# 中文自动分词若干关键问题研究

（申请清华大学工学博士学位论文）

培 养 单 位 ：计算机科学与技术系
学　　科 ：计算机科学与技术
研　究　生 ：乔 维
指 导 教 师 ：孙 茂 松 教 授

二〇一〇年六月

# Research on Several Key Issues in Chinese Word Segmentation

Dissertation Submitted to

**Tsinghua University**

in partial fulfillment of the requirement

for the degree of

**Doctor of Engineering**

by

**Qiao Wei**

**( Computer Science and Technology )**

Dissertation Supervisor : Professor Sun Maosong

**June, 2010**

# 关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

**（保密的论文在解密后应遵守此规定）**

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

# 摘 要

中文自动分词是中文信息处理中的经典问题。中文自动分词的任务是由机器在中文文本中自动识别词边界，是中文信息处理的第一道工序。目前，中文分词研究领域的几个关键问题依然没有完全解决：一是语言资源建设；二是中文分词歧义问题；三是未登录词识别。本文围绕这些关键问题开展研究，取得了具有一定理论意义和实用价值的成果。

词频信息在语言资源建设方面有着重要作用，本文提出了一种基于多种类型语料库的中文词频估计方法。该方法综合利用人工切分语料、自动切分语料和生语料在词频估计方面的优势，结合语言学知识给出一种新的中文词频近似策略。从统计观察和实际应用两个角度进行的实验表明，该策略优于现有的词频估计策略。

在中文分词歧义处理方面，基于大规模通用语料库和专业领域语料库对最大交集型歧义字段(MOAS) 的统计特性进行了系统地考察，包括通用语料库抽取的MOAS 在通用语料库和专业领域语料库上的统计特性和专业领域抽取的MOAS 在专业领域语料库上的统计特性。确定了由7,000个高频MOAS 组成的MOAS核心集合，能够稳定覆盖汉语真实文本中42% 以上的交集型歧义字段。基于这一核心集合制定的消歧策略可纠正2% 的交集型歧义切分错误。

在未登录词识别方面，提出了一种基于Web 搜索和有监督机器学习方法相结合的中文分词框架。网络文本弥补了训练语料规模小、覆盖面不够的问题。与有监督的机器学习方法相结合保证了已知词的识别准确率。在公开数据集及未登录词比例较高的测试集上进行的实验表明，该方法能显著提高未登录识别率，从而提高中文自动分词系统的精度和适应性。针对未登录词中的命名实体，提出一种基于最大间隔马尔可夫网络模型（$M^3N$）的中文分词和命名实体识别一体化方法。在公开数据集上进行的实验表明，基于$M^3N$的分词效果优于基于条件随机场模型的分词方法（提高0.3%~2.0%），分词与命名实体一体化方法能同时提高中文分词（1.5%~5.5%）和命名实体识别（5.7%~7.9%）的性能。

**关键词**：中文自动分词；中文词频估计；分词歧义；未登录词识别

# Abstract

Chinese word segmentation(CWS) is one of the classic problems in the field of Chinese information processing. The task of Chinese word segmentation is to automatically identify the word boundary in Chinese running text. It is the first step of Chinese language processing. By now, there are still several key issues which haven't got satisfactory solution: the first one is the construction of language resources; the second one is segmentation ambiguity problem; the third one is out-of-vocabulary word recognition problem. In this thesis, we focus on these aspects and achieve some results with a certain theoretical and practical value.

Word frequency information is very important for language resource construction. In this thesis, a scheme called multi-type corpora based Chinese word frequency approximation is proposed. By taking into account the characteristics of the Chinese language, we jointly use corpora of different type (raw corpora, MM-segmented corpora and manually segmented corpora) to approximate word frequencies in Chinese. Experiments are performed from both statistical and application-oriented perspectives, indicating that the proposed scheme is the most effective one among the approaches under consideration.

For ambiguity problem in CWS, based on a very large balanced general-purpose Chinese corpus and two domain-specific corpora, the statistical properties of Maximal Overlapping Ambiguity String (MOAS) are systematically studied from two aspects: the statistic characteristic of MOAS extracted from general corpora in both general corpora and domain-specific corpora; the statistic characteristic of MOAS extracted from domain-specific corpora in domain-specific corpus. A core set of MOAS which contains 7,000 high frequent MOAS is determined. A disambiguation strategy for overlapping ambiguities is proposed consequently, suggesting that over 42% of overlapping ambiguities in Chinese running text could be covered. Preliminary experiments show that about 2% of pseudo ambiguities which are mistakenly segmented by state-of-art

segmenters can be properly treated by the proposed strategy.

For Out-of-Vocabulary (OOV) problem, a framework which combines using web search technology and machine learning method is proposed. By taking this framework, the problem of lacking knowledge in small sized training data can be solved by incorporating large scaled web data as knowledge supplement while the supervised machine learning model insures high precision on in-vocabulary words. Evaluations performed on both public evaluation datasets and high out-of-vocabulary testset show that it achieves significant improvement on both OOV recall rate and $F_1$-measure. The CWS system turns to be more precise and robust under this scheme. As for named entity which is one of the special types of OOV, we bring forward a $M^3N$-based joint CWS and named entity recognition (NER) scheme in which joint training and testing are performed. Experimental results show that, $M^3N$-based word segmenter outperforms CRFs-based word segmenter with 0.3%~2.0% improvement on final score. Integrated CWS and NER scheme can benefit both of the two tasks. The improvement on CWS is 1.5%~5.5% and is 5.7%~7.9% on NER.

**Key words:** Chinese word segmentation; Chinese word frequency estimation; segmentation ambiguity; Out-of-vocabulary words recognition

# 目 录

# 主要符号对照表

| | |
|---|---|
| CWS | 中文自动分词 (Chinese word segmentation) |
| NER | 命名实体识别 (Named Entity Recognition) |
| $P$, $R$, $F_1$ | 准确率 (Precision)、召回率 (Recall)、综合评价 ($F_1$-measure) |
| SIGHAN | 中文处理专业委员会的简称 |
| OOV | 未登录词 (Out of Vocabulary) |
| IV | 词典中的词 (In Vocabulary) |
| MM | 最大匹配法 (Maximum Matching) |
| MOAS | 最大交集型歧义字段 |
| PMOAS | 最大交集型伪歧义字段 |
| TMOAS | 最大交集型真歧义字段 |
| MOAS_Token | 最大交集型歧义字段的一次出现 |
| PMOAS_Type | 最大交集型歧义字段类型 |
| CRF | 条件随机场 (Conditional random fields) |
| $M^3N$ | 最大间隔马尔可夫网络 (Max-Margin Markov Network) |
| MC | 人工切分语料库 |
| RC | 生语料库 |
| SCRC | 斯皮尔曼秩相关系数 (Spearman Correlation Rank Coefficient) |

# 第 1 章 引言

## 1.1 研究背景及意义

随着以互联网为基础的信息技术的快速发展，越来越多的信息以电子文档的形式存在，世界迈入了一个“信息爆炸”的时代。电子信息的急速增加一方面使得互联网及各种信息媒体成为富含宝藏的人类知识的矿山，而另一方面，也为快速而准确地挖掘并获取有效信息增加了难度。对信息知识的获取及利用水平正日益深刻地影响着人们的生活，也逐渐成为衡量一个国家和民族科技竞争力的重要指标。对于使用人数最多的语言中文而言，近年来中文网页数目急剧增加，中文电子图书及出版物迅速普及。中文信息处理技术对大到国家信息安全、小到百姓的日常生活都有着重要意义。在这一形势下，以非受限文本为主要对象的中文信息处理研究的重要性日益显著。

近年来，中文信息处理技术得到了快速发展，但从实用角度而言还远远达不到人们的需求，构建高效的中文信息处理系统仍是一个亟待解决而且相当困难的任务。中文信息处理的对象是中文文本，对中文而言，“词”是承载语义的最小单元，由词构成语句，再由语句构成篇章。因此，中文信息处理大致包括自下而上的三层：词法层、句法层和语义层。与西方语言不同（如英文）的是，中文文本是由连续的字序列组成的，词与词之间缺少天然的分隔符，因而中文信息处理在词法层比英文等西方语言多一步工序，即识别词边界。本文所研究的中文自动分词[①] 任务，就是由机器在中文文本中自动识别词边界，通俗地说就是要由机器在词与词之间自动加上分隔符，是词法层的一个主要研究内容。

中文分词的任务描述起来是简单清晰的，一个中文分词系统的输入是一个中文句子，如“今天天气晴朗。”，输出是带有分隔符的切分文本“今天 | 天气 | 晴朗 | 。”这里，我们给出的中文分词的形式化描述[1,2]：

**定义 1.1：** 由一个中文字序列$(C_1, C_2, \ldots, C_M)$ 构成的句子$S$。一个由$S$ 中第$i$ 个字

① 文内此后的“中文分词”均指中文自动分词 (automated Chinese word segmentation)。

开始的连续$n$ 个汉字组成的字符标记，也就是中文词，是一个有序的$n$ 元组形式$(C_i, C_{i+1}, \ldots, C_{i+n-1})$。句子$S$ 的划分$W$ 将该句中所有汉字分隔为长度不同、互不交叠的中文词$((C_1, C_2, \ldots, C_{i_1-1}),\ (C_{i_1}, C_{i_1+1}, \ldots, C_{i_2-1}), \ldots, (C_{i_{N_1}}, C_{i_{N_1}+1}, \ldots, C_M))$。我们把$W$中的元素记作$(W_1, W_2, \ldots, W_N)$。中文分词的任务就是找到句子$S$ 的划分$W$。

中文分词处于中文信息处理的底层，是公认的中文信息处理的第一道“工序”，在中文信息处理的诸多应用领域（中文信息检索、文本分类及摘要，中英机器翻译，中文句法分析等）中扮演着极为重要的角色[3–6]。在中文分词上造成的切分错误将向上层传递并逐级放大，对高层应用造成严重影响并难以纠正。因此，中文信息处理对中文分词的精确度和速度要求很高，只有跨越中文分词这一障碍，各种高层的语言分析才能够顺利进行。

中文分词是中文信息处理中的经典问题，自80年代初期开始已历经近二十年的研究，吸引了大量优秀的中外研究机构和学者。国外进行该研究的主要包括斯坦福、麻省理工、新加坡国立、香港中文、台湾中央研究院等知名学府和研究机构。国内进行中文分词研究的高校和研究所主要包括中科院、哈工大、清华和北大等。在工业界，谷歌、百度、雅虎等搜索引擎公司对中文分词技术都非常重视，搭建了自己的中文分词系统。

尽管中文分词在学术界和工业界都得到了广泛的重视，但到目前为止还未有一个经得起考验的，被广泛认可的中文分词系统。与此形成鲜明对照的是，日语同样也存在分词问题，但已经有了较为广泛认同的日语分词系统。目前中文分词领域尚未攻克的关键问题包括以下三点：

1. 中文语言资源建设不够充分。中文分词规范和通用词表的缺失对构建一致性好的大规模中文标注语料库造成很大困扰。

2. 中文自动分词中的切分歧义问题。目前虽然已有针对交集型歧义的探索性研究，但还并不完善。歧义问题仍是困扰中文分词的一大难题。

3. 中文自动分词中的未登录词（Out-of-Vocabulary word，简称OOV）识别问题。OOV 是造成中文分词精度下降的主要因素，国际中文分词评测结果显示，目前未登录词的识别率仅在70%~80%[7–10]。而且当面对未登录词比例更高的开放环境时，这个结果将更差。

上述三个关键问题是中文分词研究的重点和难点，也是本文研究的主要目标，其中切分歧义和未登录词几乎是中文分词研究中一个“永恒”的话题。虽然近二十年的研究离圆满解决还有很大差距，但也取得了诸多可喜的进展，尤其是自2003年7月首届国际中文分词评测活动开展以来，中文分词领域的研究更是取得了长足进步，主要包括：

1. 针对中文“词”的界定不清这一困扰中文分词的问题，近年来，学者们通过在分词规范和词表方面的努力，结合分词语料库，给出了词语可计算的定义，这成为了实现中文自动分词和性能评测的基础。

2. 针对中文自动分词中的切分歧义问题，对歧义的主要类型——交集型歧义，进行了深入地考察，朝着解决歧义问题的方向迈进了一步。

3. 基于统计机器学习的分词方法自90年代初期开始获得了较多关注，为中文分词的研究开辟了新的道路。比起基于人工规则的分词方法，基于统计学习的分词方法将中文分词的性能提高了一个层次。

4. 研究表明，未登录词造成的精度损失比分词歧义大5倍以上[11]，未登录词识别性能逐渐成为衡量一个分词系统性能的主要指标。一个能够大幅提高未登录词识别性能的分词系统往往具有更好的性能。

5. 基于字序列标注、利用分类模型进行分词的方法优于以往的基于词（或词典）的方法，使中文分词系统的精度达到了新的高度。

6. 虽然迄今为止还没有一个公认的中文分词系统，但微软、中科院、海量等机构已经推出了各自的中文分词系统，能够一定程度上服务于中文信息处理的一些高层应用。

综上所述，在中文电子信息的数量急剧增加的大背景下，人们对中文分词技术的需求越来越急切。虽然中文分词的研究极具挑战性的，但二十余年的研究积累为该领域奠定了坚实的基础，同时也为我们呈现了解决该问题的曙光。本文的工作正是基于以上所提出的几个关键问题，进行了一些尝试和探索，希望能为推动这个领域的研究贡献自己的一份力量。

## 1.2 评测数据集及性能评价体系

一个完善的、有效的评价体系对中文分词的研究至关重要。在2003 年以前，

中文分词的评测还局限于国家863 和973 的内部评测，一直缺少一个权威化、公开化的评价体系。

表 1.1 历届SIGHAN中文分词评测数据集一览表[7–10]

| 提供单位 | 语料库 | 编码 | 训练集词次数 | 测试集词次数 | OOV率* |
|---|---|---|---|---|---|
| 台湾中央研究院（AS） | AS2003 | Big5 | 5.8M | 12K | 0.022 |
| | AS2005 | | 5.45M | 122K | 0.043 |
| | AS2006 | | 5.45M | 91K | 0.042 |
| | AS2007 | | 7.22M | 91K | 0.074 |
| 香港城市大学（CityU） | CityU2003 | | 240K | 35K | 0.071 |
| | CityU2005 | | 1.46M | 41K | 0.074 |
| | CityU2006 | | 1.64M | 220K | 0.040 |
| | CityU2007 | | 1.09M | 236K | 0.082 |
| 美国宾州大学（CTB） | CTB2003 | GB | 250K | 40K | 0.181 |
| | CTB2006 | | 508K | 151K | 0.088 |
| | CTB2007 | | 642K | 81K | 0.056 |
| 微软亚洲研究院（MSRA） | MSRA2005 | | 2.37M | 107K | 0.026 |
| | MSRA2006 | | 1.26M | 100K | 0.034 |
| 北京大学（PKU） | PKU2003 | | 1.1M | 17K | 0.069 |
| | PKU2005 | | 1.1M | 104K | 0.058 |
| 中国国家语委（NCC） | NCC2007 | | 913K | 152K | 0.047 |
| 山西大学（SXU） | SXU2007 | | 528K | 114K | 0.051 |

*：在中文分词中，OOV 泛指新词、专名等未登录词。这里的OOV 是指在一种语料库的测试集中出现，但未在其对应的训练集中出现的词。OOV 率是指测试集的未登录词出现次数在该测试集总词次数中所占的比率。显然，通常来讲OOV 率越高的测试集，其切分的难度也越高。

2003年，首届国际中文自动分词评测于日本札幌举行[7]。国际中文自动分词评测（简称SIGHAN评测）① 不再采用一个单一的数据集，而是采用多个由不同机构提供的数据集进行评测。每个机构提供的数据集都包括训练语料、测试语料和标准答案，参评者可以自由选择一种或多种数据集进行评测。SIGHAN 采用多数据集评测的原因就是分词规范短期内难以确立，使用唯一的数据集难免造成偏颇。在几个不同的数据集上进行评测不失为一个折中的解决方式。而且

① SIGHAN 是国际计算语言学会(ACL)下属的“中文处理专业委员会”的简称。

训练语料本身就体现了该数据集的分词规范和标准，绕开了分词标准不一致的问题。

SIGHAN 评测包括两种可选评测方式：开放测试和封闭测试。封闭测试要求参评者只允许使用与测试语料出处相同的训练语料，不可使用其它任何资料。开放测试则不受限于这一约束，可以使用任意其它语料库、拥有专利权之词典以及全球互联网等语料。

SIGHAN 评测为中文分词研究提供了统一的性能比较平台，大大提高了中文分词系统性能的可比性。解决了之前中文分词各家评各家，难做公开的横向比较等问题。这一评测体系的建立对之后的中文分词研究起到了重要的推动作用，在这之后进行的中文分词研究大都采用SIGHAN 评测数据集进行对比实验。

SIGHAN 评测自2003年开始以来至今已经举办了四届：2003年在日本札幌举行的第一届Bakeoff，目前已经成为了中文分词性能评价的一项重要指标；2005年在韩国济州岛举行的第二届Bakeoff[9]，评测结果充分展示了中文分词领域研究中的一些重要进展和发展动向；2006年在澳大利亚悉尼举行的第三届Bakeoff[8]，这次评测在前两届的基础上加入了中文命名实体识别评测；2007年在印度举行了第四届Bakeoff[10]，本次评测项目除了中文自动分词和中文命名实体识别外，还增加了中文词性标注的评测。

下面，我们将介绍SIGHAN 数据集和中文分词的性能评价指标。

### 1.2.1 国际中文自动分词评测数据集

迄今为止，四届SIGHAN评测共计提供了17个不同的中文分词评测数据集。表 1.1给出了17个中文分词评测数据集的统计信息。

### 1.2.2 中文自动分词评价指标

中文分词的评价体系包括精度和速度两个方面。在中文分词精度的评价上，SIGHAN评测采用五个指标衡量中文分词系统的精度：整体的准确率、召回率、综合评价值以及未登录词的召回率和已知词的召回率。下面给出这五种评价指标的形式化计算式。

### 准确率 (Precision) 和召回率 (Recall)

在中文分词中，准确率是切分正确的词数与切分出来的总词数的比值。显然，它代表了分词系统的准确性。召回率是指切分正确的词数与标准答案中切分出来的词数的比值。显然，它代表了分词系统识别词边界的广度。下面给出准确率和召回率的形式化计算式：

$$Precision = \frac{\text{正确切分出的词的数目}}{\text{切分出的词的总数}}, \tag{1-1}$$

$$Recall = \frac{\text{正确切分出的词的数目}}{\text{应该切分出的词的总数}}. \tag{1-2}$$

### 综合性能指标$F_1$-measure

准确率和召回率一般成反比的关系。通过某些方法提高准确率，会导致召回率的下降，反之亦然。为了均衡考虑两方面的因素，给出分词系统的综合性能评价。研究者采用调和平均综合两个指标，即$F$-measure。$F$-measure是一个综合评价，反应了准确率和召回率的一个折中，一个分词系统只有在Precision和recall都得到提高，综合评价$F$-measure才会升高，否则单方面提高某一个评价值而伤害到另一个，最终的评价指标将不会提升，反而有可能下降。下面给出$F$-measure的形式化计算式：

$$F_\beta = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{1-3}$$

上式中的$\beta$为权重因子。如果将准确率和召回率同等看待，取$\beta = 1$，即得到最常用的 $F_1$-measure：

$$F_1 - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{1-4}$$

#### 未登录词召回率(R_OOV) 和词典中词的召回率(R_IV)

SIGHAN评测还包括专门针对未登录识别效果的未登录词召回率(R_OOV)和在词典中的词的召回率(R_IV) 两个指标：

$$R_{OOV} = \frac{\text{正确切分出的未登录词的数目}}{\text{标准答案中未知词的总数}}, \tag{1-5}$$

$$R_{IV} = \frac{\text{正确切分出的已知词的数目}}{\text{标准答案中已知词的总数}}. \tag{1-6}$$

在中文自动分词的速度上，一般以每秒钟处理的文本的规模(KB/s或M/s)来评价。目前学术界的研究大都集中在对中文分词精度的提高上。SIGHAN评测也主要考察分词的精度，对速度未做考察。本文的工作也是着眼于提高中文分词精度进行的。在工业界，如搜索引擎公司，对中文分词系统的速度要求很高，大约需要达到1M/s以上的处理速度才能满足搜索引擎实时搜索的需求。

## 1.3 中文自动分词研究中的关键问题

本节详细描述在第 1.1节中提出的中文分词研究中的三个关键问题：语言资源建设、切分歧义的消解和未登录词识别。

### 1.3.1 语言资源建设

构建中文分词系统离不开语言资源的支持，语言资源建设是中文分词研究中不可或缺的基础环节。这里提到的语言资源包括：中文分词规范及通用词表、已进行切分并带有词性标注的语料库（简称标注语料库，也称熟语料）和大规模未经处理的汉语真实文本（也称生语料）。

一个合理、可实施性强的中文分词规范是保证语料库切分一致性的重要前提，而中文通用词表是分词规范的一个重要组成部分。在中文信息处理中，各种基于词典的中文分词方法以及中文分词歧义的静态分布等，都很大程度上取决于所采用的中文词表。对于大规模标注语料库而言，其重要作用不言而喻。词频信息、统计模型（如隐马尔可夫模型）的参数估计以及各种基于词的中文分词模型都需要大规模标注语料库的支持。同时，标注语料库也是中文分词系统性能

评价中不可缺少的资源。大规模生语料是以上各种语言资源建设的基础，各种语言资源都需要在其基础上进行构建。另外，在信息处理中，一些全局统计量的获得需要依赖极大规模生语料库（如互信息、串频信息等）。

在分词规范方面，我国从1988年开始着手于《信息处理用现代汉语分词规范》的国家标准的制定，并于1993年出台了初步的分词规范——《信息处理用现代汉语分词规范》。但由于中文语言现象极为复杂，很难有不出现例外的规则，这个分词规范提出了“结合紧密，使用稳定”的原则作为判定是否可以作为分词单位的准则。但是，显然这个原则不够具体，可实施性较差，实行起来往往见仁见智。虽然有了一个分词规范，但仍然解决不了不同系统中分词单位不一致的问题。

在中文词表方面，我国研制了《信息处理用现代汉语常用词表》。在1993年分词规范发布后，有学者建议在规范之外，根据规范制定一个词表，采用“规范+词表”的策略，以有利于对规范的说明和实施。1994年，该规范的主要制定者刘源教授等人根据现代汉语词频统计的结果，公布了《信息处理用现代汉语常用词表》，该词表共计包含43,570个词。然而，受限于分词规范固有的问题，这个词表对于规范中存在的一些难点，仍然没有作出令人满意的处理，很难形成公认的词表。

在标注语料库建设方面，国家语言文字工作委员会于1991年开始着手建立国家级大型中文语料库，规模计划达7,000万汉字，为了加工这个国家级语料库，国家设立了社科重大项目“信息处理用现代汉语词汇研究”，其中一个子课题就是信息处理用现代汉语分词词表，目前这项工作还在进行。

在大规模生语料库的建设方面，自1992年以来，各大研究中文信息处理的单位，包括《人民日报》光盘数据库，北京大学计算语言学研究所，北京语言文化大学，清华大学等，都建立了自己的语料库，并在各自语料库基础上进行各种加工。但由于这些语料库建立的目的不同（有些为研究切分歧义，而有些是为了进行短语结构分析），且加工深度不同、标准也不同，这些语料库很难合并起来构建一个各方公认的语料库。

与国内情况形成对比，国外的英文语言资源建设取得了很大成就。由美国Brown大学建立了BROWN语料库（布朗语料库），由英国Lancaster大学、挪威Oslo大学与Bergen大学联合建立的LOB语料库以及美国宾州大学建立的LDC

（Linguistic data Consortium）语料库。其的规模和质量受到学术界的广泛认可，成为欧美研究者公认的研究资源和平台。欧美各国学者利用这两个语料库开展了大规模的研究。大大推进了英文信息处理的研究工作。

综上，国内在语料库的建设方面作了大量的工作，但是关于分词规范、中文通用词表以及标注语料库依然未能形成像欧美国家那样公认的语言资源。目前在语言资源建设上面临几个非常困难的问题，这些困难都源自中文语言学研究中的一些“经典”问题，比如词、语素及短语的界限不清晰，词类划分体系以及词的归类没有一个可依据的标准等。这些悬而未决的问题给中文语言资源的建设造成极大的障碍。以分词规范的建立为例，目前虽然已经有了国家标准[12,13]，甚至有的单位也制定了自己的规范[14,15]，但这些规范的可操作性都不够强，据之构造出一致性好的中文通用词表和标注语料库具有很大的难度[16]。

### 1.3.2　中文自动分词中的歧义问题

中文自动分词中歧义问题① 在中文分词里是一个重点和难点。首先举两个例子说明中文分词中的歧义现象。例句1a和例句1b给出了三字串“其次要”在不同上下文环境下的不同切分形式。例句2a和2b给出了“火把”在不同上下文环境下的不同切分形式。

【例1】“其次要”

1a.先解决其主要问题，再解决其次要问题；

1b.首先要关注整体，其次要注意细节。

【例2】“火把”

2a.天色暗下来，人们举起了火把。

2b.这场突如其来的大火把村庄烧成了废墟。

中文分词中歧义产生的原因主要有：

第一，大多数中文字（汉字）可以在不同词里的不同位置上出现，虽然中文汉字的个数是有限的，但是由于其出现位置非常自由，我们无法通过其出现位置确定词边界。例如例1中的“次”字，它可以是“其次”的右边界，也可以是“次要”的左边界。

① 本文后续均简称“中文分词歧义”

第二，某些汉字在某种上下文环境中，作为一个复合词的组成部分出现，而在其它一些情况下，这个汉字本身又可独立成词（单字词）。例如例2 中的“火”字。

第三，许多汉字有不唯一的发音，一个具有多种发音的汉字虽然字形相同，在意义上却相差甚大，不同的发音决定了这个汉字在词里的从属关系，从而为词边界的确定带来了困扰。

人脑之所以能够判别例2中“大火把”的切分形式应为“大火 | 把”而不是“大 | 火把”，是因为人类大脑有一个庞大的“语言知识库”作支撑，根据上下文的语境和语言“常识”判别出正确的切分形式。但是，由于语言现象非常复杂，很难通过制定一系列的人工规则让机器习得人脑具备的“语言知识库”，所以这个看似简单的任务对机器来说是一个非常困难的问题。

### 中文分词歧义的定义及基本类型

最早对中文分词歧义进行系统考察的是梁南元，他定义了两种基本的切分歧义类型[17]：

**定义** 1.2：【交集型切分歧义】汉字串$S = s_1, s_2, s_3$ 被称作交集型切分歧义，如果满足$s_1, s_2$，$s_2, s_3$ 同时为词（$s_1$、$s_2$和$s_3$均为汉字串）。汉字串$S$ 被称作交集串。

**定义** 1.3：【多义组合型切分歧义】汉字串$S = s_1, s_2$ 被称作多义组合型切分歧义，如果满足$s_1, s_2$，$s_1$以及$s_2$同时为词。

文献[18] 指出关于多义组合型切分歧义的定义中存在容易混淆的地方，对其定义进行了补充：

**定义** 1.4：【多义组合型切分歧义*】汉字串$S = s_1\ s_2$ 被称为多义组合型切分歧义，如果满足：（1）$s_1, s_2$，$s_1$以及$s_2$同时为词；（2）中文文本中至少存在一个前后语境C，在C的约束下，$s_1$以及$s_2$在语法和语义上都成立。

孙茂松、黄昌宁等认为，关于“多义组合型切分歧义”改为“覆盖型切分歧义”更为恰当[19]。本文后续部分均采用“覆盖型切分歧义”这一名称描述类似例2所示的中文分词歧义。

综上，中文分词歧义的类型主要分为两种：交集型切分歧义（简称交集型歧义）和覆盖型切分歧义（简称覆盖型歧义）。孙和邹的工作给出了更为详细的关于中文分词歧义类型的介绍[20]，这里不再赘述。

孙茂松、左正平将中文分词歧义里的“交集型歧义”进一步分为“真歧义”和“伪歧义”[21]，真歧义是指没有上下文环境时无法判别其切分形式的歧义现象，这个上下文环境可以是发生歧义的字串附近的文本（局部信息），也可以是跨越到句子以外的文本（全局信息）。

例1的“其次要”就属于真歧义，但判别它只需要局部的上下文信息。

这里再给出一个需要全局信息的真歧义的例子:

【例3】乒乓球拍卖完了。

在这一句话里，对“乒乓球拍卖完了”的两种可能切分:“乒乓 | 球拍 | 卖 | 完 | 了”和“乒乓球 | 拍卖 | 完 | 了”，我们只能从句子以外的篇章里寻找答案。

通过上述对歧义现象的产生原因以及对歧义类型的描述，对中文分词歧义有了大概的认识。下面我们将介绍现有的中文分词歧义的处理方法。

### 中文分词歧义的检测和消解

对中文分词歧义的处理从逻辑上可分为切分歧义的检测和切分歧义的消解两部分[20]。已有的处理方法大致来讲可分为基于词典的、基于统计的以及其它一些混合的方法。

首先介绍基于词典的歧义检测方法：

最早出现、也是最基本的中文分词方法——“最大匹配分词法（Maximum Matching，简称MM）”，也称为“最长词优先匹配法”，是典型的具备歧义检测能力的中文分词方法。

刘源、梁南元首次将MM应用到中文分词系统中[22]。MM的基本思想是给定一个词典，从句子的一端开始扫描，尽可能地匹配词典中最长的词，如此循环下去直至扫描至句子的另一端。MM算法详细的工作原理的形式化解释见文献[1]。最大匹配分词方法具有实现简单、时间复杂度低的优点[23]。

由MM的工作原理可知，MM实际上是一种将切分歧义的检测与消解两个步骤同时进行的中文分词方法，对有歧义的句子只输出唯一的切分结果。

根据扫描方向的不同，最大匹配分词法可以进一步分为正向最大匹配法（Forward MM，简称FMM）和逆向最大匹配法（Backward MM，简称BMM）。从FMM和BMM的工作原理可知，通过比较FMM和BMM的输出结果即可检测出歧义现象的发生。例如，假设“其次”和“次要”都是事先给定的词典里的词。那么FMM将把“其次要”切分为“其次|要”，而BMM将切分为“其|次要”。文献[18]指出，对90.0% 左右的中文句子，FMM和BMM的切分结果完全相同且与标准答案一致。有9.0% 左右的句子FMM和BMM 给出的切分结果不同，但其中有一个是正确答案，这种情况即是成功检测到歧义的地方。只有不到1.0% 的句子FMM和BMM的切分结果一致但结果是错误的，或者切分结果虽然不同但两种切分均是错误的，这种情况即是歧义检测失败。因此，这种基于双向最大匹配分词法在歧义检测上是存在“盲区”的。

关于基于词典的中文分词歧义的检测还有两个相关的工作：

一是“最少分词法”[24]，这种分词方法在歧义检测效果上较双向最大匹配法稍好，能够产生的可能的切分候选个数略有增加。

另一个是“全切分法”[25]，全切分法顾名思义，是穷举所有可能的切分形式，这个方法能够实现完备的切分歧义检测，但是同时也引入了大量的“噪音”。

在中文分词歧义消解方面，包括神经网络[26]、有限状态机[27]、隐马尔可夫模型[28,29] 在内的大量人工智能领域的算法和模型都曾被尝试用于歧义消解问题上。随着近年来统计机器学习方法的兴盛，有监督和无监督的机器学习方法也被广泛应用于中文分词歧义的处理上（这些方法并不针对歧义消解进行，但方法本身已具备了对歧义的处理能力）。从对语言知识的利用由简单到复杂的演变来看，这些方法经历了一个由浅入深的过程，这体现在所利用的语言信息的复杂性上：从最早期的利用词频及语素（如松弛法等）[25,30]，到运用音节信息[31]，到利用词法和句法规则（包括扩充转移网络[32]、短语结构文法、专家系统[33]、“Brill式转换法”[34]等），再到利用句法统计[19,28]，甚至运用更为复杂的句法分析[5]。

可以看到，在文献[5]之后，使用句法层信息进行歧义消解的工作鲜有出现。原因是，相对于词法层句法分析显然更为困难。用一个更为困难的任务解决一个相对简单的任务存在着矛盾。况且，在可预期的将来，鲁棒的句法分析器几乎没有实现的可能。

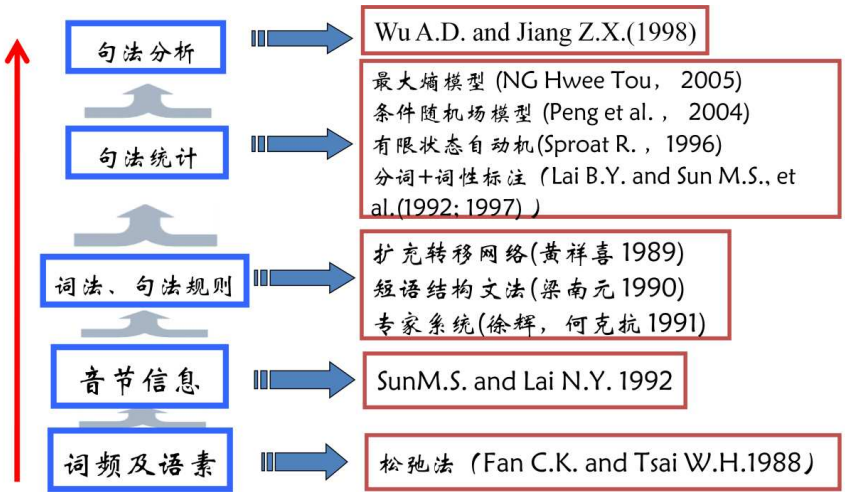



图 1.1 中文分词歧义消解方法一览

随着统计机器学习方法在中文信息处理中的应用越来越广泛，包括隐马尔可夫模型在语音识别中的应用、支持向量机和最近邻法在文本分类中的应用等，后续的研究集中在统计机器学习方法在中文分词的应用上，取得了许多有价值的研究成果，这包括：有监督机器学习方法的利用[35–38]，无监督机器学习方法的使用[39,40]以及一些混合使用基于统计和基于规则的方法[29,39,41]。

按照所使用的语言信息和方法的不同，图 1.1对二十余年来所出现的中文歧义消解方法进行了归类。

利用的语言信息越多，方法本身对标注语料库的质量和规模等的要求就越高（如进行句法分析就需要高质量树库的支持），同时，算法本身的时间和空间复杂度也越高。已有的研究表明，使用的语言信息越复杂、越深入，并不一定带来更好的歧义消解效果。例如，近年来出现的基于字序列标注的中文分词方法仅使用简单的字信息就可以取得比使用句法分析更好的分词效果。然而，对于某些歧义现象，句法等更高层的语言信息显然对歧义消解有着积极的、不可替代的作用。目前，对于中文分词歧义的消解是否需要利用更为复杂的句法信息目前还是学术界悬而未决的问题。

按照歧义类型的不同，歧义的检测消解方法也有所不同。在两类中文分词歧义——交集型歧义和覆盖型歧义中，交集型歧义占所有歧义的90%以上[17]，覆盖型歧义所占比例较小。迄今为止，对中文分词歧义的研究较多集中在交集型歧义上。其中，值得一提的工作是，孙茂松、左正平等基于一个1亿字的生语料，

