\text{正确分词个数}}{\text{正确结果包含的总词个数}} \tag{3-14}$$

$$F = \frac{2 * P * R}{P + R} \tag{3-15}$$

除了以上三个指标以外，汉语分词还有两个重要的指标，那就是交叉歧义和组合歧义。但在具体的评测过程中对组合歧义的评测有一定的难度，这主要来源于两个方面：1、一些名实体：例如人名，时间词等没有收录在词典中，分词的时候没有对名实体作特殊的识别，所以通常会被当成组合歧义来计数；2、分词的不一致问题，训练语料和测试语料中一些伪组合歧义词存在多种切分方式。例如：词“一个”和“一/个”都被认为是对的，以此来评价平滑算法将会引入关于分词规范的一些争论，为避免这些问题，本文只针对交叉歧义进行了评测。

试验所用词典包含122664个词，所用的语料来源于1998年前半年人民日报，前5个月为训练语料，测试采用开放测试，测试语料为第6个月语料，包含13万个句子和120万个词。在第2.4.1节中已经证明了改进Katz平滑算法在交叉上量度上要优于Abs平滑和W-B平滑，这里也利用改进Katz平滑方法处理零概率的问题，试验的最后结果见表3-13：

表3-13不同平滑算法的试验结果比较

Table 3-13 Experiment result of different smoothing algorithms

	P	R	F	交叉歧义
Bi-gram模型	0.9447339	0.960808	0.95270	979
Tri-gram模型	0.9438467	0.9613521	0.95251	954

从表3-13给出的结果中可以看出，One-best分词结果不能有效处理一部分分

词歧义。所以本章对分词歧义作进一步研究，主要包括分词歧义的识别和消解。

3.6.1.2 分词歧义识别和消解试验结果

首先，用前文中的例句计算对应的K-best汉语分词输出结果，结果列于表3-14。然后从人民日报语料中挑选了200个包含歧义的例句，K-best的结果能够识别出98.5%的歧义字段。

表3-14 例句前四个结果

Table 3-14 Four best segmentation results of example sentence

	不同的分词结果	$L^k(S)$
1	市场/中/国有/企业/才/能/发展	36.623
2	市场/中国/有/企业/才/能/发展	42.482
3	市场/中/国有/企业/才能/发展	43.579
4	市场/中国/有/企业/才能/发展	49.439

表3-15给出分词歧义消解的试验结果：包含1个交叉歧义“从小学”和6个组合歧义，它们都是真歧义，存在两种合理的切分方式。为构建这个评测语料，首先从1974-1984十年的人民日报抽取所有包含以下歧义字段的句子，再经过人工检查，给出每一个句子的正确切分。

表3-15 分词歧义消解结果

Table 3-15 Disambiguity result of word segmentation

歧义词	切分方式	切分句子数	训练句子数	测试例句数	正确率	平均正确率
从小学	从/小学	117	90	20	90%	90%
	从小/学	386	90	20	90%	
才能	才能	879	200	20	90%	92.5%
	才/能	7894	200	20	95%	
人才	人才	112	90	20	90%	87.5%
	人/才	324	90	20	85%	
上来	上来	567	200	20	100%	97.5%
	上/来	453	200	20	95%	
个人	个人	5643	200	20	95%	92.5%
	个/人	798	200	20	90%	
总会	总会	383	200	20	95%	97.5%
	总/会	3667	200	20	100%	
一起	一起	160	140	20	95%	87.5%
	一/起	3194	140	20	80%	

表3-15中“切分句子数”代表不同的切分方式在语料库中出现的次数。从不同的切分例句中随机地选取平衡的训练和测试例句。这样做的目的是因为我们想考察模型在没有加入先验分布知识的基础上，只利用特定句子的上下文特征信息对歧义词处理的准确度。如第3.3节所述，先验概率可以很方便地加入到模型中，这样在处理真实语料的时候，系统的性能还会进一步地提高。“平均正确率”给出两类切分正确率的平均值。从试验结果上看，最后的歧义消解结果为92%。我们可以发现，当利用最大熵模型对分词歧义进行消解的时候，左右两个词配合长距离触发对就可以基本给出正确的分类。

3.6.2 人名识别试验结果

建立多源知识表只是系统完成的第一步，第二步是如何使计算机高效地表示和使用这个知识表。以名用字为例，分为三层结构，可以用哈夫曼编码来表示这种结构，如0代表专用，1代表通用，10代表通用名中的竞争名用字，110代表非竞争名用字中的名尾字，111代表非竞争名用字中的非名尾字。这样可以用一个枚举结构enum来表示这四种分类，枚举的内部取值分别为enum Flag{0,2,6,7}分别代表0，10，110，111这四种情况。weight为使用概率和出现概率乘积后再取对数的值。基于以上表示，定义结构如下：

```
Struct NameItem
{
    Enum    flag    //姓(名)用字的分类
    Float    weight //姓(名)用字的权重
}
```

当利用这些资源的时候，希望能够通过一个字，快速的找到相对应的NameItem，同时还要满足对资源不断地增、删、改等要求，对集合进行遍历的操作较少，所以比较理想的结构是哈希(hash)数据结构。哈希结构的关键字采用姓(名)用字。

利用多源知识表分别进行了两个不同的实验。第一个实验采用概率阈值的方法。对于基于统计的人名识别系统，通过建立一个不变的阈值来决定候选的字符串是否是个人名。实验语料为865句包含632个人名的句子。当采用单阈值时，分别以高阈值V_H和低阈值V_L进行两次试验，然后再进行一次对常用姓采用低阈值V_L和对非常用姓采用高阈值V_H的双阈值试验。试验结果如表3-16所示：

可以看出固定阈值的选择有一个矛盾的地方，那就是精度升高时召回率却会下降。而基于多源知识表的双阈值试验却可以避免这个问题，从试验结果可以看出：系统精度提高的同时召回率基本不变，从而取得了最好的F量度结果。

表3-16 单阈值和双阈值试验结果对比

Table 3-16 Comparsion result between single threshold value and double threshold values

	高阈值	低阈值	双阈值
召回率	93.9%	86.1%	93.7%
正确率	84.7%	86.8%	86.8%
F量度	89.3%	86.4%	90.3%

表3-17给出了采用多源知识表以后得到纠正的部分识别错误例句以及纠正的原因，从表3-17中可以看出，多源知识表有效地解决了统计语言模型中的姓(名)用字竞争问题。

表3-17 采用多源知识表后纠正的部分错误

Table 3-17 Partial correction of error after application of multi-source table

错误类型	错误例句	纠正原因
召回错误	#和玉井#/先生/的/声音/相/呼应	解决了姓(名)用字的竞争问题
	#宋健向#/#川达迪#/详细/介绍	
	/#贝格在#/宴会/上/祝酒/时/表示	
精度错误	套/三/#房一#/厅/新/住房/，/	提高非常用姓用字识别阈值
	/#官多#/增加/了	
	/#万余#/斤/粮食/被盗/	

3.7 本章小结

本章主要介绍了基于REA算法的K-best分词模型。在理论上证明了K-best分词模型所用的REA算法在汉语分词领域要优于其它的用于图K-shortest路径寻优的算法。除了汉语分词问题外，很多问题都可以归结为在图中找寻最优路径的问题，如词性标注和音字转换。词性标注是一个多状态图，并且图中状态的数量通常很小，这是因为汉语中句长一般小于100个词。REA算法对这种多状态图也是一种有效的算法。所以，在汉语词法分析中，K-best分词模型的思想可以应用到更多的应用领域中。同时，在实践上也证明了K-best分词模型可以有效的识别分词的歧义，这就为分词歧义消解打下了坚实的基础。在此基础上，本章利用最大熵模型对分词歧义消解进行了研究。利用最大熵模型的优点，加入了长距离和下文信息，大幅度提高了分词歧义消解的正确率。

在中文人名识别领域，姓(名)用字多源知识表配合统计方法使用，取得了良好的结果。这说明统计与规则方法相结合，互为补充，通常会产生更为理想的结果。对于人名和地名识别来说，常用的基于统计的方法也有很多，但笔者认为名实体识别过程中是不存在“银弹”的，即不要希望能够以一种方法解决所有的问题。人名识别需要多种知识源，需要很细致很复杂的分析，单凭一种方法和一种知识源都不能很好地解决这些问题。即使是多知识源，也不存在那种知识源会起到完全的确定作用。所以，对于人名识别来说，建立符合语言规律的多层次，细致的语言资源，并能对这些资源进行良好的应用，是至关重要的。本章中只对姓(名)用字资源进行了布尔分类，一个更为理想的方法是利用模糊集的方法进行模糊分类，这也是本文未来的研究方向。模糊分类需要充足的资源来保证分类充分拟合现实情况，但目前还不具备这种条件。本章首先采取布尔分类这种“硬”分类的方法来提高现有系统的性能，待加工出一定数量的训练语料后，再进行模糊分类方向的研究。

第4章 基于最大熵模型的词性标注研究

4.1 引言

汉语词性标注的研究起步较晚，基本的方法和理论大都借鉴英语的相关研究。与英语词性标注相比，汉语词性标注主要存在以下的困难：首先，汉语单词缺少象英语单词那样对确定未登录词词性有明显提示作用的形态信息，如大写、词尾变化和词缀等信息；其次，汉语的表达更多依赖于词义，而较少依赖于固定的句法规则。这就造成了汉语的词序相对英语自由，汉语的词性与句法成分之间不象英语那样存在简单的一一对应关系。汉语的以上特点给其词性标注的研究带来了较大的挑战。

随着自然语言处理的研究从基于规则的方法向基于统计的方法转变，HMM模型在处理音字转换和词性标注问题时与传统的规则方法相比获得了一定程度的成功。但是HMM模型依然面临两个主要的问题：首先，HMM模型本身利用联合概率来模拟求解一个条件概率的问题；其次，HMM模型只能加入彼此独立的特征。以上两个问题将在4.2节中给予详细地描述。针对HMM模型以上两个问题，本章首先采用了支持向量机和最大熵两种统计语言模型对HMM模型不能正确标注的复杂兼类词进行了研究。同时，采用融合了转换触发对的最大熵模型进行基于句子的汉语词性标注的研究。作为一种统计语言模型，最大熵模型符合目前自然语言处理研究的主流方向；同时它用指数的形式计算条件概率，这更加符合词性标注问题的本质；最为重要的是，它可以利用二值特征函数来融合上下文中不需要满足独立条件约束的各种信息。本文充分利用了这一优点，2.3.2节中介绍的长距离约束特征—转换触发对加入到最大熵模型中，有效地解决了HMM模型不能包含语言中长距离约束的问题。最后本章利用词性标注的研究成果对音字转换开展了相关的探索性工作。

本章的主要内容组织如下：4.2节简单介绍了传统HMM词性标注模型的两个问题；4.3节分别利用支持向量机模型和最大熵模型对复杂兼类词进行了标注；4.4节给出了融合转换触发对的最大熵词性标注模型；4.5节借鉴词性标注的研究成果进行了音字转换的研究；4.6节给出了本章所有的试验结果。4.7节是本章的小结。

4.2 传统HMM词性标注模型的问题

为说明HMM模型在处理词性标注中的问题，本节首先对HMM词性标注模型作简要地介绍。词性标注、音字转换和语音识别这三个问题都可以被认为是信息论中噪声信道的解码问题。首先，一个数据源以概率 $p(T)$ 产生一系列词性标记 T，这一系列分离的词性标记 T 通过一个有噪声的信道，信道以条件概率 $p(W|T)$ 输出一系列分离的词 W。词性标注的目的就是把输出的一系列词 W 解码成原始输入的一系列词性标志 T，这可以通过寻找一个使得后验概率 $p(T|W)$ 最大的 $\hat{T}$ 来解决。由于计算 $p(T|W)$ 比较困难，应用贝叶斯定理交换 T 和 W 的顺序，见式(4-1)：

$$p(T|W)=\frac{p(T,W)}{p(W)}=\frac{p(W|T)p(T)}{p(W)} \tag{4-1}$$

当去掉常量 $p(W)$ 后，$\hat{T}$ 可以用式(4-2)来表示：

$$\hat{T}=\arg\max_{T} p(W|T)p(T) \tag{4-2}$$

式(4-2)中 $p(W|T)$ 可以利用HMM中发射概率计算，$p(T)$ 可以利用HMM中的转移概率计算，如式(4-3)所示：当 n 分别取2和3时即是经常使用的Bi-gram和Tri-gram模型。由于存在数据稀疏问题，Four-gram模型基本上不会被用到。

$$p(T)=p(t_1,\cdots,t_{i-1})\cdot\prod_{i=n}^{N} p(t_i|t_{i-n+1},\cdots,t_{i-1}) \tag{4-3}$$

综合转移概率和发射概率，式(4-2)可以用式(4-4)表示：

$$p(W|T)\cdot p(T)=p(t_1,\cdots,t_{i-1})\cdot\prod_{i=n}^{N} p(w_i|t_i)\cdot p(t_i|t_{i-n+1},\cdots,t_{i-1}) \tag{4-4}$$

整个计算过程较简单，只需将每一个状态的转移概率和状态到观察的发射概率相乘即可。但如果直接对每个路径都采取这种办法计算最后的结果，将会需要 $(2T+1)\times N^{T+1}$ 乘法，其中 T 为状态的数量，N 为每一状态发射出观察的平均数量，这种指数级的复杂度使得我们无法处理实际应用中所面临的任何真实问题。