对交集型歧义进行了统计观察，并提取了4,619个高频交集型歧义字段。作者发现这些高频交集型歧义字段具有很好的覆盖率。并在此基础上提出了基于个例的交集型歧义消解策略[42]。文献[39] 将孙和左的工作以预处理的方式应用到所建立的分词模型中，实验结果表明孙和左的工作对交集型歧义的消解有很好的效果。但文献[42] 和文献[39]的工作均是基于新闻语料得出的结论，而且语料库规模相对较小，得出的结论需要在更大规模的平衡语料上做进一步地论证。另外上述工作均是在通用语料库上的观察，目前还缺少基于专业领域语料库的考察，通用语料库得到的高频交际型歧义字段在专业领域上的特性是否会发生变化还有待进一步考察。本论文第三章的工作便是基于孙和左的工作，在更大规模、平衡性更好的通用语料库上对这个结论进行了验证，给出了具有稳定覆盖率的高频7,000个交集型歧义[43]，并且基于两个专业领域语料库对这部分高频交集型歧义字段在专业领域语料库上的统计特性进行了深入考察。并且对从专业领域语料中抽取的交集型歧义字段的统计分布及真、伪歧义的变化情况进行了考察[44]。

上面阐述了已有的交集型歧义的消解方法，对于大多数交集型歧义，可以根据交集型歧义字段本身的信息或使用句法为主的局部上下文信息进行消歧。而对于覆盖型歧义，往往需要更大范围的上下文或更高层的语义知识进行消歧。已有的工作主要依靠一些人工规则进行排歧，另外还有一些利用语境、词义等进行消歧的方法，如曲等人提出了一种基于语境信息的覆盖型歧义消解方法，利用“相对词频”的概念和歧义字段前后语境信息对覆盖型分词歧义进行消解[45]。肖和孙等人提出一种将覆盖型歧义消歧问题看作词义消歧的等价问题，借鉴词义消歧取得的成果应用于覆盖歧义消解[46]。对于覆盖型歧义，其切分形式无一例外地完全取决于上下文环境，不存在可以直接确定切分形式的覆盖型歧义。因此，与对交集型歧义不同，对覆盖型歧义的统计观察结论很难达到有效消除歧义的作用。受以上因素的制约，目前对于覆盖型歧义的研究和统计观察都相对比较薄弱。

### 1.3.3 中文自动分词中的未登录词问题

中文分词中基于词典的方法需要一个事先定义好的词典。词典是一个由有限个中文词构成的列表，是一个封闭的静态系统，但是新词的产生是动态的。在

词典确定后新出现的，无法被词典所涵盖的词被称为未登录词（OOV）。这些未登录词给中文信息处理中基于词典的应用（如信息检索[47]，中文分词等）带来很多困扰，在中文分词研究中，称对未登录词的检测和识别为中文分词中的未登录词问题。

统计数据表明，每年有超过1,000个新的中文词产生。这些新词有很多是具有时效性和领域特性的，是在某段特殊时间针对某个领域产生的，如“蓝筹股”、“非典”、“海归”等。未登录词不断产生且难以识别的原因是，几乎所有的中文字都是一个词素或者一个词，而且大多数词素是一词多义的，通过词素或者词的简单拼接即可形成一个新词。从语言学构词法的角度来讲，未登录词的产生可分大致为两种成因：复合成词和加缀成词[48]。复合词是指由已有的几个词或词素通过各种中文复合词构词法构成的新词。比如“网民”是由“网”和“民”两个名词组合成词的。加缀成词是指一个词干与一个前缀或者后缀词素相结合构成的新词。

由上述分析可知，中文新词的创造非常随意和简单，加上中文一词多义，很难通过提炼人工规则对其进行识别。在下面的小节里，将逐一阐述未登录词的类型、对中文分词精度的影响以及现有的未登录词识别方法。

### 未登录词的分类

未登录词有多种分类体系，文献[49]中将未登录词分为五类：

第一类和第二类分别是由未登录词的两个产生原因定义的两类未登录词类型：复合词和派生词。

第三类：缩略语。如“世博会（世界博览会）”、“奥运会（奥林匹克运动会）”、“西交大（西安交通大学）”等。

第四类：专有名词（也称命名实体）。一般而言，专有名词进一步分为三个常见子类别——人名、地名和机构名。

第五类：数字类复合词。这类复合词是由数字和汉字共同构成的。如日期、地址、时间等。

与上述分类相似，文献[41]将未登录词分为四类。较上面的五分类体系，作者将一般的复合词和缩略语统一归为“其它未登录词”。另外，将数字类复合词和其它与英文混用的未登录词统一称为虚构词（factoid）。文章给出了这四类未

表 1.2 使用FMM在SIGHAN_2003上测得的未登录词对分词性能的影响

| 语料库 | 顶线$F_{top}$ | 基线$F_{base}$ | $F_{top} - F_{base}$ | $1 - F_{top}$ | 比率 |
|---|---|---|---|---|---|
| AS2003 | 0.992 | 0.915 | 0.077 | 0.008 | 9.6 |
| CITYU2003 | 0.989 | 0.867 | 0.122 | 0.011 | 11.1 |
| CTB2003 | 0.985 | 0.725 | 0.260 | 0.015 | 17.3 |
| PKU2003 | 0.995 | 0.867 | 0.128 | 0.005 | 25.6 |

登录词在中文文本中所占比例：派生词4%、命名实体24%、虚构词41%、其它类31%。

文献[20]根据未登录词的可否预知将其分为两大类, 第一类是新涌现的通用词或专业术语等，如“非典”这样的在特定的时间段产生的新词。这种未登录词在“理论上”是可预期的，可以预先加入词表中（实际很难实现）。第二类是专有名词（命名实体），如人名、译名、地名、机构名等。这种未登录词是不可预期的，无论词表的规模多么庞大，也无法涵盖所有的专有名词。

### 未登录词对分词精度的影响

文献[18]指出,“真实文本中（即便是大众通用领域），未登录词对分词精度的影响超过了歧义切分”。为了检测未登录词对分词精度的影响，国际中文分词评测一般用FMM对语料库进行带有未登录词的基线（Baseline）和不含未登录词的顶线（Topline）两种切分。其中，$F_{base}$ 和$F_{top}$ 分别表示基线和顶线的$F_{measure}$ 值。文献[50] 用$(F_{base} - F_{top})$ 表示未登录词单独给分词系统带来的精度失落。文献[11] 进一步使用$(1 - Ftop)$ 表示分词歧义单独造成的分词精度失落，在2003与2005年国际中文分词评测数据集上系统地对未登录词在造成精度损失的因素里所占比例进行了定量评估，表1.2给出了SIGHAN_2003 提供的四个语料库不含和含有未登录词时的FMM分词性能对比。

数据显示，在2003年SIGHAN评测给出的四个语料库中，未登录词造成的分词精度失落比歧义切分造成的精度失落至少大10倍左右（比率范围为9.6 ~ 25.6）。

SIGHAN_2005的四个语料库的测试结果，见表 1.3，由SIGHAN_2005的四个语料库统计数字可知，未登录词造成的分词精度失落比歧义切分造成的精度失落大5.6 ~ 14.2倍之间。

表 1.3 使用FMM在SIGHAN_2003上测得的未登录词对分词性能的影响

| 语料库 | 顶线$F_{top}$ | 基线$F_{base}$ | $F_{top}-F_{base}$ | $1-F_{top}$ | 比率 |
|---|---|---|---|---|---|
| AS2005 | 0.982 | 0.882 | 0.100 | 0.018 | 5.6 |
| CITYU2005 | 0.989 | 0.833 | 0.156 | 0.011 | 14.2 |
| CTB2005 | 0.991 | 0.933 | 0.058 | 0.009 | 6.4 |
| PKU2005 | 0.987 | 0.869 | 0.118 | 0.013 | 9.1 |

综上，未登录词是影响中文分词精度的主要因素，提高未登录词识别率对中文分词系统性能的提升具有重要作用。接下来，我们将介绍现有的针对未登录词的处理方法。

### 未登录词的识别

在介绍未登录词识别方法方面，我们采用孙等提出的粗分类体系——1）新涌现的通用词或专业术语等可预期的未登录词；2）专有名词等不可预期的未登录词。这两类未登录词的识别在处理方法上有较大差别，下面我们将逐一介绍。

【针对第一类未登录词的处理方法】

对第一种“可预期”的未登录词，已有的识别方法大多是基于极大规模中文真实文本中提炼出的$n$ 元汉字串（$n \geq 2$）的分布，这些方法的区别在于使用的语言信息的种类不同：

1、字串频、词频及相对词频

频度是常见的用于检测未登录词，抽取未登录词串候选集合的信息量。一个高频字串或者同现频率较高的字串经常被作为未登录词的候选。但是对于本身出现频度较低的未登录词，这种方法就会失效。刘挺、吴岩等的工作利用了相对简单的字串频信息[51]，利用串频的高低判断未登录词。

2、基于汉字“粘着力”的一类方法

用于定量估计“粘着力”的典型统计量是互信息，研究者通过互信息判断两个汉字之间的结合力，进而判断一个字串是否构成词[52,53]。类似的统计量是$T$-测试差和$\chi^2$–统计量，文献[54] 提出了汉字间$T$-测试差的概念作为互信息的有益补充。文献[55]引入了“四分联立表”及检验联立表独立性的$\chi^2$–统计量，对长度为2、3和4字的任意汉字串做内部关联性分析，获得候选词表。除此之外，使用左右熵[56]和语境依赖[57]也可归入这类方法中。

上述方法均是通过衡量当前字串与其临近的上下文的“粘着”程度来识别词边界。

3、构词形态特性

另外一种判断未登录词的方式是利用中文的构词形态特性。已有的方法包括使用字在词内部（非词边界）出现的概率以及构词模式等信息量等。Chen 等使用字在词内部出现的概率[58]，对初始分词结果中产生的相邻的单字词，如果它们在词内部出现的概率高于设定的阈值，则对其进行合并。Wu 等及Fu等除了使用字在词内出现概率以外，还使用了构词模式信息[59,60]，用以刻画了一个字在词的某个特定位置出现的概率。

4、独立成词概率及停用词表

对于由多个字组成的未登录词，早期的许多研究者试图用下面的方法进行识别[59]：在进行初步的分词和命名实体识别后，若文本中存在连续几个单字散落的现象（既不互相粘连也不与临近的词粘连），那么这些单字连接起来构成词的概率很大。基于这一假设，研究者们做了许多深入的研究。但是，并不是所有这样的单字串组合起来都能成词。只有在一个连续的单字串中，不存在能够构成词典中已有词的字串，这个连续的单子串才是作为未登录词的合适候选。为了滤除不合格的候选串，研究者引入了独立成词概率和停用词表[61,62]：

独立成词概率是一个刻画单字或字串成词倾向性的信息量。具体来讲，单个字独立成词概率是指单个汉字在文本中作为单字词的似然度，其数学定义为：

$$WP(C) = \frac{N(W_C)}{N(C)}, \tag{1-7}$$

其中，$C$为一个中文字符，$N(W_C)$是该字在给定的文本里以单字词形式出现的次数。$N(C)$为该字在文本里的字频。显然，两者的比值刻画了$C$在文本中独立成词的概率。

停用词是指一些单字介词、副词和连词等。显然，候选字串中包含这些停用词的应该被滤除。文献[63]提出使用垃圾串检测的方法进行候选集合的过滤。作者设计了几种不同的过滤机制，从含垃圾串的候选集合中分离出真正的未登录词。从语料库中自动学习出垃圾串的识别特征，如垃圾串的首尾标识、常用垃圾词等。

上面提到的四类统计量都是依赖于极大规模语料库的，我们称之为使用全

局统计量进行未登录词识别的方法。

【针对第二类未登录词的处理方法】

对第二类未登录词，其识别方法与第一类未登录词有较大不同。这是因为专有名词的构词方式较第一类未登录词有更多规律可循。一些前缀和后缀对判断这类专有名词有一定的帮助，比如中文的常用姓氏是有限的（常用的大约为300个），这些前缀为人名识别提供了很大的帮助。对于地名而言，常利用“市”、“区”等后缀帮助判断。机构名则有“公司”、“所”、“部”等后缀信息帮助识别。

专有名词识别方法通常是：首先从各类专有名词库中统计出姓氏表、地名表等信息，并通过制定一些人工规则，给出专有名词构词上的结构规则。根据这些信息提取可能成为专有名词的候选汉字串，利用前面所列举的有标识意义的局部上下文信息以及其它全局和局部统计量对候选集合进行过滤，达到识别专有名词的目的。由于专有名词附近的文本为识别专名提供了很大帮助，针对专名识别的方法大都依赖专名附近的上下文信息，我们称之为使用局部统计量①的未登录词识别方法。沈达阳等特别强调了局部统计量在未登录词处理中的价值[64]。

由于未登录词的种类较多，且每类专名的识别标识差别也很大，因此，有一些工作并不寻求一种统一的解决方案，而是集中于特定种类的未登录词进行深入研究（在信息抽取研究领域，专名识别本身就是单独进行的一项任务）。这类工作涉及以下四种常见的专有名词的识别：

1、中文人名的识别

中文人名由单字或二字的姓和名组成，根据人民日报语料库上的统计结果，中文人名占左右未登录词的四分之一。是专名识别中关键的一类。已有的研究方法大多采用姓名的局部上下文作为识别依据。从标注语料库中统计“中文姓氏表”、“单字名表”、“双字名的起始字表”、“姓名前缀表”等信息，这些表及其统计信息（如频度）被用于构建基于人工规则[65–68]或模式匹配[69]的人名识别系统。在基于统计的人名识别系统里[41,70,71]，这些信息则作为训练特征使用。

2、外国译名的识别

① 局部统计量是相对全局统计量而言的，是指从当前文章得到且其有效范围一般仅限于该文章的统计量（通常为字串频）。

外国译名识别相对中文人名困难[72]，因为它可以由不同长度的字串构成，且一般是音译过来。但是还是有一些规律可循，比如英文译名用字往往是那些没有具体意义的汉字“尔”、“姆”、“斯”等。一些研究者从发音着手，使用音素（phoneme）信息[73,74]，或使用中英文平行语料库提取信息[75,76]。

3、中国地名的识别

地名识别最简单的方法是建立地名册，通过查词典的方式，结合语料库的统计信息进行识别[77]。如“中国地名集”、“中华人民共和国地名词典”、“中国古今地名大词典”等。虽然已经有如此多的地名信息，已有的研究表明在真实文本处理中还是有30%左右的地名未被涵盖。近来基于统计机器学习的地名识别方法被应用于地名的识别，文献[78]提出了一种先使用N-gram进行地名候选检测，然后使用最大熵模型进行挑选的方法。该方法未使用地名词典，在最大熵模型训练特征选取上使用了语义概念。其报告的识别效果为封闭测试88.49%，开放测试84.19%。

4、机构名的识别

机构名的识别比起前面几类专名更为困难[79]，这是因为机构名长度不定，其中可能存在缩略语或其它命名实体。对机构名的研究大多致力于提取判断机构名的特征，如机构名左边界等。还有一些工作是基于机构名的内部构词模式，比如“专有名词+机构名类型”，通过人工规则进行匹配识别。此外，机构名的构词特点决定了它不可能含有一些特定的“停用词”，如“其它”、“失败”等。研究者基于上述信息提出了一系列基于规则[80–82]和基于统计的机构名识别方法[83,84]。

SIGHAN评测自2006年开始增加了命名实体识别评测（主要针对人名、地名、机构名），SIGHAN命名实体测试报告的最好结果总结如下：

1、人名识别率（包括外国译名在内），2006年在数据集LDC上测试结果较低，封闭测试和开放测试都只有78%左右；其它测试集上的结果为：封闭测试在90%左右，开放测试在微软数据集（MSRA）上能达到96%以上。2007年SIGHAN报告的结果为封闭测试CityU达到89%，MSRA则高达95%，开放测试CityU为96%，MSRA则达99%以上。

2、地名识别率，2006年的三个数据集：在CityU和MSRA上封闭测试结果大致在85%左右，开放测试在90%左右。LDC上测得的地名识别效果则非常

低。2007年的两个数据集：CityU上测试的最好结果为封闭测试88%，开放测试93%，在MSRA数据集上最好结果达到了封闭测试94%。

3、机构名识别率，2006年CityU和MSRA在机构名识别上大致在80%到83%，开放测试大致为86%。LDC上的评测结果则很低，只有50%多。2007年CityU上封闭测试为72%，开放测试75%。MSRA上封闭测试为88%，开放测试90%左右。

从已有工作报告的识别效果来看，构词规律越明显的，识别效果也越好：人名、译名的识别效果最好，地名次之，机构名的识别最差。这一方面是由于机构名前后缀的规律性较差，另一方面，机构名的缩写也给识别造成很大困难。

上述按照未登录词类型的不同，对未登录词处理方法进行了分类总结。在中文分词的研究中，早期的研究一般将未登录词识别作为单独的模块进行处理，处理方式一般是先做通用的分词，再在初步分词的文本上进行未登录词的识别([49,59]。后期的中文分词研究显示，中文分词与词法层的其它任务，同步处理比分步处理的效果要好，如词性标注[28,85–87] 和命名实体识别[41,88,89]。这是因为语言处理中很多任务之间都是互相耦合，互相影响的。割裂开来不利于互相提供信息和依据。作为未登录词而言，通用词的切分同样影响着未登录词边界的确定，未登录词的边界又反过来影响通用词的切分。在这一思想指导下，出现了很多有价值的研究。Peng等提出了一种循环检测的方法，检测出来的未登录词将被加入原始词典中和已知词一起参与下一个循环的切分过程，加入了新的信息的分词系统会提高切分的准确率，同时，更为准确的切分结果也能够提高下一轮未登录词识别的效果。近年来统计机器学习模型成为研究热点，自2003年Xue提出将中文分词任务转换为基于字的序列标注任务以来，包括支持向量机、最大熵模型、条件随机场模型等在内的统计机器学习模型（分类模型）已被用于中文分词任务上。这类方法也属于将通用分词和未登录词识别同步进行的方法[38,62,90]，这类方法目前是在未登录词识别上最有效的方法。

## 1.4 中文自动分词的研究现状

本章介绍中文自动分词领域代表性的研究方法和目前能够达到的最好性能。

关于中文分词的研究方法方面，第1.3.2和第1.3.3小节中已经涵盖了一些中

文分词领域的研究方法。本节我们将选取一些代表性工作对二十年来中文分词领域的研究方法进行系统地介绍。

中文分词的研究始自上世纪 80年代。在过去的二十几年间，中文分词一直是中文信息处理领域的研究热点和难点，国内外许多研究机构和学者在这一问题上进行了深入的研究，取得了很多有意义的成果。二十年间，中文分词研究方法从基于规则的[34,91]到基于统计的[27]；从基于词的到基于字的[92]，以及字词结合的[93–95]；从基于有监督机器学习的[38,50,90]到基于半监督机器学习[96]和基于无监督机器学习的[40,97–99]，另外还有结合规则与统计方法的以及结合有监督和无监督机器学习的混合方法[100]。

这些方法的发展历程可以大致分为三个阶段：

第一个阶段是从80年代初到90年代中期，以基于人工规则和词典的方法为主。

第二个阶段是从90年代中期至2003年，以基于语料库统计学习方法为主。

第三个阶段是从2003年至今，中文分词领域的研究发生了从基于词的方法到基于字的方法的转变，2003年后的研究大都集中在训练分类器，对字在词中的位置进行分类标注上。

图 1.2给出了中文分词研究方法随时间推移，其性能提升情况及每个阶段的代表性方法。图中用红线分隔为两部分，左边部分是基于规则及词典的，右边部分是基于大规模语料库或标注语料的。

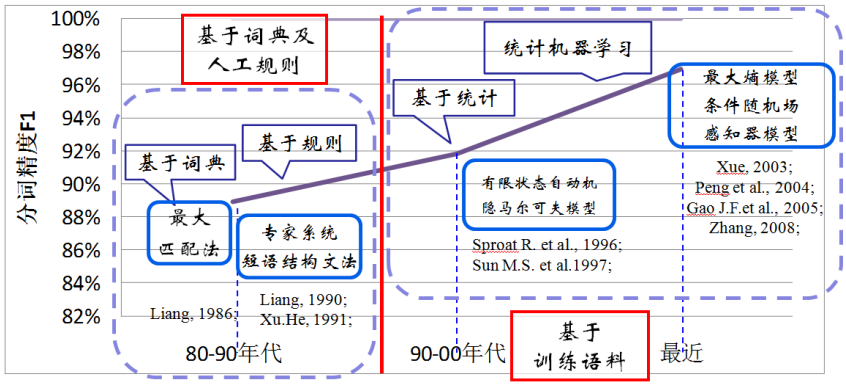



图 1.2　中文分词研究方法一览

这里，我们按照中文分词研究的阶段性进展，分三个部分对现有的分词方法

进行介绍（由于这一领域有大量的研究成果，这里，我们仅仅介绍每个阶段的代表性工作）。

### 1.4.1 基于词典及规则的方法

【基于词典】

1986年梁南元提出的最大匹配分词法（MM）是典型的基于词典的分词方法。第 1.3.2节关于分词歧义检测与消解中已经对MM及在其基础上衍生出的前向最大匹配（FMM）、后向最大匹配（BMM）以及双向最大匹配方法做了描述。

这类方法具备一定的处理中文歧义的能力，但对未登录词几乎没有处理能力。而且由于是基于词典的，这类方法的性能受词典质量的影响很大。实验结果显示，FMM的错误率为1/169左右，BMM的错误率为1/245左右。但由于其速度快，实现简单，目前仍是中文分词常用的方法之一。

【基于人工规则】

规则的方法可以是说一种确定性的“演绎推理”方法，对于简单的现象，用规则的方法的效率很高，但规则的方法无法表示小粒度的知识，对于不确定性的知识表达能力较弱，而且当处理复杂现象是，往往面临规则过多而造成规则间冲突增多，难以保证规则的一致性。基于规则的分词系统出现于90年代初期，包括专家系统[33]、短语结构文法[101–103]等。由于中文语言现象非常复杂，往往很难制定出涵盖所有语言现象的人工规则，得到的效果也较差，因此这类方法在后期逐渐淡出了人们的视线。

### 1.4.2 基于统计的方法

【基于简单统计信息的分词方法】

这类方法最常用的统计信息量是互信息，两个汉字$C_1$和$C_2$的互信息$I(C_1, C_2)$定义为：

$$I(C_1, C_2) = \frac{P(C_1, C_2)}{P(C_1)P(C_2)}, \tag{1-8}$$

其中，$P(C_1, C_2)$ 表示$C_1$ 和$C_2$ 在语料中的同现概率，$P(C_1)$ 和$P(C_2)$ 分别表示$C_1$ 和$C_2$ 在语料中独立成词的概率。

文献[53]通过计算两个汉字之间的互信息是否高于指定的阈值来判断它们是分开还是连接关系。该方法按照互信息由高到低的顺序，对文本中的二字串，选取当前互信息最高的一对进行捆绑，依次进行直到所有高于阈值的二字串都被处理。Sun等人在这一工作的基础上进行了扩展[54]，在互信息的基础上增加了汉字间*t*- 测试差的概念，用两种统计量衡量字与字的结合紧密程度。

另外一个值得一提的工作是文献[104]，该工作提出了一种基于期望最大化（Expectation Maximization，简称EM）算法的概率模型，算法初始为每个候选串的出现概率赋一个初始值，用这些串对待切分文本进行切分，切分后重估参数值，重估后再进行切分，依次迭代直至算法收敛。

【统计方法与词典相结合的分词方法】

文献[27]是这类方法的代表性工作，作者提出一种带加权的有限状态机进行中文分词的方法，这种方法将词典表示为一个状态转换机，每个词都是由一系列状态转移弧来表示，词串所带的权重代表了状态转移所需的代价。转换机从一个状态到另一个状态的转移概率是在一个20M字规模的语料库上统计串频得到的。算法选择一条代价最低的转移路径作为切分结果。词串的权重由下式计算得到：

$$Cost(w) = -log\frac{F(w)}{N}, \tag{1-9}$$

其中$F(w)$ 为词串$w$ 在语料中的频度，N为语料库的规模。

显然，这一方法存在如何估计未知词（语料库中未出现过的词）的概率问题，Church和Gale 曾使用Good-Turing方法进行估计。估计未知词概率的方式还有很多，这里不再阐述。

这类方法的另一个代表性工作见文献[105]，其中使用EM算法的一种变换形式进行中文分词，该方法同时维护一个核心词典和一个候选词典，核心词典包含真实的词，候选词典则包含其它所有不在核心词典出现的多元串。EM算法用于最大化训练语料的似然度，并为核心词典推荐新词的候选。一旦推荐的新词加入核心词典，EM算法将重置，并使用新的核心词典指导下一次的切分。

文献[104] 指出，利用互信息剪枝法能够得到显著的性能提升，但无论使用什么统计手段，所使用的基本的词典的质量好坏对系统性能的影响，可能要大于选择何种统计手段所带来的影响[27]。

上面介绍的两类方法都用到了词典，这类方法的性能受词典的覆盖率和平

衡性影响很大，其报告的分词准确率大致在90%左右。下面我们介绍一些摆脱了词典依赖的一些统计方法。

【一些复杂的基于词的统计机器学习方法】

基于有限状态机模型的中文分词模型提出以后，各种基于统计的方法被大量用于中文分词的研究中。随着已标注的训练语料越来越多，许多有监督的机器学习方法开始被用于中文分词任务上[34,106,107]。典型的基于统计机器学习的方法有以下三种：

a. 基于转换的分词方法

基于转换的分词方法（TBL）[108]可以说是第一个用于中文分词的机器学习方法。这一方法需要一个标注语料和一个初始分词器。初始的分词器可以是任意的一种，将标注语料恢复为未处理的文本，用初始分词器对这一文本进行切分，算法的第一次循环将切分结果与原始标注语料的切分进行比对，得出一个规则，如果使用该规则可以得到用某种目标函数定义的最大增益（比如错误率的下降），则初始分词器将加入这一规则继续上述环节，直至得到的增益低于设定的阈值。算法给出一个排好序的规则集。将这些规则集用于测试文本的切分。

b. 基于隐马尔可夫模型的方法（Hidden Markov model，简称HMM）

隐马尔可夫模型（HMM）是用来描述一个含有隐变量的马尔可夫过程。与一般的马尔可夫模型不同，在HMM中，状态并不直接可见，但受这个隐状态影响的某些变量则是可见的。每个隐状态在可观测的输出上有一个概率分布，因此通过可观察到的输出，可以得到隐含状态序列的一些信息。在中文分词中，HMM被用于在已知模型参数的情况下，寻找以最大概率产生某一特定输出序列的隐状态序列，通常使用维特比（Viterbi）算法进行。HMM在自然语言处理中的应用始自20世纪70年代的语音识别，文献[109] 首次利用HMM解决中文分词问题。这一方法将中文分词和词性标注同步进行。在之后还有许多基于HMM的工作[28,110]。基于隐马尔可夫模型的中文分词方法的性能大致在90%~ 92%。

c、基于信源信道模型的方法（source-channel）

文献[111]提出了Source-Channel模型。文献[41] 基于该模型建立了中文分词与命名实体处理系统，文中提出中文分词技术在不同应用下要求的颗粒度不同，该系统的特点之一就是可以输出不同颗粒度的分词结果。Source-Channel模型的雏形是由线性混合模型演化而来的，它可以使用极大规模的语言学和统计学信

息作为特征进行学习，从而提高分词的性能。

### 1.4.3 基于字序列标注的方法

在2003年之前，中文分词方法大多是基于词的——或者基于词典、或者使用词作为特征。2003年，文献[92]首先提出了基于字的中文分词方法，这种方法将中文分词任务转换为字序列标注问题。首先，为每个字打上位置（词中的位置）标签，然后根据这些位置标签把字序列转化成词序列，从而得到切分结果。显然，基于这一机制，中文分词问题可以看作一个分类问题，所有可以完成分类任务的机器学习方法都可以用于中文分词任务。

基于字的中文分词方法在2005年的国际中文分词评测中显示出优越的特性，文献[38]在最大熵模型基础上实现了基于字的分词方法，得到三个评测集的第一名。基于字的方法从此成为中文分词的研究热点。近年来，除了之前提到的最大熵模型外，许多结构更为复杂的统计机器学习模型被广泛用于中文分词任务上，如支持向量机（SVM）[112]、线性链条件随机场模型（CRF）[36]、最大间隔马尔可夫网络模型（$M^3N$）[113]等。文献[90] 首次将线性链条件随机场模型用于中文分词，其报告的效果超过了基于最大熵模型的分词方法。在此基础上许多研究者进行了大量扩展性工作，包括提高CRF的训练速度、降低训练CRF模型的空间代价[114]以及通过多层CRF的应用进一步提高分词的效果等[93,95]。

无论选用的机器学习模型是什么，基于字的分词方法都需要确定标注方式和训练模型所需的特征。文献[11]的研究结果显示，六词位标注（在四词位基础上，增加了两个中间位标识）能够获得更好的分词结果。目前常用的标注方式还有二词位（是否是词边界）和四词位（左边界、右边界、中间位和单字）。在使用的特征方面，常用的特征还仅仅限于字本身以及字的局部上下文（一般采用五字滑动窗口，即当前字的左右两个字）。远距离的全局信息使用较少。

有监督机器学习方法的性能受限于训练语料的规模和平衡性，且面临训练代价较大（占用内存较多以及训练时间长）的问题，因此工业界目前未未能采用这类方法，而是使用最简单的分词算法以得到高效率的处理。

以上提到的各种方法中，基于词典的方法最简单易行，且能够保证较高的已知词识别准确率。但是由于对未登录词的处理能力较差，召回率一般都较低。基于统计学习的分词方法需要标注语料库进行模型参数的训练，一般而言，训练语

料规模越大模型性能越好，但同时训练代价也成倍增长。因此各种混合模型也被广泛研究。试图博众家之长，扬长避短，达到最好的分词效果。另外，近年来的一个研究趋势是将中文分词和中文信息处理的其它任务，如词性标注、命名实体识别、句法分析等同步进行。研究结果表明，同步进行的方式往往由于分步进行，能同时提高子任务的效果。本文第五章的工作也是沿着这一思路对同步处理方法进行的探索性研究。

### 1.4.4 中文自动分词的现有水平

SIGHAN为中文分词的性能评价提供了一个公开、权威的平台。很大程度上反映了中文分词研究的最新进展，体现了当前中文分词系统能够达到的最好性能。表 1.4 ~ 1.7给出了四届SIGHAN报告的最佳评测结果（$F_1$）：

表 1.4 SIGHAN_2003（日本）

| 语料源 | 测试类型 | $F_1$ |
|---|---|---|
| 香港城市大学 | 封闭测试 | 0.940 |
| | 开放测试 | 0.956 |
| 台湾中央研究院 | 封闭测试 | 0.961 |
| | 开放测试 | 0.904 |
| 宾夕法尼亚大学 | 封闭测试 | 0.881 |
| | 开放测试 | 0.912 |
| 北京大学 | 封闭测试 | 0.951 |
| | 开放测试 | 0.959 |

表 1.5 SIGHAN_2005（韩国）

| 语料源 | 测试类型 | $F_1$ |
|---|---|---|
| 香港城市大学 | 封闭测试 | 0.943 |
| | 开放测试 | 0.962 |
| 台湾中央研究院 | 封闭测试 | 0.952 |
| | 开放测试 | 0.952 |
| 微软亚洲研究院 | 封闭测试 | 0.964 |
| | 开放测试 | 0.972 |
| 北京大学 | 封闭测试 | 0.950 |
| | 开放测试 | 0.969 |

由四届SIGHAN的评测结果可知，近年来虽然中文分词的研究取得了很多有意义的工作，但在准确度上并无明显的提高。而且SIGHAN的测试规模相对较小，未登录词比例也较低（2%~8%），在开放环境下得到的中文分词系统的性能将远低于这一水平，图 1.3给出了在现有的公开分词系统——ICTCLAS1.0 上，随意输入一段网络文本的测试结果。

从图中可以看到，现有的中文分词系统对未登录词（红色标记）、交集型歧义（绿色标记）、命名实体（蓝色标记）的处理还存在很多问题，中文自动分词系统的准确性和鲁棒性亟待提高。也正是本文研究的主要内容。

表 1.6 SIGHAN_2006（澳大利亚）

| 语料源 | 测试类型 | $F_1$ |
|---|---|---|
| 香港城市大学 | 封闭测试 | 0.972 |
| | 开放测试 | 0.977 |
| 台湾中央研究院 | 封闭测试 | 0.958 |
| | 开放测试 | 0.959 |
| 微软亚洲研究院 | 封闭测试 | 0.963 |
| | 开放测试 | 0.979 |
| 宾夕法尼亚大学 | 封闭测试 | 0.933 |
| | 开放测试 | 0.944 |

表 1.7 SIGHAN_2007（印度）

| 语料源 | 测试类型 | $F_1$ |
|---|---|---|
| 香港城市大学 | 封闭测试 | 0.951 |
| | 开放测试 | 0.970 |
| 台湾中央研究院 | 封闭测试 | 0.947 |
| | 开放测试 | 0.956 |
| 山西大学 | 封闭测试 | 0.962 |
| | 开放测试 | 0.974 |
| 国家语委 | 封闭测试 | 0.941 |
| | 开放测试 | 0.976 |

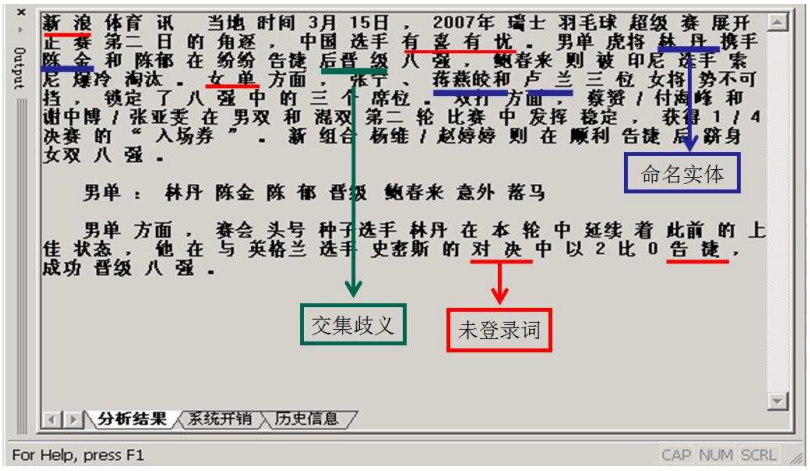



图 1.3 目前的中文分词系统在开放环境下的切分结果

在实用系统方面，目前公开的分词系统有中科院计算所开发的ICTCLAS[115]、微软亚洲研究院开发的MSRSeg[41]，以及海量分词系统等。其它还有一些单位和机构开发的内部的分词系统：清华大学SEGTAG系统，哈工大统计分词系统（报告的速度为100KB/s），北大计算语言汉语文本分析系统，国家语委文字所应用句法分析技术的汉语自动分词系统等。

在分词精度方面，ICTCLAS 最新版本报告的分词精度为98.45%，MSRSeg在SIGHAN 评测上报告的分词精度大致在97%左右。海量分词系统在人民日报语料库（与SIGHAN中北京大学提供的数据集(PKU)同源）上报告精度高达99.5%，但没有报告在SIGHAN 评测语料上的性能。在中文分词的速度方

面，ICTCLAS 最新报告的分词速度为单机996KB/s[①]。哈工大统计分词系统报告的速度为100KB/s。这些分词系统的速度还很难满足搜索引擎等的实际应用。海量分词系统报告的速度单机版大约为700KB/s，非单机版超过了MB/s 级，已经在为某些搜索引擎提供服务。学术界目前公认的准确度最好的基于有监督机器学习方法（如ME、CRF等）的中文分词速度大约在600KB/s。

## 1.5　本文的研究重点与内容安排

由前面的综述可知，中文自动分词作为一个经典问题已经获得较多关注，近年来的研究热点主要集中在有监督的机器学习方法上，在方法框架方面鲜有进展。本文针对中文自动分词研究领域的几个关键问题，试图填补一些空白或研究得较不充分的问题。

图 1.4 给出了本文的研究内容框架图，由图可知，本文主要的研究内容是通用词表建设以及基于大规模语料库的未登录词识别和分词歧义消解。通用词表的建设是以支撑中文分词规范和大规模语料库建设为目标的。对分词歧义消解和未登录词识别问题，本文的工作是基于大规模语料库进行研究的。