为解决HMM模型在进行路径寻优时面临的计算复杂度问题，通常采用一种基于动态规划的方法来进行路径寻优的计算。其实质就是通过保存每一步计算

的中间局部结果以避免重新进行计算。这种方法同时被应用在计算语言学中的分词和线性表句法分析器(Chart Parser)等领域。为了能保存每一步的中间结果，需要一个网格形式的数据结构来保存终止于某个特定状态的最优子路径的概率值。这样，全局路径的概率值就可以通过不同阶段的最优子路径的概率值计算得到。这种网格形式的数据结构通常是一个状态转移图，其中包含转移概率和发射概率。下面以汉语词性标注中的一个例句来进行说明，对例句“首/*a*　位/*q*　的/*u*　工作/*vn*”。整个词性标注过程分为三步：第一步是根据输入的词串在词典中查找每个词对应的词性来构建一个词性网格，见图4-1：这个词性网格中的词性节点对应HMM模型中的不同状态。第二步需要计算HMM模型的发射概率和转移概率。通常转移概率通过Bi-gram或Tri-gram计算得到。例如。词性节点“*a*”和“*q*”间的转移概率在Bi-gram模型中通过条件概率 $p(q \mid a)$ 来计算。词性标注中的发射概率需要计算已知词性而发射出不同词的条件概率 $p(W \mid T)$，发射概率通常在带有词性标注的训练语料中训练获得。第三步就是求解词性网格中最优的路径，针对上面的例句最优的路径在图4-1中用黑色加粗的箭头线表示，路径寻优可以看成是HMM模型中已知观察序列而求状态转移序列的经典问题。

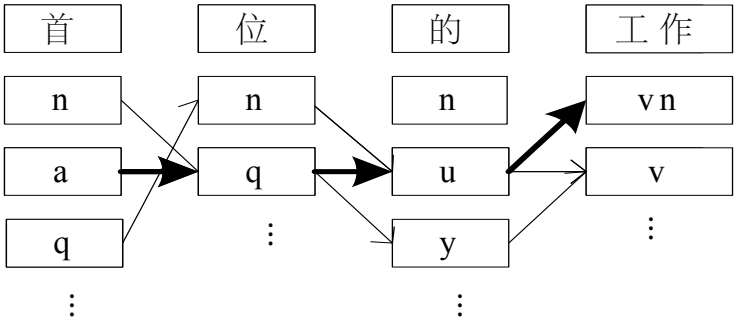

图4-1 用于词性标注的词性网格

Figure 4-1 POS lattice for POS tagging

在HMM模型中，广泛使用的动态规划算法是Vertibi算法，它主要包含三个部分：初始化、正向迭代和反向状态序列求解，具体步骤可参考相关文献[126]。

以上简要论述了HMM模型在汉语词性标注中的理论推导过程和具体的计算步骤。下面给出HMM模型在处理汉语词性标注问题时面临的主要问题。词性标注从本质上说是已知一系列词 W 而求一系列词性 T 的条件概率 $p(T \mid W)$ 问题，但HMM在处理时却用贝叶斯定理转换成式(4-1)中的 $p(T,W)$ 联合概率，也

就是说，HMM用联合概率来模拟求解一个条件概率的问题，这并不符合汉语词性标注和音字转换这两个问题的本质特征；另外，在2.3.2节中已经介绍过词性标注过程中存在长距离的约束词特征，如例句“植物学/*n* 上/*f* 应/*v* 称/*v* 它/*r* 为/*v* 复叶/*n* ，/*w*”中的“称→为/*v*”，但是在传统的HMM模型中却不能加入这些有效的词特征信息。

4.3 复杂兼类词标注

英语的词性标注研究侧重于对未登录词的识别，这是因为未登录词一般有比较明显的词法特征，如大小写，后缀等。而汉语的未登录词识别通常认为是分词中名实体识别中的一个任务；同时，由于汉语是一种重词义而轻语法的语言，这就造成了汉语中存在大量的词性兼类的现象。本章所指的复杂兼类词主要指满足以下三个条件的词：首先，它的出现频度很高；其次，它有两个以上的词性；最后，每个词性的出现频度也比较高。以北京大学计算语言研究所提供的1998年上半年人民日报标注语料为例进行说明，抽取了第6个月部分语料共106604个句子。表4-1给出了部分复杂兼类词及其对应的词性和频度信息。

表4-1 部分复杂兼类词词性与频度信息

Table 4-1 Partial complex POS and frequent informations

词	词性1（频度）	词性2（频度）	词	词性1（频度）	词性2（频度）
中	j(864)	f(2615)	来	f(421)	v(698)
为	p(2108)	v(1400)	改革	v(445)	vn(638)
与	c(1041)	p(1569)	到	p(486)	v(1725)
发展	v(1222)	vn(1668)	在	d(229)	v(237)
组织	n(472)	v(378)	建设	v(374)	vn(1484)

据统计，汉语句法分析有80%以上的错误与兼类处理错误有关[127]。例如“安装”一词，它有两个词性，分别为*v*和*vn*。给出以下两个短语“安装/*v* 电灯/*n*”和“安装/*vn* 工程师/*n*”，只有正确地标注“安装”的词性，才能保证这两个短语根据规则“*vp→v+n*”和“*np→vn+n*”被正确的标注为动词短语*vp*和名词短语*np*。这里可以看出，正确地标注兼类词性可以为下一步的语言理解打下坚实的基础。

我们用传统的三阶HMM为例对上面抽取出的第6个月部分语料进行了标注试验，以前5个月为训练语料。采用改进Katz平滑算法处理统计模型的零概率问题。表4-2给出了部分复杂兼类词的最后标注结果。

表4-2 基于HMM模型的复杂兼类词标注结果
Table 4-2 Tagging result of complex POS based on HMM

词	正确词性	错误标注	数量	词	正确词性	错误标注	数量
中	j	f	842	组织	n	v	381
为	v	p	714	来	v	f	285
为	p	v	478	改革	v	vn	243
与	p	c	452	到	p	v	242
发展	v	vn	440	在	d	p	227

从表4-2中可以看出HMM模型不能有效地处理这类复杂词性兼类词。这是因为HMM模型性能依赖于发射概率和状态转移概率，状态转移概率主要通过前一个或两个词性进行计算，词性的约束能力较弱，不能给出有效的消歧信息。当复杂兼类词的发射概率近似相等的时候，通常就会产生错误的标注结果。另外从表4-2中也能发现词性标注错误分布的不均匀性。语言本身满足zip'f定律的约束，使得词性标注错误分布也很不均匀。最后结果中共有9379个词被错误的标注，前100个词占全部错误的28.80%。前500个占到全部错误的51.76。所以给我们一个提示，那就是可以重点解决某些复杂兼类词的标注错误，虽然它们个数较少，但使用频度很高，正确的标注它们可以有效地减少整体的标注错误。

目前对兼类词词性标注的研究主要有人工神经网络[128]的方法。本文主要利用最大熵模型和支持向量机模型对其进行了研究。在图4-2中给出了面向对象设计的两种语言模型的UML图，图中只给出了核心的对象以及每一个对象的最重要的数据结构和函数。这种设计方法可以利用面向对象设计的封装、继承、多态三个特点，增加代码的复用能力。例如：无论是最大熵模型还是支持向量机模型，都需要对输入的句子按一定的窗口大小进行扫描，同时将扫描后的结果转换为特征的形式，这样就可以引入CWindow和CFeature两个类。同时，两个模型都需要对特征进行过滤，这样可以把过滤特征这种公共操作放到基类CFeatureLib中，而对于不同的问题，特征模版和收集特征的方法都是不一样的，这样就可以从CFeatureLib中派生出针对不同问题的子类，并重载其中的虚函数ExtractFeatureFromWindow(CWindow & cwindow)。接口CPredicate有两个子类，当分类的类别较多时，如词性标注问题，一个特征可以对应数十个词性，那么就选用CPredicateMap类，其中数据结构hash_map<string, map<int, double> >的含义如下：string代表特征CFeature中type加上data联合构成的一个值，map<int,double>中的int为CFeature中的tag，double中保存的是这个特征对应的权重。如果分类的类别较少，如分词歧义消解，只有两类，那么可以使用

CPredicateVector类。可以这样认为CPredicate是CFeature的一种索引方式，因为在标注的过程中，CFeature中的tag是未知的，利用CPredicate这种索引方式就可以找到CFeature中type加上data联合构成的一个值对应的所有tag以及对应的权重了。

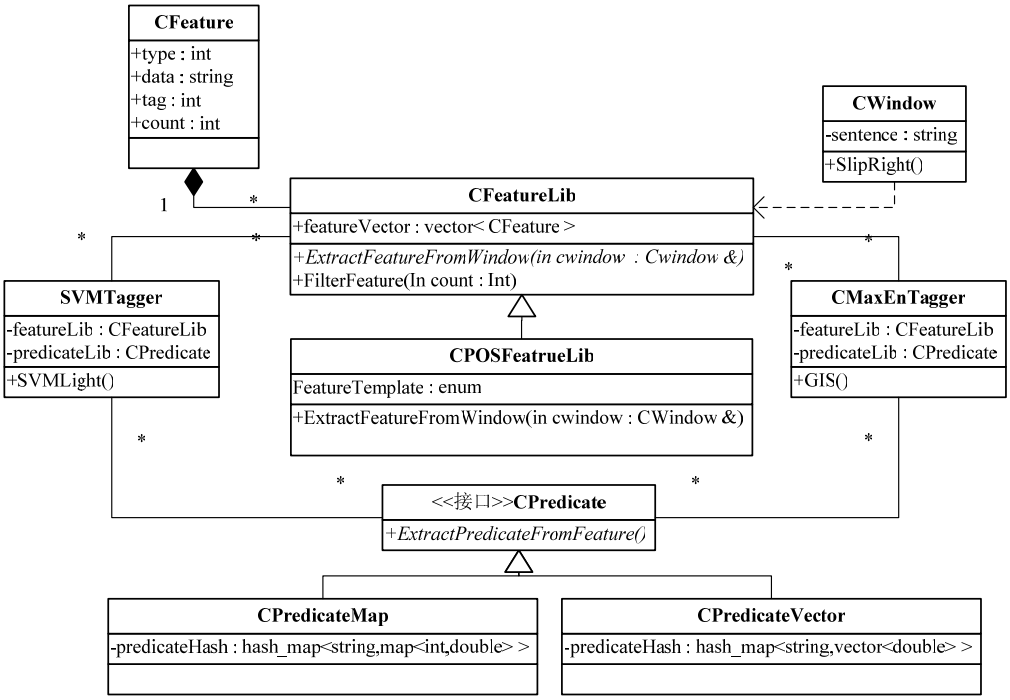

图4-2 最大熵模型和支持向量机模型的UML图
Figure 4-2 UML of Maximum Entropy and Support Vector Machine

基于以上的定义，下面简要给出支持向量机模型的具体应用过程，给定两个例句：“首位/的/工作/*vn*”和“应该/努力/工作/*v*”。以标注词“工作”的词性为例，如果取前两个词为特征。整个过程如图4-3所示：图中w-2代表待标注词的前面第二个词。可以通过CFeatureLib子类中的featureTemplate这个枚举值enum将其转换为CFeature中的type,“首位”代表CFeature中的data。第二步将特征映射到特征的编号，在支持向量机模型中代表空间的维。针对多词性问题，本章采用One-versus-rest的方法，对每一个词性训练一个二值分类器，最后集合所有的分类器进行最后的分类。最后就是将第二步的结果利用Joachims开发的SVMlight程序[129]进行训练。

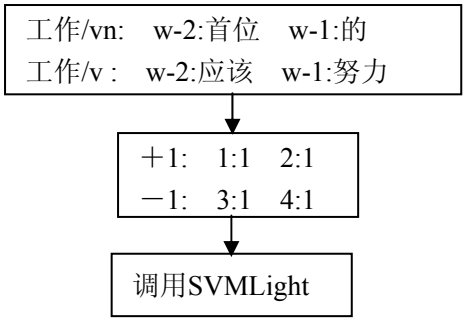

图4-3 基于支持向量机的词性标注训练过程

Figure 4-3 Training process of POS tagging based on SVM

复杂词性兼类词的标注试验采用语料来源于1998年前半年人民日报，从前5个月语料中抽取包含每一个复杂兼类词的窗口大小为5的文本片断进行训练，测试采用与HMM试验相同的测试语料。针对每一个复杂兼类词分别构建最大熵模型和支持向量机模型进行训练和标注，选取的特征分别为前两个词、后两个词以及前两个词性。结果如表4-3所示：图4-4中给出了两种模型的图形化结果。

表4-3 基于最大熵模型和支持向量机模型的复杂兼类词标注结果

Table 4-3 Tagging result of complex POS of words based on Maximum Entorpy and SVM

词	正确词性	错误词性	最大熵结果	支持向量机结果	词	正确词性	错误标注	最大熵结果	支持向量机结果
中	j	f	16	19	组织	n	v	36	52
为	v	p	173	148	来	v	f	8	5
为	p	v	129	146	改革	v	vn	58	49
与	p	c	248	173	到	p	v	172	195
发展	v	vn	97	92	在	d	p	132	186

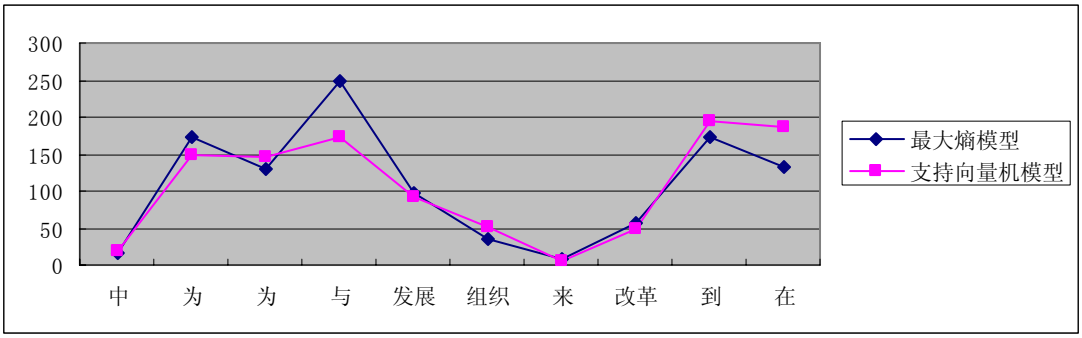

图4-4 复杂兼类词标注结果图形显示

Figure 4-4 Illustration of tagging result of complex POS of words

表4-3中的结果可以发现与HMM相比，无论是最大熵模型还是支持向量机模型都大幅度地减少了复杂兼类词的错误。图4-4中给出两种模型标注结果的图形化显示，可以看出最大熵模型和支持向量机模型的结果大致相同。

与支持向量机模型相比，最大熵模型可以直接处理多分类问题，所以下面利用最大熵模型研究针对句子的词性标注过程。与复杂兼类词的标注不同，对一个句子进行词性标注是一个序列分类模型。

4.4 融合转换触发对的最大熵语言词性标注模型

最大熵模型可以集成很多异构信息，这些信息被隐含地描述成特征函数，而且特征函数彼此没有独立性假设，并允许人们使用关于何种信息是较重要的先验知识。最大熵模型可以使我们不对未知的情况作任何假设，尽量保持未知的原始状态。这符合哲学对于世界的认识，即如果没有任何约束，那么事物总是向熵最大的方向发展。由于最大熵模型可以包含很多异构的信息，所以，如果将HMM模型中的观察和前一个状态看成是两个不同的特征，那么可以将两个分离的转移概率 $p(s_t \mid s_{t-1})$ 和发射概率 $p(o_t \mid s_t)$ 最终以一个条件概率 $p(s_t \mid s_{t-1}, o_t)$ 表示，如图4-5所示：HMM模型中观察只是依赖于当前的状态，而在最大熵模型中，观察除了依赖于当前状态外，还依赖于以前的状态。通过最大熵模型就可以构造出条件概率模型，这符合词性标注问题的本质特征。

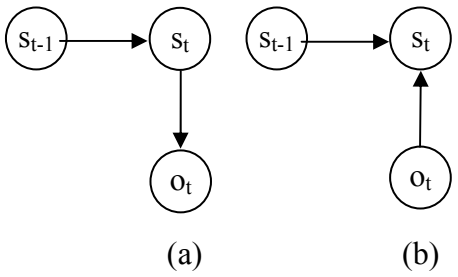

图4-5 HMM模型的依赖关系(a)和ME依赖关系(b)

Figure 4-5 The dependency graph of HMM(a) and ME(b)

最大熵模型在词法分析领域通常定义在 $H \times T$ 上，H 代表上下文中所有特征的集合，T 代表所有可能的标记集合。给定一个上下文环境 h_i 环境，标记 t_i 的条件概率计算与处理分词歧义时类似，本文在第3.3节中给予了介绍，这里不再详述。本章重点研究了两个问题：其一为特征选择问题，在词性标注过程中，

上下文环境h_i中包含较多的特征，除了词特征以外，还包含词性特征。这样我们就面临特征选择的问题，即上下文环境h_i中那些特征对正确地标注词性比较重要。过多的特征通常不仅增加了系统的时间和空间复杂度，同时包含的噪声会降低系统的正确率。第二个问题是HMM可以对一个随机过程进行模拟，但是最大熵模型从本质上说是一个统计分类模型。为了能够处理词性标注和音字转换这样的序列分类问题，需要能够对每一步分类的结果进行路径搜索。下面在4.4.1节介绍了局部特征和长距离特征的抽取方法；4.4.2节给出了应用于序列分类的路径寻优Beam Search算法。

4.4.1 特征选择

最大熵模型训练过程中，训练语料格式如下：“植物学/*n* 上/*f* 应/*v* 称/*v* 它/*r* 为/*v* 复叶/*n* ，/*w*”。特征信息从训练语料每一个事件$(h_i, t_i / w_i)$中产生。定义局部特征局限在窗口为5的上下文环境中，也就是说当前词的前两个和后两个词的范围内。从例句中可以看出，局部上下文中可用的特征包括前两个词和词性以及后两个词(标注的过程采用从前向后进行标注)，从描述能力上看，词的描述能力要远远大于词性的表述能力。为了避免训练语料中存在的标注不一致现象，采用简单的特征模板，这也可以同时避免复杂特征模板带来的数据稀疏问题。变量X，Z的取值从训练语料中自动地获得，特征模板如表4-4所示：

表4-4 词性标注特征模板

Table 4-4 POS tagging feature template

$t_{i-2} = X \ \& \ t_i = Z$
$t_{i-1} = X \ \& \ t_i = Z$
$w_i = X \ \& \ t_i = Z$
$t_{i+1} = X \ \& \ t_i = Z$
$t_{i+2} = X \ \& \ t_i = Z$
$w_{i-1} = X \ \& \ t_i = Z$
$w_{i-2} = X \ \& \ t_i = Z$

以上面例句中的词“为”为例，分别以t_{i-1}和w_{i-1}为例产生以下两个特征函数，见式(4-5)和(4-6)：

$$f(h_i, t_i) = \begin{cases} 1 & \text{if } ((t_{i-1} = r) \ \& \ (t_i = v) \ \& \ (w_i = \text{"为"})) \\ 0 & \text{otherwise} \end{cases} \tag{4-5}$$

$$f(h_i,t_i)=\begin{cases}1 & \text{if } ((w_{i-1}=\text{"它"})\,\&\,(t_i=v)\,\&\,(w_i=\text{"为"}))\\ 0 & \text{otherwise}\end{cases} \tag{4-6}$$

利用最大熵模型的优点，2.3.2节介绍的转换触发对也可以当成一种特征加入到模型中去，特征函数见式(4-7)：

$$f(h_i,t_i)=\begin{cases}1 & \text{if } (w_{target}=\text{"为"}\&w_{\text{trigger}}=\text{"称"}\text{且出现在}h_i\text{中}\&t_{target}=v)\\ 0 & \text{otherwise}\end{cases} \tag{4-7}$$

由于自然语言受Zip'f定律约束，所以自然语言处理领域中的统计语言模型都会面临数据稀疏的问题，最大熵模型也不例外。在上面的例句中，单词“统称”、“简称”和“称呼”与单词“称”具有相同的词义，所以对正确标注词性“为/*v*”有相同的贡献。但数据稀疏问题使得我们不能在训练语料中获得全部这些转换触发对。解决数据稀疏问题的一个有效手段是词语聚类，目前一共有三种方法来进行词语聚类。第一种方法是利用1.4.1节介绍的同义词词典；第二种方法主要是采用可计算的词义词典，如HowNet，本文在第2.2.4节中进行了较为详细地介绍；第三种方法是5.2节介绍的矢量空间模型进行自动词语聚类。纵观这三种方法，每一种方法都有其鲜明的特点。第一种方法只局限于同义词的范围，优点是精度较高，缺点是收录的词典数目较少。第二种方法可以得到包含同义、反义和对义等多种关系的词语集合，任意给出一对词，可以计算出它们介于0和1之间的词义相似度；第三种方法自动化程度较高，聚类后的词不仅局限于词义相同的词，也包括词义相关的词。但缺点是需要较大的训练语料才能保证结果的有效性。本文综合这三种方法，进行必要的综合和优化，配合必要的人工检查，生成最后的词语聚类$S=\{w_1,w_2,w_3,\ldots,w_n\}$，这个结果可以用来建立聚类转换触发对。假设S为词“称”的词语聚类集，聚类转换触发对特征函数如式(4-8)定义：

$$f(h_i,t_i)=\begin{cases}1 & \text{if } (w_{\text{target}}=\text{"为"}\&\exists w_{trigger},w_{trigger}\in S\text{且出现在}h_i\text{中}\&\,t_{\text{target}}=v)\\ 0 & \text{otherwise}\end{cases} \tag{4-8}$$

4.4.2 序列分类的Beam Search搜索算法

给定一个句子，包含n个词，分别为$\{w_1,\ldots,w_n\}$，一个对应的词性标记序列$\{t_1,\ldots,t_n\}$的条件概率见式(4-9)：

$$p(t_1 \ldots t_n \mid w_1 \ldots w_n) = \prod_{i=1}^{n} p(t_i \mid h_i) \tag{4-9}$$

其中，h_i 实第 i 个词 w_i 所对应的上下文环境。

从式(4-9)可以看出，基于句子的词性标注是一个序列分类问题，需要枚举出对应句子的所有词性标注序列的候选，并且输出概率值最大的一个词性序列作为答案。常见的搜索算法主要有Vertibi算法，另外就是Beam Search算法。本章主要选用Beam Search算法。Beam Search算法其实质是一个宽度优先搜索(Breadth First Search)；为了避免搜索过程中的组合爆炸问题，对每一步后续的所有候选中，只对前 K 个最优的候选进行扩展。其它的通过剪枝处理掉。类似于一个人在走夜路，用一束光照亮前面的路，然后只沿着光照到的地方向下走，其他的地方不进行尝试。这也是为什么这种算法叫做Beam Search的原因。

我们用符号 s 代表词性的标注序列，包含 $|s|$ 个词性标记 $\{t_1,\ldots,t_{|s|}\}$。算法的主要过程包含三个函数，它们分别为 *advance*，输入的是词性的标注序列 s，输出 m 个新的词性序列 $s_1 \ldots s_m$，其中每一个序列的长度都为 $|s|+1$。第二个函数为 *insert*，主要的功能就是将词性序列 s 插入到一个堆 *hp* 中，堆 *hp* 为一个先入先出的线性数据结构，主要配合进行图的先宽搜索。第三个函数为 *extract*，输入的是堆 *hp*，输出堆 *hp* 中具有最高分数的词性标注序列 s 并从堆中去掉这个 s。Beam Search算法如算法4-1表述：

算法4-1 Beam Search 算法

输入：包含 n 个词的句子。

输出：n 个词对应的 n 个词性标志，保存在堆 *hp* 中。

n 为句子的长度

$K = 10$

for $i = 0$ *to* $n-1$

  $sz = \min(K, |hp_i|)$

  for $j = 1$ *to* sz

    $s_1 \ldots s_m = advance(extract(hp_i))$

    for $p = 1$ *to* m

      $insert(sp, hp_{i+1})$

return $extract(hp_n)$

下面分析这个算法的时间复杂度，我们需要向堆 hp 中插入 KT 个句子，每一次插入需要耗时 $O(\log KT)$，其中 T 为词性标记集的数量，K 通常取10，K 过大对提高系统性能提升不大，但是却增加了搜索过程的时间复杂度。整个的时间复杂度为 $O(nKT\log(KT))$。

4.5 音字转换的研究

与基于词形的汉字输入法相比，基于拼音的输入法以其记忆简单和易于掌握等优点被97%的计算机汉语用户广泛使用。但是汉语中存在很多的同音字，如拼音“*zhi*”对应着100余个汉字。平均来说，6700余个汉字只对应着410个没有音阶的拼音，众多的同音字给音字转换带来了较大的困难。音字转换的方法主要可以分为基于规则和基于统计两种。在基于规则的方法中，粗糙集理论被用来从语言信息表中抽取出用于音字转换的规则[130]。基于规则的方法可以用较简单和较直接的方式自动地获得语言学知识。在统计方法的研究中，广泛使用的统计语言模型是HMM[7]。

音字转换也可以看成是噪声信道解码问题的特定应用。用于解码的统计语言模型和算法与词性标注的一致，所不同的是词性标注和音字转换各自对应的输入和输出。词性标注的输入为一系列独立的词组成的词串 W，输出的是词串 W 中每一个词所对应的正确的词法分类标记——词性串 T；音字转换的输入为一系列分离的拼音组成的拼音串 Y，输出的是拼音串 Y 中每一个拼音所对应的正确的汉字——汉字串 C。音字转换的发射概率为已知字而发射出不同拼音的条件概率 $p(Y|C)$。如果不考虑汉语中多音字的现象，那么发射概率 $p(Y|C)=1$。所以在音字转换中只需考虑转移概率即可。这里可以看出，音字转换可以看成是词性标注的一个特例来处理。所以上面论述的用于词性标注的最大熵模型方法完全可以用于音字转换问题。

在音字转换的模型训练过程中，训练语料格式如下：“*yi*/一 *zhi*/枝 *mei*/美 *li*/丽 *de*/的 *xian*/鲜 *hua*/花”。特征信息从训练语料中每一个事件 $(h_i, y_i/c_i)$ 中产生。与词性标注不同的是，上下文环境 h_i 中主要包含拼音特征 y_{i-2}^{i+2}，而词性标注中主要特征为 w_{i-2}^{i+2}，词的描述和约束能力要比拼音强。所以在音字转换任务中，本文定义了两个不同的模板，分别是表4-5定义的简单模板和表4-6定义的复杂模板。变量 X，Y，Z 的取值从训练语料中自动地获得。

表4-5 简单特征模板
Table 4-5 Simple feature template

$y_{i-2}=X\ \&\ c_i=Z$
$y_{i-1}=X\ \&\ c_i=Z$
$y_i=X\ \&\ c_i=Z$
$y_{i+1}=X\ \&\ c_i=Z$
$y_{i+2}=X\ \&\ c_i=Z$
$c_{i-1}=X\ \&\ c_i=Z$
$c_{i-2}=X\ \&\ c_i=Z$

表4-6 复杂特征模板
Table 4-6 Complex feature template

$y_{i-1}=X\ \&\ c_i=Z$
$y_{i-2}y_{i-1}=XY\ \&\ c_i=Z$
$y_i=X\ \&\ c_i=Z$
$c_{i-1}=X\ \&\ c_i=Z$
$c_{i-2}c_{i-1}=XY\ \&\ c_i=Z$

在音字转换问题中，对于拼音“*mei*”到字“美”的转换，分别用待转换拼音前面的字作为特征，简单特征模板产生式(4-10)定义的特征函数，复杂特征模板产生式(4-11)定义的特征函数。

$$f(h_i,c_i)=\begin{cases}1 & \text{if } ((c_{i-1}=\text{"枝"})\ \&\ (c_i=\text{"美"})\ \&\ (y_i=\text{"}mei\text{"})) \\ 0 & \text{otherwise}\end{cases} \tag{4-10}$$

$$f(h_i,c_i)=\begin{cases}1 & \text{if } ((c_{i-2}=\text{"一"})\ \&\ (c_{i-1}=\text{"枝"})\ \&\ (c_i=\text{"美"})\ \&\ (y_i=\text{"}mei\text{"})) \\ 0 & \text{otherwise}\end{cases} \tag{4-11}$$