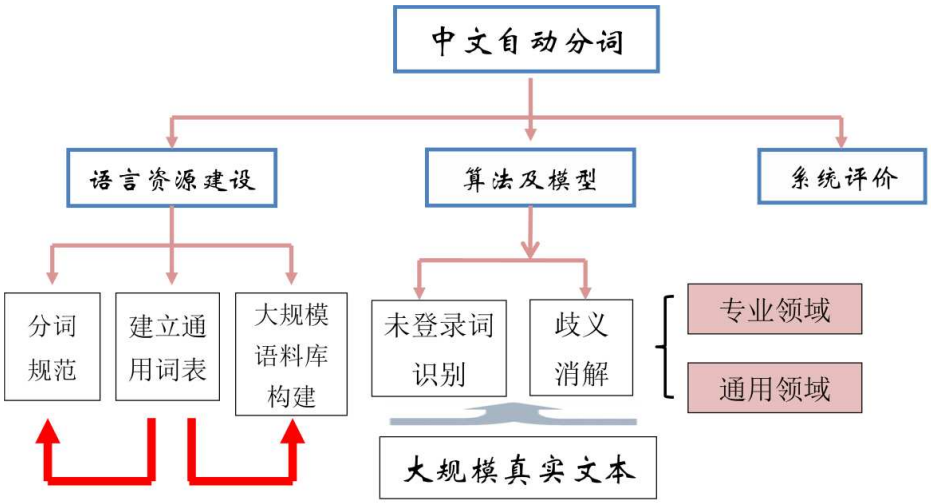



图 1.4　本文研究的主要内容

本文后续部分的内容安排如下：

① http://www.ictclas.org/

第二章探讨中文分词中语言资源建设问题，提出一套基于多语料混合策略的中文词频近似估计方法。

第三章主要针对中文分词中的交集型歧义，在通用语料库和专业领域语料库上研究交集型歧义的统计分布及确定核心交集型歧义字段。

第四章针对中文分词中的未登录词的问题，提出了利用搜索技术处理中文分词中未登录词问题的方法。

第五章提出了基于最大间隔马尔可夫网络的中文分词与命名实体识别一体化方法。实验结果表明，将命名实体识别和中文分词同步进行有助于同时提高分词和命名实体识别的效果。

第六章对本文的全部研究工作做讨论式的总结，并展望在中文自动分词领域今后可能的研究方向。

# 第 2 章　基于多类型语料的中文词频近似方法[①]

本章研究中文分词在语言资源建设上面临的一个重要问题——中文词频估计方法。第 1.3.1节已经对语言资源建设及中文词表的重要性作了介绍，现安排各节内容如下：

第 2.1 节简述中文词频估计问题的背景及研究意义；第 2.2节介绍基于多语料混合策略的中文词频近似估计方法；第 2.3 节介绍参数调整和实验部分所用到的语料库；第 2.4 节介绍基于遗传算法的参数调整；第 2.5 节是实验及结果分析；第 2.6对本工作进行总结讨论。

## 2.1　研究背景

建立具有广泛适用性及高覆盖率的中文通用词表是建设一致性好的标注语料库的必要条件，准确的中文词频是建立通用词表的重要信息。除此之外，中文词频信息在中文信息处理中有着广泛的应用，比如信息检索、对外汉语教学，基于词频的分词模型[58]等。

对于英文而言，词频信息可以从一个大规模（理论上可以任意大）的英文文本中，通过统计词的出现次数来得到较为精确的词频估计结果。然而中文词频信息的获得却比英文要困难和复杂得多，这是由中文语言本身的特殊性决定的。中文里没有显式的词边界，即词与词之间没有天然的隔断，我们缺少一个天然的“完美”（完全正确的已经切分好的）语料库来统计词频。因而如何准确而高效地获得中文词频信息成为中文信息处理中的一个研究问题，我们称之为中文词频估计。中文词频估计不仅是一项工作量巨大的工程，同时也涉及到许多其它研究领域，如统计学、语言学理论和计算机技术等的很多问题。它与抽样理论、中文“词”的定义和语义理解等都有直接的关系[22]。目前中文词频估计在中文信息处理里依然是一个严峻的挑战，如何利用现有的语言资源及工具得到准确的词频信息，是一项具有重要研究价值且颇具挑战性的工作，也是本章研究的

① 本工作的主要部分以全文形式发表在国际期刊Journal of Quantitative Linguistics, 国际会议 ICCPOL’06以及国内会议SWCL’06上，见发表的学术论文 [1,4,7]。

主要内容。

如之前所提到的，通常而言要得到中文词频信息，我们需要有一个完全正确的已经切分好的中文语料库进行词频统计[116]。但是要获得这个语料库并非易事，我们主要面临以下两个基本的困难：

一，即使采用相同的分词标准，几个人工标注语料库之间甚至同一语料库内部依然存在严重的标注不一致性[16]。这是由于中文的构词特点[117,118]造成的：尽管“词”的定义在语言学家的角度看来非常清晰[119,120]，然而中文里存在一批词，我们既可以把它们看作是一个复合词[117,118]，比如“猪肉”，也可以把它们看作是一个由两个一字词“猪”和“肉”组成的短语。因此当我们统计“猪肉”这个词的词频时，如果按照前一种理解，那么该词的词频会非常高。然而，如果按照后者理解，“猪肉”在语料库中出现次数就是零。对“词”的界定的模糊性使得语料库的一致性很难保证。

二，由于长尾效应的存在，对一个中等规模的词表的词频信息进行估计需要上亿字（而不仅仅是上百万字）的已切分语料库的支持。但是人工进行切分费时费力，目前还没有一个规模足够大的准确无误的人工切分语料库。而且如此大规模的切分语料库在短期内是不可能得到的。

鉴于一个“完美”的切分语料库目前很难得到（但即使是“不完美”的人工切分语料库对中文词频估计也有其可利用之处），我们不得不考虑使用可获得的几种类型语料库进行词频估计的可能性。下面分析目前可用于进行中文词频估计的几种类型的语料。

第一种是*“完美”的自动切分语料*：用一个高质量的分词器对一个语料库进行自动切分，从而得到词频的估计。理论上讲，这个方法无疑是最好的。然而这样的一个“完美”分词器目前是不存在的。尽管在过去的二十年里，研究者们在这方面作了很多的努力，然而中文自动分词的性能依然不尽如人意。2003年SIGHAN[7]评测在四个小规模测试集上（具体规模见第一章的介绍）进行的开放测试结果显示，$F_1$-measure值最高分别达到95.9%，95.6%，90.4%和91.2%。而在第后面几届分词比赛中[8–10]，分词的性能虽然在小范围内有了一定的提高（大致在97%左右），但并未从本质上解决这一难题。分词的性能依然不能满足实际需要，尤其是未登录词出现时，分词的精确度会受到严重的影响。

第二种是*最大匹配自动切分语料*：这里我们选取最基本的分词方法–最大匹配（Maximum Matching）分词法对语料进行切分。然后从经过最大匹配切分的语料中直接统计词频，从而得到所需的词频估计值。刘和梁的工作第一次将最大匹配切分方法用于处理大规模的文本[22]。根据扫描顺序的不同，最大匹配切分可以进一步分为两种，一是由前向后扫描文本中的最长匹配串，称为前向最大匹配（Forward MM）切分，简称FMM；二是由后向前扫描，称为后向最大匹配（Backward MM）切分，简称BMM。梁南元的研究结论显示，最大匹配分词法具有速度快（只需查词表）且易于实现等优点，且其分词精度可达90%以上[17]。孙的工作表明[18]，基于最大匹配切分语料得到的词频估计结果是相对准确的，另一方面，使用基于最大匹配分词的另一个优点是能够保证分词结果具有较高的一致性。使用最大匹配分词和其它分词方法一样，不可避免地存在分词错误，尤其是处理存在大量未登录词的文本时，最大匹配方法几乎不具备未登录词识别的能力，性能下降明显。

第三种是*生语料*。这里我们考虑用串频近似地代替词频的方法[54]。串频信息可以直接从任意一个未经人工处理的生语料库中统计得到。很明显，对任意一个词，统计出的串频值都会高于该词在语料中真实的词频值。对某些词，这一方法会出现“过估计”现象。对于单音节词，这一现象尤其严重。但是这种机制、有两个优点：首先该方法避免了切分错误的影响。其次，生语料容易获得，理论上来说其规模可以任意大。

综上所述，可将能够用于中文词频估计的各类型语料及各自的优缺点总结为图 2.1所示。

基于上述可能用作词频估计的语料，目前已有一些使用生语料、最大匹配切分语料进行中文词频估计的研究（某些工作是在基于词频的分词模型中使用的）：

文献[54]和[27] 采用串频的方法近似地估计词频并用于中文分词任务。这种方法的优点是能够实现无监督的中文分词——词频信息可以从未加工过的生语料中直接获取。显然，用串频估计词频，对某些词而言（尤其是单音节词），这种策略会造成严重的词频过估计问题。日本学者Nagata提出了一种最大匹配串频的改进方法[121]，该方法的优点是能避免仅靠串频出现的过估计问题，但却容易产生数据稀疏的问题：假设我们有两个词：$w_i$和$w_j$，$w_i$是$w_j$的子串。而在训练语

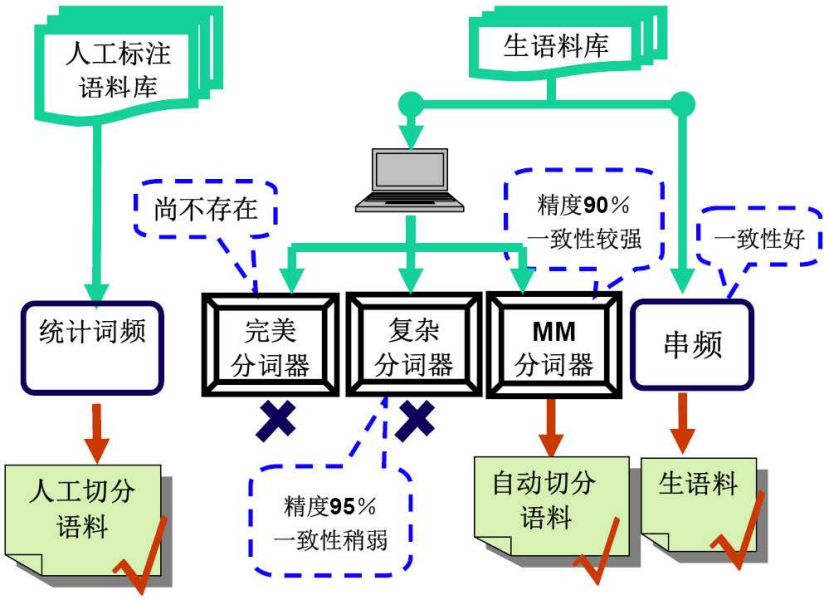



图 2.1　可以用作中文词频估计的语料类型

料中，$w_i$ 以$w_j$ 的子串形式出现在语料库里的概率很高，则$w_i$ 的词频$Pr(w_i) = 0$。这显然是错误的估计。文献[122]提出了仅使用生语料进行中文词频估计的方法，该方法中文语言特性考虑进去，通过假设检验给出了最大匹配分词（包括正向和反向）的结果和串频结果的最优组合方式，并用它进行词频估计。该方法也是一种无指导的词频估计。但这种方法对与短词的估计仍不够准确。除此之外，文献[123]曾采用汉字字串构词能力的方法近似估计词频。

综上，对于中文词频估计任务，无论是通过人工切分还是使用机器自动切分，我们所需要的“完美”的切分语料目前无法获得。我们能够利用的是存在切分错误和不一致的规模较小的人工切分语料和未经处理的生语料。在已有的研究工作中，对生语料的利用较为充分，但对人工切分语料的利用是一个空白，虽然如我们之前的分析可知，人工切分语料存在各种缺陷，但其作用是显然存在的。因此，本文的研究目标是如何最大限度地使用所有类型的语料，以期达到最准确的词频估计结果。

我们在考虑用这些语料库进行词频估计时，每类语料各自己的优势和劣势，如图 2.1所示。鉴于它们中的任何一个都很难单独完成中文词频估计的任务，我们不得不考虑一个折中的策略，尝试使用目前可以获得的多种不同类型的语料——人工切分语料、最大匹配切分语料以及生语料，希望能发挥各自之所长，

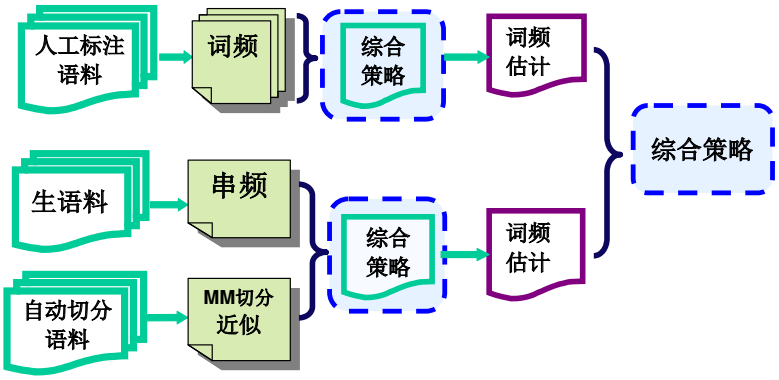



图 2.2　基于多类型语料的词频估计策略框架示意图

这里我们并不直接进行词频估计，而是尝试得到词频的一个近似值。

下面我们将详细介绍本文提出的基于多类型语料的中文词频近似估计方法。

## 2.2　中文词频近似方法框架

为了有效地结合已有的几类可用语料进行中文词频估计，我们分为三步进行：首先考虑生语料及MM切分语料的综合策略；其次，考虑如何结合人工标注语料的估计结果；最后综合考虑前两步得到的估计结果，进行合并，得到最终的词频近似估计方法。我们将在接下来的小节逐步进行介绍。为描述清楚，整个词频估计策略的框架见图 2.2所示。

### 2.2.1　整合生语料库和最大匹配自动切分语料库的词频估计结果

首先，假设我们有一个定义好的词表（简称WL）和一个生语料库$C$。使用前向最大匹配切分和后向最大匹配切分对$C$进行切分,可以得到两个最大匹配切分语料库。对于词表WL里的每个词$w_i$，我们可以得到以下三个统计量：

$f_{FMM}(w_i, C)$ ：通过FMM切分从生语料库$C$中得到的$w_i$的词频估计；

$f_{BMM}(w_i, C)$ ：通过BMM切分从生语料库$C$中得到的$w_i$的词频估计；

$f_{RAW}(w_i, C)$ ：从生语料库$C$中获得的$w_i$的串频。

对上述所得的三种词频估计值，我们需要一个策略对它们进行有效的综合，使之取长补短。而前向最大匹配、后向最大匹配以及串频估计在中文词频估计

中的效果，已有的研究工作[122]通过假设检验，得出了这三种方法的比较结果：

对于1到4字词，取$f_{FMM}(w_i, C)$和$f_{BMM}(w_i, C)$的平均值作为$w_i$的词频最逼近真实值,对于五字词则是$f_{FMM}(w_i, C)$给出的估计结果最好, 对于大于五字的词，串频$f_{RAW}(w_i, C)$给出的估计最为准确。

这里我们直接采用这一结论。

用$F_{RFB}(w_i, C)$表示综合考虑$f_{RAW}(w_i, C)$，$f_{FMM}(w_i, C)$和$f_{BMM}(w_i, C)$得到的综合词频估计结果，则根据上面的结论可得：

对于词长为1~4的词：

$$F_{RFB}(w_i, C) = \frac{1}{2}\left[f_{FMM}(w_i, C) + f_{BMM}(w_i, C)\right]. \tag{2-1}$$

对词长为5的词：

$$F_{RFB}(w_i, C) = f_{BMM}(w_i, C). \tag{2-2}$$

对词长超过5的：

$$F_{RFB}(w_i, C) = f_{RAW}(w_i, C). \tag{2-3}$$

通过整合生语料库估计结果和最大匹配自动切分语料库估计结果得到词频估计的方法记作$RFB$.

### 2.2.2　整合人工切分语料库的词频估计结果

这里我们用$MC$表示一个包含$N$个不同的人工切分语料库语料库的集合，这$N$个人工切分语料库表示为$MC_j$ $(j = 1, 2, \ldots, N)$。通过直接统计词频，我们可以从每个$MC_j$中得到WL中每个$w_i$在该语料库中的词频，表示为$f_{MS}(w_i, MC_j)$。如果用$F_{MS}(w_i, MC)$表示由人工切分语料库得到的$w_i$的词频估计结果，$F_{MS}(w_i, MC)$的值可由公式2-4进行计算。

$$F_{MS}(w_i, MC) = \sum_{j=1}^{N}(f_{MS}(w_i, MC_j)). \tag{2-4}$$

通过整合不同的人工切分语料库得到词频估计的方法记作$MS$。

### 2.2.3　整合$F_{RFB}(w_i, C)$ 和$F_{MS}(w_i, MC)$

综合考虑 2.2.2和 2.2.3小节，对于WL中的每个词$w_i$，我们都可以得到两个不同的词频估计结果：$F_{RFB}(w_i, C)$ 和$F_{MS}(w_i, MC)$。为了进一步对得到的两个词频估计结果进行合理的合并，我们需要考虑以下问题：

第一，这两个统计量是由两个规模不同的语料库得到的：生语料库$C$ 的规模，记作$N_C$，通常比人工切分语料库$MC$ 的规模$N_{MC}$ 大很多。因此，这里设置参数$\alpha$ 来平衡语料库规模的差异。直观上讲，我们可以简单的将$\alpha$ 的值设置为$C$ 和$MC$ 的比值。我们将在第 2.4节讨论如何通过遗传算法(GA)自动调节参数$\alpha$ 的值。

这里，考虑将生语料库的规模缩小至与人工切分语料库相同的规模，使得合并后的新的语料库规模为$2N_{MC}$。将平衡语料库规模的参数$\alpha$ 看作是生语料库缩小的倍数，则可得生语料库的规模$C$ 缩小至$N_C/\alpha$。相应的，人工切分语料库的规模应为$N'_{MC}$：

$$N'_{MC} = 2N_{MC} - N_C/\alpha. \tag{2-5}$$

相应地，$F_{RFB}(w_i, C)$应变为$F'_{RFB}(w_i, C)$：

$$F'_{RFB}(w_i, C) = F_{RFB}(w_i, C)/\alpha. \tag{2-6}$$

继而，$F_{MS}(w_i, MC)$应变为：

$$F'_{MS}(w_i, MC) = F_{MS}(w_i, MC) \times \frac{2N_{MC} - N_C/\alpha}{N_{MC}}. \tag{2-7}$$

第二个问题是基于对中文词的一个观察结论：越短的词，用熟语料估计的词频准确度越高，即更加可靠。因此，对于词长越短的词我们希望增加通过人工切分语料得到的估计值的权重。为了将这一因素考虑进最后的综合估计策略中，我们引入参数$\beta$ 用于权重的调整。词长越短的词给予$F_{MS}(w_i, MC)$ 的权重应该越大，即$\beta$ 的值越大。根据词长的不同，我们将中文词分为四类：一字词，二字词，

三字词以及四字以上词。对每类词，设置参数$\beta$ 如下：

$$\beta = \begin{cases} \beta_1 & \text{对一字词} \\ \beta_2 & \text{对二字词} \\ \beta_3 & \text{对三字词} \\ \beta_4 = 0 & \text{四字及四字以上词} \end{cases}$$

由之前的讨论可知，$\beta$的值应满足约束：$\beta_1 \geq \beta_2 \geq \beta_3 > \beta_4$。引入参数$\beta$后，公式2-6应改为：

$$\begin{aligned} F''_{RFB}(w_i, C) &= F'_{RFB}(w_i, C) \times \frac{1}{1+\beta} \\ &= F_{RFB}(w_i, C) \times \frac{1}{\alpha(1+\beta)}. \end{aligned} \tag{2-8}$$

同样地，公式2-7应改为：

$$F''_{MS}(w_i, MC) = F_{MS}(w_i, MC) \times \frac{2N_{MC} - \frac{N_C}{\alpha(1+\beta)}}{N_{MC}}. \tag{2-9}$$

综合式2-8和式2-9，对于WL中的每个词$w_i$，可以得到最终的基于多类型语料的词频估计近似值$F_{RFB+MS}(w_i, C+MC)$：

$$\begin{aligned} F_{RFB+MS}(w_i, C+MC) &= F''_{MS}(w_i, MC) + F''_{RFB}(w_i, C) \\ &= F_{MS}(w_i, MC) \times \frac{2N_{MC} - \frac{N_C}{\alpha(1+\beta)}}{N_{MC}} \\ &\quad + F_{RFB}(w_i, C) \times \frac{1}{\alpha(1+\beta)}. \end{aligned} \tag{2-10}$$

这种通过合并$F_{RFB}(w_i, C)$和$F_{MS}(w_i, MC)$得到最终词频估计的方法记作$RFB+MS$.

## 2.3　数据集

本节介绍在参数调整及实验部分所用到的语料库。

首先介绍两个人工切分语料库：第一个是由清华大学和语言文化大学构建的，规模为1,040,190词，1,763,762字的平衡语料库，记作HUAYU。第二个是北京

大学构建的新闻语料库，其规模为7,286,870词，13,030,237字，记作BEIDA。综上，目前已有的两个人工切分语料总规模为8,327,060词，14,793,999字。

其次是标准语料库：我们采用由国家语委构建的经人工校对的平衡的中文切分语料库，记作YUWEI，作为评价词频估计策略的标准语料库。YUWEI的规模为25,000,309 词，51,311,659字。这里我们将YUWEI作为评价标准，是因为YUWEI是一个由国家权威机构构建的规模较大的平衡语料库。从YUWEI中，可以提取到一个带有词频信息的词表。将词频小于4次的词去除，得到本节后面所用的标准词表，记作YWL。YWL中共计99,660个词。

第三个语料库是生语料库：我们构建了一个大规模的生语料库，记作RC，它包含1,019,986,721 字。从RC中，我们可以得到YWL中每个词对应的串频。

最后，是最大匹配切分语料：将YWL作词表，分别在RC上作前向最大匹配切分和后向最大匹配切分，可以相应地得到两个最大匹配切分语料，记作RC_FMM 和RC_BMM。

综上，已有的可用作词频估计的语料库有：两个规模适中的人工切分语料库HUAYU和BEIDA，我们将这两个语料库统一记作HB；一个规模很大的生语料库（RC）；两个MM切分语料库(RC_FMM and RC_BMM)；一个用于参数调整和实验比较的标准语料库(YUWEI)。

## 2.4　参数调整

本节使用遗传算法（Genetic Algorithm，简称GA）[124]，在标准语料库YUWEI上进行实验，从而确定RFB+MS词频估计算法中的$\alpha$ 和$\beta$（$\beta_1$,$\beta_2$ 和$\beta_3$）参数的值。

这里使用的GA是由作者本人开发实现的，鉴于GA有许多方法上的变种，我们详细描述本文用到的GA类型：

所用的GA是基于非重叠种群的，即没有两代是完全相同的。在每步的迭代中，变异概率设置为一个最小值为0.03的变量。交叉概率设置为0.75。

我们将YUWEI语料随机地分为两个部分，分别用于训练和测试，记作YUWEI_1 和YUWEI_2。YUWEI_1 的规模为28,536,843字，YUWEI_2 为22,774,816字。YUWEI_1 作训练集使用，而YUWEI_2 则作为测试集。

从YUWEI_1和YUWEI_2中分别抽取出一个词表，词表中包括每个词在该语料库中出现的次数。将两个词表中词频小于4的词去除，可以得到下面我们要使用的两个词表：YWL_1和YWL_2。YWL_1共包含83,101个词，YWL_2包含76,514个词。下面我们将采用GA算法在YUWEI_1上优化所需的参数。

将WL_1中83,101个词按词频由高到低的顺序排列，会得到一个词序列以及由排序序号构成的序列，记作$R_{YW1}$。根据$F_{RFB+MS}(w_i, RC + HB)$可得词序列中的每个词使用RFB+MS 方法得到的词频估计值。保持词序列中词的顺序不变，根据$F_{RFB+MS}(w_i, RC + HB)$，每个词都会得到一个新的排序序号，这个使用RFB+MS方法得到的新的排序序列记作$R_{RFB+MS}(YWL_1)$。

下面，我们定义适应度函数。这里引入两种用于度量词频估计方法效果的方法。

第一种：度量由一种词频估计策略衍生出的排序序列和已知的标准序列$R_{YW1}$的相似程度。两个序列近似度越高，说明词频估计的结果越准确，估计策略也越有效。使用相似度度量的出发点是，虽然即便是使用两个规模相当的平衡语料库得到的词频统计，每个词在整个语料库中按照词频高低排序的结果也不完全相同，但是，如果将中文词按照词频高低分为高频、中频和低频词，则根据词频进行排序得到的中文词序列可视作遵循一个大致的分布。因此，如果词频估计较为准确，其得到的排序序列应与标准序列有较高的相关性。换句话说，词频估计的准确程度可以通过排序序列与标准序列的相关性的高低进行判断。将中文词序列固定，根据不同的词频估计值，可以得到同样的词序列对应的不同的排序序列。为了衡量任意两个排序序列的相关性，采用斯皮尔曼秩相关系数（Spearman Coefficient of Rank Correlation，简称SCRC）来计算两个序列的相似程度。斯皮尔曼秩相关是一种研究两个变量间相关关系的方法。是利用两变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，适用范围较广，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量的观测资料转化得到的等级资料，无论两个变量的总体分布形态、样本容量的大小如何，均可使用斯皮尔曼秩相关来进行研究。

假设对于一个给定的词表WL，有两个相应的排序序列$R_1$ 和$R_2$。则$R_1$ 和$R_2$的相似程度，即两者的SCRC值可由下式计算得到：

$$SCRC_{WL}(R_1, R_2) \equiv 1 - 6\sum_{i=1}^{N} \frac{d_{w_i}^2}{N(N^2-1)} \ ,$$

这里$d_{w_i}$是某一个词$w_i$在两个排序序列$R_1$和$R_2$中对应的排序位置的差值，$N$是WL的规模。显然，SCRC是一个范围在0到1之间的实数。

SCRC值反映了两个序列的相似程度，虽然词序列根据词频高低并不严格遵循一个排序，但是从高、中、低频角度来看，词序列的排序整体服从一个大致的趋势。这个趋势将从SCRC值上得到体现。

为了进一步证明将SCRC作为一种衡量机制的有效性，我们用两个英文语料作验证实验。由于英文词频较易准确获得，从两个语料库中统计得到的英文词频，并对英文词序列进行排序，如果两个序列的相似性得到证实（有较高的SCRC值），则用SCRC衡量词频估计的效果是可靠有效的。

这里我们选取两个常用的英文语料库：一是LOB语料库，包含53,823个英文词。另一个是Brown语料库，它包含55,734个词。求交集后，得到27,807个在两个语料库中都出现的英文词集合。对这27,807个词在LOB和Brown里进行词频统计，得到两个相应的排序序列，计算得到两个序列的SCRC值为0.72。去掉词频小于4次的词后，共计11,357个词，对应的两个排序序列的SCRC值为0.79。因此，SCRC从一定程度上反映了根据词频高低得到的排序序列是遵循一个大致的趋势的。通过与标准排序序列相比较，词频估计的准确程度可以由排序后序列与标准序列的SCRC值得到反映。

根据第一种衡量机制，GA算法要优化的问题可以表示为：

$$\arg\max_{\alpha,\beta} SCRC_{YWL\_1}(R_{YW1}, R_{RFB+MS}(YWL\_1)).$$

这里我们将第一种机制定义的适应度函数记作$FIT_1$。

第二种：由一种词频估计策略得到词频估计后，根据词频高低可以排序得到一个词序列$R$，度量$R$中前$N$个高频词对语料库$C$的覆盖率的高低。覆盖率越高说明该估计策略越准确。这一衡量机制可表示为：Coverage($R$,$N$,$C$)。由下式计算得到：

$$\text{Coverage}(R, N, C) = \frac{\text{R 中的前N个高频词覆盖C中的汉字数目}}{\text{语料库C中的总字数}}.$$

表 2.1　由GA得到的优化参数值（采用最大匹配切词）

| | $\alpha$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | SCRC | 覆盖率 |
|---|---|---|---|---|---|---|
| $FIT_1$ 确定的参数 | 15.8 | 0.9 | 0.6 | 0.3 | 0.748 | 96.3% |
| $FIT_2$ 确定的参数 | 17.0 | 1.2 | 0.4 | 0.2 | 0.743 | 96.6% |
| 最终参数值 | 16.4 | 1.0 | 0.5 | 0.3 | 0.746 | 96.5% |

根据第二种度量机制，GA算法要优化的目标问题可表示为：

$$\arg\max_{\alpha,\beta} Coverage(R, N, C),$$

这里需要说明的是，我们在YUWEI_2上进行的统计结果显示，高频前50,000个词对语料库的覆盖率已经达到98%。由于高频50,000个词已经达到较高的覆盖率，在下面的实验中，上式中的参数$N$设置为50,000。由第二种度量机制——覆盖率所定义的GA算法的适应度函数，记作$FIT_2$。

由于两种机制是从两个不同的角度对词频估计策略进行评估。这里在用GA调整参数时，同时考虑两种策略作为优化目标。因此，我们分别用$FIT_1$和$FIT_2$作为GA的适应度函数，相应地可以得到两组优化参数值，在实验中，迭代次数选择150次。最终参数值以及每组参数下计算得到的SCRC值和覆盖率值见表 2.1所示。

比较表 2.1中第一组参数和第二组参数的结果可知，第一组参数得到了较高的SCRC 值，但覆盖率却较低。而第二组参数的情况则相反，覆盖率较高而SCRC较低。由于我们希望同时兼顾两种策略，因此，这里简单地对两组参数进行折中，取两组参数的平均值作为最终的$\alpha$ 和$\beta$ 的值，如表 2.1所示。

上述实验都是在分词器选用简单的最大匹配分词法进行的，那么选用当前公认的更为准确的分词器结果会如何呢？不同的分词器对参数的确定又有什么影响？为了回答上面两个问题，检验参数值对选用的分词器的敏感程度，我们选用了一个公认的性能较好的中文分词器，由中科院开发的ICTCLAS1.0[①] 对生语料进行切分，代替最大匹配切分得到的词频结果，用GA重新优化参数，以考察参数值是否因分词器的不同而发生较大改变。

由于在ICTCLAS1.0中，不再存在前向、后向匹配的问题，我们保持其它已

① ICTCLAS 1.0: http://www.nlp.org.cn

表 2.2　由GA得到的优化参数值（采用ICTCLAS1.0分词器）

| | $\alpha$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | SCRC | 收敛速率 |
|---|---|---|---|---|---|---|
| $FIT_1$确定的参数 | 16.2 | 1.0 | 0.5 | 0.2 | 0.738 | 95.8% |
| $FIT_2$确定的参数 | 16.8 | 1.2 | 0.5 | 0.2 | 0.724 | 96.9% |
| 最终参数值 | 16.5 | 1.1 | 0.5 | 0.2 | 0.734 | 96.5% |

经定义的表达式不变，加入下面针对ICTCLAS1.0切分结果的定义：

$f_{ICTCLAS}(w_i, C)$ ：使用ICTCLAS1.0分词器对生语料库$C$进行切分后，得到的词表中每个词$w_i$的词频。

相应地，词频估计策略RFB的表达式随之改变为：

对词长为1-5字的词：

$$F_{RFB}(w_i, C) = f_{ICTCLAS}(w_i, C). \tag{2-11}$$

对词长为5字以上词：

$$F_{RFB}(w_i, C) = f_{RAW}(w_i, C). \tag{2-12}$$

用ICTCLAS1.0代替最大匹配切分，由RFB策略得到新的由生语料估计出的词频，使用GA优化得到的新一组参数以及对应的SCRC值、覆盖率值见表 2.2。

与表 2.1相比可知，改变分词器并未使得最终确定的参数发生大的改变，同时，注意到使用公认的高质量的分词器并没有使得SCRC值和覆盖率得到提高。这一实验结果进一步证实了之前的分析，即使用最大匹配分词器是适合我们这里提出的词频估计策略的。

至此，我们提出的基于多类型语料的综合词频估计策略中，两个参数（$\alpha$和$\beta$）的准确值已经确定下来，词频估计策略也已完成。在后面的一系列实验中，我们将考察所提出的词频估计策略的实用性和有效性。

## 2.5　实验结果及分析

在实验部分，我们使用上一节中介绍的YUWEI_2作为标准语料库，在其上比较所提出的基于多类型语料的词频估计策略（RFB+MS）与其它词频估计策略的优劣。除了与第 2.2节介绍的RFB、MS两种策略比较外，考虑到提

出的策略RFB+MS中，RFB中使用的是简单的最大匹配分词，为了验证使用其它分词器产生的影响，在实验比较中，加入与使用其它分词器得到的词频估计策略的比较。这里我们采用基于线性链条件随机场（简称CRF）[90]的分词器代替RFB+MS中使用的最大匹配分词（简称MM）。这种词频估计策略记作CRF+MS。

本实验中的CRF分词器是基于Taku Kudo开发的CRF++ 0.53开源软件包① 实现的。中文分词可以转换成基于字的序列标注问题，即对文本中每个字标注其在词中的位置，线性链CRF模型适于这一多分类任务。采用四标注集标注字中词中的位置，设置为：'S'（单字词），'L'（词的左边界）,'M'（非词边界），'R'（词的右边界）。训练语料方面，使用HUAYU和BEIDA作为训练语料训练基于CRF的分词器。由于内存的限制，最后的训练语料是从两个语料库中随机抽取其中的一部分句子得到的，抽取的训练语料的规模为15MB，在训练特征的选择方面，窗口大小设置为5，即考察当前字的前后两个字作为特征。在特征模板的选择上，直接采用文献[38]中介绍的特征模板：

(a) $C_n, n = -2, -1, 0, 1, 2$

(b) $C_nC_{n+1}, n = -2, -1, 0, 1$

(c) $C_{-1}C_1$

(d) $Pu(C_0)$

(e) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

这里$C_n$ 代表一个中文汉字，$n$ 代表相对于当前字$C_0$ 的位置偏移量。比如$C_1$ 表示紧邻$C_0$ 的下一个字，$C_-1$ 就表示当前字$C_0$ 的前一个字。$Pu(C_0)$ 表示当前字是否是一个标点。$T(C_n)$ 代表了中文字$C_n$ 所属的类型(Type)。我们将中文字分为四类：第一类是阿拉伯数字；第二类是日期（中文的“日”,“月”,“年”）；第三类是英文字母；其它的归为第四类。详细的介绍参见文献[38]。

综上，在实验部分，我们将把提出的词频估计策略RFB+MS与下列几种策略进行比较：

1、RFB：用生语料库和最大匹配切分语料库得到词频估计的策略；

2、MS：仅用人工切分语料库得到的词频估计策略；

3、CRF+MS：类似于RFB+MS，只是在中文分词器的选择上用CRF 替代最

① http://chasen.org/taku/software/CRF++/

表 2.3　四种词频估计策略在YWL_2上的SCRC值比较

| | $(R_{YW2}, R_{MS}(YWL\_2))$ | $(R_{YW2}, R_{RFB}(YWL\_2))$ | $(R_{YW2}, R_{RFB+MS}(YWL\_2))$ | $(R_{YW2}, R_{CRF+MS}(YWL\_2)$ |
|---|---|---|---|---|
| $SCRC_{YWL\_2}$ | 0.66 | 0.73 | 0.77 | 0.73 |