与词性标注同理，也可以加入2.3.2节中得到的转换触发对特征，见式(4-12)：

$$f(h_i,c_i)=\begin{cases}1 & \text{if } (y_{target}=\text{"}zhi\text{"}\ \&\ y_{\text{trigger}}=\text{"}hua\text{"}\ \text{且出现在}\ h_i\ \text{中}\ \&\ c_{target}=\text{"枝"}) \\ 0 & \text{otherwise}\end{cases} \tag{4-12}$$

4.6 试验结果

4.6.1 词性标注试验结果

词性标注试验需要特定的词性标注集。汉语语言学界认为，划分词性的依据有三种：分别为形态标准、意义标准和分布标准。对于汉语来说，形态标准和意义标准都是行不通的，所以只能根据词在句法结构里所担当的句法功能，也就是分布标准进行分类。同时，也要使标注集满足三个要求：分别为标准性、兼容性和扩展性。标准性是指尽量采纳当前应用较广的词性标准或正在成为词性标准的分类体系和标记符号，没必要从头做起。兼容性是指尽量使标注集的

表示与已经存在的标注集可以相互进行转化。扩展性是指对未解决的遗留问题或是未来可能的技术发展方向充分地加以考虑，以便可以加以扩充和修改，并使扩充和修改对系统的整体影响代价最小。

目前，比较有影响的词性标注集有“八五”汉语语料库给出的24个词性标记和北京大学计算语言研究所定义的39个词性标记。同时还有各个科研院所自行定义的词性分类体系。如哈尔滨工业大学机器翻译教研室定义的含有42个词性标记的词性标注集[131]和清华大学定义的含有111个词性标记的词性标注集[132]。应该说明的是，这些标注集在大类上差别不大，各个标注集只是对某个大类进行了不同的更细微的语法划分。本文将以北京大学计算语言研究所标准为例进行介绍，见表4-7：这里需要强调的是，词性标注集目前还没有国家制定的标准。不同的词性分类体系基本上是来源于不同语言学家对语言现象的认识。在计算语言学不同的领域，人们对词性依赖的程度不同，处理的精度不同，所以对词性分类的颗粒度定义也不同，在不同的应用中应该定义适用于自己系统的词性标注集。

表4-7 北京大学的汉语词性标注集

Table 4-7 Chinese POS tagging set of Peking university

ag	形语素	j	简称略语	r	代词
a	形容词	k	后接成分	s	处所词
ad	副形词	l	习用语	tg	时语素
an	名形词	m	数词	t	时间词
b	区别词	ng	名语素	u	助词
c	连词	n	名词	vg	动语素
dg	副语素	nr	人名	v	动词
d	副词	ns	地名	vd	副动词
e	叹词	nt	机构团体	vn	名动词
f	方位词	nz	其他专名	w	标点符号
g	语素	o	拟声词	x	非语素字
h	前接成分	p	介词	y	语气词
i	成语	q	量词	z	状态词

在最大熵词性标注模型中，分别给出了局部、融合转换触发对和融合聚类转换触发对三种最大熵模型的标注结果。其中聚类转换触发对特征由转换触发对配合词语聚类获得。由于考虑到训练语料库的规模和人工标注存在的不一致性，本文并没有选用复杂的复合特征。只选用了七个简单的特征，定义如下：$h_i \overset{def}{=} \{t_{i-1}, t_{i-2}, w_{i-1}, w_{i-2}, w_i, w_{i+1}, w_{i+2}\}$。其中包含前后各两个词、当前词和前两个

词性共7个简单特征。在局部最大熵模型的基础上，融合了长距离的转换触发对信息，进一步应用词语聚类的结果构建了基于聚类转换触发对的最大熵模型。为了和HMM模型做比较，采用相同的训练和测试语料。语料来源于1998年前半年人民日报，前5个月为训练语料，测试采用第6个月部分语料，包含大约10万个句子。标注结果如表4-8所示：

表4-8 HMM模型和最大熵模型词性标注比较结果

Table 4-8 Comparsion between POS tagging results based on HMM and ME

模型	标注准确率
HMM(ABS Smoothing)	94.427%
ME局部模型	95.425%
融和转换触发对的ME	95.692%
融和聚类转换触发对的ME	96.323%

从表4-8中可以看出，词性标注的错误率下降了34%。在表4-9中分别给出三种模型对测试语料中两个例句的处理结果。在HMM中都是错误的，融合转换触发对的最大熵模型纠正了一个，融合聚类转换触发对的最大熵模型得到全部正确的结果。

表4-9 三个模型对两个例句的词性标注结果

Table 4-9 Result of POS tagging of two example sentences based on three models

HMM	称/v 其/r 为/p “/w 第一/m 位/q 的/u 工作/vn ”/w ，/w
	被/p 简称/v 为/p “/w 牡康/nz ”/w 。/w
融合转换触发对的ME	称/v 其/r 为/v “/w 第一/m 位/q 的/u 工作/vn ”/w ，/w
	被/p 简称/v 为/p “/w 牡康/nz ”/w 。/w
融合聚类转换触发对的ME	称/v 其/r 为/v “/w 第一/m 位/q 的/u 工作/vn ”/w ，/w
	被/p 简称/v 为/v “/w 牡康/nz ”/w 。/w

4.6.2 音字转换试验结果

4.6.2.1 基于HMM模型的音字转换试验结果

在基于HMM的音字转换试验中，在不同的训练语料规模上分别试验了Bi-gram和Tri-gram两种模型。训练语料由1998年前5个月的人民日报语料组成。测试语料随机从1998年第6个月中抽取，包含大约2万个句子，38.6万个字符。转换的准确率在图4-6中给出，参数的数量在表4-10给出。

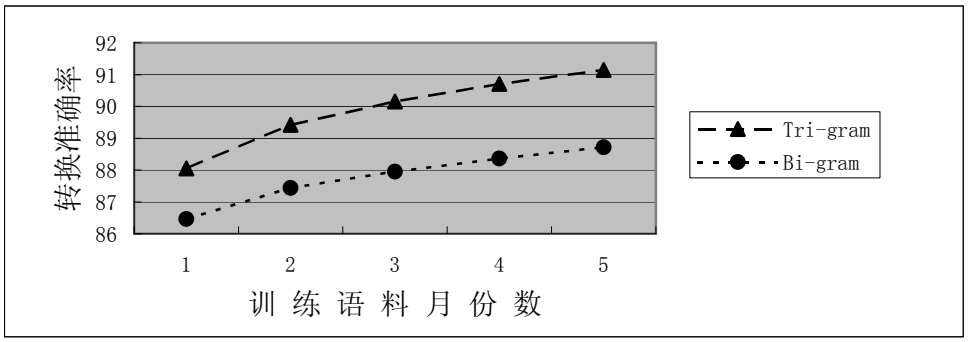

图4-6 基于HMM模型的音字转换正确率
Figure 4-6 Hit Rate of character to Pinyin experiment based on HMM

表4-10 HMM模型部分试验参数
Table 4-10 Partial parameter of experiment based on HMM

	1个月	2个月	3个月	4个月	5个月
Uni-gram参数数量	4565	4955	5125	5306	5426
Bi-gram参数数量	269462	395503	483835	571660	641699
Tri-gram参数数量	772491	1321140	1776372	2262063	2680167

从HMM的结果可以看出，转换的准确率随着训练语料的增长而增长。但是增长的趋势并不是非常的明显。与此相反，参数的数量增加了4倍，这就给运行的时间和内存造成了巨大的压力。试验结果表明：单纯通过增加训练语料的尺寸来提高音字转换的正确率是行不通的。

4.6.2.2 基于最大熵模型的音字转换试验结果

在基于最大熵模型的音字转换试验中，训练语料包含随机从1998年前5个月的人民日报语料选择的10万个较长的句子，包含140万个字符。在此基础上，本文尝试了从上下文中抽取特征的不同策略。首先分别基于简单和复杂模板建立了两个ME模型。另外，建立了融合3万转换触发对的ME模型，试验的结果如表4-11所示：

表4-11 ME拼音转换结果
Table 4-11 Result of Pinyin-to-charater conversion based on ME

模型	描述	正确率
HMM	Tri-gram配合绝对平滑	88.645%
ME1	基于简单特征模板的ME	85.961%
ME2	基于复杂特征模板的ME	89.018%
ME3	在ME2基础上融合转换触发对特征	89.142%

从表4-11中可以看出，在一个较少的训练语料上，音字转换的错误率减少了4%，结果还是比较理想的。另外，我们对音字转换的错误进行了必要的分类，

以使便对错误作进一步地分析。第一种错误被表述成“A-B”型，指的是正确的部分是单字，错误的部分也是单字。第二种错误被表述成“A-NotB”型，指的是正确的部分是单字，但是错误地转换成了一个词。第三种错误被表述成“NotA-B”型，指的是正确的部分是一个词，但是错误地转换成了多个单字。详细的例子见表4-12。

表4-12 三种类型错误对应的例句
Table 4-12 Examples of three type of error

A-B	Correct	经 济 (再) (罩) (阴) (云)
	Error	经 济 (在) (着) (因) (云)
A-NotB	Correct	对(通信) 的 广 泛 要 求
	Error	对(同)(新) 的 广 泛 要 求
NotA-B	Correct	一 (支)(由) 几 个
	Error	一 (只有) 几 个

表4-13 模型结果对比
Table 4-13 Comparsion between models

	HMM	ME3
A-B	32271	31308
A-NotB	7847	7898
NotA-B	3822	2810
SUM	43940	42016

从表4-13中，我们发现ME模型有效地减少了“A-B”和“NotA-B”两种类型的错误，这是因为ME模型可以更多地利用上下文中的特征信息。而“A-NotB”这种类型的错误来源于训练语料规模过小，例如“通信”一词并没有出现在训练语料中。作为有指导学习的HMM和ME模型来说，它们都不能有效地处理。从最后的整体结果上看，与HMM相比，ME更有潜力。

4.7 本章小结

本章首先简要介绍了HMM模型在处理词性标注和音字转换任务时的基本步骤，针对HMM模型面临的问题，采用融合长距离约束—转换触发对的最大熵模型进行了研究。在最大熵模型中主要研究了局部特征选择、聚类转换触发对生成和用于序列分类的Beam-Search搜索算法三个部分。试验结果证明，融合了转换触发对的最大熵模型在词性标注领域减少了34%的错误率，音字转换的错误率在一个较少的训练语料上减少了4%。

标注和转换错误率的下降可以归结为以下两个方面的贡献：首先，在自然语言处理任务中，充分利用上下文中的特征可以有效提高统计语言模型的质量，而最大熵模型可以通过特征函数做到这一点，最大熵模型一个最主要的优点就是通过特征函数可以包含各种重叠的、长距离的、颗粒度很细的基于词的特征。这种能力使得最大熵模型更加适合对自然语言建模。其次，特征的质量对于最大熵模型至关重要。正确的特征可以提高模型的质量，与此相反，大量包含噪声和不确定性的特征通常会减少模型的质量。针对音字转换任务，由于拼音的

描述能力不及词，所以采用了复杂特征模板。通过转换触发对加入长距离的约束信息。同时对转换触发对“$w_A \rightarrow w_B / t_B$”中的$w_A$进行了词语聚类，建立了聚类触发对特征，有效解决了转换触发对特征的稀疏问题。以上处理方法充分保证了特征的质量。

值得一提的是，本章以词性标注和音字转换为例进行了研究，但是研究的思路可以直接应用到自然语言处理词法分析中的其它问题中，如汉语分词和词义消歧等，同时也可以应用到语音识别等应用系统中。

第5章 基于矢量空间模型的词义相似度计算研究

5.1 引言

词义问题是自然语言处理中的核心问题，尤其在汉语这种轻结构而重意义的语言中更是如此。同时，它也是词法分析领域中最大的没有得到有效解决的问题。这不仅源于语言本身的复杂性，同时词义的研究也需要将抽象的词义具体地表示出来，以便能够被一种自动的系统高效地处理。但词义的表示和量化依赖于一些核心的人工智能问题：如知识的获取与表示等，这些领域目前还只停留在探索的阶段。所以，本章并没有对词义问题作全面的探讨，而是主要探讨了单义词的词义相似度计算[2]以及基于词义相似度计算的词语聚类。它们可以认为是词义问题的一个子问题。

词义问题的另外一个子问题是多义词的词义消歧，它可以认为是多义词在特定上下文中的词义标注(Word Sense Tagging)问题。由于以下两个原因，本章没有对多义词词义消歧进行研究。其一是：对词义的定义和描述还没有一个成熟、确定、被学术界广泛接受的规范，例如：在本文第1.4.1节中至少介绍了三种不同的词义定义方法。这样就导致了目前的词义标注研究都是基于若干个常用的词义歧义词，然后进行自定义的手工歧义标注，通过选取某一个统计语言模型如最大熵模型、支持向量机模型等进行有指导的学习和分类。这种研究方法最大的问题在于没有办法进行横向评测，这是因为每个研究人员建立的词义评测集合从数量上还是标注规范上都是不同的。其二是：词义标注所用的基本统计分类模型已经在第4章中进行了详细的研究，不需重复。

我们认为，对词义的研究在没有一个清晰问题描述的前提下，应该面向特定的应用进行研究，所以本章遵循“从应用中来，又回到应用中去”的词义问题研究思路。其理论根据在于词义问题本身是一种中间性的任务[133]，它最终要用在信息检索、句法分析、机器翻译等应用领域中。本章从词语聚类中提取出词义相似度计算这个词义问题。在充分借鉴前人工作的基础上，选用了具有鲜明特点的矢量空间模型进行了研究。首先，在充分利用矢量空间模型优点的基础上，利用分辨力指标对坐标轴词进行筛选，同时，应用词触发对信息减少了