大匹配分词。

下面的实验观察从五个不同的角度对上述几种词频估计策略进行性能评价。

### 2.5.1　观察角度1：斯皮尔曼秩相关系数

在这一小节中，我们从由词频得到的排序序列的相关度的角度比较词频估计策略的优劣。我们用之前介绍的斯皮尔曼秩相关系数（Spearman Correlation Rank Coefficient，简称SCRC）作为衡量序列相似程度的度量。使用类似获得$R_{YW1}$ 和$R_{RFB+MS}(YWL_1)$ 的方法，从YUWEI_2 中可以得到YWL_2 中76,514个词对应的排序序列，记作$R_{YW2}$。以YWL_2作为词表，基于$F_{MS}(w_i, HB)$，$F_{RFB}(w_i, RC)$，$F_{RFB+MS}(w_i, RC + HB)$和$F_{CRF+MS}(w_i, RC + HB)$，可以得到对同一词序列的四种不同排序序列：$R_{MS}(YWL\_2)$，$R_{RFB}(YWL\_2)$，$R_{RFB+MS}(YWL\_2)$ 和$R_{CRF+MS}(YWL_2)$。然后分别计算$R_{YW2}$ 和其它四个序列的SCRC值，表 2.3给出了计算结果。

由表 2.3可知，我们提出的RFB+MS策略得到了最高的SCRC值，说明与其它词频估计策略$R_{MS}(YWL_2)$、$R_{RFB}(YWL_2)$和$R_{CRF+MS}(YWL_2)$相比，由该词频估计策略得到的排序序列$R_{RFB+MS}(YWL_2)$与标准序$R_{YW2}$更为相似。即新的词频估计策略得到的词频估计结果更为准确。

表 2.3是在整个YWL_2上的观察，为了观察各种策略在不同词频段上的性能，我们将YWL_2中词频≥ 10次和≥ 200次的词提取出来构成YWL_2的子集，在这两个子集上进行SCRC的比较实验，表 2.4给出了实验结果。

表 2.4表明，无论考察高频词还是考察除去非常低频词后的情况，与其它策略相比，我们提出的新的词频估计策略都得到更高的SCRC值。

我们总结了在YWL_2的不同的子集下得到的SCRC值比较结果，见表 2.5所

表 2.4　四种估计策略在YWL_2上≥ 10次和≥ 200次的词集合上SCRC值的比较

| | $SCRC(R_{YW2}, R_{MS}(YWL\_2))$ | $SCRC(R_{YW2}, R_{RFB}(YWL\_2))$ | $SCRC(R_{YW2}, R_{RFB+MS}(YWL\_2))$ | $SCRC(R_{YW2}, R_{CRF+MS}(YWL\_2))$ |
|---|---|---|---|---|
| 词频≥ 10 | 0.67 | 0.75 | 0.79 | 0.76 |
| 词频≥ 200 | 0.71 | 0.79 | 0.83 | 0.80 |

表 2.5　YWL_2的不同词频区间段上SCRC值的提升幅度比较

| | 词数 | $RFB+MS$ $VS.MS$ | $RFB+MS$ $VS.RFB$ | $RFB+MS$ $VS.CRF+MS$ |
|---|---|---|---|---|
| 词频≥4 | 76,514 | 0.11 | 0.04 | 0.04 |
| 词频≥10 | 47,270 | 0.12 | 0.04 | 0.03 |
| 词频≥200 | 5,725 | 0.12 | 0.04 | 0.03 |

示。

由表 2.5可知，尽管人工标注语料库的切分更为准确，但规模较小，且由不同机构构建的语料库也存在大量的切分不一致问题。这导致了仅用人工切分语料得到的估计结果并不理想。另外，需要说明的是，虽然CRF+MS的效果略优于RFB，但考虑到CRF分词训练时间的代价远高于MM分词，以及当需要切分的生语料库规模非常大（比如将全部中文网页文本作生语料库），MM分词器的高效优势将更加适合做词频估计任务。这些从另一个角度体现了提出的词频估计策略的优点。

### 2.5.2　观察角度2：排序序列差值

本小节从另一个角度——序列差，进行实验观察。序列差是指对两组排序序列，计算每对序列号的差值的绝对值，进行累加。累加得到的值就是两组序列的序列差。在本试验中，其中一个序列设置为标准答案序列。显然，这个值越小，说明与标准序越接近，即词频估计越准确，反之则与标准序差距大，词频估计不准确。为描述方便，将序列差记作$\sigma$。

由某一种词频估计策略（RFB+MS, MS, RFB, CRF+MS）得到的排序序列记作$R_{scheme}$，则标准序列$R_{YW2}$ 和$R_{scheme}$ 的序列差，记作$\sigma_{R_{scheme}}$，可由

表 2.6　不同词长下对应的序列差比较

| 词长 | $\frac{\sigma_{RFB+MS}-\sigma_{MS}}{\sigma_{MS}}$ | $\frac{\sigma_{RFB+MS}-\sigma_{RFB}}{\sigma_{RFB}}$ | $\frac{\sigma_{RFB+MS}-\sigma_{CRF+MS}}{\sigma_{CRF+MS}}$ |
|---|---|---|---|
| 1 | -23.5% | -18.5% | -17.2% |
| 2 | -20.8% | -16.2% | -16.5% |
| 3 | -15.9% | -9.1% | -8.6% |
| 4+ | 10.2% | 11.5% | 10.9% |

式$\sum_i |R_{scheme}(w_i) - R_{YW2}(w_i)|$ 计算得到。

这样，可以相应地计算$\sigma_{MS}$，$\sigma_{RFB}$，$\sigma_{RFB+MS}$ 和$\sigma_{CRF+MS}$ 的值。

使用不同的词频估计策略可以得到不同的一组$\sigma_{MS}$，$\sigma_{RFB}$，$\sigma_{RFB+MS}$ 和$\sigma_{CRF+MS}$ 的值，$(\sigma_{RFB+MS} - \sigma_{MS})/\sigma_{MS}$，$(\sigma_{RFB+MS} - \sigma_{RFB})/\sigma_{RFB}$ 和$(\sigma_{RFB+MS} - \sigma_{CRF+MS})/\sigma_{CRF+MS}$ 表示了与所提出的词频估计方法相比，其它策略的$\sigma$ 值的变化率。变化率为负数则表示提出的词频估计策略优于与之比较的策略。为正则反之。根据词长不同，可将YWL_2 划分为四个子集（一字词、二字词、三字词和四字以上词），表 2.6给出了四个集合上序列差值的变化率。

由表 2.6可知，对于一字词、二字词和三字词，我们提出的词频估计策略优于其它策略，但是对四字及四字以上词，实验结果显示估计效果变差。

为了得到更详尽的观察结果，我们将YWL_2 中的词按照高中低频进一步细分，观察哪一部分的词造成了效果变差。图 2.3给出了高频前N个词对YUWEI_2 的累积覆盖率曲线，用以确定高、中、低频词集合的范围。

根据图 2.3，我们设置HM 点作为区别高频和中频段的分界线，设置ML 作为中频和低频段的分界线，最终有：

高频段：前6,384个高频词(1 ~ HM)，这些词的词频> 174，对YUWEI_2 的覆盖率达80%；中频段：第$6,385^{th}$ 到$31,794^{th}$ (HM ~ ML)，这些词的词频> 18，高频词加上中频词对YUWEI_2 的累积覆盖率达95%；低频段：剩余的所有词(ML ~ 76,514)，词频> 3。

表 2.7，表 2.8和表 2.9分别给出了在YWL_2的高、中、低频段上考察得到的序列差：

由表 2.7，2.8，2.9可知，大多数情况下（12个中的10个），提出的词频估计策略得到了最好的实验结果。但对于低频段的一字词和4+字词估计效果变差。正是低频段的词使得在4+字词上得到的序列差实验结果变差（见表 2.6）。这一结

表 2.7　YWL_2 中高频段不同词长对应的序列差比较

| | 一字词 | 二字词 | 三字词 | 四字及以上词 |
|---|---|---|---|---|
| $\frac{\sigma_{RFB+MS}-\sigma_{MS}}{\sigma_{MS}}$ | -45.0% | -38.6% | -67.8% | -89.0% |
| $\frac{\sigma_{RFB+MS}-\sigma_{RFB}}{\sigma_{RFB}}$ | -34.8% | -32.9% | -58.7% | -82.0% |
| $\frac{\sigma_{RFB+MS}-\sigma_{CRF+MS}}{\sigma_{CRF+MS}}$ | -32.5% | -31.0% | -56.3% | -76.8% |
| RFB+MS是最优策略？ | √ | √ | √ | √ |

表 2.8　YWL_2 中中频词段不同词长对应的序列差比较

| | 一字词 | 二字词 | 三字词 | 四字及以上词 |
|---|---|---|---|---|
| $\frac{\sigma_{RFB+MS}-\sigma_{MS}}{\sigma_{MS}}$ | -32.4% | -15.8% | -8.3% | -14.2% |
| $\frac{\sigma_{RFB+MS}-\sigma_{RFB}}{\sigma_{RFB}}$ | -19.2% | -8.2% | -10.1% | -11.2% |
| $\frac{\sigma_{RFB+MS}-\sigma_{CRF+MS}}{\sigma_{CRF+MS}}$ | -32.5% | -7.0% | -10.0% | -10.8% |
| RFB+MS是最优策略？ | √ | √ | √ | √ |

表 2.9　YWL_2 中低频段不同词长对应的序列差比较

| | 一字词 | 二字词 | 三字词 | 四字及以上词 |
|---|---|---|---|---|
| $\frac{\sigma_{RFB+MS}-\sigma_{MS}}{\sigma_{MS}}$ | 25.8% | -25.3% | -19.2% | 36.8% |
| $\frac{\sigma_{RFB+MS}-\sigma_{RFB}}{\sigma_{RFB}}$ | -4.2% | -12.4% | -5.7% | 12.9% |
| $\frac{\sigma_{RFB+MS}-\sigma_{CRF+MS}}{\sigma_{CRF+MS}}$ | -3.5% | -12.0% | -5.3% | 8.2% |
| RFB+MS是最优策略？ | × | √ | √ | × |

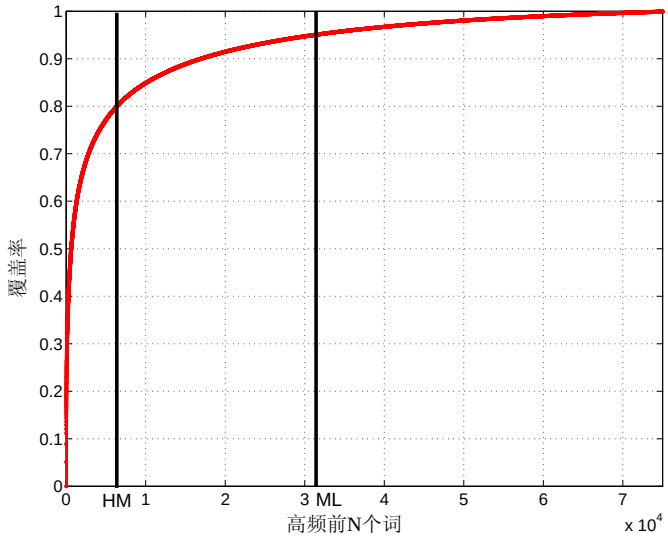



图 2.3　高频前N个词对YUWEI_2的覆盖率

表 2.10　四种策略下，高频前50,000词对YUWEI_2的覆盖率

| 词频估计策略 | $MS$ | $RFB$ | $RFB+MS$ | $CRF+MS$ |
|---|---|---|---|---|
| 覆盖率 | 94.0% | 95.0% | 96.9% | 94.8% |

果可能部分是由于使用的标准评测语料YUWEI 引起的。YUWEI 是一个相对而言规模较大而且平衡的语料库，它给出的词频统计结果较为可靠，但是由于中文大量低频词的存在，使得从YUWEI 得到的低频词的词频统计结果可能产生偏差。YUWEI 语料库的规模还不足以大到可以完全精确估计低频词的程度。因此，我们观察了一些实例，这里我们给出一些与$R_{YW2}$ 相比，在RFB+MS 策略中被调整到高频段的词:“友”,“渡”,“晴”,“深圳证券交易所”,“摩托罗拉公司”,“国务院新闻办”等。

### 2.5.3　观察角度3：对语料库的覆盖率

词频估计的结果越准确，由其给出的高频词部分对语料库的覆盖率越高。这里，根据四种词频估计策略（MS，RFB，RFB+MS和CRF+MS）得到四个排序序列，我们分别选取其各自对应的高频前50,000词，考察其对YUWEI_2的覆盖率。表 2.10给出了覆盖率测试的结果。

由表 2.10可见，与其它三种词频估计策略，MS、RFB和CRF+MS相比，我们

表 2.11　正例样本在高、中、低频段上的分布

| 词频区间 | 词数 | 合理调整样例数目 | 比例 |
|---|---|---|---|
| 高频词 | 6,384 | 5,024 | 78.7% |
| 中频词 | 25,410 | 17,914 | 70.5% |
| 低频词 | 44,720 | 28,129 | 62.9% |

表 2.12　负例样本在高、中、低频段上的分布

| 词频段 | 总词数 | 不合理调整样例数目 | 比例 |
|---|---|---|---|
| 高频词 | 6,384 | 1,360 | 21.3% |
| 中频词 | 25,410 | 7,496 | 29.5% |
| 低频词 | 44,720 | 16,591 | 37.1% |

提出的词频估计策略RFB+MS 在覆盖率上分别提高2.9%，1.9%和2.1%。

### 2.5.4　观察角度4：实例分析

将$R_{YW2}$ 作为标准序，与由人工切分语料库得到的排序序列$R_{MS}(YWL_2)$ 相比，共计51,052个词在$R_{RFB+MS}(YWL_2)$ 中被正确调整（这些词的排序与标准序$R_{YW2}$ 更相近），我们称这些被合理调整的词为正例样本；另有25,447个词被错误调整（这些词的排序更远离标准序$R_{YW2}$），我们称之为负例样本；还有15个词在$R_{MS}(YWL_2)$ 和$R_{RFB+MS}(YWL_2)$ 中有着相同的排序。表 2.11和表 2.12给出了正例样本和负例样本在不同频段（高、中、低频）上的分布情况。

我们对得到的正例样本进行了观察，诸如“生物技术”,“知识经济”,“信息高速公路”和“温室效应”这样的词被排在了更靠前的位置，这些词是近年来新产生的术语和名词，越来越广泛地在人们的生活中使用，因此通过新的词频估计策略得到较靠前的排序是合理的。我们同样给出一些负例样本，如“邱吉尔”,“中央红军”和“使女”。这些词现在已经很少使用，但我们的词频估计策略得到了较高的词频估计，使得这些词的排序更加靠前。造成这一错误估计的原因是这些词在历史上一度被频繁使用，我们使用的大规模生语料库RC 的时间跨度较大，比HUAYU，BEIDA 和YUWEI 更加集中反映了这一语言现象。因此这类曾经是高频的词在现今却变为了低频词，从而造成估计的偏差。

### 2.5.5　观察角度5：在中文分词任务上对词频估计策略的评测

前面的四小节都是从统计观察的角度对词频估计策略进行的评价。这一小节我们将提出的词频估计策略应用到中文分词中，从实际应用的角度出发，检验所提出的词频估计策略的有效性，使得实验结果更有说服力。

中文分词是中文信息处理的基本技术，这里，我们采用基于Unigram模型的中文分词作为应用，评测词频估计策略的优劣。Chen 等就是用基于词的Unigram模型进行中文分词的[58]。采用Unigram模型的分词作为评价词频估计策略优劣的原因是，基于Unigram的分词性能仅取决于词频估计的准确度。因此词频估计策略的优劣可以通过分词性能体现出来。

这里我们简单介绍一下Unigram模型，Unigram模型是N-gram 模型的一个特例。N-gram模型定义了当前状态$x_i$ 只与其前面的n个状态$x_{i-1}, x_{i-2}, \ldots, x_{i-n}$ 有关。用概率形式可以表示为$P(x_i|x_{i-1}, x_{i-2}, \ldots, x_{i-n})$。Unigram模型表示了每个状态只与其自身有关，各个状态之间是独立的关系。假设给定一个句子$S$ 和与$S$ 对应的切分形式，表示为$n$ 个词组成的序列：$w_1, w_2, \ldots, w_n$。由Unigram模型给出的句子$S$ 在该切分形式下成立的概率可由下式表示：

$$Prob(S) = \prod_{i=1}^{n} Prob(w_i),$$

这里$Prob(w_i)$ 表示词$w_i$ 在中文真实语料中出现的概率。$Prob(w_i)$ 的准确值无法得到，所以在实际应用中，一般用训练语料中$w_i$ 在训练语料中出现的次数与语料库中总词数的比值代替。对于给定的句子$S$，可能产生由不同的词序列构成的切分候选集合，通过计算这些切分候选对应的$Prob(S)$ 的值，选取$Prob(S)$ 最大的切分形式作为最终$S$ 的切分结果。动态规划算法可用于计算最优的切分形式。所以，在基于Unigram模型的中文分词中，分词性能取决于对$w_i$ 的词频估计的准确程度。由于在有监督学习方法里，词频估计来源于训练语料，所以训练语料的规模及覆盖领域的平衡性将影响该方法的性能。这里我们用提出的词频估计策略给出$w_i$ 的词频，代替通过训练语料统计词频的方法。从而检验词频估计策略的有效性。分词结果越好，说明词频估计越准确。需要说明的是，这一实验并非旨在提高分词的性能，而是以分词任务为依托，评价词频估计策略的优劣，从应用角度出发检验提出的词频估计策略的有效性，进行评价。

表 2.13　SIGHAN_2005数据集的统计信息

| 语料库 | 训练集规模(KB) | 测试集规模(KB) | 词表规模 |
|---|---|---|---|
| MSRA2005 | 12,542 | 368 | 74,608 |
| PKU2005 | 5,769 | 336 | 51,113 |
| AS2005 | 26,988 | 413 | 141,338 |
| CityU2005 | 6,085 | 133 | 69,085 |

表 2.14　基于unigram模型使用SIGHAN_2005的训练语料得到词频和词表的分词效果

| 语料库 | MSRA2005 | PKU2005 | AS2005 | CityU2005 |
|---|---|---|---|---|
| $F_1$(%) | 91.4 | 88.4 | 88.3 | 86.1 |

在下面的实验中，使用中文分词的综合评价值$F_1$-measure（$F_1$）衡量中文分词的效果（详见第一章关于评价体系的介绍）。

实验采用第三届SIGHAN评测（SIGHAN_2005）[9]提供的所有四个分词数据集进行评测（第一章已进行了相关介绍）：MSRA2005；PKU2005；AS2005；CityU2005。四个语料库的统计信息见表 2.13。使用SIGHAN提供的标准评测程序测试分词结果。

使用Unigram 模型，在SIGHAN_2005四个数据集上进行分词实验，表 2.14给出了实验结果。

为了使得实验比较更全面，在分词实验比较中，除了前面提到的四种词频估计策略：MS，RFB，RFB+MS 和CRF+MS，另外加入了仅使用生语料进行估计的策略，即直接采用串频近似代替词频的策略，记作STR。所以下面的实验中，我们同时比较五种策略：STR，MS，RFB，RFB+MS 和CRF+MS。

由SIGHAN的四个训练集可以得到四个对应的词表，为了进行公平的比较，我们选取这四个词表中词的交集构建一个公用词表，记作CL，用于后面的实验比较。CL中共计包含62,639个词。

前面的四种角度的比较中YUWEI 语料库是作为标准语料库使用，在中文分词的实验中，由于使用中文分词的结果进行评价，不需要设置标准语料库作参照，因此我们将YUWEI 作为和HUAYU、BEIDA 相同的人工切分语料使用，相应地，由人工切分语料库得到词频估计的策略MS，根据式 2-4的定义，由原来的$F_{MS}(w_i, HB)$ 变为$F_{MS}(w_i, HB + YUWEI)$。同时，按照式 2-10中的定义，最终的

表 2.15　以CL为词表，五种词频估计策略在SIGHAN_2005数据集上得到的$F_1$值比较(%)

| 语料库 | STR | MS | RFB | RFB+MS | CRF+MS |
|---|---|---|---|---|---|
| MSRA2005 | 85.2 | 85.6 | 85.4 | 86.3 | 85.9 |
| PKU2005 | 85.4 | 86.8 | 85.8 | 87.4 | 87.2 |
| AS2005 | 78.3 | 78.8 | 78.8 | 80.4 | 79.0 |
| CityU2005 | 77.4 | 78.4 | 77.9 | 80.1 | 79.2 |

词频估计策略应改为$F_{RFB+MS}(w_i, RC + (HB + YUWEI))$。

以CL作词表，使用五种词频估计策略对CL中的词进行词频估计，并通过Unigram模型进行分词，得到五种策略对应的分词结果。表 2.15 给出了比较结果。

由表 2.15可见，与其它四种词频估计策略相比，我们提出的词频估计策略RFB+MS得到了最高的$F_1$值。说明RFB+MS的估计策略优于其它四种。然而，我们也可以看到，由RFB+MS得到的$F_1$值低于表 2.14的结果。考虑到表 2.14的结果是对每个测试集都使用与之对应的训练集获取词表及对应的词频估计得到的。而这里我们的结果是对四个测试集使用一个相同的词表得到的。由于使用的词表不同，因此这一比较并不是在完全公平的条件下进行的。

表 2.15的结果显示RFB+MS是已有几种词频估计策略中效果最好的一个，下面我们去除由于词表不同对最终的分词结果带来的影响，在相同的条件下比较RFB+MS与表 2.15中给出的结果。对每个数据集，在RFB+MS中不再使用CL作为词表，而是对不同的测试集使用不同的训练语料得到的词表，只是词频使用RFB+MS策略估计得到的结果代替。由于CL是四个训练集得到的词表的交集，故对于未在CL中出现的词，采用拉普拉斯平滑(Laplace smoothing)方法处理零值问题。由SIGHAN_2005的四个训练集可以相应地得到四个词表，表 2.16的第一行为四个词表和CL的交集部分的词数。第二行给出了在使用相同词表的情况下，RFB+MS得到的分词结果。

由表 2.16可知，使用相同的词表进行实验，分词结果与表 2.14所示的结果具有可比性（仅有0.1%~0.3%的差距）。同时，注意到在RFB+MS策略中，有50%的词是在CL中未出现的，使用平滑策略进行补救的。平滑策略得到的词频与真实频度有不小的差异，如果这些未出现的词的频度由SIGHAN训练集得到的词频进行补充，分词效果应该会得到提高。按照这一策略进行实验，得到的分词结果

表 2.16　使用SIGHAN训练集词表，RFB+MS 在SIGHAN_2005数据集上测得的分词结果（未出现词使用平滑处理）

| 语料库 | MSRA2005 | PKU2005 | AS2005 | CityU2005 |
|---|---|---|---|---|
| 交集部分词数 | 38,992 | 32,439 | 42,374 | 30,299 |
| $F_1$(%) | 91.3 | 88.3 | 88.0 | 85.9 |

表 2.17　使用SIGHAN训练集词表，RFB+MS在SIGHAN_2005数据集上测得的分词结果（未出现词使用SIGHAN训练集的词频替代）

| 语料库 | MSRA2005 | PKU2005 | AS2005 | CityU2005 |
|---|---|---|---|---|
| $F_1$(%) | 92.3 | 90.3 | 89.8 | 87.4 |

见表 2.17所示。

由表 2.17可知，在SIGHAN评测集上使用本文提出的RFB+MS策略估计词频（未出现的词使用SIGHAN训练集词频信息），可使基于Unigram的中文分词效果分别提高0.9%，1.9%，1.5%和1.3%。这一结果从应用角度验证了本文提出的基于多种类型语料库的中文词频近似策略的有效性和良好的适应性。

## 2.6　本章小结

中文词频估计在自然语言处理的各个领域中都有着重要的应用，如中文词表的构建等。中文词与词之间没有天然分隔的特点使得中文词频估计依然是一个严峻的挑战。在本文中，我们首先分析了现有的几种中文词频估计方法，指出了目前可以用作中文词频估计的语料类型。基于这些分析我们提出了“折中”的策略，对中文词频进行近似，希望通过一种合理有效的策略，对已有的几种类型语料库进行融合，取长补短。宏观统计与在中文分词上进行的实验结果都表明，与现有词频估计策略相比，所提出的基于多类型语料的混合策略得到了最好的词频估计结果。同时，我们也正在根据这一词频估计策略，构建中文信息处理用汉语分词词表。这一工作为词表的建立奠定了坚实的基础。

本章在方法上的创新包括以下几点：

1. 针对目前可以利用的三类语料：生语料、人工标注语料、MM 切分语料，分析了包括使用串频近似、正向最大匹配、反向最大匹配切词后进行统计三种词频估计策略的优缺点。基于以上分析，提出了综合利用几种不同类型语料库

进行中文词频近似的框架。

2. 将中文语言特性，如词长、语料库规模等因素综合进行考虑，体现在综合策略中。通过实验给出了参数选择的方法。

3. 设计了宏观的统计实验，分析并验证了该策略的有效性。实验设计包括与标准词序列的相关度比较、序列差比较以及高频词对语料的覆盖率。

4. 从实际应用的角度，进一步测试并验证了该策略的有效性。我们将该策略下得到的词频估计结果，应用于Unigram中文自动分词模型中，成功验证了所提出的基于多类型语料的词频近似方法的性能。

# 第 3 章　中文分词中的交集型歧义[①]

上一章节我们介绍了语言资源建设中的一个关键问题中文词频估计，本章我们将关注中文自动分词中的切分歧义问题。

第 1.3.2节已经对中文分词歧义作了介绍，第 1.5节对本章的研究内容进行了简述。本章各节内容安排如下：

第 3.1 节简述问题的提出及研究背景；第 3.3节介绍中文交集型歧义的相关概念；第 3.4节介绍通用语料库中最大交集型歧义字段在通用语料库及专业领域语料库的统计特性；第 3.5节介绍专业领域语料库的最大交集型歧义字段在专业领域语料库里的统计特性，并分析了其与通用语料库的异同点；第 3.6节是根据前几节的统计观察提出的消歧策略及效果分析；第 3.7对本工作进行总结讨论。

## 3.1　研究背景

中文自动分词是制约中文信息处理发展的瓶颈之一，其中，中文分词歧义是影响分词精度的重要因素[125]。中文分词歧义主要分为交集型歧义和覆盖型歧义两种，而交集型歧义切分字段占所有歧义字段的90%以上[17]。1998年中国863智能计算机主题对国内一些具有代表性的中文分词软件的评测结果显示，交集型歧义的切分正确率仅为78%[126,127]。学者们对交集型分词歧义进行了很多研究（详见第 1章第 1.3.2节），已有的工作大致可分为基于规则和基于统计的两类：

最早的最大匹配法（MM）是公认的最简单的基于规则的交集型歧义消解策略。黄昌宁（1997）指出，MM对交集型歧义的解决只有97.3%[125]。郑和刘制定了一系列人工规则进行歧义消解[128]，其报告的结果是81.0%。文献[129]用基于词典的方法得到95.0%的准确率。

基于统计方法的消除交集型歧义的方法可大体分为两类：

一种是使用的是字或词的N-gram模型，或者是词性标注的N-gram模型来找

① 本工作的主要内容发表在国际会议TSD’08和中文核心期刊《中文信息学报》上，见发表的学术论文[3,6]。

到待切分句子的候选切分集合。例如孙提出一种利用汉字二元语法关系解决汉语自动分词中的交集型歧义的方法[19]，对交集型歧义消解的正确率为90.3%。

统计学习方法的另一个模式是将中文分词作为二分类问题。例如李蓉等使用字的互信息作为特征，训练基于支持向量机的分类器来进行分词[35] 报告的结果为92.0%。另外，孙等人提出了一种结合了规则学习方法与统计方法的，针对三字长交集型分词歧义的消解策略，达到92.0%的准确率[29]。

虽然交集型歧义已经被研究了很多，但根据我们对现有分词软件的考察，结果显示交集型分词歧义的切分效果依然不能完全令人满意，尤其是当处理专业领域文本时，其性能将进一步下降。要解决好交集型分词歧义问题，需要深入了解交集型歧义字段在中文真实文本中的分布规律及统计特性，以便有针对性地制定解决策略。文献[21,42]基于一个规模约为1亿字的新闻语料库对交集型歧义切分字段进行了详尽的、多角度的考察，确定了4,619个高频最大交集型伪歧义字段（这些伪歧义字段占所考察语料库中全部交集型歧义字段的53.35%），继而提出了一种基于记忆的消歧策略，可快速准确地处理这一部分交集型分词歧义。文献[39]从规模为6.5亿字的人民日报语料库中，抽取到约73万个交集型歧义切分字段，确定了约4万个高频交集型伪歧义字段，并利用文献[21,42] 的策略对它们进行消歧，有效提高了中文分词的准确率。还有一些工作是循着这个思路进行了进一步的研究[130,131]。这些工作均表明，交集型歧义切分字段的统计分布对分词歧义消解有重要作用。

本章的研究是针对文献[42]和文献[39]尚未解决的三个基本问题进行的：

**问题一**：文献[42]和[39] 进行的观察都是基于新闻语料进行的。显然，文章中得到的观察结论应该在更为“合适”的语料上进行确认。否则这些结论仍然不具备足够的说服力。

**问题二**：即便上述观察在更为可信的语料上得到确认，我们仍然需要确定几个重要参数，尤其是确定交集型伪歧义字段核心集合，测试这个核心集合的覆盖率，以及在中文真实文本上检测其稳定性，不仅仅是针对通用语料，还需要在专业领域语料库上进行检测。

**问题三**：迄今为止，针对交集型分词歧义还没有基于专业领域语料库的相关考察。专业领域中交集型歧义字段的统计特性及其与在通用语料库中的异同，从通用语料库得到的高频交集字段的一些结论是否在专业领域语料库中仍然适

用，都需要进一步的考察。

下面我们将在一个大规模通用语料库和两个专业领域语料库上，对上述问题进行考察。第 3.2节将详细介绍所用到的语料库及词表。

## 3.2　语料库

本文所利用的语料库均为未经人工加工的生语料库。其中：通用语料库（记作CBC）取材于小说、新闻、图书、文摘以及网络文本等，年代跨越了1920年至今，规模为929,963,468字；两个专业领域语料库分别是大百科语料库（记作Ency55）和我们自行人工构建的网络语料库（记作Web55）。Ency55来源于《中国大百科全书》电子版，规模为90,023,253字，按大百科的分类体系分为55个专业领域（大百科分类体系见表3.1）；Web55由抓取的20,000篇网页文本构建而成，并按大百科的分类体系对它们人工进行类别标注，规模为54,974,094字。

本文利用的词表来自北京大学的《现代汉语语法信息辞典》[132]，共计74,191个词（记作PKWL）。

本文基于文献[42]提出的最大交集型歧义字段抽取算法，依据PKWL 从CBC中共抽取到733,066个MOAS_Type，对语料库的覆盖率为4.24%；类似地，依

表 3.1　大百科分类体系

| 共55 类 | | | | |
|---|---|---|---|---|
| 『文物、博物馆』 | 『哲学』 | 『新闻出版』 | 『心理学』 | 『教育』 |
| 『图书情报档案』 | 『军事』 | 『语言文字』 | 『经济学』 | 『体育』 |
| 『中国历史』 | 『电影』 | 『机械工程』 | 『政治学』 | 『法学』 |
| 『外国历史』 | 『戏剧』 | 『戏曲曲艺』 | 『物理学』 | 『农业』 |
| 『生物学』 | 『宗教』 | 『中国文学』 | 『社会学』 | 『民族』 |
| 『音乐舞蹈』 | 『电工』 | 『外国文学』 | 『天文学』 | 『纺织』 |
| 『中国传统医学』 | 『化工』 | 『中国地理』 | 『地质学』 | 『轻工』 |
| 『现代医学』 | 『美术』 | 『世界地理』 | 『地理学』 | 『力学』 |
| 『大气、海洋、水文科学』 | | 『土木工程』 | 『化学』 | 『水利』 |
| 『建筑、园林、城市规划』 | | 『环境科学』 | 『数学』 | 『交通』 |
| 『电子学与计算机』 | | 『航空航天』 | 『考古学』 | 『矿冶』 |
| 『财政、税收、金融、价格』 | | 『固体地球物理学、测绘学、空间科学』 | | |
| 『自动控制与系统工程』 | | | | |


图 3.1　交集型歧义字段示例

据PKWL 从Ency55 和Web55 中分别抽取到168,478 和119,663个MOAS_Type，对各自语料库的覆盖率分别为4.20%和3.70%。

在对交集型歧义字段进行观察前，我们首先在第 3.3节介绍交集型歧义的几个相关概念。

## 3.3　交集型歧义的相关概念

首先，介绍交集型歧义字段的概念：设$S$ 是一个中文字符串，$D$ 是一个中文词表，且$S$ 不属于$D$。如果至少存在一个词序列$(w_1, w_2, ..., w_m(m > 2))$，恰好覆盖字符串$S$，且每两个邻接的词互相交叉但又不互相覆盖，则称为交集型歧义切分字段（OAS）；图3.1给出了一个OAS的实例：“其次要”是一个OAS，因为“其次”和“次要”都是词典$D$ 中的词，且他们互相交叠于“次”字。

文献[21]中定义了段长、链长、交集因子、耦合长度、宏结构五个基本概念，用来刻画最大交集型歧义切分字段的细部特征，这里简单描述如下：

***段长***：OAS包含的中文字符个数。例如图3.1中，“其次要”的段长为三。

***交集因子***：在交集字段$S = c_1, c_2, \ldots, c_n$ 中，$W = c_i, \ldots, c_j$ 是$S$ 的一个子串，且满足：

(1) $W$是词；

(2) $S$中不存在包含$W$的词。

则称$W$是$S$的交集因子。例如图3.1中，“其次”和“次要”就是交集字段“其次要”的两个交集因子。显然，一个交集字段至少包含两个交集因子，每个交集因子长度至少为2，因而交集字段的长度至少为3。

***链长***：一个交集字段的交集因子个数称为其链长。图3.1给出的交集字段链长为2。

***耦合段***：交集字段中两个相邻交集因子的相交部分称为它们的耦合段，耦合段的长度称为耦合长度。图3.1给出的交集字段的耦合段为“次”，耦合长度为1。

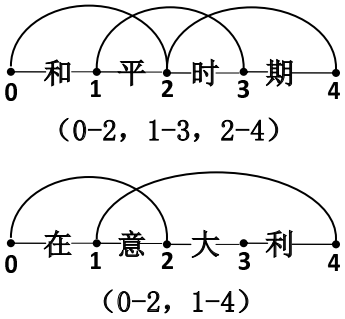



图 3.2　交集型歧义字段的宏结构


图 3.3　最大交集型歧义字段示例

***宏结构***：交集型字段的所有交集因子构成该交集字段的宏结构。定义交集型字段的宏结构是因为，长度相同的最大交集字段的内部结构可能会大相径庭。如“和平时期”和“在意大利”长度均为4，但前者的交集因子为（和平, 平时, 时期）而后者的交集因子则是（在意，意大利），如果用数字标识出交集因子的起始位置，则两者的宏结构可分别表示为(0–2,1–3,2–4) 和(0–2,1–4)，见图3.2所示。

下面详细阐述三个重要的概念：

1、***最大交集型歧义字段***

如果在一个句子（或句子片断）$S_1$ 中,不存在包含$S$ 的更大的交集型歧义切分字段，则称$S$ 为关于$S_1$ 的最大交集型歧义切分字段（MOAS）；

例如在句子“他为推广普通话费尽心血”中,“普通话费”和“普通话费尽心血”均为交集字段，如图3.3所示。

由图3.3可知,“普通话费”包含于“普通话费尽心血”中，故“普通话费”不是最大交集字段。而“普通话费尽心血”不为任何交集字段所包含，所以是最大交集字段。将最大交集歧义字段与交集字段区分开来的好处在于：最大交集字段具有较强的独立性，不再与周围的任何文字形成新的交集因子，这使得该字段在上下文环境中形成一个孤岛，将它们分离出来单独进行研究成为可能。