2 为论述统一，本文将通常的语义相似度称为词义相似度，二者指同一个概念

“词袋”[3]效应带来的噪声，最后在这个高质量的矢量空间中进行词义相似度的计算。这充分体现了“从应用中来”的特点。与此同时，利用矢量空间模型提供的词义相似度计算功能进行了词语聚类的研究，词语聚类可以直接应用于查询扩展，构建基于类的语言模型或者解决传统统计语言模型的稀疏问题等应用中。这充分体现了“回到应用中去”的特点。这样，对词义研究的评测可以直接放到实际应用中去，这种评测方法可以充分借鉴相对成熟的应用系统研究成果，可以跨语料、大规模地对词义的研究进行评测。整个研究内容如图5-1所示：

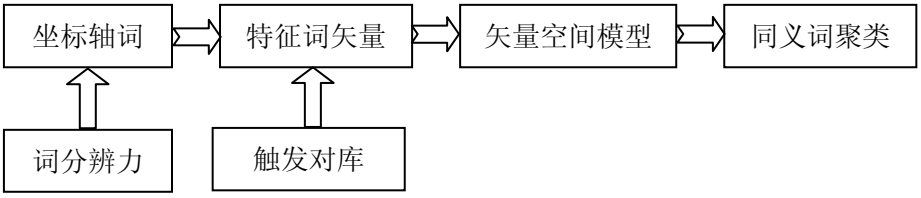

图5-1 本章研究内容图示

Figure 5-1 Illustration of content of this chapter

本章的主要结构如下：5.2节建立了基于触发对的矢量空间模型并进行了词语聚类的研究；5.3节给出了试验结果；最后是本章的小结。

5.2 基于矢量空间模型的词语聚类的研究

传统的语言学认为词之间具有两种重要的关系：聚合关系和组合关系。聚合关系定义了语言单位之间的相似性，也称为词义相似度(Word Semantic Similarity)。它是一种可称作“替代式”的相似度，两个具有聚合关系的词在特殊的上下文中可以相互替代，而不影响句子意义的合法性。如：“我使用电脑”和“我使用计算机”，这两句话中“电脑”和“计算机”具有相同的词义。

词义相似度计算方法主要可以分为两种：基于词典的方法和基于统计的方法。词义词典在1.4.1节中进行了详细地描述，这里不再介绍。需要指出的是，词义词典中包含的词条通常较少，且更新较慢；同时，词义词典通常由语言学家们手工完成，包含了语言学家对客观世界内省的认识，这种认识是否可以在自然语言处理中的所有问题都普遍适用还有待推敲。另外一种计算词义相似度的方法是基于统计的方法，在统计方法中，通常一个词的词义通过它的上下文来表示。两种主要的表示方法是：将上下文表示为一个概率分布或表示为矢量

3 词袋类似于词的集合，但是允许有重复的词存在。这里词袋包含上下文中所有的词，不包含位置信息。

空间中的一个矢量。第一种方法中，两个词对应的概率分布分别视为两个词的上下文概率分布，即：$p(\cdot|x)$ 和 $q(\cdot|y)$。词对(x,y)之间的词义相似度可以通过计算 $p(\cdot|x)$ 和 $q(\cdot|y)$ 之间的分布差异进行度量。通常采用Kullback-Leibler距离来度量两个概率分布之间的差异，见式(5-1)：其中 z 为词所在的上下文。

$$D(p\|q)=\sum_{z}p(z|x)\cdot\log\frac{p(z/x)}{q(z/y)} \tag{5-1}$$

本章采用矢量空间模型进行词义相似度的计算。矢量空间是指一个高维、离散、基于词的空间，空间中的每一维是从词典中根据特定量度选出来的一个词，称为坐标轴词。在矢量空间中，每一个特征词表示为空间中的一个矢量。根据特征词与文本中的每个坐标轴词之间的同现频度来定义该特征词，这样每一个特征词的词义可以通过坐标轴词来表示。

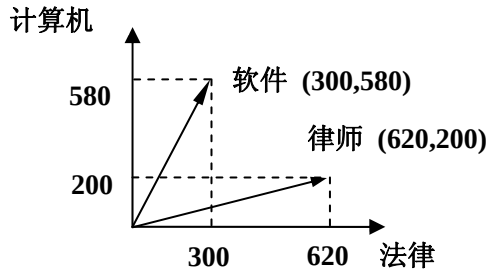

图5-2一个两维的矢量空间模型
Figure 5-2 Vector space model with two dimensions

图5-2表示在特定的上下文范围内，特征词“软件” 与坐标轴词“法律”和“计算机”分别共现了300次和580次。所以，特征词“软件”可以被表示为矢量 $\vec{A}=\{300,580\}$。这种表示方法充分利用了语言中“观其伴、知其义”的本质特征。每个特征词都可以被看成特征空间中的一个矢量。两个特征词之间的词义相似度 $Sim(\vec{A},\vec{B})$ 可以根据对应矢量之间的正则相关系数(夹角余弦)来计算，见式(5-2)：

$$Sim(\vec{A},\vec{B})=corr(\vec{A},\vec{B})=\frac{\sum_{i=1}^{N}\vec{A}_i\vec{B}_i}{\sqrt{\sum_{i=1}^{N}\vec{A}_i^2\sum_{i=1}^{N}\vec{B}_i^2}} \tag{5-2}$$

要建立这样的一个矢量空间需要考虑两个问题：首先，哪些词应该是这个空间的坐标轴词；其次，如何避免“词袋”效应而引入的大量噪声。针对这两

个问题，本章主要的创新工作包括：首先，用坐标轴词在特征词上分布的方差来作为分辨力的量度，优先选择有高分辨力的词作为坐标轴词；其次，利用触发对信息从上下文中获取语言中的结构信息来避免“词袋”效应带来的噪声。

5.2.1 坐标轴词的选择

坐标轴词应该是较重要的词。通常一个词在文档中出现的频度是该词重要性的标志之一。如果一个文档中词 A 出现的频度大于词 B 的频度，那么词 A 对于描述该文档内容的重要性应该比 B 大。因此，一个简单的确定一个词的重要程度的方法就是利用该词在文档中的频度。

然而，词的频度量度并不十分完善，主要表现在词在某文档中的频度不能够完整地表达该词的分辨能力。从整个数据全集的角度来看，一个词如果出现的频度很高，那么它对于表示某特定文档的内容帮助不大，也就是说，它不能有效地区分相关文档和不相关文档。例如，在关于计算机的数据全集中，“计算机”一词的频度很高，但它对表达文档内容的重要性应该比“硬盘”、“显示器”等词汇要小；反之，如果一个词在整个数据全集中出现频度很低，它应该是反映包含该词的文档内容的重要词汇。因此，一个词的重要性量度还应与该词所在的文档的总数成反比或者近似反比的关系，本章利用倒文档频度(Inverse Document Frequency, IDF)来描述这种关系，定义见式(5-3)：

$$IDF = \log(\frac{N}{N_A}) \tag{5-3}$$

其中，N 为数据全集中文档的总数，N_A 为包含词 A 的文档总数。这一量度反映了包含该词的文档区别于其它文档的程度，是一个词在整个数据集合中的重要性的全局性统计特征。

本章充分利用频度和倒文档频度两个指标对坐标轴词进行初选。首先，通过IDF，过滤出包括助词、叹词等的所有虚词。然后，在剩下的实词中选择出2000个高频实词作为坐标轴词。在此基础上，本文同时从基本实词中选择出14000个高频实词作为特征词。下一步的工作就是构建一个14000×2000的二维矩阵。矩阵中每一行代表矢量空间中的一个矢量。矢量中的每个分量分别纪录着这个特征词在一定尺寸的上下文窗口中与每个对应坐标轴词共现的频度。与计算平均互信息相同，在统计共现频度的时候，本章考虑特征词的左右两面各12个词。

上面已经提到过，当选择坐标轴词的时候，除了词频信息以外。必须同时考虑坐标轴词在处理词义问题时分辨力的强弱。倒文档频度可以对不同词性的词的分辨能力作出区分，但针对相同的词性，这种量度比较粗糙。本章采用了另外一种概率分布方差量度，描述如下：如果一个坐标轴词几乎与所有的特征词共现，那么它几乎没有什么分辨力，如果一个坐标轴词经常与某些特征词共现，而与另外的一些特征词从不共现，那么它本身就有很强的分辨力。在上文建立的二维矩阵中,每一列都可以看成是这个坐标轴词在14000个离散值上的离散概率分布，这个分布的方差越大，说明这个坐标轴词的分辨能力越强；反之，则越弱。从表5-1可以看出，坐标轴词“人们”,“看到”，基本上可以与任何词共现，当进行词义消歧的时候，它们不能够给出足够的信息，而坐标轴词“气温”基本上只与气候的词义有关，而与其它的词义无关，所以它的分辨能力很强，比较适合作坐标轴词。

表5-1 高分辨力词和低分辨力词列表

Table 5-1 List of words with high differentiation and low differentiation

小方差	人们	所有	看到	全部	理由	突然
大方差	冷空气	政治局	联赛	立案	冠军	气温

5.2.2 基于触发对建立词矢量空间模型

当统计共现频度的时候，通常把特征词上下文中的所有坐标轴词当成一个“词袋”来处理，这种简化的处理方式丢失了包括语序和位置信息在内的结构信息，所以不可避免地会引入大量的噪声。通过句法分析器可以获得语言的结构信息并用来进行词义消歧[134]，但是句法分析需要树库进行训练。由于构建树库的代价比较昂贵，所有目前树库的规模普遍较小；同时，句法分析也依赖于词义问题的根本解决，这样就形成了一个环状问题。本章利用词触发对来部分模拟一个依存句法分析器，它的主要优点是触发对的获取只依赖于经过分词的语料库而不再需要昂贵的树库，同时还可以提供很多蕴涵于触发对中的搭配句法信息。

第2.3节中已经详细介绍了如何采用平均互信息来抽取长距离的词触发对，本章不再详述。在建立共现矩阵时，首先，将坐标轴词和特征词看成是一个词触发对候选，如果出现在2.3节建立的200万词触发对库中时，将触发对对应的平均互信息的值映射到0-1之间，类似于模糊集理论中的隶属度，将这个隶属度当成权重。然后，将共现频度乘以这个权重，写入到共现矩阵中。通过试验发现，这种方法只是小幅度地提高了模型的质量，但是却增加了计算的时间和空

间复杂度（由整型变为浮点型运算）。共现频度和AMI虽然不完全对应，但是却有较强的线性对应关系，所以本章采用bool型的权重方式，也就是说，当出现在触发对库中时，直接统计其对应的共现频度，否则，不对其进行统计。具体的试验结果见5.3节。

5.3 试验结果

选择坐标轴词所需的IDF值从2000年人民日报统计得到，共现频度与触发对都是从1974-1984年10年的人民日报统计得到。本章分别建立了三个矢量空间模型，第一个模型选用了2000个高频实词作为它的坐标轴词。第二个模型首先选择了2500个高频实词，构建了一个14000×2500的二维矩阵，然后根据5.2.1节中介绍的方法，利用分辨力量度选出前2000个高分辨力的词作为坐标轴词。第三个模型和第二个模型选用相同的坐标轴词，但利用上面的方法基于触发对库建立特征词矢量。本章选择7对共14个同义词。其中6对为名词，1对为形容词。根据夹角余弦量度分别在3个模型上计算它们彼此的词义相似度。限于篇幅，只给出模型一和模型三的结果。如表5-2和表5-3所示：

表5-2 七组同义词的相似度矩阵(2000高频实词为轴)

Table 5-2 Similarity matrix of seven pairs of synonyms (2000 highest frequence word axis)

	衬衣	外套	面包	馒头	电脑	计算机	工业	农业	宾馆	饭店	法规	法律	高兴	愉快
衬衣	1.00	.823	.357	.282	.163	.139	.222	.170	.183	.170	.082	.093	.193	.157
外套	**.823**	1.00	.283	.252	.119	.088	.095	.069	.152	.146	.073	.086	.175	.140
面包	.357	.283	1.00	.755	.266	.238	.382	.300	.329	.343	.128	.162	.249	.202
馒头	.282	.252	**.755**	1.00	.166	.129	.182	.156	.268	.282	.074	.097	.215	.163
电脑	.163	.119	.266	.166	1.00	.609	.305	.217	.236	.229	.137	.176	.151	.145
计算机	.139	.088	.238	.129	**.609**	1.00	.393	.278	.230	.227	.176	.219	.146	.137
工业	.222	.095	.382	.182	.305	.393	1.00	.792	.276	.270	.218	.231	.203	.171
农业	.170	.069	.300	.156	.217	.278	**.792**	1.00	.193	.189	.222	.232	.182	.137
宾馆	.183	.152	.329	.268	.236	.230	.276	.193	1.00	.925	.143	.179	.280	.285
饭店	.170	.146	.343	.282	.229	.227	.270	.189	**.925**	1.00	.145	.183	.239	.241
法规	.082	.073	.128	.074	.137	.176	.218	.222	.143	.145	1.00	.855	.110	.103
法律	.093	.086	.162	.097	.176	.219	.231	.232	.179	.183	**.855**	1.00	.162	.161
高兴	.193	.175	.249	.215	.151	.146	.203	.182	.280	.239	.110	.162	1.00	.562
愉快	.157	.140	.202	.163	.145	.137	.171	.137	.285	.241	.103	.161	**.562**	1.00

表5-3七组同义词相似度矩阵（应用触发对进行筛选）

Table 5-3 Similarity matrix of seven pairs of synonyms (selected by trigger-pairs)

	衬衣	外套	面包	馒头	电脑	计算机	工业	农业	宾馆	饭店	法规	法律	高兴	愉快
衬衣	1.00	.901	.070	.035	.010	.000	.024	.009	.000	.000	.034	.027	.043	.027
外套	**.901**	1.00	.000	.000	.000	.000	.006	.005	.000	.000	.029	.022	.029	.024
面包	.070	.000	1.00	.666	.028	.035	.174	.144	.037	.060	.003	.009	.030	.005
馒头	.035	.000	**.666**	1.00	.022	.001	.011	.009	.034	.055	.000	.003	.017	.002
电脑	.010	.000	.028	.022	1.00	.456	.071	.043	.018	.062	.012	.038	.047	.009
计算机	.000	.000	.035	.001	**.456**	1.00	.219	.159	.025	.057	.026	.057	.051	.013
工业	.024	.006	.174	.011	.071	.219	1.00	.720	.056	.069	.056	.081	.130	.043
农业	.009	.005	.144	.009	.043	.159	**.720**	1.00	.039	.054	.091	.118	.134	.035
宾馆	.000	.000	.037	.034	.018	.025	.056	.039	1.00	.877	.015	.023	.084	.066
饭店	.000	.000	.060	.055	.062	.057	.069	.054	**.877**	1.00	.024	.037	.060	.016
法规	.034	.029	.003	.000	.012	.026	.056	.091	.015	.024	1.00	.812	.034	.010
法律	.027	.022	.009	.003	.038	.057	.081	.118	.023	.037	**.812**	1.00	.072	.021
高兴	.043	.029	.030	.017	.047	.051	.130	.134	.084	.060	.034	.072	1.00	.607
愉快	.027	.024	.005	.002	.009	.013	.043	.035	.066	.016	.010	.021	**.607**	1.00

在表5-2和表5-3中，同义词之间的距离基本不变，但非同义词之间的距离下降很大，为综合衡量这两种变化，将表5-2和表5-3中每一行看成一个14个离散值的概率分布。大的方差意味着同义词间更相关，而非同义词间更不相关。每个词的分布在不同的模型中的方差值在图5-3中给出，最后一列代表全体的平均值。可以看出与第一个模型相比，模型二和三的质量分别提高了7%和38%。

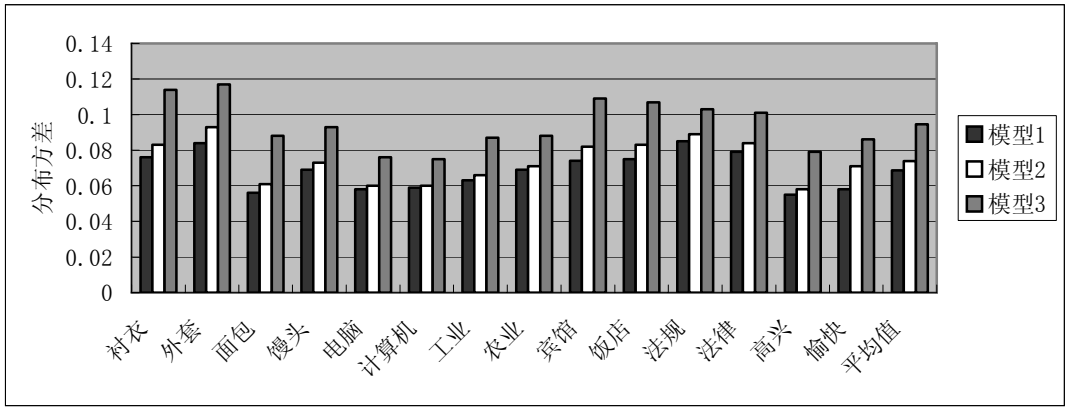

图5-3 词的相似度在三个模型中的分布方差

Figure 5-3 Distributing variance of similarity in three different models

矢量空间模型提供了一种有效的词义相似度计算方法。在此基础上，我们进行了词语聚类的研究。限于篇幅，本文给出6个名词的聚类结果。在6个词对应的每一列中，左侧给出相同聚类中的每一个词，右侧给出对应的词义相似度，词义相似度采用矢量夹角余弦来计算。从表5-4中可以看出词义相似度高的并不局限于同义词的领域，例如：“计算机”一词并不是“智能”的同义词，但是它与这个词相关。同理，“法律”和“公民”有关，但它们也不是同义词。试验结果证明模型不仅能发现同义词，还可以发现领域相关的词。这个特性对信息检索应用中的查询扩展有很大的帮助，同时也可用于建立基于类的语言模型等多种应用中。

本章的主要试验围绕名词展开，对于动词和形容词，词语聚类并没有取得非常理想的结果。从词义的角度来说，动词和形容词是“狭义”的，它们通常与上下文中的某个固定词形成固定的搭配，这就决定了它们的词义被上下文中某些固定的词语所限制和描述。而名词的词义是“广义”的，通常蕴含在一个较广的上下文中，被这个上下文所限制和描述。矢量空间模型恰好可以处理这种分布上较广的词义现象，所以它在名词的试验上获得了较理想的结果。

表5-4 词语聚类部分结果

Table 5-4 Partial result of words clustering

衬衣		面包		计算机		宾馆		法律		工业	
上衣	.925	方便	.813	软件	.807	客房	.886	法规	.812	重工业	.809
毛衣	.923	巧克	.806	微电子	.757	饭店	.877	宪法	.730	轻工业	.804
大衣	.905	饼干	.777	芯片	.706	酒店	.766	规定	.727	生产	.738
外套	.900	糕	.728	终端	.694	旅馆	.715	条文	.715	农业	.720
棉袄	.878	美味	.723	智能	.694	招待所	.681	法令	.671	消费品	.701
衣衫	.832	糖果	.691	并行	.686	餐厅	.668	公民	.671	产值	.664
衣服	.765	糕点	.682	微机	.684	酒楼	.660	适用	.646	手工业	.660
棉衣	.764	月饼	.682	联网	.678	旅店	.636	规章	.645	畜牧业	.645
西服	.747	馒头	.666	集成电路	.657	大堂	.620	通则	.641	原料	.644

5.4 本章小结

在词义领域：针对单义词构建了基于触发对的矢量空间模型用来进行词义相似度的计算，对传统的矢量空间模型进行了改进，并以改进的矢量空间模型为基础进行了词语聚类的研究。结果表明：词语聚类的结果满足进一步应用的需要。

本章的研究始终以应用为驱动，无论是词义问题的提出，还是模型的评测，都根植于特定的应用中，这充分体现了研究的应用价值，同时也保证了评测本身的有效性。在词语聚类问题中，名词取得了很好的结果，这是因为名词的词义更多地被篇章所限制，也就是说，只凭上下文中的某个词，我们很难精确描述出名词的词义，所以用矢量空间模型，这种模型的特点是可以采集到上下文中很多的相关词，而每个词对词义的贡献大致相等，这种蕴含于篇章中的特征信息正好符合名词的词义特点。本文的研究间接验证了模式识别中“没有免费的午餐”定理。即如果没有先验知识，任何一种机器学习算法都不比另外一种好。本文中试验的成功并不是矢量空间模型要好于其它的模型，而是因为这模型的能力适合于要解决的问题。也就是说，在词义问题上，根据这个词本身的特点来用一个适合的模型来处理，要比单纯的提高某一种机器学习方法，或试图证明某一种模型要比另外的模型在处理所有词义问题上都要好，要有意义的多。

结　论

词法分析是自然语言处理环节中最基础的部分，本文研究的汉语词法分析主要包括分词、词性标注和词义相似度计算三个部分，主要采用统计语言模型对其进行研究。统计语言模型中主要利用词法信息改进了N-gram模型的平滑算法。利用基于REA算法的K-best分词模型对分词歧义进行了识别，同时利用了最大熵模型对分词歧义进行了消解。在最大熵框架下利用Beam Search搜索算法对词性标注和音字转换进行了研究。利用基于触发对的矢量空间模型对词义相似度进行了计算。利用以上的研究成果实现了INSUN-LEX汉语词法分析软件。主要研究成果如下：

一、在N-gram模型的平滑方面，本文利用词性信息提出了新的Katz平滑折扣系数，不仅有效解决了概率折扣的问题，在语言模型交叉熵量度上也取得了比Abs平滑和W-B平滑更好的结果。在此基础上，利用HowNet词义词典提供的词义相似度计算功能，提出了融合词义信息的Uni-gram平滑算法。与传统的Uni-gram平滑算法相比，新的平滑算法不仅在语言模型交叉熵量度上获得了部分的降低，而且还可以用在其它的高阶平滑算法中。证明在平滑算法中加入词法信息是有效的。利用平均互信息抽取了长距离的触发对，结果证明：触发对可以有效承载长距离的语言约束关系。在此基础上，针对特定的汉语词法分析问题，提出了转换触发对的概念。

二、针对分词问题，提出了基于REA算法的K-best分词模型。与其它的K-shortest路径搜索算法相比，REA算法更加适合于汉语分词问题。同时提出了K值的计算方法。结果证明这种方法可以有效地识别出大部分的真歧义字段。同时利用最大熵模型对分词消歧进行了研究，试验证明：局部特征配合触发对，消解的正确率达到了92%。针对人名识别问题，综合统计与语言学知识，建立了人名识别用多源知识表，通过利用知识表，配合对应的统计方法，可以保证在不降低召回率的情况下，提高识别的准确率。

三、针对词性标注问题，首先利用最大熵模型和支持向量机模型对复杂兼类词标注进行了研究。试验证明：加入上下文中的词特征可以有效降低词性标注的错误。在此基础上，利用最大熵模型对基于句子的词性标注进行了研究，主要研究了Beam Search方法和基于转换触发对“$w_A \rightarrow w_B / t_B$”的聚类触发对的加入。最后，将音字转换看成词性标注的特例，分别研究了简单特征模板和

复杂特征模板，试验表明：音字转换中复杂特征模板取得了更好的结果；同时针对词性标注问题，融合了聚类转换触发对的最大熵模型与HMM模型相比错误率减少了三分之一。

四、针对词义问题，利用矢量空间模型对词义相似度进行了计算。利用词分辨力量度选择坐标轴词；同时为解决“词袋”效应带来的噪声，利用触发对来模拟一个依存句法分析器，使矢量空间模型可以包含语言中的结构信息，从而提高矢量空间模型的质量。通过对传统的矢量空间模型的以上两个改进，新的矢量空间模型在词语聚类上获得了比较理想的结果。

本文的研究只获得阶段性的成果，未来的研究工作包括：

一、在词法分析领域，如果将分词和名实体识别分开来处理，分词的错误将会扩散到名实体识别的过程中。一个更为理想的办法就是建立一个统一的基于类的语言模型，将这两部分融合在一起进行处理。

二、通过使用粗糙集来对具有噪声的数据进行特征信息的提取，利用上近似的概念完成知识的挖掘。这是因为当我们尝试获得上下文特征信息的时候，需要一些先期的处理步骤，如分词等，这些不能保证完全正确的先期处理步骤不可避免地会引入一些噪声。

参考文献

1 黄昌宁，高剑峰，李沐. 对自动分词的反思. 哈尔滨，全国第七届计算语言学联合学术会议，2003:26-38

2 L. Coin, A. Bateman and R. Durbin. Enhanced Protein Domain Discovery by Using Language Modeling Techniques from Speech Recognition. Proc Natl Acad Sci USA, 2003:4516-4536

3 M. Ganpathiraju, D. Weisser and R. Rosenfeld et al. Comparative N-gram Analysis of Whole-Genome Protein Sequences. Proceedings of the Human Language Technologies Conference, San Diego, 2002 http://www.cs.cmu.edu/~madhavi/publications/Ganapathiraju-HLT2002.pdf

4 J. J.Feng, A. Sears. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. ACM Transaction on Computer-Human Interaction. 2004, 11(4):329-356

5 Y. Q. Gao, B.W. Zhou et al. MARS: A Statistical Semantic Parsing and Generation-Based Multilingual Automatic Translation System. Machine Translation. 2002, 17(3):185-212

6 Alceu de S. Britto Jr. A Two-Stage HMM-Based System for Recognizing Handwritten Numeral Strings. Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001:396-404

7 X. L. Wang, D. Yeung and X. Wang. Chinese Intelligent Input Method. In Proceedings on the International Conference on Artificial Intelligence, Las Vegas, USA, 2000:1203-1208

8 Z. Chen, K. F. Lee. A New Statistical Approach to Chinese Pinyin Input. ACL-2000, Hong Kong, 2000:241-247

9 J. Ponte, W. B. Croft. A Language Modeling Approach to Information Retrieval. In proc. 21st Int. Conf. Research and Development in Information Retrieval(SIGIR'98), 1998:275-281

10 E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics. 1995, 21(4):543-565

11 J. W. Grzymala-Busse, L. J. Old. A Machine Learning Experiment to Determine Part of Speech From Word-endings. In: Z.W.Ras, A.Skowron (Eds.), ISMIS'97. Springer-Verlag, Berlin, Germany, 1997: 1-630.

12 Q. C. Chen, X. L. Wang et al. A Word Sense Disambiguation Approach Based on Rough Set , Journal of Harbin Institute of Technology. 2002, 9(2):201-204.