2、***真歧义***

如果某个MOAS 至少有两种可能出现的切分形式在汉语真实文本中均得到一定程度上的实现，则称其为最大交集型真歧义切分字段（True MOAS，简

称TMOAS）。例1给出了交集型歧义字段“其次要”在不同上下文环境下的不同切分形式。

【例1】

(a) 先解决其主要问题，再解决其次要问题。

(b) 首先要关注整体，其次要注意细节。

3、*伪歧义*

与真歧义相对的，如果某个MOAS 只有一种可能的切分形式，则称其为最大交集型伪歧义切分字段（Pseudo MOAS，简称PMOAS）。例如，“部长篇小说”的两种可能的切分形式——“部长| 篇小说”和“部|长篇小说”中，只有后者可能在真实文本中出现，因此是PMOAS。

下文中，称每一个不同的最大交集型歧义字段为MOAS的段型，记作MOAS_Type。一个MOAS_Type 可能在文本中出现多次，MOAS_Type 的一次出现称为段次，用MOAS_Token 表示。

## 3.4　MOAS关于通用语料库的统计特性

### 3.4.1　最大交集型歧义字段的抽取

孙等人于1998年给出了最大交集型歧义字段的抽取算法[21]，本文的工作延用这一方法提取MOAS。为了论文的完整性，这里引自孙等人对于该算法的描述，见图 3.4所示。其中，算法主体程序ExtractMaxCrossSeg，里面包含一个子程序OneLoc。

### 3.4.2　MOAS的统计分布

针对问题一和问题二，我们设计构建了一个大规模平衡语料库（Chinese Balance Corpus，简称CBC）。CBC语料的内容极为丰富，例如它包含了始自1920年的文学著作。同时它涵盖了类型广泛的文学内容，如小说、散文、新闻、娱乐、健康、幼儿教育以及网络文字信息等。CBC的总规模为929,963,468字。本章通篇采用由北京大学[132]构建的中文词表，从大规模语料中抽取最大交集型歧义字段，该词表共计包含74,191个词，这里将该词表记做CWL。基于CWL，通过查词典的方法，我们自动抽取了所有CBC中的最大交集型歧义字段，共计

**算法：ExtractMaxCrossSeg**

```
BEGIN
  nLoc = 0;
  WHILE(nLoc 未到 S 的末尾)
  BEGIN
    nLoc = nLoc+OneLoc(nLoc);
  END
END
```

**算法：OneLoc**

```
输入：   nLoc
输出：   如果存在从 S 的第 nLoc 位置开始的最大交集字段则返回字段长度
         否则返回从 nLoc 开始的最长词长度
BEGIN
  如果不存在从 nLoc 开始的词，
       返回 1;
  nWordLen = nLen = MaxWordLen(nLoc);
  nCur = nLoc + 1;
  WHILE(nCur < nLoc + nLen)
  BEGIN
    nLen = MAX{nLen, nCur + MaxWordLen(nCur)-nLoc};
    nCur=nCur+1;
  END

  如果 nLen 大于 nWordLen，
    则：找到一个从 nLoc 开始的长度为 nLen 的最大交集字段，并记录之
  否则：不存在一个从 nLoc 开始的交集字段;
  返回 nLen;
END
```

注：算法 OneLoc 中, 记 MaxWordLen(nLoc)为从 S 的第 nLoc 位置开始的按最大匹配得到的最长词长度

图 3.4　最大交集型歧义字段的抽取算法

得到733,066个不同的MOAS_Type，这些MOAS_Type组成本章将要考察的完整的MOAS集合，记作CS-MOAS。这些MOAS_type共计在CBC中出现11,103,551次，共覆盖39,432,267个中文字，占CBC语料库的4.24%。

下面我们将考察交集型歧义字段在汉语真实文本中的分布，从而系统地考察这些交集型歧义字段的统计特性。

### 最大交集型歧义与Zipf定律

图3.5 给出了在CS-MOAS 上考察得到的MOAS_Type与它所对应的MOAS_Token之间的关系，使用对数坐标系（log–log）。从图中可以看到两者之间的关系大致服从Zipf 定律，也即$rank \times TF = C$，这里的常量$C$ 为$0.1 \times \sum TF$（$TF$的累加和为11,103,551）。

图3.6给出了高频前N 个MOAS_Types 对CBC 中的MOAS_Token（也即CS-MOAS在CBC中的所有出现次数）的覆盖率。图中，MOAS_Type 按其对应

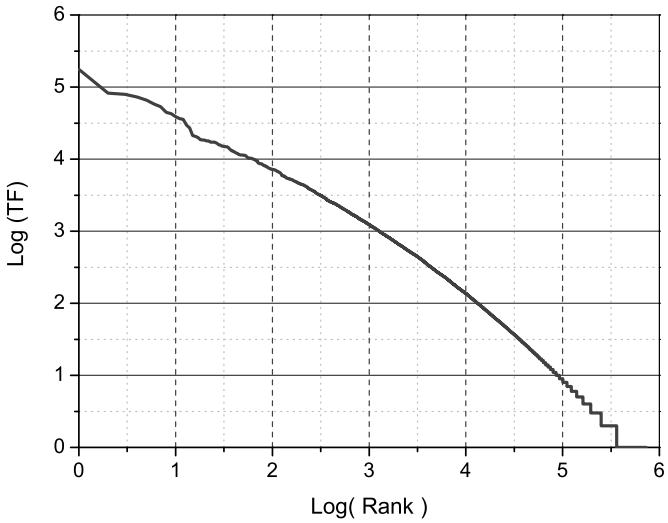



图 3.5　MOAS的分布与Zipf定律的关系

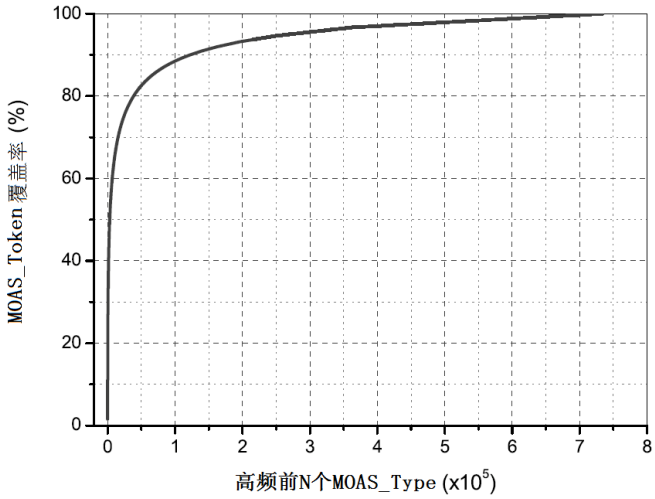



图 3.6　高频前N个MOAS的覆盖率

的MOAS_Token 出现的数目进行排序。由图可知，如我们所期待的一样，高频前3,500，7,000 和40,000个MOAS_Types 对CBC 中出现的MOAS_Token 的覆盖率分别达到50.78%，60.43% 和80.39%。

### MOAS在语料库中的详细分布情况

***观察角度1***: MOAS的段长在CBC上的统计分布（见表 3.2）

表 3.2　MOAS的段长在CBC上的分布

| 长度 | MOAS类型数 | 对MOAS_Type的覆盖率 | 对MOAS_Token的覆盖率 |
|---|---|---|---|
| 3 | 211,270 | 28.82% | 55.22% |
| 4 | 348,065 | 47.48% | 36.86% |
| 5 | 108,786 | 14.84% | 5.83% |
| 6 | 52,233 | 7.13% | 1.81% |
| 7~12 | 12712 | 1.74% | 0.28% |
| 总数 | 733,066 | 100.00% | 100.00% |

表 3.3　MOAS的链长在CBC上的分布

| 链长 | MOAS_type的个数 | 对MOAS_Type的覆盖率 | 对MOAS_Token的覆盖率 |
|---|---|---|---|
| 2 | 269,717 | 36.79% | 64.49% |
| 3 | 383,794 | 52.35% | 33.68% |
| 4 | 53,748 | 7.33% | 1.27% |
| 5 | 23,171 | 3.16% | 0.51% |
| 6~10 | 2,636 | 0.35% | 0.05% |
| **总数** | 733,066 | 100.00% | 100.00% |

由表 3.2可知，段长≤ 4 的MOAS，其对MOAS_Type 和对MOAS_Token 的覆盖率分别为76.30% 和92.08% 而对于段长≤ 6 的MOASs，其覆盖率高达98.26%和99.72%。明显地，这些MOAS_Type 在中文交集型歧义中占有重要分量。

***观察角度2***: MOAS的链长在CBC上的统计分布（见表 3.3）

由表 3.3可知，链长为2和3的MOAS 对MOAS_Type 和对MOAS_Token 的覆盖率分别达到89.14% 和98.17%。链长小于5的MOAS 对MOAS_Type 和对MOAS_Token 的覆盖率更高达99.63% 和99.95%。从链长的分布可以看到，相比较于表 3.2给出的段长分布，链长的分布具有更明显的集中性。从另一个角度进一步表明MOAS 在语料库中的分布具有集中性。

***观察角度3***: MOAS的耦合长度在CBC上的统计分布（见表 3.4）

相比之下，MOAS的耦合长度的分布更为集中，耦合长度为1的MOAS所占比例高达99.66%。

***观察角度4***: MOAS的宏结构在CBC上的统计分布（见表 3.5）

表 3.4　MOAS的耦合长度在CBC上的分布

| 耦合长度 | MOAS_Type的个数 | 对MOAS_Type的覆盖率 | MOAS_Token的个数 | 对MOAS_Token的覆盖率 |
|---|---|---|---|---|
| 1 | 1,300,736 | 99.66% | 15,224,314 | 99.39% |
| 2 | 4,408 | 0.34% | 92,434 | 0.60% |
| 3 | 30 | 0.00% | 1,060 | 0.01% |
| **总数** | 1,305,174 | 100% | 15,317,808 | 100% |

表 3.5　MOAS的宏结构在CBC上的分布

| 宏结构 | MOAS_Type的个数 | 对MOAS_Type的覆盖率 | 对MOAS_Token的覆盖率 |
|---|---|---|---|
| 0-2,1-3,2-4 | 306,627 | 41.83% | 29.63% |
| 0-2,1-3 | 211,270 | 28.82% | 55.22% |
| 0-2,1-3,2-4,3-5 | 40,482 | 5.52% | 0.94% |
| 0-3,2-4,3-5 | 27,064 | 3.69% | 1.28% |
| 0-3,2-4 | 25,846 | 3.53% | 5.08% |
| 0-2,1-3,2-5 | 22,197 | 3.03% | 1.54% |
| 0-2,1-3,2-4,3-5,4-6 | 18,549 | 2.53% | 0.42% |
| 0-2,1-4 | 14,947 | 2.04% | 1.89% |
| 0-2,1-3,2-6 | 13,644 | 1.86% | 0.53% |
| 0-2,1-5 | 8,735 | 1.19% | 0.43% |
| 其它 | 43,705 | 5.96% | 3.04% |
| **总数** | 733,066 | 100% | 100% |

MOAS的宏结构实际上是由链长与交集因子个数的综合表示。例如，OAS“其次要”的宏结构表示为(0-2,1-3)（注：*i*-*j* 表示对于给定的OAS，存在一个链，起始位置为*i*，结束位置为*j*）。我们基于CS-MOAS从CBC中共计抽出了97种不同的宏结构。表 3.5列出了前10种主要的结构类型。

### 3.4.3　MOAS核心集合的稳定性

第 3.4.2小节给出的MOAS在语料库中的各种角度的统计分布显示，MOAS的分布在我们关注的几个方面一致表现出显著的集中性。这些统计特性提示我们，在汉语真实文本中，MOASs的一个核心集合（由数目相对较少的一部分MOASs组成的，覆盖率较高的MOAS集合）有其存在的可能性。图 3.6显示，高

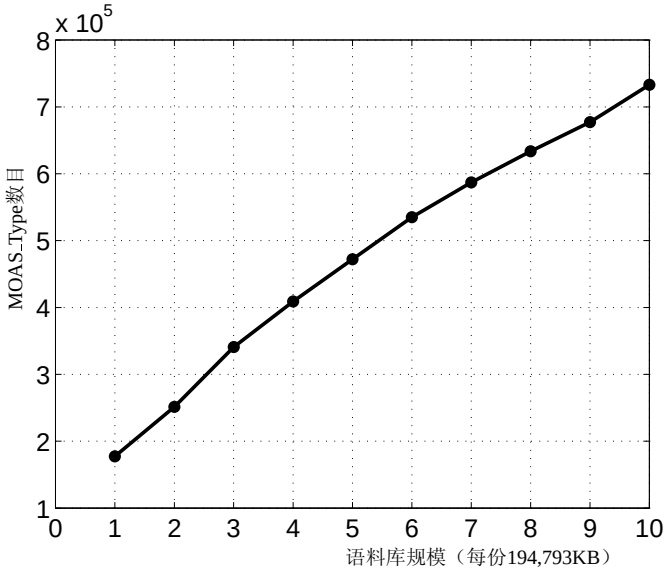



图 3.7　MOAS_Type的个数随语料库规模增长的变化情况

频前3,500，7,000以及40,000个MOAS有可能成为我们所寻找的核心集合候选。

这里遇到的问题是：这些候选MOAS集合是否在汉语真实文本中具有足够的稳定性，换句话说，它们是否具有普适性？

我们从下面两个角度对三个MOAS候选核心集合的稳定性进行观察实验。

**观察角度1：语料库规模对稳定性的影响**

首先，观察语料库规模对候选MOAS核心集合的稳定性的影响。实验设计如下：

我们随机地将CBC分为10个规模相同的子语料库，每一个子语料库的规模为194,793KB。实验从其中随意一份子语料库开始进行，随后逐步增大语料库的规模，每轮实验增加一份子语料库。直至规模达到整个CBC的规模。观察随着语料库规模的增加，MOAS_Type的数目的增长情况以及候选核心集合中的高频MOAS_Type的数目增长情况。

图 3.7显示，子语料库抽取到的MOAS_Type的数目随着子语料库规模的增大而呈线性上升态势。变化非常明显。

相应地，我们对高频前3,500，7,000和40,000个MOAS_Type也进行同样的实验，观察高频MOAS_Type在子语料库中数目的变化情况。其数目随语料库规模

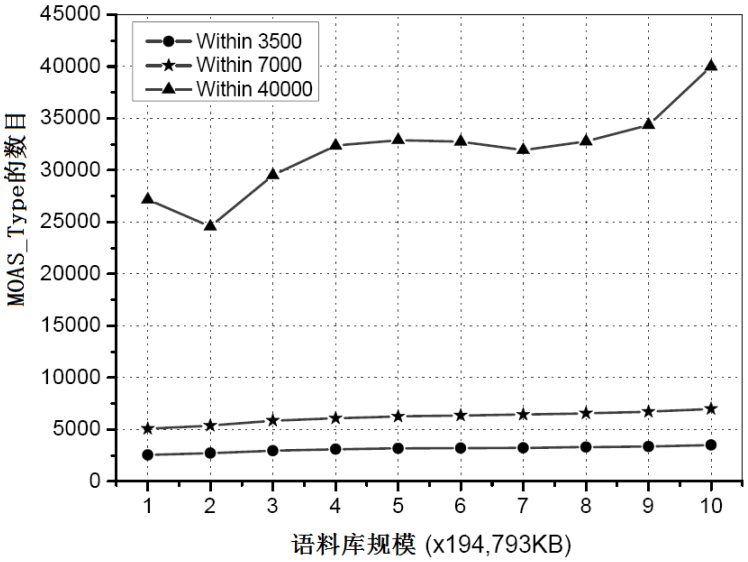



图 3.8　高频前N个MOAS_Type的个数随语料库规模增长的变化情况

增长的变化情况见图 3.8，图中纵坐标表示高频前N个MOAS_Type中，既出现在当前子语料库中，也出现在CBC中的那一部分MOAS_Type的数目。

由图 3.8可见，高频3,500 和7,000 的变化曲线比较平坦，高频部分的MOAS_Type 并未因语料库规模的增长而发生显著变化。高频前40,000的MOAS 集合则相对变化剧烈。

### 观察角度2：专业领域语料库上的稳定性

观察角度1的考察是在通用语料库上进行的，那么这部分高频MOAS在通用语料库中的特性在专业领域语料库上是否也适用呢？为此，我们在设计构建的两个专业领域语料库：Ency55和Web55上考察高频部分MOAS的稳定性。

首先，我们取通用语料库中高频3,500, 7,000和40,000的MOAS集合与Ency55和Web55抽取出的MOAS的交集。每种情况下的交集部分MOAS数目见表 3.6，结果显示高频3,500和高频7,000的MOAS集合在专业领域语料库中依然大量存在。

交集部分的MOAS数目从一个角度反映出CBC中高频MOAS集合在专业领域语料中依然存在，但存在不一定说明其大量稳定地出现在语料中，因此下面我们测试这些交集部分对专业领域语料的覆盖率情况。图 3.9，图 3.10和图 3.11给出了测试结果。如图所示，高频3,500、7,000和40,000的MOAS集合对专业领域语

表 3.6　通用语料库高频前N个MOAS与专业领域语料库高频前N个MOAS的交集部分数目。

| CBC中高频前N个MOAS数目 | CBC与Ency55高频MOAS的交集数目 | CBC与Web55高频MOAS的交集数目 |
|---|---|---|
| 3,500 | 3,135 | 3,439 |
| 7,000 | 6,004 | 6,718 |
| 40,000 | 26,677 | 29,363 |

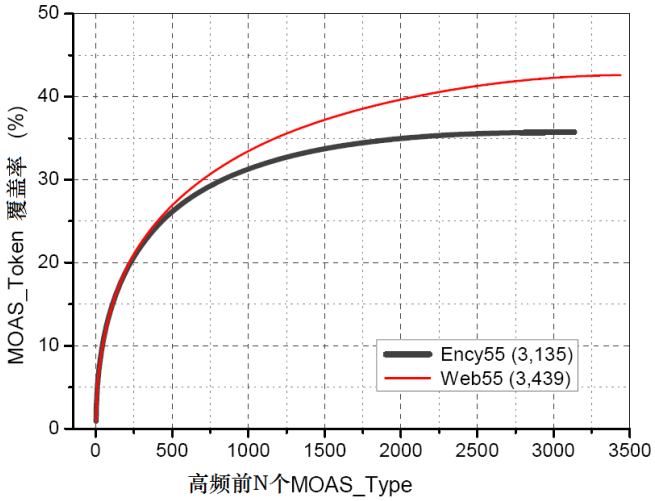



图 3.9　高频3,500MOAS集合对Ency55和Web55的MOAS_Token覆盖率

料的MOAS_Token覆盖率仍然高达45%，50%和70%，与在CBC上所达到的覆盖率（50.78%, 60.43%和80.39%）相比，只有大约5% ~ 10%的下降。这一结果从另一个角度证实了MOAS核心集合存在的可能性。

由上述观察结果可知，高频前7,000 个MOAS 能够覆盖CBC 中60.43%的MOAS_Token，对两个专业领域语料库Ency55 和Web55 中的MOAS_Token 的覆盖率超过42.00%。而且在语料库规模变化的情况下，这部分高频MOAS 具有很好的稳定性。因此，我们最终从核心集合的三个候选中选取高频前7,000个MOAS_Type 作为核心MOAS 集合。这个核心集合将是后面章节研究的基础。

## 3.5　基于专业领域语料库对MOAS统计特性的考察

针对交集型分词歧义，基于通用语料库的考察目前已有不少，然而，如果我们把注意力从通用语料库转移到专业领域语料库，情形会怎样呢？专业领域语

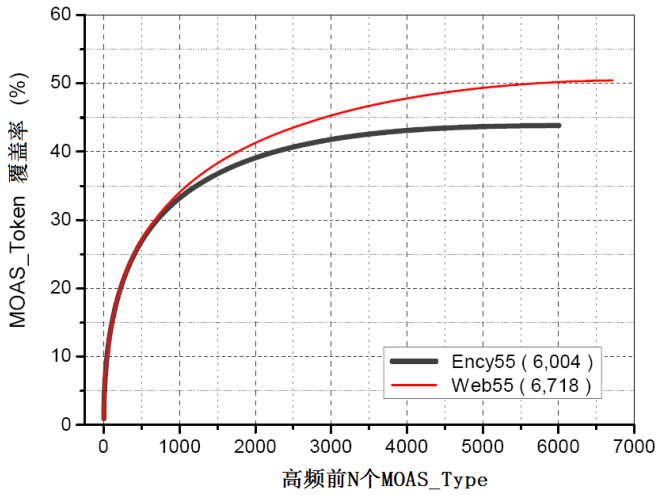



图 3.10　高频7,000MOAS集合对Ency55和Web55的MOAS_Token覆盖率

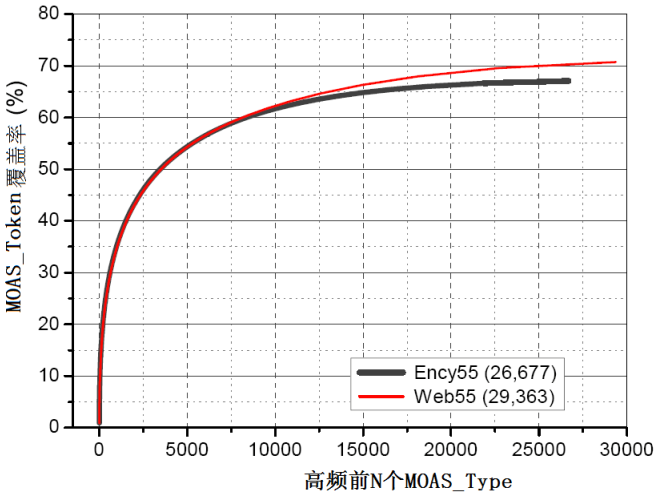



图 3.11　高频40,000MOAS集合对Ency55和Web55的MOAS_Token覆盖率

料库具有更多的专业词汇，这一特殊性会对上述观察结果产生何种影响呢？

本节我们将基于专业领域语料库，对MOAS的统计特性进行考察，这包括：

1、通用语料库中高频MOAS在专业领域语料库中的统计特性（第 3.5.1小节）；

2、专业领域语料库中高频MOAS在专业领域语料库中的统计特性（第3.5.2小节）。

在介绍上述工作前，我们不妨先观察例2中，包含交集型歧义字段“分析出”

的一组例句，以对专业领域语料的特殊性有所认识:

【例2】

(a)“研究人员发现从指纹中可分析出人的生活方式。”（通用领域）

(b)“化工产品是否能分析出具体成分呢?”（专业领域：化工）

(c)“影响煤粉挥发分析出量的因素探讨。”（专业领域：化工）

(d)“煤中的水分和挥发分析出后,所剩下的固体物质称为焦炭。”（专业领域：化工）

在通用领域中,“分析出”基本上都取切分形式“分析 | 出”，可以认为它是一个伪歧义字段（例2 (a)）；在专业领域（化工）中，这种切分形式仍然保存（例2(b)），但会引入另外一种切分形式“分 | 析出”（例2(c)）。

实际上，如果没有专业知识，人一下子也很难搞懂例2中(c)句的意思，切分会出现困难，看过例2的(d)后才恍然大悟。因此,“分析出”在化工领域中转变为真歧义字段。

迄今为止，针对专业领域的汉语交集型分词歧义的考察还是一个空白。本文根据通用词表并基于大规模通用语料库和专业领域语料库，对从通用语料库中抽取的高频交集型歧义切分字段在专业领域语料库中的统计特性，以及从专业领域语料库中抽取的交集型歧义切分字段关于专业领域的统计特性进行了系统考察。

这里需要再次说明的是，这里的观察实验是基于通用词表PKWL来抽取语料库中的最大交集型歧义切分字段的，词表本身未涉及专业领域词汇，因而观察还并不全面。

### 3.5.1　通用语料库中高频MOAS在专业领域语料库中的统计特性

#### 通用语料库中高频MOAS的覆盖率及其歧义类型分布

CBC 中，前7,000 个MOAS_Type 对该语料库中MOAS_Token 的覆盖率达到60.43%。我们对这7,000 个高频MOAS_Type 人工进行了TMOAS 与PMOAS 的判断，表 3.7 给出了PMOAS 和TMOAS 的数目及其对CBC 中全部MOAS_Token 的覆盖率情况，与文献[21]的观察基本一致。

表 3.7　CBC高频7,000个MOAS_Type真、伪歧义分布

| | PMOAS | TMOAS | 高频前7,000 |
|---|---|---|---|
| MOAS_Type数目 | 5,507 | 1,493 | 7,000 |
| 对MOAS_Token的覆盖率 | 52.73% | 7.7% | 66.43% |

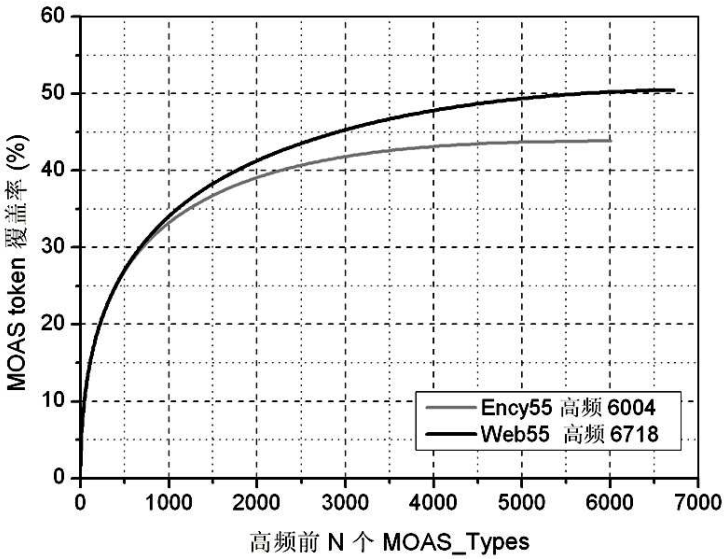



图 3.12　高频MOAS_Type 交集部分对CBC 的MOAS_Token 累积覆盖率

### 通用语料库中高频MOAS在专业领域语料库中的覆盖率

将CBC中前7,000个高频MOAS_Type分别与Ency55和Web55中抽取的MOAS_Type取交集，得到6,004和6,718个MOAS_Type。图 3.12给出了这些MOAS_Type对Ency55与Web55中MOAS_Token的累积覆盖率情况，显示两个交集的MOAS_Type仍能在专业领域语料库中保持较高的覆盖率，达到了45.00%和50.10%。

孙和左的工作指出，PMOAS对于歧义消解有着良好的性质[21]，在通用语料库中有着较高的MOAS_Token覆盖率。那么，它们对专业领域语料库的MOAS_Token覆盖率是否依然可以保持在一个较高水平？我们分别求取了CBC中的5,507个PMOAS与两个专业领域语料库中MOAS_Type的交集。表

表 3.8　CBC中高频PMOAS对专业领域语料库MOAS_Token的覆盖率

| | Ency55 | Web55 |
|---|---|---|
| 交集中的PMOAS数目 | 4,342 | 5,079 |
| 对MOAS_Token的覆盖率 | 45.21% | 47.89% |

3.8给出了交集部分PMOAS的数目及对各自语料MOAS_Token的覆盖率情况，显示它们在专业领域语料库中依然有超过45%的MOAS_Token覆盖率。

### 对通用领域中高频MOAS在专业领域中是否发生歧义类型转变的考察

例2中提示我们，通用语料库中的MOAS_Type（本文只关注高频MOAS_Type）的真伪性在专业领域语料库中可能会发生转变，而可能的转变方式不外乎以下三种：

（1）通用领域中的伪歧义字段在专业领域中转变为真歧义字段；

（2）通用领域中的真歧义字段在专业领域中转变为伪歧义字段；

（3）未发生转变。

第1种方式通常是因为MOAS_Type 所包含的某些词语在专业领域中具有通用领域中所没有或罕见的义项或很少出现的用法，从而使通用领域的PMOAS 在专业领域中转变为TMOAS。我们对CBC 中的5,507 个PMOAS 逐一排查，发现仅有很少量的PMOAS 在特定专业领域中发生了这种转变，如表 3.9所列（其中【】内为在某个专业领域中具有新增义项，或者在通用领域中出现频度很低但在专业领域相对通用的词语）。

其中有些专业领域新增切分形式比较流畅，如“分【析出】”、“之【间架】”、“【应力】求”，有些则受到很大局限，如“【大数】量”、“【和声】音”、“【男女双】方”。需要指出的一点是，判断一个MOAS_Type 所包含的词语是否在专业领域中具有通用领域所没有或罕见的义项，或者判断其是否只在特定专业领域出现，有时很难严格界定，例如就不太好说“男女双方”中的“男女双”一词一定专属于体育领域。

第2种方式意味着通用领域TMOAS 所包含词语的某个义项或用法在专业领域中消失了，从而转变为PMOAS。鉴于本文关注的只是高频MOAS_Type，所以这种情形不太可能发生。例如，“人参与”在通用领域中属真歧义字段，在专业领

表 3.9　通用领域中的PMOAS在专业领域中的转变为TMOAS的实例

| MOAS | 通用领域例句 | 专业领域新增切分形式 | 专业领域例句 |
|---|---|---|---|
| 分析出 | 从Skype聊天中分析出你是个什么样的人 | 分【析出】 | 城市生活垃圾可燃组分挥发分析出动力学预测（化工） |
| 之间架 | 心灵之间架桥梁 | 之【间架】 | 颜体书法之间架结构（美术） |
| 地支配 | 怎样合理地支配财富 | 【地支】配 | 根据天干化五运地支配六气的关系，可以推算出每年的运气（文物博物馆） |
| 使女性 | 睡眠不足易使女性发胖 | 【使女】性 | 此使女性温和，容貌姣好（中国文学） |
| 应力求 | 补贴应力求雪中送炭 | 【应力】求 | 局部薄蒙应力可用等效均布应力求出（物理学） |
| 没有理解 | 没有沟通就没有理解 | 没【有理解】 | 这道题如果没有理解，那无理解该怎么求呢？（数学） |
| 大数量 | 求购大数量特殊视频插头 | 【大数】量 | 生活中的大数量并不少（数学） |
| 和声音 | 用照片和声音留存记忆 | 【和声】音 | 这种和声音很美（音乐舞蹈） |
| 男女双方 | 结婚必须男女双方完全自愿 | 【男女双】方 | 由于队员们的顽强拼搏，我国乒乓球男女双方得以初战告捷（体育） |
| 部长石 | 外经贸部部长石广生就外经贸工作答记者问（注：涉及专名） | 部【长石】 | 连接部长石含量达42%～52%，其中以钾长石为主（地质学） |
| 发病原因 | 医生指出了他的发病原因 | 发【病原】因 | 原发病原因药物得到了控制（现代医学）（注:”原发”在通用领域中一般认为非词） |

域“中国传统医学”中两种切分形式依然可能实现，虽然“人参”的出现频度会有所增加，但那也只是程度上的差异而已。我们的初步考察结论是方式2实际上并不存在。

其余情形都属于第3种方式。与方式2所讨论的情形类似，会在程度上体现出一定差异，但不会产生质变。例如，“以北约”在通用领域、专业领域中两种切分形式都可能实现，区别在于，“北约”在专业领域“军事”中的出现频度也许会较通用领域有所提升。

### 3.5.2　专业领域语料库的MOAS在专业领域语料库上的统计特性

第 3.5.1小节考察了通用语料库高频MOAS 在专业领域语料上的统计特性。本小节将对专业领域语料库中的MOAS 关于专业领域的统计分布进行考察。

#### 专业领域语料库中MOAS的统计分布

我们将Ency55 与Web55 视作一个整体，从中抽取了238,477 个MOAS_Type。我们从段长、链长、耦合长度、宏结构这四个角度考察了这些MOAS_Type 的统计分布，分别列于表 3.10、表 3.11、表 3.12和表 3.13。

表 3.10至表 3.13显示，从专业领域语料库中得到的MOAS的统计分布与孙等（1999）给出的在通用语料库上的观察基本一致。

#### 专业领域语料库中高频MOAS的覆盖率

专业领域语料库（Ency55 + Web55 ）前N 个高频MOAS_Type 关于MOAS_Token 的累积覆盖率如图 3.13所示。

由图 3.13可见，前7,000个高频MOAS_Type的累积覆盖率为58.06%。这与通用语料库中前7,000个MOAS_Type对CBC中MOAS_Token的覆盖率60.43%相比，下降了2.37%。

#### Ency55与Web55中的MOAS关于各专业领域的统计分布及其一致性分析

本小节从以下三个角度来考察，从来源迥异的Ency55 和Web55 分别抽取出来的MOAS 关于55个专业领域的统计分布是否呈现出一定的规律性。

(1) MOAS_Type 关于覆盖领域数目的分布

表 3.10　专业领域语料库中MOAS的段长分布

| 段长 | MOAS_Type 数目 | 对MOAS_Type 的覆盖率(%) | 对MOAS_Token 的覆盖率(%) | 实例 |
|---|---|---|---|---|
| 3 | 84,376 | 35.38 | 53.55 | 中指出、芳香味、等差别 |
| 4 | 110,315 | 46.26 | 38.22 | 道路面积、冰冷藏车、经纪人为 |
| 5 | 29,159 | 12.23 | 6.08 | 松花江流域、蛋白质变性 |
| 6 | 12,070 | 5.06 | 1.88 | 北京大学历史、脊椎动物化石 |
| 7 12 | 2,557 | 1.07 | 0.27 | 丹江口水利枢纽、清代书法理论著作、主要领导人流亡国外、提高人民生活水平方面具有 |
| 合计 | 238,477 | 100.00 | 100.00 | |

注：段长为MOAS所含字数。

表 3.11　专业领域语料库中MOAS的链长分布。

| 链长 | MOAS_Type 数目 | 对MOAS_Type 的覆盖率(%) | 对MOAS_Token 的覆盖率(%) | 实例 |
|---|---|---|---|---|
| 2 | 105,450 | 44.22 | 63.63 | 输出使、出自然界 |
| 3 | 114,818 | 48.15 | 34.40 | 妇女人数、促进化学反应 |
| 4 | 12,504 | 5.24 | 1.41 | 从中国民族、抬高地下水位 |
| 5 | 5,203 | 2.18 | 0.52 | 水平方向导数 |
| 6 11 | 502 | 0.21 | 0.04 | 世界大战前日本心理学、人口学分为人口学分析 |
| 合计 | 238,477 | 100.00 | 100.00 | |

注：“输出使”对应（输出–出使），故链长为2。

表 3.12　专业领域语料库中MOAS的耦合长度分布

| 耦合长度 | MOAS_Type数目 | 对MOAS_Type的覆盖率(%) | 对MOAS_Token的覆盖率(%) | 实例 |
|---|---|---|---|---|
| 1 | 394,067 | 99.00 | 99.00 | 密电阻、个子系统 |
| 2 | 1,989 | 1.00 | 1.00 | 结晶体管、<br>阴雨连绵不绝 |
| 3 | 14 | 0.00 | 0.00 | 热核反应堆、<br>绝妙不可言喻 |
| 合计 | 396,070 | 100.00 | 100.00 | |

注：下划线部分所含字数即为耦合长度。

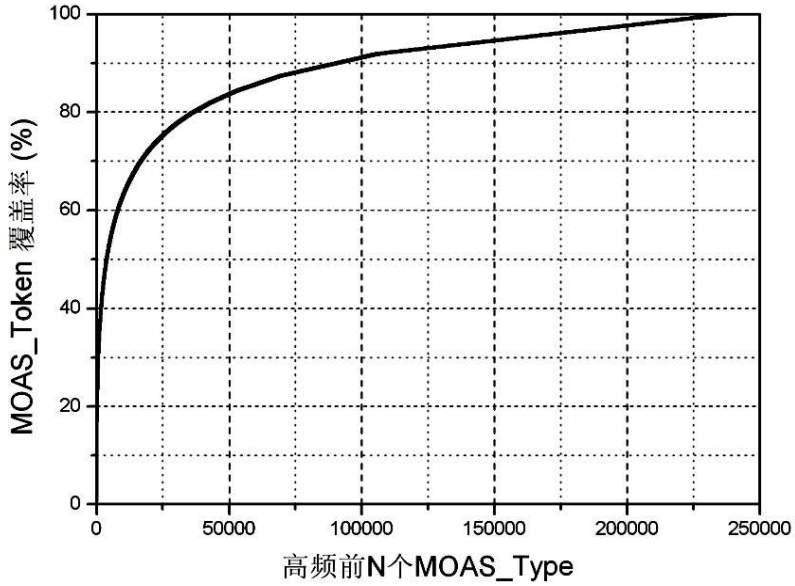



图 3.13　专业领域语料库前N个高频MOAS_Type的累积覆盖率

表 3.13　专业领域语料库中MOAS的宏结构分布。

| 宏结构 | MOAS_Type 数目 | 对MOAS_Type 的覆盖率(%) | 对MOAS_Token 的覆盖率(%) | 实例 |
|---|---|---|---|---|
| (0-2,1-3) | 84,376 | 35.38 | 53.55 | 山丘陵、同形式 |
| (0-2,1-3,2-4) | 93,914 | 39.38 | 30.01 | 实验方法、胚胎发育 |
| (0-3,2-4) | 10,490 | 4.40 | 5.92 | 输电线路、粘土矿物 |
| (0-2,1-3,2-4,3-5) | 9,487 | 3.98 | 1.05 | 法国民法典、面目标定位 |
| (0-3,2-4,3-5) | 7,730 | 3.24 | 1.45 | 磷酸盐矿物、联合国会费 |
| (0-2,1-3,2-5) | 6,758 | 2.83 | 1.75 | 基本地形图、高级数据链 |
| (0-2,1-4) | 5,574 | 2.34 | 1.93 | 冰冷藏车、中共价键 |
| (0-2,1-3,2-4,3-5,4-6) | 4,189 | 1.76 | 0.41 | 国家政治保安、具有劳动能力 |
| (0-2,1-3,2-6) | 2,477 | 1.04 | 0.41 | 传统计划经济、全国有色金属 |
| (0-4,3-5,4-6) | 1,758 | 0.74 | 0.38 | 农民起义战争、浪漫主义气质 |
| 其它 | 11,724 | 4.91 | 3.14 | 横截面积分布、生物化学风化作用 |
| 合计 | 238,477 | 100.00 | 100.00 | |

注：“山丘陵”的宏结构，即(山丘,丘陵), 可表示为(0–2,1–3)。

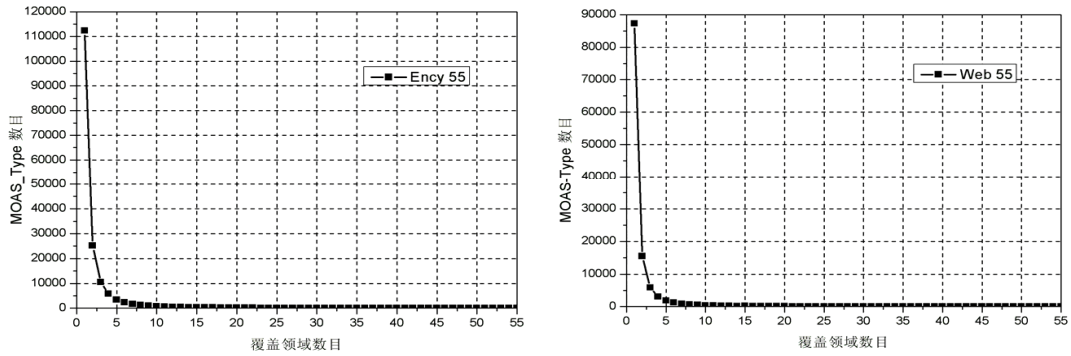



图 3.14　Ency55与Web55中MOAS_Type关于覆盖领域数目的分布

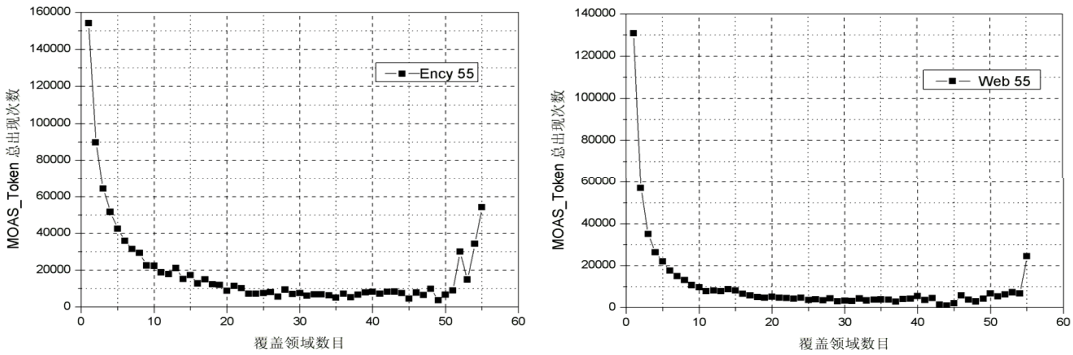



图 3.15　Ency 55与Web55中MOAS_Token关于覆盖领域数目的分布

图 3.14显示了Ency55 和Web55 中覆盖个领域的MOAS_Type 数目（横坐标为覆盖领域数，纵坐标为MOAS_Type的数目）。可以看到，能够覆盖全部55个领域的MOAS_Type 数目很少，大多数MOAS_Type 覆盖的领域数小于20个。当覆盖领域数减少到一定程度后，MOAS_Type 的数目急速增加。这说明专业领域中的大部分MOAS_Type 是领域相关的，所有领域都通用的MOAS_Type 很少。

显然，图 3.14的两个统计分布之间存在高度的一致性。我们采用相关度来衡量一致性。任意两个维数为$n$ 的数值向量$x$ 和$y$ 之间的相关度$R$ 由下式给出：

$$R(x,y)=\frac{\sum_{i=1}^{n}x_i\cdot y_i-n\cdot\overline{x}\cdot\overline{y}}{\sqrt{(\sum_{i=1}^{n}(x_i)^2-n\cdot(\overline{x})^2)(\sum_{i=1}^{n}(y_i)^2-n\cdot(\overline{y})^2)}},$$

其中，$\overline{x}$ 和$\overline{y}$ 分别为$x$ 和$y$ 的平均值。$R$ 越大说明$x$ 与$y$ 的相关度越高。图 3.14的两个统计分布的相关度为0.996。

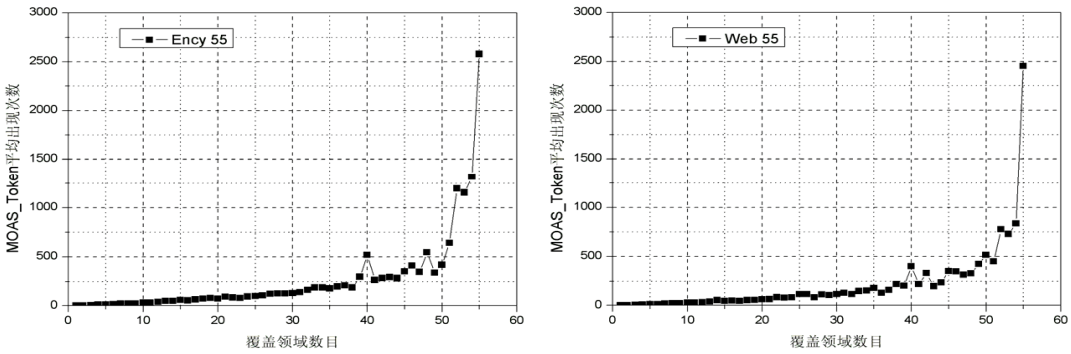



图 3.16　Ency 55与Web55中MOAS_Token平均出现次数关于覆盖领域数目的分布

(2) MOAS_Token关于覆盖领域数目的分布

接下来考察MOAS_Token与所覆盖的领域数目的关系。从图 3.15可以看出，较之图 3.14，覆盖领域数目较多的(≥50) MOAS_Type 所对应的MOAS_Token 数目明显上升，这说明领域覆盖能力强的MOAS_Type 数目虽少，但在各个领域中均有较高的出现频度。图 3.15的两个统计分布的相关度为0.960。

(3) MOAS_Token平均出现次数关于覆盖领域数目的分布

图 3.16显示了MOAS_Token平均出现次数与所覆盖领域数目的关系，两个统计分布的相关度达0.970。它们与图 3.15相互补充，且其趋势与图 3.14恰好相反。

本小节上述三组曲线之间均存在很高的相关度，显示这些统计特性对专业领域是比较稳定的，具有一般性。

### Ency55与Web55中MOAS的交集在专业领域语料库中的统计分布

在上一小节的基础上，我们进一步考察Ency55 与Web55 中MOAS的交集的统计分布，以判断在专业领域中是否存在一个相对稳定的MOAS_Type 集合。我们将Ency55 和Web55 中抽取的MOAS_Type 求交集后，共得到49,664个MOAS_Type，记作C_MOAS_Type。

(1) C_MOAS_Type关于覆盖领域数目的分布

我们将覆盖领域数分为5个区间：覆盖领域数在45～55之间称高覆盖度区间，在35～45和25～35之间称中覆盖度区间，在15～25和0～15之间称低覆盖度区间。

表 3.14　C_MOAS_Type覆盖领域的区间分布

| 领域数目区间 | 45～55 | 35～45 | 25～35 | 15～25 | 0～15 |
|---|---|---|---|---|---|
| Ency55 | 182 | 446 | 944 | 2,351 | 45,741 |
| Web55 | 109 | 258 | 563 | 1,344 | 47,390 |

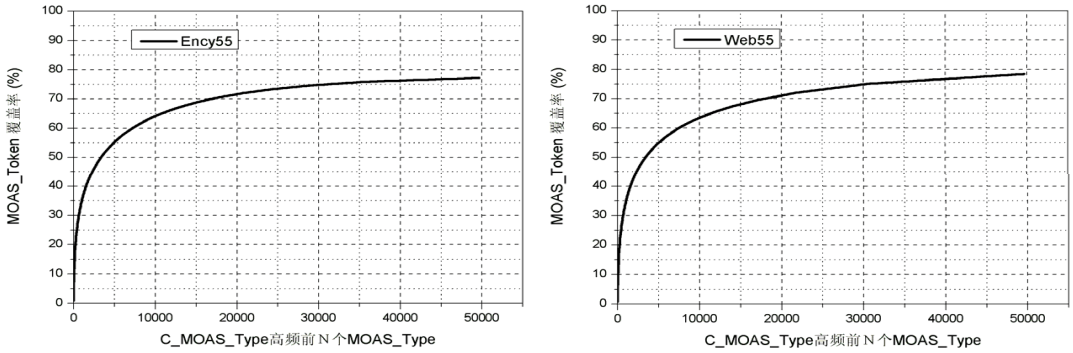



图 3.17　Ency 55与Web55中高频C_MOAS_Type的累积覆盖率

表 3.14给出了每个区间对应的MOAS_Type 数目，可以看到：90%以上的MOAS_Type均分布在低覆盖度区间内，且随着覆盖领域数的增加，MOAS_Type数目逐渐减少。这与前一小节中图 3.14的观察大体一致。

(2)前N个高频C_MOAS_Type的覆盖率

图 3.17给出了C_MOAS_Type 中前N 个高频MOAS_Type 对Ency55 和Web55中MOAS_Token 的累积覆盖率。我们更从C_MOAS_Type 中分别抽出Ency55和Web55 相关的前7,000个高频MOAS_Type，则它们对各自语料库中MOAS_Token 的覆盖率分别达到了61.4%和60.8%。将这两个高频MOAS_Type集合取并集后，再取其10,000个高频MOAS_Type，则对两个语料库MOAS_Token的覆盖率分别达到了62.0% 和61.3%。这显示专业领域语料库中应该存在一个比较稳定的MOAS_Type 集合。

我们将专业领域语料库中确定的这10,000个高频MOAS_Type 与通用语料库中高频7,000个MOAS_Type 取交集，得到了4,420个MOAS_Type。一个观察到的合乎逻辑的现象是，有相当多的通用语料库中高频MOAS_Type，如“今天下午”、“今天上午”、“于今年”、“去年底”、“昨天下午”、“想起来”等，在专业领域语料库变为中低频，而专业领域语料库中一些较高频MOAS_Type，如“潮剧

表 3.15　MOAS核心集合里的PMOAS和TMOAS对CBC的MOAS_Token覆盖率

| | PMOAS | TMOAS | 高频7,000 |
|---|---|---|---|
| MOAS_Type的数目 | 5,507 | 1,493 | 7,000 |
| 对MOAS_Token的覆盖率 | 52.73% | 7.7% | 60.42% |

团”、“纳米线”、“物理学报”、“相关系数”等，在通用语料库中则变为低频。

## 3.6　消歧策略

最终确定的高频7,000个MOAS 可分为PMOAS 和TMOAS（见第 3.3节的介绍），有的情况下，判断一个MOAS 是PMOAS 还是TMOAS 是非常困难的。以“出国门”这个MOAS 为例，在几乎所有情况下其切分形式都应该是“出|国门”，然而仍然不能排除有极少的情况下，该字串具有“出国|”这种切分形式，比如在“他想出国门都没有”这句话里，其切分形式应该是后者。实际上，如果将“出国门”送入搜索引擎（如谷歌），所有检索结果无一例外都是前者的情况。但是后种情况也无法完全排除其出现的可能性，因此按照真歧义的定义，“出国门”是一个TMOAS。但是在实际应用中，无疑应该作为一个PMOAS 对待，否则，对任何分词器而言，与简单地将其作为PMOAS 相比，坚持两种可能的切分都存在会使得切分性能下降的风险更大。

因此，我们在这里放宽了判定为PMOAS 的条件：对于那些在实际应用中，在绝大多数情况下只存在一种切分形式的TMOAS，这里将其作为PMOAS 对待。

基于上述对PMOAS 的判断条件，我们对高频7,000个MOAS 进行了人工判别，共计5,507个MOAS 为PMOAS。这些PMOAS 和TMOAS 对CBC 中所有MOAS 的MOAS_Token 的覆盖率见表 3.15所示。

为了进一步研究这些高频PMOAS的稳定性，我们将这些PMOAS与专业领域语料库Ency55和Web55抽取到的MOAS取交集，表 3.16给出了交集部分对Ency55和Web55的MOAS_Token覆盖率。

表 3.16表明，通用语料库高频7,000个MOAS 里的PMOAS 依然能够覆盖专业领域语料库中45.21%（Ency55）和47.89%（Web55）的MOAS。这说明这部分PMOAS 具有较好的稳定性。因此，如果这部分PMOAS 的切分形式能够确定

表 3.16　PMOAS的交集部分对Ency55和Web55的MOAS_Token覆盖率

| | Ency55 | Web55 |
|---|---|---|
| 交集部分MOAS数目 | 4,342 | 5,079 |
| MOAS_Token覆盖率 | 45.21% | 47.89% |

下来，则在汉语真是文本中，有大约45%的交集型歧义字段的切分就可以得到圆满解决。

由于PMOAS 具有唯一的切分形式，不受上下文的影响，因此这里可以给出一个简单的消歧策略：

对高频PMOAS，以词典方式记录其唯一的切分形式，使用时通过简单的查表方法即可。本质上讲，这是一种基于个体的方法，该方法具有下面的优点：

1. 对汉语真实文本中的MOAS有满意的覆盖率；
2. 可以保证对PMOAS的切分完全正确；
3. 时间和空间复杂度均较低，该策略只需要大约100KB的内存来存储表，通过折半查找可快速定位查询串；
4. 高频PMOAS核心集合是一个潜在的可用于消歧的语言资源。

这一方法本身可以进一步扩展到基于相似度匹配的消歧方法。假设“今天下午”的切分已经存在表中，其切分形式为“今天|下午”如果文中出现“今天上午”，该字串并未记录在表中，我们可以通过相似度匹配扩展出“今天上午”的切分形式。

这里，大家关心的问题是该策略是否对现有分词系统有积极的意义，换句话说，如果现有分词系统已经可以很好地处理这些PMOAS，就不需要这一策略。因此，我们针对现有的公认的性能较好的分词系统——ICTCLAS1.0[115]和MSRSeg1.0[41]进行测试，这两个系统分别在2003和2006的SIGHAN分词评测中得到第一名。对5,507个PMOAS，我们从网络文本中为每个PM 随机选取了一个包含该PMOAS 的例句，显然，使用提出的查表策略，我们可以完美地解决这些PMOAS 的切分。相比之下，ICTCLAS1.0分词系统有2.6%的PMOAS 被错误切分，MSRSeg1.0则出现2.3%的错误切分。

这里，我们给出一些ICTCLAS1.0 中出现的切分错误实例（下划线部分为错误部分）：

核电站 的 特殊性 质

认为 生产 经营 已 经过 时

反 而是 大幅 增加 的

产品 将 在 本届 展 会上 亮相

保险 公司 应 退还 投 保人 相应 的 保险费 及 利息

从 一个 设想 到 开 发出 一 台 样机

他 先后 两 次 在 大 会上 宣读 论文

烧烤 鸡 、 冻 鸡 、 苹果 及 苹 果汁 等

同样，我们也给出MSRSeg1.0中出现的切分错误实例：

联合国外交部 长 安南

我们 来到 了 高尔夫球 场 打球

公安局 长 是 主管 这一 事故 的

我们 应该 要求 他们 开发 票

机动车 辆 应该 按 要求 年检

领导人 员 应该 以身作则

上海市 民 对 这 件 事 不满

第一部 分 讲述 了

税务局 长 被 罢免 了

他 的 心灵 是 富 有的

能够纠正2.3%~2.6%的PMOAS分词错误看上去只是一个很小的提高，但是需要说明的是，这部分提升是在当前最好的分词系统上得到的，这些分词器（ICTCLAS1.0和MSRSeg1.0）大都集中了非常细致和复杂的切分算法，甚至可能还包含了专门应对分词歧义的消歧模块。相对的，这里提出的消歧策略能够简单且无误的解决这些。值得一提的是，该策略带来的性能提升是一种净收益，不会对其它部分的切分带来困扰。这5,507个PMOAS 将会作为公开数据资源供研究使用。

## 3.7　本章小结

交集型分词歧义是汉语自动分词中的主要歧义类型之一。现有的汉语自动

分词系统对它的处理能力尚不能完全令人满意。针对交集型分词歧义，本章进行了两个方面的工作：

一、以前的工作是基于新闻语料进行的考察，规模也较小。本章在规模更大、平衡性较好的大规模语料库上对已有结论进行了论证，并进一步给出了在专业领域语料上的考察，确定了具有满意覆盖率的高频最大交集型歧义字段集合。经实验验证，这一核心集合能稳定覆盖汉语真实文本中42%以上的交集型字段。基于这一核心集合制定的消歧策略，作为预处理运用到目前公认的性能最好的分词系统中，可以进一步纠正2%的因交集型歧义引起的切分错误。这一核心集合的确定，为中文信息处理提供了处理交集型歧义问题的重要资源。

二、基于通用语料库的考察目前已有不少，但还没有基于专业领域语料库的相关考察。本文基于一个含74,191词的汉语通用词表、一个规模约为9亿字的通用语料库和总规模约为1.4亿字的两个专业领域语料库，考察了从通用语料库中抽取的高频最大交集型歧义切分字段在专业领域语料库中的统计特性，发现在这个过程中仅有极少数发生了歧义类型转变，且高频伪歧义字段在专业领域依然具有较强的覆盖能力。我们还考察了从专业领域语料库中抽取的最大交集型歧义切分字段关于专业领域的统计特性，初步揭示了它们关于领域分布的一般规律，发现其中的高频歧义字段在专业领域也具有较强的覆盖能力。本文给出的观察结果对交集型分词歧义的处理（尤其是面向专业领域的交集型分词歧义），具有参考价值。后续工作将深入考量专业词表的引入对通用语料库和专业领域语料库中交集型分词歧义统计特性的影响，以使观察更加全面。

# 第 4 章　搜索与有监督机器学习相结合的中文分词方法 ①

上一章对中文分词中的歧义问题进行了研究，本章我们将针对中文分词的未登录词问题进行研究。

第 1.3.3节已经对中文分词中的未登录词问题及已有的一些处理方法作了介绍，第 1.5节对本章的研究内容进行了简述，现安排本章的各节内容如下：

第 4.1节阐述本章工作问题的提出背景及出发点；第 4.2节讨论目前基于有监督机器学习的中文分词方法的优缺点及可能的提升空间；第 4.3节介绍本文提出的网络搜索技术与有监督机器学习相结合的中文分词框架；第 4.4 节对所提出的中文分词框架进行性能评价及结果分析；第 4.5节对本工作进行总结。

## 4.1　问题的提出和出发点

根据SIGHAN的评测结果报告[8,9]，中文分词在开放测试上达到的最好性能为97.9%（$F_1$-measure），然而对于未登录词的召回率仅为84%。需要指出的是，SIGHAN评测的测试语料的未登录词率(OOV率)仅为2%~8%。因此，这一成绩是在OOV率较低的情况下得到的，当处理的对象是OOV率相对较高的中文真实文本时，分词系统的性能将显著下降。文献[11]指出，在造成中文分词精度下降的因素中，未登录词引起的精度下降是中文分词歧义引起的精度下降的五倍以上。因此能大幅度提高未登录识别率的中文分词方法一般也会提高分词系统的整体性能。

已有的研究结果表明，基于统计机器学习方法的中文分词方法在未登录词的识别方面具有显著的优势，也使得中文分词的精度达到了一个新的高度。在基于统计机器学习方法的中文分词方面，涌现出大量优秀的研究工作，详见第一章的介绍，这里不再赘述。2003年，Xue首次提出将中文分词任务转换为字序列标注问题[92]，基于字标注的中文分词方法的提出，使基于有监督机器学习方法的中文分词成为研究热点。这类方法从训练语料中提取机器学习所需要的语言知识（特征），学习后的模型用于对测试语料进行标注，再转换为分词结果。有

① 本工作的主要部分以全文的形式发表在 PACLIC 2009国际会议上，见发表的学术论文 [5]。

监督机器学习方法的性能很大程度上受训练语料的规模及平衡性的影响。由于训练语料的不易获得，其规模一般都较小，平衡性也较差，通常情况下无法覆盖足够的语言现象，因此很难训练出精确、鲁棒的分类器。而中文真实文本中存在大量的未登录词，如何解决训练语料语言信息不足对有监督机器学习方法造成的限制，是目前面临的一大问题。

近年来，中文网页数目急剧增加。根据中国互联网络信息中心于2010年1月发布的《第25次中国互联网络发展状况统计报告》[133]，中文网页的数量已经超过了336亿，而且其中87.8%以文本形式存在。网络文本可以看作一个巨大的知识库，其规模理论上可以无穷大，且更新很快，可以认为是一个不存在未登录词串的极大规模生语料库。因此，可以考虑的一个解决方法就是利用网络文本作为训练语料的知识补充，以提高未登录词的识别率。当前，利用网络文本或者维基百科（wikipedia）作为资源进行研究越来越受到学界的关注，与本文工作最为相关的是文献[134]，作者提出了一种基于搜索的中文分词方法。这一方法将中文分词转换为一种搜索任务，直接利用搜索引擎在网络文本上寻找切分结果。首先，利用标点等句子中的显示分隔符，将整句拆分为较短的子句。然后，这些子句被作为用户查询串送入搜索引擎中进行搜索。最后，搜索引擎返回的查询结果中的高亮（标记为红色，也称“标红”）部分被用于构建子句的切分结果。在SIGHAN测试集上得到的实验结果显示，这一方法在未登录词的召回率上有一定的提高。这种方法的优点是它是一种完全无监督的分词方法，但其报告的$F_1$-measure仅为87%，这一结果远低于有监督机器学习方法得到的结果（97%以上）。该工作表明，完全基于搜索的方法对未登录词有效，但由于网络文本存在大量噪音，造成了已知词的精确率大幅下降。因而整体性能并不理想。

基于以上出发点，本章的工作考虑如何有效地将网络搜索和有监督机器学习方法进行结合，发挥各自的优势，从而提高中文分词系统的性能。

## 4.2　基于有监督机器学习方法的中文分词

第一章给出了Xue的字标注方法的详细描述，这里用表 4.2给出字标注方法的一个示例。

如表 4.2所示，通过对每个字标注其在词中的位置（L：左边界；M：中间位

表 4.1　字序列标注示例

| 北 | 京 | 中 | 华 | 世 | 纪 | 坛 | 礼 | 花 | 齐 | 放 | 。 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | R | L | R | L | M | R | L | R | L | R | S |

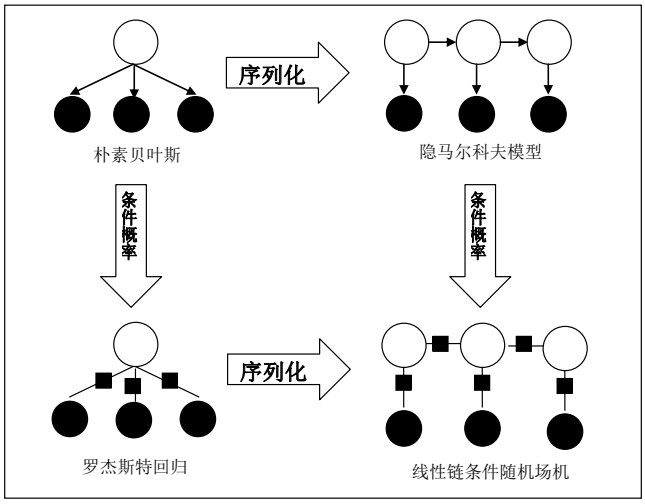



图 4.1　线性链条件随机场模型与其它机器学习方法的关系

置；R：右边界；S：单字），得到待切分句子的分词结果“北京 | 中华 | 世纪坛 | 百花 | 齐放 | 。”。

Xue的工作使得各种有监督机器学习方法广泛应用于中文分词任务上，其中以最大熵模型（Maximum Entropy，简称ME）、多分类SVM、条件随机场（Conditional Random Field，简称CRF）模型为代表。这里我们介绍本文采用的线性链条件随机场模型。

条件随机场模型于2001年提出[36]，属于无向的概率图模型，它可以看作是在罗杰斯特回归（Logistic Regression）基础上经过序列化演变而来的。它与常见的朴素贝叶斯（Naïve Bayes）和隐马尔可夫模型的关系可由下图 4.1表示：

CRF模型训练的目标是最大化全局图结构的条件概率，线性链是图结构的一种最简单形式，适合于中文分词这样的序列标注问题。给定一组观察序列$W$，参数为$\Lambda = \lambda_1, \ldots, K$的CRF模型定义了这组观察序列的预测标记序列$Y$的条件概率，如下式所示：

$$P(Y|W) = \frac{1}{Z(W)} exp(\sum_{t \in T} \sum_{k} \lambda_k f_k(Y_{t-1}, Y_t, W)), \tag{4-1}$$

其中，$f_k$ 为特征函数（一般为二值函数），$\lambda_k$ 为特征 $f_k$ 的权重，$T$ 是预先指定的标记集合，$Z_W$ 是归一化因子。模型参数确定后，使用维特比算法可以有效给出任一观测序列$W$ 的具有最大概率的标记序列$Y^*$：

$$Y^* = \arg\max_Y P(Y|W). \tag{4-2}$$

与传统的HMM 相比，HMM 属于产生式模型（基于联合概率分布的），CRF和ME 则属于判别式模型（基于条件概率的），HMM 过强的独立性假设使得模型无法使用观察序列的多重特征。从模型结构而言，对中文分词这种序列标注问题，CRF 和ME 均优于HMM。与ME相比，ME 模型里标记序列里标记与标记之间是互相独立的关系，而CRF 模型考虑了标记与标记之间的依赖关系，中文分词的标记序列之间存在很强的约束性，如词的左边界后面不可能跟随一个左边界，因此CRF在中文分词上的性能优于ME。另外一点是，CRF 作的是全局归一化，即全句扫描完成才作归一，而ME则是局部归一化。这也从另一个角度说明了ME 性能不如CRF 的原因。

线性链CRF 模型在2004年首次被用于中文分词任务并表现出优越的性能[36,90]。其后又有许多基于CRF 的扩展工作[114]，但大都在这个学习框架之下。研究结果显示，CRF模型是目前被应用于序列标注问题上最为有效的一种机器学习模型。因此，在本章的工作中，我们选取基于CRF 的分词系统作为基本的基于有监督机器学习的分词器，并将基于CRF 的分词性能作为比较的基准。在4.2.1节，我们将介绍文中使用的基于CRF 的分词器的具体实现。

### 4.2.1　基于CRF的中文分词系统的实现

本文使用的基于CRF的分词器，是在CRF++软件包0.53版[①] 基础上实现的。在标注集方面，使用四标注集体系：S（单字词），L（词的左边界），M（词的中间字），R（词的右边界）。特征集合方面，采用五字滑动窗口提取特征，即考虑当前字的标注时，与该字相邻的四个字（左相邻及右相邻各两个字）被抽取作为特征进行训练。特征模板方面，这里我们直接采用文献[38]中给出的特征模板：

(a) $C_n, n = -2, -1, 0, 1, 2$

(b) $C_nC_{n+1}, n = -2, -1, 0, 1$

① http://chasen.org/taku/software/CRF++/

(c) $C_{-1}C_1$

(d) $Pu(C_0)$

(e) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

这里，$C_n$ 代表一个中文字符，$n$ 表示与当前字符$C_0$的相对偏移量。例如，$C_1$ 表示的是$C_0$右边紧邻的字符，$C_{-1}$ 则表示$C_0$ 左边紧邻的字符。$Pu(C_0)$ 表示当前字是否是标点。$T(C_n)$ 表示字符$C_n$ 所属的类别。我们将中文字分为数字、表示日期的字符（“日”,“月”,“年”）、英文字符和其它字符四类。更详细的说明可参见文献[38]。

在 4.2.2小节，我们将对基于CRF的分词结果进行错误分析，确定主要影响其性能的因素，考察可能的提升空间。

### 4.2.2　基于CRF的中文分词系统的切分错误分析

第二届SIGHAN评测（SIGHAN_2005）共提供了四个中文分词评测集，我们采用微软亚洲研究院数据集MSRA2005（见第一章关于评测数据集的介绍）进行分词错误分析。

在MSRA2005的训练集上训练基于CRF的分词器，在MSRA2005的测试集上进行测试，共计出现2,908个切分错误。

为了确定造成切分错误的主要因素，通过对切分结果的分析，这里将中文分词错误粗分为A~D四类：

A类：由OOV引起切分错误；

B类：由切分歧义引起的切分错误；

C类：切分结果虽然与标准答案不符，但仍可以接受；

D类：标准答案本身错误。

基于上面的分类体系，对所有的2,908处分词错误进行人工分类。表 4.2给出了基于CRF的分词系统的分词错误类型分布。

由表 4.2可知，基于CRF的分词器产生的切分错误中，有超过54%的是由OOV问题造成的。OOV问题一直是影响中文自动分词精度的主要因素，通过上述测试可见，对于基于CRF的分词器而言，OOV仍然是精度下降的主要因素。那么基于CRF的分词是否还有提升的可能性和空间呢？上述测试结果是仅考虑概率最高的一种切分形式时得到的，那么如果考虑多个切分候选时情况会如何？

表 4.2　基于CRF的分词系统的分词错误的类型分布（MSRA2005上测得）

| 错误类型 | A | B | C | D |
|---|---|---|---|---|
| 切分错误数目 | 1,581 | 897 | 357 | 73 |
| 错误比例 | 54.4% | 30.8% | 12.3% | 2.5% |

是否有可能最优的切分形式存在于其它切分候选中？换句话说，是否低概率的切分候选实际上是更接近标准答案的切分形式？如果这一假设成立，那么如果我们可以通过重排序（Re-ranking）选出最佳切分，就可以提升现有的基于CRF的分词系统的性能。

为了回答这些问题，在 4.2.3小节，我们将记录置信度最高的前N个切分候选，考察可能带来的提升空间。

### 4.2.3　考虑前N个切分候选带来的提升空间

对基于CRF的分词系统，通过设置输出参数，可以得到带概率值的前N个切分候选。本节测试基于CRF的分词系统在考虑多个切分候选时可能的性能提升。实验设置如下：

对由CRF给出的切分结果，不直接选取概率最高的一种，而是选取概率最高的前N个切分候选，如果标准答案落在这N个候选中，即记作系统给出了正确切分。在具体实验中，通过与标准答案进行比对，从前N个候选中选取与标准答案最相近的一个作为最终的切分结果。这样，考虑N个候选得到的分词性能构成了使用基于CRF的分词系统所能达到的性能上界。也是我们考虑多个候选可能得到的提升空间的上限。

基于上述实验设置，我们在SIGHAN_2005提供的四个中文分词数据集（MSRA2005，PKU2005，CityU2005以及AS2005）[①]上进行实验。这里，关于候选数目N的选取，实验数据表明，考虑更多的切分候选（20个或更多）并不能得到性能的提升，反而会带来噪音，给Re-ranking增加难度。因此，N的取值考虑10个切分候选以内。如何根据待切分句子的不同，有针对性地确定N的数值将在第 4.4 节详细介绍。

① 数据集的统计信息参见第一章的相关介绍

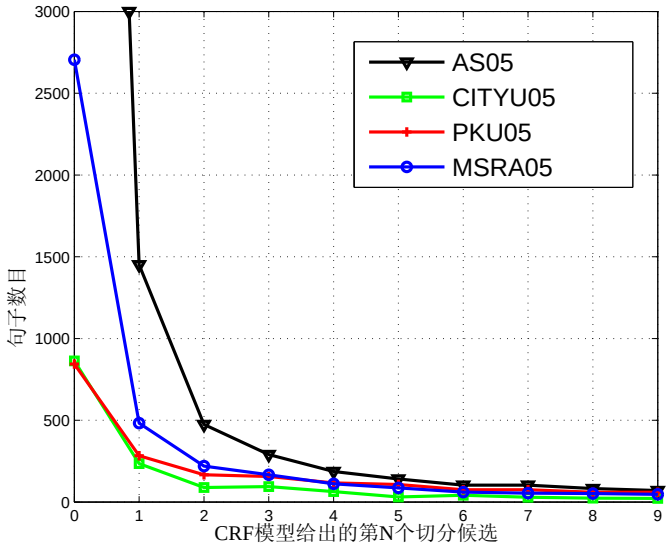



图 4.2　标准答案在10个候选上的分布情况。

表 4.3　考虑10个切分候选带来的性能提升(%)

| | OOV率 | R_OOV 1 best | R_OOV 10 best | R_OOV 提升 | $F_1$ 1 best | $F_1$ 10 best | $F_1$ 提升 |
|---|---|---|---|---|---|---|---|
| MSRA2005 | 2.6 | 75.6 | 89.1 | 13.5 | 96.6 | 98.9 | 2.3 |
| PKU2005 | 5.8 | 76.8 | 84.1 | 7.3 | 94.4 | 96.4 | 2.0 |
| CityU2005 | 7.4 | 78.5 | 91.1 | 12.6 | 95.4 | 98.4 | 3.0 |
| AS2005 | 4.3 | 71.3 | 80.2 | 8.9 | 95.0 | 97.1 | 2.1 |

表 4.3给出了考虑10个切分候选与仅考虑1个切分候选时的性能比较，评价指标使用整体评价值$F_1$-measure（简称$F_1$）和未登录词召回率R_OOV。

由表 4.3可知，考虑10个切分候选时，在$F_1$上的提升大约为2%，在未R_OOV上的提高幅度达到7.3% ~ 13.5%。这一提升具有统计显著性，也是在基于CRF的分词上可以得到的提升上限。