13 关毅. 基于统计的汉语语言模型研究. 哈尔滨工业大学工学博士学位论文. 1999:3-6

14 S. Martin， C. Hamacher et al. Assessment of Smoothing Methods and Complex Stochastic Language Modeling. In 6th Eruopean Conference on Speech Communication and Technology, Budapest, Hungary, 1999:1939-1942

15 J. T. Goodman, A Bit of Progress in Language Modeling. Computer Speech and Language. MSR-TR-2001-72

16 D. Ron, Y. Singer, N. Tishby. The Power of Amnesia. in Advances in Neural Information Processing Systems 6, J. Cowan, G. Tesauro, and J. Alspector, Eds. San Mateo, Morgan Kaufmann, CA, 1994:176–183

17 T. Niesler, P.Woodland. Variable-length Category n-gram Language Models. ICASSP 96, 1996:164-171

18 M. Siu, M. Ostendorf. Variable n-grams and Extensions for Conversational Speech Language Modeling. IEEE Transactions on Speech and Audio Processing. 2000, 8(1):63-75

19 G. D. Zhou, K. T. Lua. Interpolation of n-gram and Mutual Information Based Trigger Pair Language Models for Mandarin Speech Recognition. Computer Speech and Language. 1998, 12:125-141

20 R. Lau, R. Rosenfeld and S. Roukos. Trigger-Based Language Model: A Maximum Entropy Approach. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Minneapolis,

1993:45-48

21 H. Ney, U. Essen and R. Kneser. On Structuring Probabilistic Dependences in Stochastic Language Modeling. Computer Speech and Language. 1994, 8:1-38

22 I. J. Good. The Population Frequencies of Species and the Estimation of Population Parameters, Biometrika, 1953, 40(3):237-264

23 F. Jelinek and R. L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In proceedings of the workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May. 1980:381-397

24 M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE transactions on Acoustics, Speech and signal Processing. 1987, ASSP-35(3):400-401

25 I. H. Witten, T. C. Bell. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Transactions on Information Theroy. 1991, 37(4):1085-1094

26 R. Kneser, H. Ney. Improved Backing-off for n-gram Language Modeling.In Proceedings of the IEEE International Converence on Acoustics, Speech and Signal Processing, 1995, 1:181-184

27 L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. 1989, 77(2):257-286,

28 A. B. Poritz. Hidden Markov Models: A Guided Tour. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, New York Hilton, New York City, April, 1988:7-13

29 E. T. Jaynes. Information Theory and Statistical Mechanics. Physics Reviews. 1957, 106:620-630

30 S. A. DellaPietra, V. J. DellaPietra et al. Adaptive Language Modeling Using Minimum Discriminant Estimation. In Proceedings of the International Conference on Acoustics, Speech and Signal

Processing, San Francisco, March, 1992:633-636
31 J. N. Darroch, D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models, The Annals of Mathematical Statistics. 1972, 43(5):1470-1480
32 R. Rosenfeld. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. Ph.D. thesis. Carnegie Mellon University. 1994
33 A. L. Berger, S. A. Della Pietra, V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics. 1996, 22(1):39-72
34 A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity Resolution, Ph.D. thesis, University of Pennsylvania. 1998
35 Kamal Nigam, J. Lafferty, et al. Using Maximum Entropy for Text Classification. In Proceedings of the IJCAI-99 workshop on information filtering, Stockholm, SE, 1999:61-67
36 A. Borthwick. A Maximum Entropy Approach to Named Entity Recognition,Ph.D dissertation New York University, September,1999
37 A. Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In Proceedings of Conference on Empirical Method in Natural Language processing, university of Pennsylvania, 1996:133-141
38 S. D. Pietra, V. D. Pietra and J. Lafferty. Inducing Features of Random Fields. IEEE Trans. On Pattern Analysis and Machine Intelligence. 1997, 19(4):380-393
39 R. Rosenfeld, L. Wasserman et al. Interactive Feature Induction and Logistic Regression for Whole Sentence Exponential Language Model. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, Dec, 1997:230-237
40 C. Cortes, V. Vapnik. Support Vector Networks. Machine Learning, 1995, 20:273-297
41 V. Vapnik. The Nature of Statistical Learning Theory. Springer. 1999
42 T. Joachims. Text Categorization with Support Vector Machines:

Learning with Many Relevant Features. In Proceedings of the 10th Europen Conference on Machine Learning, 1998:137-142

43 T. Kudoh, Y. Matsumoto. Use of Support Vector Learning for Chunk Identification. In Proceedings of the Fourth Conference on Computational Natural Language Learning, 2000:142-144

44 C. Bahlmann, B. Haasdonk and H. Burkhardt. On-line Handwriting Recognition with Support Vector Machines--a Kernel Approach. In Proceedings of the 8th IWFHR, 2002:49-54

45 T. Nakagawa, T. Kudoh and Y. Matsumoto. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, 2001

46 C. Cabezas, P. Resnik. Supervised Sense Tagging using Support Vector Machines. In Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France, 5-6 July 2001

47 J. Weston, C. Watkins. Support Vector Machines for Mutli-Class Pattern Recognition. In Proceedings of the Seventh European Symposium, 1999

48 J. Platt. Probabilistic Output for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers. MIT Press

49 H. Schütze. Word Space. In Stephen J. Hanson, Jack D. Cowan, C.Lee Giles. Ed. Advances in Neural Information Processing Systems 5. Morgan Kauffman, San Mateo, CA, 1993:895-902

50 H. Schütze. Automatic Word Sense Discrimination, Computational Linguistics. 1998, 24(1):97-124

51 陈清才，王晓龙. 一种基于词矢量的汉语语义量化模型. 计算机研究与发展. 2001, 38(2):207-212

52 鲁松，白硕，黄雄，张健. 基于向量空间模型的有导词义消歧. 计算机研究与发展. 2001, 38(6):662-667

53 K. S. Cheng, H. Y. Gilbert, K. F. Wong. A Study on Word-Based and Integral-bit Chinese Text Compression Algorithms. Journal of the American Society for Information Science. 1999, 50(3):218-228

54 K. J. Chen, S.H. Liu. Word identification for Mandarin Chinese sentences, In proceedings of Fifteenth International Conference on Computational Linguistics, Nantes: COLING-92, 1992:101-107.

55 J. Hockenmaier, C. Brew. Error-Driven Segmentation of Chinese. Communications of CLLIPS. 1998, 1(1):69-84.

56 D. Palmer. A Trainable Rule-Based Algorithm to Word Segmentation. In: Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, Madrid, Spain, 1997:321-328

57 A. D. Wu, Customizable Segmentation of Morphologically Derived Words in Chinese. Computational Linguistics and Chinese Language Processing. 2003, 8(1):1-28

58 R. Sproat, C. Shih and W. Gale et al. A Stochastic Finite-State Sord Segmentation Algorithm for Chinese. Computational linguistics. 1996, 22(3):377-404

59 J. F. Gao, M. Li and C. N. Huang. Improved Source-channel Models for Chinese word segmentation, 41nd Annual Meeting of the Association for Computational Linguistics. Sapporo. Japan, July, 2003:7-12.

60 N. W. Xue. Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing. 2003, 8(1):29-48

61 T. H. Chiang, J. S. Chang and M. Y. Lin. Statistical Models for Word Segmentation and Unkown Word Resolution. Proceeding of ROCLING-V, Taipei, Taiwan, 1992:18-20.

62 W. J. Teahan, Y. Wen, N. McNab and I. H. Witten. A Compression-based Algorithm for Chinese Word Segmentation, Computational Linguistics, 2001, 26(3):375-393.

63 X. Luo, M. S. Sun and K. T. Benjamin. Covering Ambiguity Resolution

in Chinese Word Segmentation base on Contextual Information. In Proceedings of 19th International Conference on Computational Linguistics, Taiwan, 2002:598-604.

64 M. S. Sun, Z. P, Zou. The Role of High Frequent Maximum Crossing Ambiguities in Chinese Word Segmentation. Journal of Chinese information processing. 1999, 13(1):27-34.

65 M. Li, J. F. Gao and C. N. Huang et al. Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Segmentation, In SIGHAN2002 Sapporo, Japan, 2003:11-12.

66 T. Brants. Tnt – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth ANLP, 2000

67 F. Jelinek, and J. Lafferty and D, Magerman. Decision Tree Parsing using a Hidden Derivational Model. In Proceedings of the Human Language Technology Workshop, 1994:272-277

68 Y. Guan and X. L. Wang. Quantifying Semantic Similarity of Chinese Words From Hownet, In: Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 2002: 234-239

69 L. Lee. Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis. Harvard University Technical Report TR-11-97

70 董振东. 知网. http://www.keenage.com, 2002

71 卢志茂, 刘挺等. 基于依存分析和贝叶斯网络的无指导汉语词义消岐. 高技术通讯. 2004, 14(2):7-11

72 D. Yarowsky. One Sense per Collocation. In: Proceedings ARPA Workshop on Human Language Technology, Princeton, 1993:266-271

73 P. F. Brown, P. S. Della et al. Word Sense Disambiguation using Statistical Methods. In Proceedings of the 29th Annual Meeting of the Association for computational Linguistics ACL, 1991:264-270

74 L. Tomaso, V. D. Dini and F. Segongd. Word Sense Disambiguation with Functional Relations. In Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC Granada, 1998:

1189-1196

75 T. Pedersen, R. Bruce, Knowledge Lean Word-Sense Disambiguation. In Proceedings of the 15th National conference on Artificial Intelligence, AAAI Press, 1998:800-805

76 D. Yarowsky, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, In: Proceedings of ACL'95, 1995:189-196

77 A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. 1977, 39(B):1-38

78 R. Garside, G. Leech and T. McEnery. Linguistic Information from Computer Text Corpora: Corpus Annotation. New York: Longman House, 1997

79 F. Jelinek. Statistical Methods for Speech Recognition. The MIT Press, 1997

80 K. Lari, S. J. Young. Applications of Stochastic Context-free Grammars Using the Inside-Outside Algorithm. Computer Speech and Language. 1991, 5(2):237-257

81 E. Dermatas, G. Kokkinakis. Automatic Stochastic Tagging of Natural Language Texts. Computational Linguistics.1995, 21(2):137-163

82 王挺，史晓东，陈火旺，杨谊. 一种用未分析语料训练文法的方法. 软件学报. 1998, 9(1):36-42

83 周强， 黄昌宁. 汉语概率型上下文无关文法的自动推导.计算机学报. 1998, 21(5):385-392

84 T. Fujisaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino. A Probabilistic Parsing Method for Sentence Disambiguation. In Proceedings of 1st International Workshop on Parsing Technologies. Carnegie Mellon University, Pittsburgh, PA, 1989: 85-94

85 A. Stolcke. An Efficient Probabilistic Context-Free Parsing Algorithm That Computes Prefix Probabilities. Computational Linguistics. 1995, 21(2): 165-201

86 T. Briscoe and T. Carroll. Generalized Probabilistic LR Parsing of

Natural Language (Corpora) with Unification-Based Grammar. Computational Linguistics. 1993, 19(1):25-59

87 朱胜火，周明，刘昕，黄昌宁. 一种有效的概率上下文无关文法分析算法. 软件学报. 1998, 9(8):592-597

88 R. F. Simmons and Y. H. Yu. The Acquisition and Use of Context-Dependent Grammar for English. Computational Linguistics, 1992, 18(4):391-418

89 周明，黄昌宁. 面向语料库标注的汉语依存语法体系的探讨. 中文信息学报. 1994，8(3):35-52

90 Y. Schabes. Stochastic Lexicalized Tree-Adjoining Grammars. In Proceedings of COLING'92. Nantes, France, 1992

91 赵铁军 等. 机器翻译原理. 哈尔滨工业大学出版社, 2001

92 倪文杰，竺一鸣，高蕴琪等. 现代汉语辞海. 人民出版社, 北京. 1994

93 梅家驹，竺一鸣，高蕴琦等. 同义词词林. 上海辞书出版社, 1983

94 M. Margaret. The Thesaurus in Syntax and Semantics. Mechanical Translation. 1957, 4:1-2

95 S. J. Karen. Synonymy and Semantic Classification. Ph.D. Thesis. University of Cambridge, Cambridge, UK, 1986

96 S. Y. Sedlow, W. A. Sedelow. Recent Model-based and Model-Related Studies of a Large-scale Lexical Resource(Roget's Thesaurus). In Proceedings of the 14th International Conference on Computational Linguistics. COLING'92. Nantes, France, August, 1992:1223-1227

97 G. A. Miller, T. B. Richard et al. WordNet: An On-line Lexical Database. International Journal of Lexicography. 1990, 3(4): 235-244

98 俞士汶, 朱学峰等. 现代汉语语法信息词典详解. 清华大学出版社, 1998

99 朱德熙. 语法问答. 商务印书馆, 1993

100 黄曾阳. HNC（概念层次网络）理论. 清华大学出版社, 1998

101 D. Biber, S. Conrad. Corpus Linguistics. The Syndicate of the Press of the University of Cambridge. 1998

102 R. Garside, G. Leech and T. McEnery. Linguistic Information from

Computer Text Corpora: Corpus Annotation. New York: Longman House, 1997

103 M. P. Marcus, B. Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics. 1993, 19(2):313-329

104 W. A. Gale, G. Sampson. Good-Turing Frequency Estimation without Tears. Journal of Quantitative Linguistics. 1995, 2(3):15-19

105 L. R. Baul, F. Jelinek, and R. L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(2): 179-190, March 1983.

106 S. F. Chen, J. Goodman. An Empirical Study Smoothing Techniques for Language Modeling. In Proceedings of the 34th Annual Meeting of the ACL, Caligornia, 1996:310-318

107 F. Smadja. Retrieving Collocations from Text: Xtract. Computational Linguistics. 1993, 19(1):143-177

108 徐志明. 面向文字识别的汉语统计模型研究. 哈尔滨工业大学工学博士论文. 2001

109 S. F. Chen. Building Probabilistic Models for Natural Language, PhD thesis. the Subject of Computer Science. Harvard University Cambridge Massachusetts, May 1996.