鉴于提升空间较大，基于CRF的分词性能可以期待进一步的提升。下面我们将针对如何对10个切分候选进行Re-ranking，以尽可能地逼近提升的上限。

首先考察最佳切分形式落在其它候选中（非CRF输出的概率最高的候选）的情况所占比例及其分布情况。图 4.2给出了在10个候选上的分布图。其中，横坐

标表示第N个候选（N=0,. . .,9），纵坐标表示由第N个候选给出最佳切分结果的句子数目。

由图 4.2可见，虽然具有最大概率的切分候选在大多情况下给出了最佳切分结果，但最佳切分形式落在其它9个候选中的情况仍然存在，且其总体数目具有一定的规模，如果能有效对10个候选进行Re-ranking，将会得到性能的提升。

## 4.3　搜索与有监督机器学习方法结合的中文分词框架

本节介绍网络搜索与有监督机器学习方法相结合的中文分词框架。这一工作的基本出发点是通过网络搜索挖掘需要的信息对基于CRF 的切分结果候选进行Re-ranking，这源于以下两方面的启示：

一方面，第4.2.2节的实验结果表明，基于CRF 的分词系统的性能受OOV问题影响较大（一半以上），对未登录处理不当是由于训练语料的规模决定了知识缺失不可避免。

另一方面，网络资源可以被视为不存在未登录词的问题（几乎任何新词、热词都会及时在网络文本中出现）。网络信息资源可以被看做一个大的知识库，如果能从中挖掘所需的用于判别未登录词信息，弥补训练语料不足的问题。将可能提高现有的有监督机器学习方法的效果。

文献[134]的方法显示，对已知词基于搜索的方法效果不佳，因此这里考虑对已知词（IV）保留原有的切分结果，而对未登录词进行再次处理。那么现在的问题是：如何有效定位OOV串的位置。

实验数据显示，当待切分句子中出现未登录词时，基于CRF的分词器分配给10个候选的概率会相对平均，或者说差异较小，因此，第1个切分候选得到的概率值相对其它不含未登录词的句子得到的概率要低，这就使得未登录词部分形成一个孤岛，从而通过孤岛现象判断句中是否有未登录存在以及确定其出现的大致位置。

为了提取句子中的孤岛部分，本文提出一种利用切分候选构建格状结构（Lattice）的方法，通过Lattice定位句子中可能存在未登录词串的位置。孤岛部分的字串被作为用户查询送入搜索引擎，构建基于搜索的切分结果。句子的其余部分则保留CRF给出的最佳切分结果。将基于搜索的切分结果与CRF给出

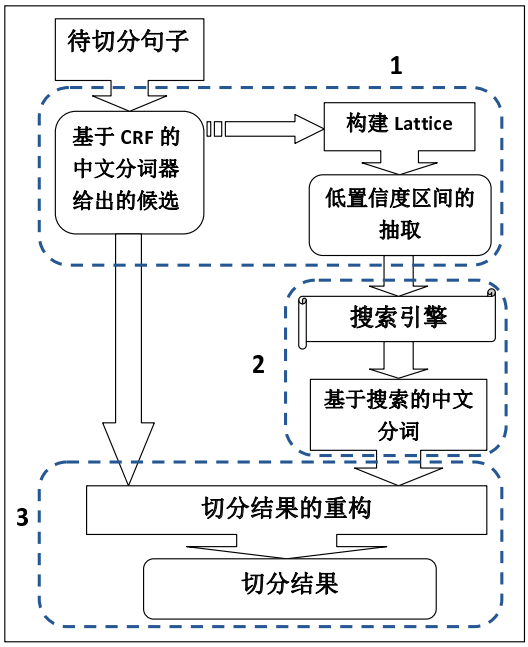



图 4.3　搜索与有监督机器学习方法相结合的分词系统框架图

的候选切分进行相似度度量，选取相似度最高的一个作为孤岛部分字串的最终的切分结果。这样，我们就完成了切分结果的重建。

为使得描述更清晰，图 4.3给出了本章提出的搜索与有监督机器学习方法相结合的分词系统结构图，它包括三个模块：

模块一：构建Lattice，抽取需进一步处理的子串（可能包含未登录词的部分）；

模块二：对抽取的字串进行基于搜索的分词，重构其切分结果；

模块三：重构整句的切分结果。

我们将在接下来的部分一一介绍。

### 4.3.1　模块1: 格状结构的构建

这里，我们给出构建Lattice的形式化描述：

给定一个长度为$l$的待切分字串$S$，则存在$l+1$个切分位置，记作$p_0, p_1, .., p_l$。$S$的可能的切分形式$s$可以用切分位置序列（$\{p_{s_0}, p_{s_1}, ..., p_{s_l}\}$）来表示，这个序列满足以下条件：

(a)　$s_0 = 0, s_l = l$

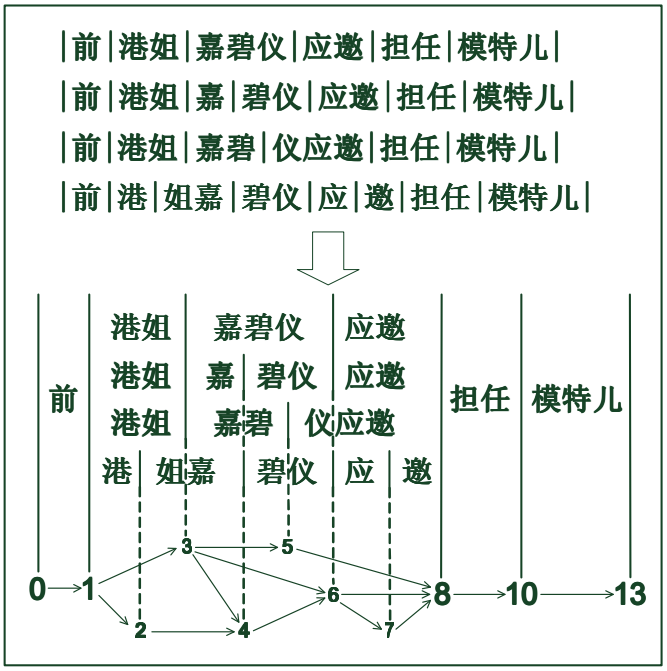



图 4.4　由切分候选构建格状结构示意图

(b) $s_i \in \{0, 1, ..., l\}$

(c) $s_i < s_{i+1}$

这里，$s$可以看作是一种全序，不同的切分形式共同构成一种偏序关系，其中$p_0$为源节点，$p_l$为终节点。以图 4.4为例，图中标号为0的为源节点，标号13的为终节点。如果所有切分候选对应的切分位置序列在某个切分位置$p_i$的结果都是一致的，则它们所定义偏序关系中，该节点为一个合节点，如图 4.4中标为粗体的1，8和10节点位。

所有合节点的位置可以从偏序关系中遍历得到，这里我们将合节点的集合记作$p_u$，如果两个合节点$p_{u_i}$和$p_{u_j}$之间的部分构成的子图是一个全序，则由$p_{u_i}$和$p_{u_j}$之间定义的子串具有唯一的切分形式。例如，图4.4中，节点0和1之间的子串“前”，节点8和10之间的“担任”以及节点10和13之间的“模特儿”。这类字串的切分形式直接通过CRF分词确定下来。

反之，如果$p_{u_i}$和$p_{u_j}$之间的子图不是一个全序，则由这两个节点定义的子串存在几种可能的切分形式，例如图 4.4中，由节点1和8定义的子串“港姐嘉碧仪应邀”。该类字串即抽取到的可能包含未登录词的字串，我们将对其进行基于搜索的分词。由这类字串组成的集合我们称为待切分字串集，记作$S_u$。

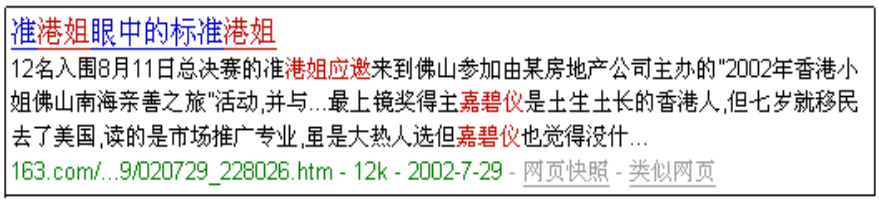



图 4.5　“港姐嘉碧仪应邀”的查询结果示意图

| 切分片段 | 频度 |
|---|---|
| 港姐 | 30 |
| 嘉碧仪 | 23 |
| 应邀 | 8 |
| 港姐应邀 | 7 |

图 4.6　“港姐嘉碧仪应邀”的切分片段集

在第 4.3.2节我们将详细描述如何对$S_u$中的字串构建基于搜索的切分。

### 4.3.2　模块2: 基于搜索的分词

基于搜索的分词是受文献[134]的启发，主要是利用搜索引擎返回的查询片段里的“标红”部分。本文所用的搜索引擎是搜狗中文搜索引擎（Sogou①）。具体的实现过程分为查询片段的收集和切分结果的重建两部分。

#### 切分片段收集

将$S_u$中的字串作为用户查询送入搜索引擎，第一步要做的是收集搜索引擎返回的查询结果片段（简称Snippet），收集与查询串相关的切分片段，这是下一步构建切分结果的基础。

图 4.5给出了当查询串为“港姐嘉碧仪应邀”时搜索引擎返回的查询结果。在查询片段中，每个被标红的字串被称为一个切分片段（简称Segment）。

对$S_u$中的每一个待切分串，记录前100个Snippet，收集切分片段并记录下每个Segment的出现次数。然后按照出现次数降序排序，如图 4.6所示，我们称这样的数据集合为一个切分片段集，每个待切分串都对应一个切分片段集。

① Sogou是搜狐公司开发的中文搜索引擎，也是国内发展最为迅速的中文搜索引擎之一。

**算法描述**

**输入**：　字串集合 P_{string}
　　查询串 S_i 及其切分片段集 Seg_i，　(i=1,...,m) 。
**输出**：　S_i 的标注序列

对每一个 P_{string}中的元素:

　　对 S_i 的每个片段 Seg_i （i=1,...,m ）:

　　　　如果 Seg_i 是 S_i 的子串：
　　　　　　将 Seg_i 从 S_i 的片段集中删除；
　　　　　　将 Seg_i 视为词，标记 S_i 中的字符；
　　　　　　如果 Seg_i 与 S_i 相同：
　　　　　　　　将 S_i 从 P_{string} 中删除；
　　　　　　否则：
　　　　　　　　将 S_i 中 Seg_i 的部分删除；
　　　　　　　　剩余字串作为新的 S_i 加入 P_{string}；
　　　　否则：
　　　　　　删除 Seg_i；

如果 P_{string} 不为空集：
　　将 P_{string} 中剩余部分在 S_i 中对应标记为单字；

输出 S_i 的标记
算法结束

图 4.7　基于搜索的切分结果构建算法

### 构建基于搜索的切分结果

对$S_u$ 里的每一个待切分串，从切分片段集中依次由高到低选择当前第一个片段作为一个切分单元，直至待切分串被覆盖或只剩单字。完成对带切分串的标注。

例如，在图 4.6中，切分片段“港姐”具有最高的出现次数，将在第一轮被挑选出来作为切分单元。按照字标注集的设置，在待切分字串“港姐嘉碧仪应邀”中，“港姐”将被标注为L 和R。然后，再继续挑选频度最高的切分片段，直至“港姐嘉碧仪应邀”被完全覆盖。为描述清晰，图 4.7给出了算法的描述。

这里，大家可能会质疑这种方法是否用到了搜索引擎背后的本地分词器，我们认为，虽然搜索引擎本身有自己的分词器，但我们用到的标红部分与其分词器输出的结果并不是一致的，我们观察了HTML源文件，搜索引擎的切分结果于标红部分大不相同。

## 4.3.3　模块3: 切分结果的重构

对于$S_u$中的字串，我们有基于搜索的切分结果，也有基于CRF的分词器给出

的切分候选。对于待切分句子，有两种最终切分结果构建方式：

1、对句中无法得到一致切分的字串（$S_u$中的字串），直接用基于搜索的切分结果替代。句子中其它部分保留原切分结果。

2、对句中无法得到一致切分的字串（$S_u$中的字串），通过相似度度量，将基于搜索的切分结果与10个候选切分进行比较，选取相似度最高的作为最终切分。句子的其它部分保留原切分。

采用第一种重构方式的分词模型记作Model_1，采用第二种重构方式的分词模型记作Model_2。在实验部分，我们将比较两种模型的优劣。

下面介绍相似度计算方法。受编辑距离的启发，本文定义切分距离（Segment Distance，简称SD）用于度量两种切分形式的近似程度。

对任意两个切分$S_1$和$S_2$，它们的切分距离由使得$S_1$变为$S_2$的从最少插入和删除的边界数目决定，其数学描述为：

$$SD(S_1, S_2) = \min\{\Sigma(Insertion(S_1 \rightarrow S_2) + Deletion(S_1 \rightarrow S_2))\}, \qquad (4\text{-}3)$$

使用动态规划算法可以方便地计算SD的值。

## 4.4　实验与结果分析

实验部分首先确定构建Lattice所需的切分候选数目，然后对所提出的搜索与有监督机器学习方法结合的分词模型的性能进行评测，并进行数据分析。

### 4.4.1　确定构建Lattice所需的切分候选数目

本小节通过实验确定构建Lattice所需要的候选切分的选择方式和数目。

在第 4.2.3节我们已经介绍，N的最大的取值为10，如果对所有的待切分句子都采用10个切分候选来构建Lattice，对于某些句子，10个切分候选的概率分布非常不均衡，后面的候选得到的概率非常小，标准答案落入该候选的几率非常小。这部分候选不会带来性能的提升，反而会使得$S_u$ 中单字待切分片段的数量增多，加大Lattice的复杂度。为了降低构建Lattice的复杂度，我们希望在选择尽量少的切分候选情况下，尽量高概率地包含标准答案。

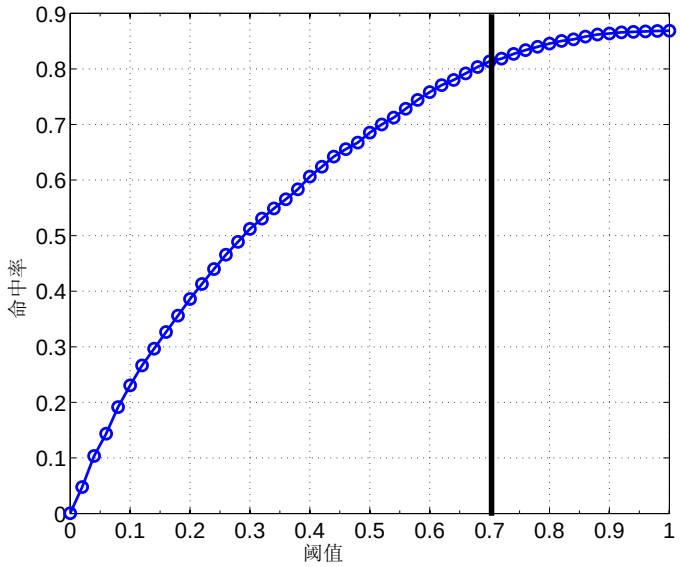



图 4.8　命中率随阈值的增大的变化情况

我们希望通过实验方式，确定一个概率阈值，当切分候选的累积概率达到该阈值时，将停止选取更多的切分候选。这里，我们定义“命中率（hit rate）”为：包含的标准切分数目除以选择的候选切分的总数目。

对10个切分候选，按照概率由高到低的顺序依次被选入，直到累积概率达到设定阈值，考察命中率与概率阈值的关系，图 4.8给出了随着概率阈值的增大，命中率的变化情况。

图 4.8的曲线显示，阈值达到0.7之前，命中率的上升较为明显，之后其变化趋势则趋于平缓。在阈值为0.7时，命中率达到80%。因此，我们选取0.7作为选择切分候选的阈值。

### 4.4.2　性能评价

#### 数据集

第 4.4.1小节确定了候选切分数目，这里我们在SIGHAN_2005提供的MSRA2005、AS2005以及CityU2005三个分词数据集[①]和SIGHAN_2006提供的两个分词数据集（MSRA2006和CTB2006）上对本文提出的分词框架的性能进

① 这里，PKU2005数据集并未应用到性能评价中，因为PKU2005数据的切分标准有所不同，对人名是姓和名分隔开的

表 4.4　SIGHAN_2005和SIGHAN_2006的分词数据集

| 语料库 | 编码 | 训练集规模（MB） | 测试集规模（KB） | OOV率(%) |
|---|---|---|---|---|
| MSRA2005 | GB | 2.37 | 107 | 2.6 |
| AS2005 | BIG5 | 5.45 | 122 | 4.3 |
| CityU2005 | BIG5 | 1.46 | 41 | 7.4 |
| MSRA2006 | GB | 1.26 | 100 | 3.4 |
| CTB2006 | BIG5 | 0.5 | 154 | 8.8 |

行评价。虽然第一章已经给出了介绍，这里为清晰起见，我们将这五个数据集的统计信息综合列于表4.4。评价指标选用$F_1$ 和R_OOV。

### 与基准系统的性能比较

首先将本文提出的分词模型（包括Model_1和Model_2两种）与基于CRF的分词性能进行比较，表 4.5给出了比较结果。

表 4.5　与基准分词系统的性能比较(%)

| 语料库 | CRF | | Model_1 | | Model_2 | |
|---|---|---|---|---|---|---|
| | R_OOV | $F_1$ | R_OOV | $F_1$ | R_OOV | $F_1$ |
| MSRA2005 | 75.6 | 96.6 | 79.6 | 97.0 | 80.9 | 97.1 |
| AS2005 | 71.3 | 95.0 | 75.2 | 95.5 | 75.8 | 95.8 |
| CityU2005 | 78.5 | 95.4 | 81.7 | 96.0 | 82.6 | 96.5 |
| MSRA2006 | 67.3 | 95.3 | 75.7 | 97.2 | 76.5 | 97.3 |
| CTB2006 | 71.2 | 93.0 | 78.3 | 94.6 | 79.5 | 94.7 |

表 4.5显示，Model_1在R_OOV和$F_1$值上均有所提高，其中R_OOV提高幅度为3.2%~8.4%，$F_1$的提升幅度为0.4%~1.9%。相比之下，Model_2在未登录召回率上得到了更好的效果。

表 4.6　与SIGHAN开放测试报告的最好结果的比较（%）

| 语料库 | 参与者 | R_OOV SIGHAN | $F_1$ SIGHAN | R_OOV Model_1 | $F_1$ Model_2 |
|---|---|---|---|---|---|
| MSRA2005 | Wei Jiang | 59.0 | **97.2** | **80.9** | 97.1 |
| | Hwee Tou Ng | 73.6 | 96.8 | **80.9** | **97.1** |
| AS2005 | Hwee Tou Ng | 68.4 | **95.6** | **75.8** | **95.8** |
| | Yaoyong Li | 68.6 | 94.8 | **75.8** | **95.8** |
| CityU2005 | Hwee Tou Ng | 80.6 | **96.2** | **82.6** | **96.5** |
| MSRA2006 | France Telecom | **83.9** | **97.9** | 76.5 | 97.3 |
| | France Telecom | **84.0** | **97.7** | 76.5 | 97.3 |
| CTB2006 | Univ. Texas Austin | 76.8 | 94.4 | **79.5** | **94.7** |

### 与SIGHAN报告的最好结果比较

与SIGHAN开放测试报告的最好结果进行比较[①]，表 4.6给出了具体数值。可以看到，在大多数情况下所提出的分词模型在$F_1$和R_OOV上都得到了显著的提升。SIGHAN评测集中未登录词所占比例相对都较小——大致在2.6%~8.8%。因此，尽管R_OOV得到了显著提高，$F_1$的提升却并不明显——仅在0.2%到0.3%之间。但是可以看到，这一提高在各个数据集上具有普遍性。

### 在高OOV率语料上的性能测试

鉴于SIGHAN的评测集OOV率普遍较低，这里通过一个高OOV率的数据集考察所提出的分词模型的有效性。我们抓取了大约100个中文网页，使用网页文本人工构建了一个小规模的高OOV率的测试集，记作*C_Web*，它包括运动、医学、机械等领域，共计4,000词。首先使用MSRA2005 的训练集训练基于CRF的分词器，再在*C_Web* 上作测试。表4.7给出了在*C_Web* 上得到的分词性能与基于CRF的分词性能的比较结果。可以看到，在高OOV的语料上，本文提出的分词模型显著提高了中文分词的未登录词识别率，在$F_1$ 上也得到具有统计显著性的提高。

① 本文的方法利用了除训练集以外的资源，因此只与SIGHAN开放测试最好成绩相比较

表 4.7　在*C_Web*上与基于CRF的分词性能的比较

| 语料库 | 未登录词比例(%) | R_OOV CRF (%) | $F_1$ CRF (%) | R_OOV CRF+Search +similarity (%) | $F_1$ CRF+Search +similarity (%) |
|---|---|---|---|---|---|
| C_Web | 21.5 | 74.8 | 92.6 | 90.3 | 97.2 |

**算法复杂度分析**

在算法复杂度方面，本算法可大致分为

1、基于CRF的分词器给出切分候选。

若只考虑基于CRF的分词器的解码过程（不考虑训练），则其复杂度等价于维特比算法的复杂度。设标记符数目为$L$（本文标记符数目为4），该字符位置的特征数为$T$，句子长度为$N$，则给出候选切分的时间复杂度为$O(L \times T \times N)$。因为本文的特征模板窗口为5，故该过程时间复杂度不超过$O(N^2)$。

2、由切分候选构成Lattice，提取待处理串。

这一过程是线性时间，扫描一遍切分候选即可得到整个句子的Lattice，即$O(N)$。

3、将待处理串作为查询串送入搜索引擎在网络文本上进行搜索，收集返回结果，统计频度信息并排序。

对某一个长度为N的待切分句子，假设产生$m$个待处理串（$m \leq N$），对前100个查询返回记录进行标红部分的串频统计是一个建表过程，时间复杂度是$O(N)$。对表中元素按照出现频度进行排序，复杂度为$O(N \log N)$，则对$m$个待处理串，其时间复杂度为$O(m \times (N + N \log N))$。

4、构建基于搜索的切分结果，用搜索结果替换从Lattice中抽取的字串部分。

构建基于搜索的切分结果是线性时间$O(N)$，替换切分结果是$O(1)$。

5、用上一步得到的切分结果与候选切分集合逐一进行相似度计算。

每个句子的切分候选数目不超过10个，相似度计算是基于动态规划算法，其复杂度为$O(N^2)$。

综上，对一个待切分句子，不考虑网络搜索下载的实际运行时间，该算法的时间复杂度上限为$O(N^2)$。

表 4.8　MSRSeg1.0和ICTCLAS1.0的典型切分错误实例

| 待切分文本 | ICTCLAS1.0和MSRSeg1.0的切分结果 |
|---|---|
| “万人迷再爆桃色” | 万 人 迷 再 爆 桃色<br>万 人 迷 再 爆 桃色 |
| “港姐嘉碧仪应邀担任模特儿” | 港 姐 嘉碧仪 应邀 担任 模特 儿<br>港 姐 嘉 碧 仪 应邀 担任 模特儿 |
| ”我们来到的士高劲歌” | 我们 来到 的士 高 劲 歌<br>我们 来到 的 士 高劲歌 |

**典型实例**

上述实验说明了新的分词模型的有效性，这里我们通过一些切分实例，与公认的性能较好的公开分词系统相比，进一步考察其在未登录词处理方面的性能。这里，我们使用ICTCLAS1.0① 和MSRSeg1.0② 两个分词系统进行比较测试，从切分结果中，选择了三个具有代表性的切分实例。

表 4.8给出了用本章提出的方法可以得到正确切分结果，而当使用ICTCLAS1.0和MSRSeg1.0时出现切分错误（下划线部分表示切分错误部分）的实例。

这三个实例分别代表着网络新词（万人迷），人名（嘉碧仪）和音译词（的士高）。这些词对于基于有监督机器学习方法的分词系统而言，处理起来是较为困难的，但借助海量的网络资源，其及时的更新性和内容的多样性，使得处理这类未登录词成为可能。

## 4.5　本章小结

本章的初步结论有：

1. 基于有监督机器学习方法的分词器，对未登录的识别能力还很大程度受限于训练语料，也是影响其性能的主要因素。但其对IV的词的识别性能还是很好的。应该充分利用。

① ICTCLAS 1.0: http://www.nlp.org.cn
② MSRSeg.v1.:http://research.microsoft.com/-S-MSRSeg

2. 网络信息对于中文分词中的一大难题——未登录词识别有着重要作用，实际的实验证明，通过利用网络信息资源，可以有效提高未登录词的识别率，从而提高分词的精度。

方法上的创新包括：

1. 探讨了基于有监督机器学习方法，尤其是基于条件随机场模型的分词器的优缺点，对其分词错误进行了分析，提出可能改进的方向。

2. 提出考虑多个切分候选进行切分结果重排的方法，并分析了可行性。

3. 提出了基于切分候选构建Lattice，定位未登录词位置的方法。

4. 提出了将搜索与有监督机器学习方法相结合的中文分词框架。

# 第 5 章　基于$M^3N$的中文分词与命名实体识别一体化方法 ①

第 1.3.3 节对未登录词解决方法的介绍中，我们提到：中文分词与词法层的其它任务，同步处理比分步处理的效果要好，如词性标注[28,85–87]和命名实体识别[41,88,89]。本章基于这一研究结果，提出一种中文分词与命名实体一体化的方法，本章各节内容安排如下：第 5.1 节介绍问题提出的背景；第 5.2节介绍最大间隔马尔可夫网络的数学模型及其优点；第 5.3节介绍基于最大间隔马尔可夫网络的中文分词与命名实体识别一体化方法；第 5.4 节进行性能评价并分析实验结果；第 5.5节对本工作进行总结。

## 5.1　问题的提出

中文自动分词和命名实体识别是中文信息处理中的两个重要环节。近年来，统计机器学习方法被广泛应用于中文信息处理中。按照学习准则的不同, 这些机器学习方法可以大致分为两类：一类是基于最大间隔学习准则的，最具代表性的是支持向量机模型（SVM）[112]，典型的应用为中文文本分类；另一类是基于最大似然估计学习准则的，如条件随机场模型（CRFs）[36]。典型应用如中文分词[90]。本文采用的最大间隔马尔可夫网络（Max-Margin Markov Networks，简称$M^3N$）[113]综合了两类学习方法的优点，既保持了最大间隔学习方法良好的分类能力及推广能力，又考虑了序列的结构信息。

近年来，统计学习方法在中文分词和命名实体识别中也得到了广泛的应用[135–138]，中文分词和命名实体识别均可以转化为基于字的序列标注问题。线性链CRFs和$M^3N$ 都适合于解决这类序列标注问题，之前已经有许多基于CRFs的工作，但基于$M^3N$ 的工作较少。与本文工作最相关的是文献[135]，作者在规模为56万词的训练集上采用$M^3N$ 模型训练中文分词器，在其构建的包含1.1万词的小规模测试集上进行的实验结果显示基于$M^3N$ 的分词器在$F_1$-measure 值上略优于基于CRFs 的分词器，但在未登录词召回率上却略低于CRFs。本文在SIGHAN 2005年（简称SIGHAN_2005）的四个数据集上比较了两个模型的性能。提出了加

① 本工作的主要部分以全文的形式发表在 清华大学学报（自然科学版）上，见已录用的学术论文 [2]。

入标记特征模版和使用动态规划算法对不合法序列进行后处理的方法。

中文分词和命名实体识别是互相影响的两个任务，已有的研究工作表明，在中文分词系统里加入命名实体识别模块，或者将命名实体作为辅助信息，可以提高分词系统的性能[41,88,89]。同样，分词信息也可以为命名实体识别系统提供帮助[138]。文献[138]的工作与本文所提出的一体化方法最相关，作者考虑了分词对命名实体识别的影响，在命名实体识别任务标记集的设置中考虑了分词的影响，在SIGHAN_2006的一个命名实体测试集上得到的$F_1$-measure 值比报告的最好结果提高了1.2%。但该工作并未考察对中文分词任务的影响，而且该工作是基于CRFs 进行的。鉴于$M^3N$ 的优良性能，本文提出基于$M^3N$ 的中文分词与命名实体识别的一体化方法，主要从以下两方面进行考察：

1、基于标记扩展的一体化方法是否能够同时提高两个任务的效果；

2、基于$M^3N$ 的一体化方法是否优于基于CRFs的一体化方法。

## 5.2　最大间隔马尔可夫网络（$M^3N$）

### 5.2.1　数学模型

同其它用于分类的机器学习模型一样，$M^3N$ 模型的学习目标是从包含$m$ 个观测实例的集合$S = \{(x^{(i)}, y^{(i)} = t(x^{(i)}))\}(i = 1, \ldots, m)$ 中学习一个从$X$ 到$Y$ 的映射函数$h : X \to Y$，该函数由包含$n$ 个权重系数的向量$w$ 所决定，每一个系数$w_i$对应一个基函数（特征函数）$f(x, y_i, y_j)$，用$f(x, y)$ 表示基函数，分类机器学习的目标就是求解一个特定的函数$h_w$：

$$h_w(x) = \arg\max_y \sum_{i=1}^{n} w_i f_i(x, y) = \arg\max_y w^T f(x, y).$$

### 5.2.2　与其它有监督机器学习方法的比较

$M^3N$ 模型是多类别SVM在结构预测学习下的一种推广，其与多类别SVM的区别主要在于损失函数的设置不同：

多类别SVM的损失函数仅与当前标记有关，用$t(x)$ 表示样本$x$ 的实际标记，其损失函数定义为：

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

表 5.1　SVM，CRF，$M^3N$模型的异同点

| | SVM | CRFs | $M^3N$ |
|---|---|---|---|
| 使用核函数 | √ | × | √ |
| 错误率理论上界保证 | √ | × | √ |
| 考虑序列结构信息 | × | √ | √ |

即损失函数取决于当前的预测标记是否与真实值一致。$M^3N$模型的损失函数则考虑一个待标记序列里的所有标记，其损失函数与整个标注序列$y$里所有错分标记相关。其损失函数定义为：

$$\Delta t_x(y) = \sum_{i=1}^{l}(\Delta t_x(y_i))$$

其中，$\Delta t_x(y_i) \equiv I\left[y_i \neq t(x)_i\right]$。即损失函数是整个预测序列里错分标记的损失总和。表 5.1从“是否使用核函数”、“是否有错误率的理论上界保证”、“是否考虑序列结构信息”三方面对SVM、CRFs以及$M^3N$ 三个模型进行了比较。

## 5.3　基于$M^3N$的中文自动分词与命名实体识别一体化方法

本节我们将从字符标注体系、特征模板设置以及对不合法序列的处理三个方面进行介绍。

### 5.3.1　字符标注体系

基于字标注的方法在第一章和第四章已经介绍了很多，这里不再赘述。本工作对于单独的分词任务所采用的标注集为四标记集（单字词：S，词的左边界：L，词的中间部分：M，词的右边界：R）。

命名实体识别任务本身也是词边界识别的过程，此外它还需要区分命名实体所属的类别，一般分为人名（简称P）、地名（简称L）、机构名（简称O）三类。如果只考虑命名实体识别任务，标记集可设置为:

非命名实体（O）, 人名/地名/机构名的首字（B-P/L/O）, 人名/地名/机构名的非首字（I-P/L/O）, 人名/地名/机构名的尾字（E-P/L/O）。

中文分词和命名实体识别都可以单独看作一个字序列标注任务，为了将两个任务统一在一个字标注框架下，我们将用于中文分词的四标记集（S、L、M、R）和命名实体类别（P，L，O）进行组合，从而使得中文分词和命名实体识别可以同步进行。表 5.2给出了经过扩展的中文分词和命名实体识别一体化的十六标记集设置：

表 5.2　中文分词与命名实体识别一体化的十六标记集设置

| 类别 | 标记及其含义 |
|---|---|
| 人名 | PS：单字人名 |
| | PL：人名左边界 |
| | PM；人名中间部分 |
| | PR：人名右边界 |
| 地名 | LS：单字地名 |
| | LL：地名左边界 |
| | LM：地名中间部分 |
| | LR：地名右边界 |
| 机构名 | OS：单字机构名 |
| | OL：机构名左边界 |
| | OM：机构名中间部分 |
| | OR：机构名右边界 |
| 非命名实体 | S：单字词 |
| | L：词的左边界 |
| | M：词中间部分 |
| | R：词的右边界 |

扩展后的十六标记集将用于标注训练语料，通过机器学习方法进行联合训练。训练得到的模型将对测试文本中的每个字进行自动标注，输出预测的标注序列。最后，我们将标注序列还原为一个标识了命名实体及词边界的文本。

这里，我们举例说明训练集的标注方式，给定训练文本中（已切分）的一句话：

国家 | 主席 | 胡锦涛 | 来到 | 四川 | 映秀镇 | 视察 | 。

用十六标记集进行标记的结果如表 5.2所示：

表 5.3　十六标注集标注序列示例

| 国家 | | 主席 | | 胡锦涛 | | | 来到 | | 四川 | | 映秀镇 | | | 视察 | | 。 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L | R | L | R | PL | PM | PR | L | R | LL | RR | LL | LM | LR | L | R | S |

在解码过程中，将以P、L、O为前缀的标记识别为不同类别的命名实体，其余标记则仅作词边界的分隔符。通过这一方式，我们可以同时得到分词与命名实体识别的结果。

在字序列标注问题中，序列标记之间存在着天然的约束关系，在我们设计的中文分词和命名实体识别一体化标记集里，L后面只可能出现M和R，而不可能出现S，同样的，(P ｜L ｜O) L之后也不可能出现(P ｜L ｜O) S。类似“(P ｜L ｜O) L-(P ｜L ｜O) S”、“S ｜M”、“R ｜M”这样的标记序列我们称之为是“不合法”序列。基于最大间隔学习准则的$M^3N$ 模型是一个广义的分类模型，可以用在很多与分类有关的任务上，模型本身并不保证预测的标记序列符合特殊的约束关系。因此，当我们将其运用到分词和命名实体任务上时，就需要考虑标记之间的约束，设计合理有效的处理方法。

为了使输出序列合法化，我们可以采用两种策略：

1、对训练过程进行干预；

2、对解码得到的不合法序列进行后处理。

接下来的两小节将逐一介绍这两种策略。

### 5.3.2　特征模板设置

已有的研究工作表明，使用当前字的左右各两个字，即窗口为5的上下文信息作为特征进行训练比较理想[11]，本文直接采用这一结论。参考已有的一些研究工作中的模版设置，我们采用下面五个基本的特征模版抽取训练特征：

(a) $C_n, n = -2, -1, 0, 1, 2$

(b) $C_nC_{n+1}, n = -2, -1, 0, 1$

(c) $C_{-1}C_1$

(d) $Pu(C_0)$

(e) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

其中，$C_n$表示中文字符，$n$表示以当前字符$C_0$ 为参照的相对位移。比如，$C_{-}1$表示当前字符左邻的第一个字符，$C_1$ 表示当前字符右邻的第一个字符，

以此类推。模版(a)~(c)表示了单字及二字串特征；模版(d)表示当前字符是否为标点符号；将中文字符分为字母类（a~z，A~Z）、数字类（0~9，零~九）、日期类（“年”、“月”、“日”）及其它，共四种类别，模版(e)表示由五字序列中每个字所属类别构成的类别序列。这五种特征模版都是基于字的，与字的标记无关。我们称之为字模版。

不合法序列的产生是因为训练得到的模型对序列之间的约束关系学习不足。为了减少不合法序列的产生，我们考虑加入表示标记之间状态转移关系的特征，使得通过训练得到的模型输出不合法序列的概率降低。因此，除了上述五种基本的字模版外，我们增加了两个与标记相关的特征模版(f) 和(g)：

(f) $Tag(C_{-1})Tag(C_0)$

(g) $C_{-1}C_0Tag(C\_1)Tag(C_0)$

其中，模版(f)表示当前字和其前一个字的标记组成的二元串，模版(g)表示当前字和前一个字以及它们对应的标签组成的四元串。由于(f) 和(g) 考虑了标记之间的关系，我们称之为标记模版。字模版和标记模版所产生的特征数目不同，分别为：