110 J. F. Gao, J. T. Goodman, M. Li., K. F. Lee. Toward a Unified Approach to Statistical Lanugage Modeling for Chinese. ACM Transactions on Asian language Information processing. 2002, 1(1):3-33

111 R. Sproat, T. Emerson. The First International Chinese Word Segmentation Bakeoff. 2003, www.sighan.org/bakeoff2003/paper.pdf

112 V. M. Jimenez, A. Marzal and J. Monne. A Comparison of Two Exact Algorithms for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition, 4th European Conference on Speech Communication and Technology, EUROSPEECH-95, Madrid, 1995:1071-1074

113 D. S. Dreyfus. An Appraisal of Some Shortest-path Algorithms.

Operations Research, 1969, 77:395-412

114 E. Q. Martins, J. L. Santos. A New Shortest Paths Ranking Algorithm, Technical Report, University de Coimbra, http://www.mat.uc.pt/~eqvm 1996

115 Eppstein, D. Finding the k Shortest Paths, SIAM Journal of Computing. 1999, 28(2):652-673

116 V. M. Jimenez, A. Marzal. Computing the K shortest paths: a new algorithm and an experimental comparison, Lecture Notes in Computer Science series, Springer-Verlag, 1999, 1668:15-29

117 Masaaki Nagata. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-best Search Algorithm, In Proceedings of COLING'94, Tokyo, Japan, 1994:201-207

118 N. Chinchor. MUC-7 Named Entity Task Definition. In proceedings Of The Seventh Message Understanding Conference, 1998

119 刘秉伟, 黄萱菁等. 基于统计方法的中文姓名识别.中文信息学报. 2000, 14(3):19-24

120 郑家恒, 李鑫等. 基于语料库的中文姓名识别方法的研究. 中文信息学报. 2000, 14(1):7-12

121 D. M. Bikel, R. Schwartz and R. M. Weischedel. An Algorithm that Learns What's in a Name, 1999, http://www.cis.upenn.edu/~dbikel/

122 Hideki isozaki, Hideto kazawa. Efficient Support Vector Classifiers for Named Entity Recognition. COLING 2002, http://acl.ldc.upenn.edu/C/C02/C02-1054.pdf

123 孙茂松，黄昌宁等. 中文姓名的自动辨识. 中文信息学报. 1995, 9(2):16-27

124 李建华，王晓龙. 中文人名自动识别的一种有效方法. 高技术通讯. 2002, (2):46-49

125 中国社会科学院语言文字应用研究所. 姓氏人名用字分析统计. 语文出版社，1990

126 付国宏. 汉语句法歧义消解的统计方法研究. 哈尔滨工业大学工学博士学位论文. 2000:52-53

127 余煊，朱凤石. 基于人工神经网络的汉语兼类处理方法的研究. 计算机研究与发展，1998，35(4):367

128 王海峰, 李生, 赵铁军. BT863-Ⅱ 汉英机器翻译系统中的兼类处理方法. 高技术通讯 2000.1

129 T. Joachims. Making Large-Scale SVM Learning Practical. MIT-Press, 1999

130 X. L. Wang, Q. C. Chen. Mining Pinyin-to-Character Conversion Rules From Large-Scal Corpus: A Rough Set Approach. IEEE Transactions on Systems. Man and Cybernetics-part B: cybernetics. 2004, 34(2): 834-843

131 张民，李生，赵铁军. 统计与规则并举的汉语词性自动标注算法. 软件学报. 1998, 9(2):134-138

132 白拴虎. 基于统计的汉语语料库词性自动标注的研究与实现. 清华大学硕士学位论文. 1992

133 Y. A. Wilks, M. Stevenson. The Grammar of Sense: Is Word Sense Tagging Much More Than Part-of-Speech Tagging?. Technical Report CS-96-05, University of Sheffield, United Kingdom, 1996

134 D. K. Lin, Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. 35th Annual Meeting of the Association for Computational Linguistics, 1997:64-71

附录 A　INSUN-LEX词法分析软件输出结果[1]

分词与词性标注结果[2]：

1. 第一/m　位/q　的/u　工作/vn
2. 编者/n　的/u　话/n　：/w　党中央/nt　国务院/nt　最近/t　召开/v　的/u　国有/vn　企业/n　下岗/vn　职工/n　基本/a　生活/vn　保障/vn　和/c　再/d　就业/v　工作/vn　会议/n　，/w
3. 提出/v　要/v　把/p　这项/r　工作/v　作为/p　当前/t　一个/m　头等/b　大事/n　来/f　抓/v　，/w
4. 并/c　做/v　了/u　全面/a　的/u　动员/vn　和/c　部署/vn　，/w
5. 为了/p　配合/v　会议/n　精神/n　的/u　贯彻/vn　落实/vn　，/w
6. 我们/r　将/d　组织/v　一/m　系列/q　报道/v　，/w
7. 多/a　层次/n　、/w　多/m　侧面/f　的/u　宣传/vn　中央/n　精神/n　，/w
8. 报道/v　各地/r　新/a　经验/n　、/w　新/a　做法/n　。/w
9. 今天/t　发表/v　的/u　是/v　第一/m　篇/q　。/w
10. 党中央/nt　国务院/nt　要求/n　，/w
11. 动员/v　全党/n　和/c　全/a　社会/n　的/u　力量/n　，/w
12. 各/r　地区/n　各/r　部门/n　将/p　其/r　当作/v　头等/b　大事/n　，/w
13. 进一步/d　加大/v　了/u　工作/vn　力度/n　，/w
14. 取得/v　了/u　不同/a　程度/n　的/u　进展/vn　。/w
15. 只要/c　在/p　思想/n　上/f　行动/vn　上/f　真正/d　将/p　其/r　当作/v　第一/m　位/q　的/u　工作/vn　抓紧/v　、/w　抓/v　实/a　，/w
16. 就/d　能够/v　抓/v　出/v　成效/n　。/w
17. 称/v　其/r　为/v　“/w　第一/m　位/q　的/u　工作/vn　”/w　，/w
18. 国有/vn　企业/n　的/u　广大/b　职工/n　，/w
19. 几十/m　年/q　来/f　为/v　国家/n　经济/n　建设/vn　、/w　改革/vn　开放/vn　和/c　国有/vn　企业/n　发展/vn　壮大/vn　，/w
20. 作出/v　了/u　重大/a　的/u　贡献/n　。/w
21. 但是/c　我们/r　正在/d　经历/v　由/p　计划经济/n　向/p　市场经济/n

[1] 在2003年10月全国863分词评测中，笔者开发的分词系统F量度为93.14%，在分词组合歧义消解上正确率为83.54%，交叉歧义消解上正确率为91.59%，均为最好成绩。

[2] 人民日报1998年第六个月语料

的/u 根本/a 转变/vn ，/w

22. 由于/c 长期以来/l 重复/vn 建设/vn 、/w 盲目/ad 建设/v 的/u 影响/vn ，/w

名实体识别结果（包含人名和因子词）：

1. 多年来 一直 与 管理所 原 所长 #孙明斋# 、 #金源# 等 保持 频繁 的 书信 来往 ，
2. 苏军 总参谋长 #阿赫罗梅耶夫# 强调 ，
3. 车间 主任 #刘春生# 、 团委 书记 #熊金虎# 走 到 #侯世杰# 跟前 说 ： “ 厂长 放心 吧 ，
4. （ #刘从礼# #董庆九# ） 军民 之间 征文 心 比 枣 儿 #红韩凤# 鹏 / #李明拴# “ 太行 的 枣 儿 ，
5. 副 总参谋长 #韩怀智# 出席 了 座谈会 。
6. 指导员 #沈庆朝# 对 武 大娘 说 ： “ 大娘 ，
7. 新华社 记者 #齐铁砚# 摄 ） 承包 经营 之后 怎么办 ？
8. 他们 是 ： 实验 核物理 专家 #胡仁宇# ，
9. 我国 热能 工程 专家 #倪维斗# ，
10. 半导体 器件 制造 专家 #叶迪生# ，
11. #张建涛# 即 勾结 #张建伟# 、 #冯长伟# 、 #乔国# 诃 等 人 从 窗口 进入 #五号库# ，
12. 新华社 北京 5月31日 电 全国 政协
13. 一九五三年 版 正面 图案 为 『 拖拉机 』 的 棕色 一角券 ；
14. 仅 １９９７年 接待 求职 人员 ８５．４６万 人次 ，
15. 为 ２５．６５万 人 介绍 了 职业 ，
16. １９９１年 和 １９９４年 ，
17. 又 相继 建成 了 荟萃 中国 ２４ 个 民族 的 民间艺术 、风俗人情 、 特色 建筑 的 “ 中国 民俗 文化 村 ”。 受到了 邓 小平 、江泽民 等 党 和 国家 领导人 以及 ５０ 多个 外国 首脑 及 著名 华侨 、 华人 的 高度 评价 。
18. 已 接待 国内外 游客 ４３００ 多 万 人 ，
19. 门票 收入 数十亿 元 ，

附录 B　基于 ME 模型的音字转换结果

1. zhe yi cheng guo yin qi tong hang men de zhong shi ，(这一成果引起同行们的重视，)
2. zhe jiu shi yi ge hen hao de shuo ming 。(这就是一个很好的说明。)
3. zhe xie shen jing yuan yi dan de dao chong fen de gong yang jiu hui chong xin fu huo 。 (这些神经愿意但得到充分的供养就会重新俘获。)
4. qi zhong duo shu shen jing xi bao huo yin shou dao bu tong cheng du de sun shang 、 huo yin de bu dao chong fen de ying yang gong ji er chu yu ban si wang huo xiu mian zhuang tai ， (其中多数神经细胞或因受到不同程度的孙上、或因的不到充分的营养共计二处于半死亡或休眠状态，)
5. bu yi ding quan bu dao zhi shen jing xi bao si wang ，(不一定全部导致神经细胞死亡，)
6. zai bing ren mian qian qie mo bei guan shi wang (在病人面前切万被官是王)
7. ru fu zheng pei ben fa 、huo xie hua yu fa 、qing re jie du fa 、tan qu shi fa 、yi du gong du fa deng 。(如扶郑培本法、活血化淤法、清热戒毒法、谭趋势法、一度攻读法等。)
8. xiao chu he bi mian wai jie zhi ai yin su dui ren ti de wei hai ； (消除和避免外界志哀因素对人体的危害；)
9. yin wei shu cai zhong han you da liang de wei sheng su 、 wei sheng su 、 wei sheng su 。 (因为蔬菜中含有大量的维生素、维生素、维生素。)
10. neng yin qi duo zhong dong wu fa sheng zhong liu (能引起多种动物发生中流)
11. bian hui zai wei nei he cheng ya xiao an ， (便会在位内和成雅小厂，)
12. yi ji yu fang shi tong guo gai bian huan jing yin su ， (以其预防是通过改变环境因素，)
13. ai zheng shi ke yi yu fang he you xiao de kong zhi de ， (爱正是可以预防和有效的控制的，)
14. ren men ying duo zhang wo fang ai de zhi shi ， (人们应多掌握妨碍的知识，)
15. chu ke xiang lin jin qi guan kuo zhan wai ， (除可向临近奇观扩展外，)
16. liang xing zhong liu yi ban zai fa sheng zhong liu de zu zhi ming cheng hou jia shang yi ge liu zi ，(两省肿瘤一般在发生中流的组织名称后家上一个陆子，)

攻读博士学位期间发表的论文

1 赵岩 王晓龙 刘秉权 关毅. 基于矢量空间模型和最大熵模型的词义问题解决策略. 高技术通讯.2004, 15(1):1-6

2 赵岩 王晓龙刘秉权 关毅. 融合聚类触发对特征的最大熵词性标注模型. 计算机研究与发展. (录用待发表)

3 Zhao Y, Wang X L, et al. Applying Class Triggers in Chinese POS Tagging Based on Maximum Entropy Model. In Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 2004:1641-1646(EI 检索)

4 Zhao Y, Wang X L, et al. Query Expansion Using Trigger-based Vector Space Model. In Proceedings of the First International Conference on AIRS, Beijin, 2004:201-204

5 赵岩 王晓龙 徐志明 刘秉权. 利用词性信息改进 Katz 平滑算法.哈尔滨工业大学学报. (录用待发表)

6 Zhao Y, Wang X L. Pinyin-to-Character Conversion Based on Maximum Entropy Markov Model. 电子与信息学报英文版. (在投)

7 Zhao Y, Wang X L. Identification and Resolution of Chinese word segmentation Ambiguity. International Journal of Chinese Language and Computing. (在投)

哈尔滨工业大学博士学位论文原创性声明

本人郑重声明：此处所提交的博士学位论文《基于统计语言模型的汉语词法分析研究》，是本人在导师指导下，在哈尔滨工业大学攻读博士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签字：　　日期：　　年　　月　　日

哈尔滨工业大学博士学位论文使用授权书

《基于统计语言模型的汉语词法分析研究》系本人在哈尔滨工业大学攻读博士学位期间在导师指导下完成的博士学位论文。本论文的研究成果归哈尔滨工业大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权哈尔滨工业大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

保密□，在　　年解密后适用本授权书。

本学位论文属于

不保密□。

（请在以上相应方框内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

致 谢

值此论文完成之际，谨向曾经给予我关心和帮助的老师、同学和亲友表示衷心的感谢。

感谢导师王晓龙教授多年来对我的关心、指导和教诲。作者博士论文的工作是在王老师的直接指导下完成的，王老师渊博的知识、敏捷的思维、平易近人的工作作风使我受益匪浅，是我永远学习的榜样。

感谢李生教授和韩继庆教授对本文的悉心审阅以及对本文提出的宝贵意见。感谢廖明宏教授、王义和教授、王宇颖教授、赵铁军教授对本文提出宝贵意见。

感谢关毅博士、刘秉权博士、徐志明博士、付国宏博士、陈清才博士、林磊博士在学习和工作中给予我特别的关心、指导和建议，我所做的工作离不开他们的帮助。

感谢自然语言处理实验室单丽莉、刘远超等老师以及赵健、宇璎、陈燕敏、徐永东、董启文、江维、王强等同学的关心和帮助，他们为我创造了一个充满自由、朝气和富于创新精神的学习和科研环境。

感谢父母与家人对我多年的教诲和关爱，感谢妻子韩丽娜对我一贯地支持、鼓励和无私奉献。没有他们在生活上的帮助，作者完成博士论文是不可能的。

个人简历

赵岩， 男， 汉族， 1974年5月出生，黑龙江省哈尔滨市人

学历：

1993年9月—1997年7月，哈尔滨理工大学电气与电子工程学院电气技术专业，毕业获学士学位；

1997年9月—1999年7月，哈尔滨工业大学电气工程与自动化学院电器与自动化专业，毕业获硕士学位；

1999年9月—2005年5月，哈尔滨工业大学计算机科学与技术学院应用软件专业，攻读博士学位；

研究方向：

自然语言处理、统计语言模型、词法分析、音字转换

攻读博士学位期间主要参加了以下科研项目：

国家自然科学基金项目“基于大规模语料库的汉语自动聚类研究”；

国家自然科学基金项目“基于粗糙集的大规模语料库语言学知识发现模型研究”；

国家863项目“智能化中文信息处理平台”；

国家863项目“面向奥运智能信息服务的平行语料库加工、文摘、自然语言检索技术研究”；