字模版所产生的特征数目为：标签类别数× 训练语料里由模版扩展出的所有特征数目；

标记模版所产生的特征数目则为：标签类别数× 标签类别数× 训练语料里由模版扩展出的所有特征数目。

这一方法对应于第一种策略，即对训练过程进行干预。我们将在实验部分观察加入标记特征模版对$M^3N$ 模型训练效果的影响。

### 5.3.3　对不合法序列的后处理

对训练过程进行干预实际上是定义凡是在训练集中出现的状态转移就是合法的，从而使得通过训练得到的模型选择不合法序列作为最终输出结果的概率降低。这种方法并不能保证不出现非法序列，而只能降低其出现概率。因此在解码后还需要对输出的不合法序列做进一步的处理。它包括两种做法：

一是直接输出模型预测出的概率最大的标记（简称为贪心解码），再通过人工规则对造成非法序列的标记进行修改，使得序列合法化。

表 5.4　通过贪心解码得到的标注序列。

| 北 | 京 | 中 | 华 | 世 | 纪 | 坛 | 礼 | 花 | 齐 | 放 | 。 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | R | L | R | L | R | R | L | R | S | R | S |

表 5.5　对不合法序列进行后处理后的标注序列。

| 北 | 京 | 中 | 华 | 世 | 纪 | 坛 | 礼 | 花 | 齐 | 放 | 。 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | R | L | R | L | R | S | L | R | S | S | S |

另一种是通过训练集得到标记之间的转移概率，在解码后所得到的候选标记集中（对一体化标注方法而言，每个字都有16种候选标记），通过动态规划算法选择合法的且得分最高的状态转移路径，得到最终的合法序列标记。

例如，通过贪心解码，我们得到如表 5.4的标注序列：

例句中，有两处不合法序列，分别是“纪坛”对应的“RR”和“齐放”对应的“SR”（标红部分）。通过人工规则或标记之间的转移概率，可以对不合法序列进行纠正，得到合法的标注序列，如表 5.5所示：

综合第 5.3.2小节和第 5.3.3小节的讨论，根据训练过程是否加入标记模版以及选择何种后处理方法，对不合法序列的处理，共有四种组合，这里归纳为以下四种（D1~D4）：

D1. 训练过程不加标记特征，采用贪心解码，作简单后处理；

D2. 训练过程不加标记特征，解码后用动态规划算法作后处理；

D3. 训练过程加入标记特征，采用贪心解码，作简单后处理；

D4. 训练过程加入标记特征，解码后用动态规划算法作后处理。

在实验部分，我们将考察上述四种方法对模型效果的影响。

## 5.4　实验及结果分析

对性能的考察，最理想的情况是在公开的数据集上和报告的最好效果做比较，但在SIGHAN评测里，分词和命名实体识别是作为单独两个任务进行评测的，这为直接评价一体化方法的性能造成困难。另外，本文工作的出发点之一是$M^3N$模型的优良性能，为了验证这一点，同时也为了使实验结果更具说服力和可比性，在实验部分，首先在中文分词任务上（命名实体任务SIGHAN训练语料未公开，故不进行单独评测）对$M^3N$ 与CRFs的性能进行比较，并考察标记特

表 5.6　SIGHAN_2005分词数据集的统计信息

| 语料库 | 训练集（词次数） | 测试集（词次数） | OOV率 |
|---|---|---|---|
| MSRA05 | 2.37M | 107K | 0.026 |
| PKU05 | 2.37M | 107K | 0.026 |
| CityU05 | 2.37M | 107K | 0.026 |
| AS05 | 2.37M | 107K | 0.026 |

表 5.7　SIGHAN_2006命名实体识别测试集信息

| 语料库 | 字数 | 实体数 | 人名数 | 地名数 | 机构名数 |
|---|---|---|---|---|---|
| MSRA06 | 172,601 | 4,362 | 1,973 | 2,886 | 1,331 |
| LDC06 | 92,264 | 1,843 | 1,028 | 106 | 881 |
| CityU06 | 364,361 | 7,511 | 4,941 | 7,450 | 4,016 |

征模版以及四种后处理方法对性能的影响。然后，基于前面的考察结论采用最优配置，在我们构建的训练集上分别训练基于$M^3N$的和基于CRFs的一体化模型，考察一体化方法与两个任务单独进行的性能差异，以及基于CRFs的一体化方法和基于$M^3N$的一体化方法的性能差异。

综上，在实验部分，我们将从以下几个方面考察基于$M^3N$模型的一体化方法的性能：

1. $M^3N$与CRF的性能比较；

2. 本文提出的标记特征模版对模型效果的影响；

3. 考察四种不合法序列处理方法对性能的影响。

然后，基于上述对$M^3N$的考察结论，在我们自己构建的标注了命名实体及词间隔的训练集上训练基于$M^3N$的一体化模型，比较一体化方法与分步方法的性能。这部分包括：

1、基于$M^3N$的一体化方法与分步方法在中文分词和命名实体识别任务上的性能比较；

2、基于CRFs的一体化方法和基于$M^3N$的一体化方法在分词和命名实体任务上的性能比较。

表 5.8　SIGHAN_2006命名实体识别测试集信息

| | 句数 | 字数 | 人名数 | 地名数 | 机构名数 |
|---|---|---|---|---|---|
| PKU_7W | 70,000 | 1,395,020 | 32,723 | 71,943 | 16,382 |

表 5.9　使用字模版所产生的特征数目

| | MSRA05 | PKU05 | CityU05 | AS05 |
|---|---|---|---|---|
| 特征数 | 9,850,456 | 6,415,956 | 8,480,448 | 16,068,952 |

### 5.4.1　数据集

本文实验部分采用SIGHAN提供的数据集进行评测。对于分词任务，我们采用SIGHAN-Bakeoff 2005（简称SIGHAN_2005）提供的四个分词数据集（含训练集和测试集）对模型性能进行评价，它们分别由微软亚洲研究院（MSRA05）、北京大学（PKU05）、台湾中央研究院（AS05）和香港城市大学（CityU05）所构建。这四个数据集的训练集及测试集的统计信息见表 5.6。

对于命名实体识别任务，我们采用SIGHAN-Bakeoff 2006（简称SIGHAN_2006）提供的三个命名实体数据集作为测试集，它们分别是由微软亚洲研究院（MSRA06）、美国语言学数据联盟（LDC06）以及香港城市大学（CityU06）提供的。三个测试集的统计信息见表 5.7。

为了实施本文提出的一体化方法，我们从北京大学计算语言所构建的《人民日报》1998年1 6月语料中抽取70,000句作为训练集（简称PKU_7W）。PKU_7W的统计信息见表 5.8。

本文实验部分中文分词与命名实体识别的性能使用精确率（$P$）、召回率（$R$）以及$F_1$-measure (简称$F_1$)进行评价，中文分词还包括未登录词的召回率（R_OOV）。关于评价体系的介绍见第一章中文自动分词评价指标一节。

### 5.4.2　在SIGHAN数据集上考察$M^3N$的性能

#### $M^3N$与CRFs的性能比较

本文的基于线性链CRFs的分词系统是基于CRF++0.53[①] 开发实现的。

① http://chasen.org/taku/software/CRF++/。

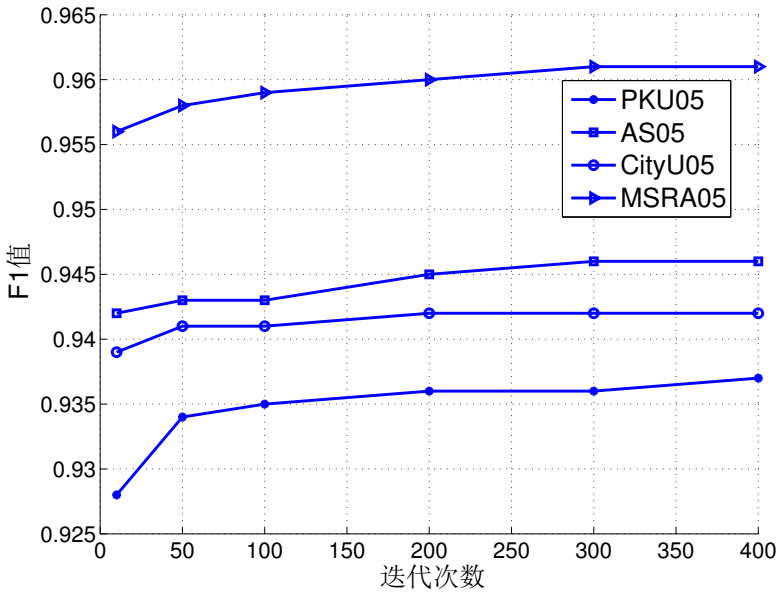



图 5.1　迭代次数对性能的影响

基于$M^3N$的中文分词是基于Egstra实现的，核函数选用线性核函数。采用SIGHAN_2005的四个分词数据集和SIGHAN_2006的三个命名实体识别数据集，选用D1方法进行训练（只使用基于字的特征模版(a)~(e)）和解码。产生的特征数如表 5.9所示：

我们考察了最小10个循环，最多400个循环其性能的变化情况。图 5.1给出了$F_1$值随迭代次数的增加的变化情况。

由图 5.1可知，$M^3N$ 模型在200个循环基本达到了较高的性能，更多的循环对效果影响并不显著。因此后面的实验我们均采用200个次迭代进行训练。

为和CRFs 进行比较，在其它设置不变的情况下，训练线性链CRFs 并使之达到收敛。图 5.2显示了两个模型得到的$F_1$ 值和R_OOV 值的比较结果。

由图 5.2可知，在仅使用字特征模版的情况下，基于$M^3N$ 的分词器比基于线性链CRFs的分词器在$F_1$ 上均有小幅提高，提高幅度分别为0.2%，0.3%，0.1%和0.6%。但在R_OOV上，基于$M^3N$ 的分词器的性能均低于基于CRFs的。

### 加入标记特征模版对训练效果的影响

标记集及后处理方式的设置同上，采用(a)~(g)特征模板，进行200次迭代训

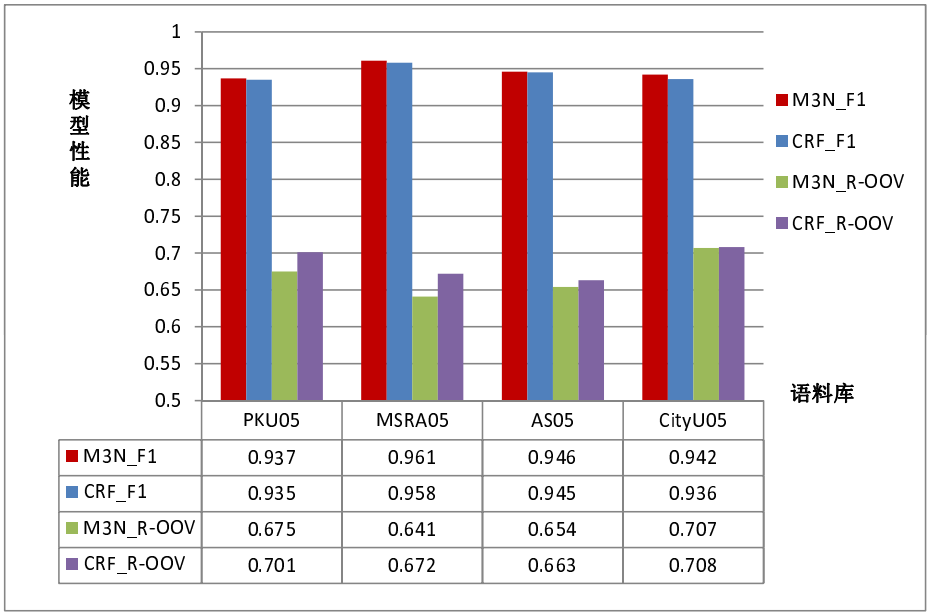



图 5.2　仅使用字模版进行训练的$M^3N$模型和CRFs模型的性能比较

表 5.10　使用标记特征模版所产生的特征数目

| | MSRA05 | PKU05 | CityU05 | AS05 |
|---|---|---|---|---|
| 特征数 | 42,332,328 | 24,799,396 | 33,732,928 | 76,180,152 |

练，表 5.10列出了所产生的特征数目：

加入标记特征训练的$M^3N$和CRFs在四个分词数据集上的比较结果见图 5.3。

比较图 5.2和图 5.3的数据可得，加入标记特征模版后$M^3N$在$F_1$上的提高幅度为1.1%， 1.0%， 0.9%和1.9%。 在R_OOV上的提高幅度为：7.9%，10.9%，6.4%和10.9%。

SIGHAN_2005封闭测试所报告的最好结果（$F_1$值）分别是PKU05：0.950，MSRA05：0.964，AS05：0.952，CityU05：0.943，由图 5.3数据可知，使用标记特征模版训练的$M^3N$分词器在其中三个测试集上超过了报告的最好效果，提高幅度分别为：MSRA05：0.7%，AS05：0.3%，CityU05：1.8%。仅在PKU05上略低0.2%。与基于CRFs的分词相比，无论考察F1还是R_OOV，$M^3N$模型均超过了CRFs给出的最好效果。在四个测试语料上的提高幅度分别为：0.8%，0.6%，0.2%和2%。这一提高在分词任务上是具有统计显著性的。

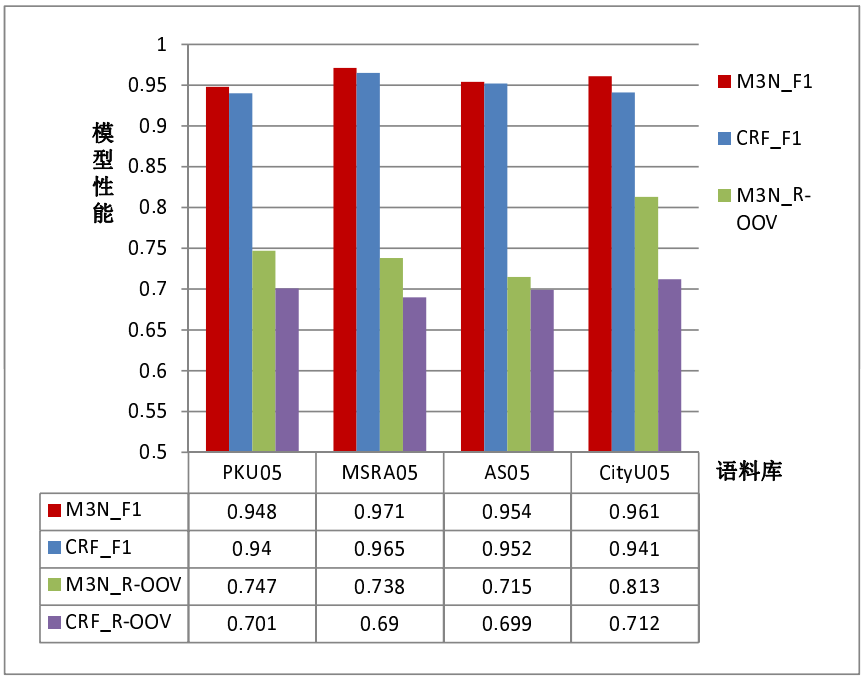



图 5.3　加入标记模版进行训练的$M^3N$模型和CRFs模型的性能比较

**不合法序列对$M^3N$性能的影响以及四种后处理方法的比较**

我们在MSRA05和PKU05两个分词数据集上比较了两种模版情况下线性链CRFs和$M^3N$模型所产生的不合法序列的个数。表 5.11给出了比较结果。

表 5.11　$M^3N$和CRFs产生的不合法序列个数比较

| 语料库 | 仅使用字特征模版 | | 加入标记特征模版 | |
|---|---|---|---|---|
| | $M^3N$ | CRF | $M^3N$ | CRF |
| MSRA05 | 3,652 | 3,389 | 14 | 2,141 |
| PKU05 | 4,443 | 4,576 | 32 | 3,520 |

由表 5.11可知，当加入标记特征后，两个模型的不合法序列数目均有所减少，但从减少的幅度上看，$M^3N$ 模型减少的更为明显，此时CRFs 产生的不合法序列个数远远大于$M^3N$ 模型。下面我们考察四种后处理方法的优劣。在其它设置均相同的情况下，后处理方法分别选用D1 ——D4。表 5.12给出了四种后处理方法的分词效果。

由表 5.12可知，较D1而言，D2方法显著提高了分词的效果。但相对于D3而

表 5.12　使用标记特征模版所产生的特征数目

| 语料库 | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| PKU05 | 0.934 | 0.940 | 0.947 | 0.948 |
| MSRA0505 | 0.958 | 0.961 | 0.970 | 0.970 |
| CityU05 | 0.943 | 0.948 | 0.951 | 0.952 |
| AS05 | 0.941 | 0.945 | 0.960 | 0.963 |

言，D4的提高并不显著，这进一步验证了在训练过程中引入标记之间的关系特征的重要性，它使得不合法序列的数目大大降低。

### 5.4.3　基于$M^3N$的一体化方法的性能

由第 5.4.2小节的实验结果可知，使用(a)~(g) 的特征模版以及D4后处理方法可使模型达到最优性能，因此本节的实验均采用这一配置。$M^3N$ 的训练均采用200次迭代，CRFs的训练至收敛。

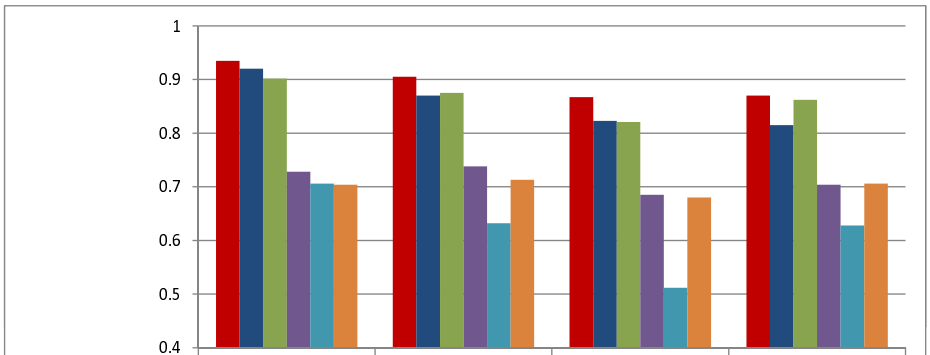
实验首先在训练特征设置完全相同的情况下（PKU_7W为训练集），分别训练基于$M^3N$的和基于CRFs的一体化模型以作比较。另外，为了考察一体化方法与单独训练模型的性能差异，我们训练基于$M^3N$ 的分词模型和命名实体识别模型（称为分步方法），与一体化方法的性能作比较。我们将分别在分词任务和命名实体识别任务上比较基于$M^3N$ 的一体化方法、分步方法以及基于CRFs的一体化方法的性能。

一体化方法和分步方法使用完全相同的特征进行训练，采用相同的后处理方法。只是标注集的设置不同：一体化方法使用表2所列的16标注集进行标注；分步方法中则对分词和命名实体识别任务进行单独训练，二者不互相提供信息，在标注集上采用第 5.3.1介绍的针对单独任务的标注方法。

图 5.4给出了在中文分词任务上一体化方法与分步方法、基于$M^3N$ 的和基于CRFs的一体化方法的性能比较（考察$F_1$ 和R_OOV）：

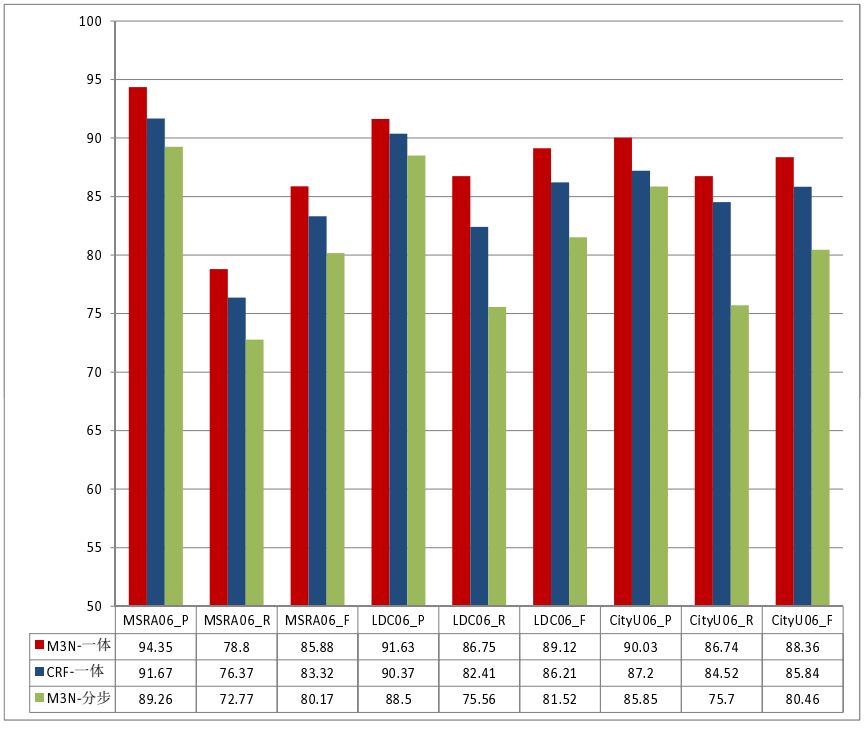
图 5.5给出了在命名实体识别任务上一体化方法与其它方法的性能比较（考察$P$，$R$和$F_1$）：

由图 5.4和图 5.5可知，基于$M^3N$的一体化方法无论在中文分词任务还是命名实体识别任务上的性能均优于分步方法，其性能也优于相同配置下的基于CRF模型的一体化方法。表 5.13给出了基于$M^3N$的一体化方法




| | PKU05 | MSRA05 | AS05 | CityU05 |
|---|---|---|---|---|
| M3N-一体-F | 0.935 | 0.905 | 0.867 | 0.87 |
| M3N-分步-F | 0.92 | 0.87 | 0.823 | 0.815 |
| CRF-一体-F | 0.902 | 0.875 | 0.821 | 0.862 |
| M3N-一体-R_OOV | 0.728 | 0.738 | 0.685 | 0.704 |
| M3N-分步-R_OOV | 0.706 | 0.632 | 0.512 | 0.628 |
| CRF-一体-R_OOV | 0.704 | 0.713 | 0.68 | 0.706 |

图 5.4　中文分词任务上的性能比较




| | MSRA06_P | MSRA06_R | MSRA06_F | LDC06_P | LDC06_R | LDC06_F | CityU06_P | CityU06_R | CityU06_F |
|---|---|---|---|---|---|---|---|---|---|
| M3N-一体 | 94.35 | 78.8 | 85.88 | 91.63 | 86.75 | 89.12 | 90.03 | 86.74 | 88.36 |
| CRF-一体 | 91.67 | 76.37 | 83.32 | 90.37 | 82.41 | 86.21 | 87.2 | 84.52 | 85.84 |
| M3N-分步 | 89.26 | 72.77 | 80.17 | 88.5 | 75.56 | 81.52 | 85.85 | 75.7 | 80.46 |

图 5.5　命名实体识别任务上的性能比较

表 5.13　基于$M^3N$的一体化方法在人名、地名、机构名上的识别结果

| 语料库 | 人名 | 地名 | 机构名 |
|---|---|---|---|
| MSRA06 | 0.9102 | 0.8810 | 0.7788 |
| LDC06 | 0.9092 | 0.9213 | 0.8048 |
| CityU06 | 0.8909 | 0.9228 | 0.7986 |

在SIGHAN2006的三个命名实体数据集上得到的人名、地名、机构名的分项识别结果（$F_1$值）。

由上表可见，总体而言人名、地名的识别效果较好，机构名的识别效果相对较差。

## 5.5　本章小结

本章提出了一种基于最大间隔马尔可夫网络（$M^3N$）的中文分词和命名实体识别一体化方法，该方法将两个任务统一在一个字序列标注框架下，进行联合训练和测试。在SIGHAN_2005的四个分词数据集以及SIGHAN_2006的三个命名实体识别数据集上进行的实验结果显示，基于$M^3N$ 的一体化方法能够同时提高中文分词以及命名实体识别的性能，且性能优于基于CRFs 的一体化方法。在分词任务上进行的实验显示，加入标记特征模版对模型性能影响显著。另外，在对不合法序列的后处理方法上，采用动态规划方法在候选标记中选择合法状态转移路径的方法能够进一步提高模型的性能。今后的工作方向将集中在如何有效选取特征模版以提高$M^3N$ 模型的训练效率以及提高未登录词识别的效果。

# 第 6 章　总结与展望

中文自动分词处于中文信息处理的底层，在中文信息处理的诸多应用中起着重要作用。中文自动分词是中文信息处理领域的经典问题，迄今已历经二十余年的研究，虽然取得了诸多可喜的进展，但中文自动分词系统的准确性和在开放环境下的自适应性还未达到实际应用的要求。在当前电子信息日益增长、网络信息挖掘技术日益成熟的背景下，基于大规模语料库及网络资源，针对中文自动分词研究中的三个关键问题展开研究：

1. 中文语言资源建设中的词频估计问题；
2. 中文自动分词中的交集型歧义问题；
3. 中文自动分词中的未登录词问题。

## 6.1　论文的主要贡献

本文的主要贡献和创新点包括以下方面：

一、在中文自动分词的研究基础——语言资源建设方面，提出了基于多类型语料的中文词频近似方法，该方法有效结合了目前可用的几种语料：生语料、自动切分语料和人工切分语料，通过对这些语料优缺点的分析，综合考虑语料库规模、词长等因素，提出了一种有效的词频近似策略。设计了多个统计观察角度，验证了该策略的有效性，并从基于Unigram模型的中文自动分词这一实际应用角度，证实了其得到的词频信息的准确性。为中文词表和中文分词国家标准的构建奠定了基础。目前我们已基于这一词频近似策略，构建了初步的中文通用词表。

二、在中文分词中的交集型歧义问题处理方面，本文基于一个含74,191词的中文通用词表、一个规模约为9亿字的通用语料库和总规模约为1.4亿字的两个专业领域语料库，考察了通用语料库中的高频最大交集型歧义字段在通用语料库和专业领域语料库的统计特性，确定了包含7,000个高频交集型歧义字段的中文交集型歧义字段核心集合，并对其区分了真歧义字段和伪歧义字段。基于其中的伪歧义字段制定的消歧策略可以作为一种预处理，提前检测中文文本中的

交集型歧义，给出正确的切分。提前解除可能对分词系统造成的干扰。鉴于专业领域词汇的特殊性，本文进一步考察了从通用语料库中抽取的高频伪歧义切分字段在专业领域语料库中的变化，发现在专业领域语料库中，仅有极少数伪歧义字段因在专业领域中发生词的义项增加导致歧义类型的转变。实验观察发现，高频伪歧义字段在专业领域依然具有较强的覆盖能力。另外，我们还考察了从专业领域语料库中抽取的最大交集型歧义切分字段关于专业领域的统计特性，初步揭示了它们关于领域分布的一般规律，发现其中的高频歧义字段在专业领域也具有较强的覆盖能力。交集型歧义字段核心集合的确立及在专业领域语料库上的观察，对交集型分词歧义的处理（包括面向专业领域的交集型分词歧义），具有参考价值。

三、在中文分词中未登录词问题的处理方面，提出了一种基于web 搜索和有监督机器学习相结合的中文分词框架。将web 作为一个极大规模的知识库，以弥补在有监督机器学习方法中面临的训练语料规模小、覆盖面不够的问题。在所提出的中文分词框架中，我们给出一种使用有监督机器学习模型输出的多个候选构建Lattice，再通过Lattice检测可能包含未登录词串的方法。对这些“可疑”的字串，通过web搜索，给出基于搜索的切分结果。本文还提出了切分相似度度量方式，用以将基于搜索的结果和切分候选进行相似度计算，得到最终的切分结果。所提出的分词框架有效利用了大规模网络文本“高覆盖率”、“高平衡性”的优势，又通过有监督机器学习方法避免了网络文本噪音的引入，在得到高未登录词识别率的同时，保证了已知词的切分准确性。从而提高了分词系统整体性能。方法本身虽然还存在很多可以改进的地方，但这一方法对分词系统在面对开放环境时适应能力的提高具有一定的参考意义。

四、未登录词问题是造成中文分词系统性能下降的主要因素。其中，命名实体构成了未登录词的主要部分。本文提出了基于最大间隔马尔可夫网络（Max-Margin Markov Network，简称$M^3N$）模型的中文自动分词与命名实体识别一体化的方法，首先基于中文分词任务，比较了$M^3N$ 模型与条件随机场模型的性能。考察了迭代次数、标记特征模板以及对不合法序列的后处理方法对$M^3N$模型性能的影响，给出了模型的最优配置。在这一基础上，使用16标记集将中文分词与命名实体识别统一在一个字标注框架下训练$M^3N$模型。实验表明，基于$M^3N$模型的一体化方法能够同时提高中文分词和命名实体识别的性能。且

其性能超过了采用相同配置的基于条件随机场模型的一体化方法。

## 6.2　进一步工作展望

在本文的研究工作基础上，以下五点值得进一步的研究和探索：

一、在本文的词频估计工作的支撑下，建立广泛认可的、高质量的中文信息处理用中文通用词表。推进中文自动分词规范及标准的建立。在此基础上，研制深度加工的平衡语料库以及极大规模通用语料库。形成权威的、公开的语言资源，为中文自动分词的研究提供扎实可靠的研究基础。

二、目前，对交集型歧义字段的研究工作是基于通用词表抽取最大交集型歧义字段进行统计观察的。由于词表本身并未涉及专业领域的词汇，所有对专业领域最大交集型歧义字段的观察还并不全面。后续的工作需要在本工作基础上通过专业领域词典的介入作进一步深入的考察。

三、基于web搜索和有监督机器学习相结合的中文分词框架有待改进。目前的工作使用搜索引擎技术得到查询串的串频，后续工作可以考虑将论文第二章提出的词频估计策略用于所提出的分词框架中。通过词频估计策略得到查询串的频度信息。相应地，也可以考虑将搜索结果融入词频估计策略，以帮助提高词频估计的效果。另外，目前对网络资源的利用还比较初步，挖掘大规模结构化文本蕴含的潜力，充分利用网络资源里的信息（如wikipedia的分类信息等），最大限度地提高中文自动分词系统对开放环境的自适应能力。

四、基于有监督机器学习（如基于CRF和基于$M^3N$的）中文分词方法存在一个共同的问题就是训练代价较高。如何在保持分词效果的同时，尽可能地降低特征的维数有待进一步的探索。这一工作可以考虑两种做法：一是通过一些统计量对海量特征进行特征选择，在降维后进行模型训练。这种方法我们已经做了初步的尝试，比如采用$\chi^2$检验和信息增益等，取得了初步的结果，有待深入的研究。二是在训练过程中筛选出有效特征，拉普拉斯$M^3N$ 是我们认为适合于这一任务的统计模型，可以在其上做进一步的研究。

五、从工业界的需求来看，目前学术界报告的准确率最高的基于有监督机器学习的分词方法在工业界还无法投入使用。原因是这类方法的训练代价较高（包括内存的占用量和训练时间），模型的可移植性较差，如果应用到某一个专业领

域语料上，一般需要重新训练模型。如何简单的通过一个领域词典，使用已有的在通用语料上训练得到的模型，在专业领域语料上得到好的分词效果是值得研究的一个方向。

六、基于统计机器学习方法的中文分词将分词精度提升了一个高度，但这种方法存在可以达到的性能上限，对于因复杂语言现象导致的中文分词问题的最终解决，离不开语言学、认知科学等学科的指导。而且，中文处理的几个层次，词法层、句法层、语义层，都是互相关联，互相影响的。后续工作的方向考虑如何融合词法、句法甚至部分语义信息，研究集统计方法与人工规则于一体的分词算法。

本文基于当前日益丰富的网络资源，在大规模语料上利用统计分析、机器学习方法、搜索技术等对中文自动分词的几个关键问题进行研究，取得了一些有意义的研究结果。本文的工作仅仅是向解决中文自动分词问题的道路上作出的努力和尝试，所得到的一些研究结论及成果虽然离解决该问题尚有很大差距，但希望能够为这项事业尽自己的一丝绵薄之力。

# 参考文献


[1] Guo J. Critical tokenization and its properties. Computational Linguistics, 1997, 23(4):569–596.

[2] Kam-Fai Wong R X, Zhang Z. Introduction to Chinese natural language processing. Morgan & Claypool, 2010.

[3] Wu Z, Tseng G. Chinese text segmentation fortext retrieval: achievements and problems. The American Society for Information Science, 1993, 44(9):532–542.

[4] Wu Z, Tseng G. ACTS: an automatic Chinese text segmentation system for full text retrieval. The American Society for Information Science, 1995, 46(2):83–96.

[5] Wu A, Jiang Z. Word segmentation in sentence analysis. Proceedings of 1998 International Conference on Chinese Information Processing, 1998. 169–180.

[6] Foo S, Li H. Chinese word segmentation and its effect on information retrieval. Information Processing and Management, 2004, 40(1):161–190.

[7] Sproat R, Emerson T. The first international Chinese word segmentation bakeoff. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, 2003. 133–143.

[8] Levow G. The third international Chinese word segmentation bakeoff. Proceedings of the 5th SIGHAN workshop on Chinese Language Processing, 2006. 108–117.

[9] Emerson T. The second international Chinese word segmentation bakeoff. Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, 2005.

[10] Jin G, Chen X. The fourth international chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging. Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, 2007. 69–81.

[11] 黄昌宁, 赵海. 中文分词十年回顾. 中文信息学报, 2007, 21(003):8–19.

[12] 国家技术监督局. 中华人民共和国国家标准:《信息处理用现代汉语分词规范》. 北京: 中国标准出版社, 1993.

[13] 刘源, 沈旭昆, 谭强. 《信息处理用现代汉语分词规范及自动分词方法》. 清华大学出版社、广西科学技术出版社, 1994.

[14] 俞士汶, 朱学锋, 段慧明. 大规模现代汉语标注语料库的加工规范。. 中文信息学报, 2000, 14(6).

[15] 黄居仁, 陈克健, 陈凤仪, 等. 资讯处理用中文分词规范. 《语言文字应用》, 1997, 1:92–100.

[16] 孙茂松. 谈谈汉语分词语料库的一致性问题. 语言文字应用, 1999, 2:87–90.

[17] 梁南元. 书面汉语自动分词系统—CDWS. 中文信息学报, 1987, 1(2):44–52.

[18] Sun M S, Tsou B K Y. Ambiguity resolution in Chinese word segmentation. Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation, 1995. 121–126.

[19] 孙茂松, 黄昌宁, 邹嘉彦, 等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展, 1997, 34(5):332–339.

[20] 孙茂松, 邹嘉彦. 汉语自动分词研究评述. 当代语言学, 2001, 3(1):22–32.

[21] 孙茂松, 左正平. 汉语真实文本中的交集型切分歧义. Proceedings of 汉语计量与计算研究, 1998. 323–338.

[22] Liu Y, Liang N Y. Counting word frequencies of contemporary Chinese–An engineering of Chinese processing. Journal of Chinese Information Processing. Vol. 0, 1986, 1:17–25.

[23] 揭春雨, 刘源, 梁南元. 论汉语自动分词方法. 中文信息学报, 1989, 3(1):3–11.

[24] 王晓龙, 王开铸, 李仲荣. 最少分词问题及其解法. 科学通报, 1989, 34(13):1030–1032.

[25] 马晏. 基于评价的汉语自动分词系统的研究与实现.《语言信息处理专论》. 2–36.

[26] 徐秉铮, 詹剑. 基于神经网络的分词方法. 中文信息学报, 1993, 7(002):36–44.

[27] Sproat R. A stochastic finite state word segmentation algorithm for Chinese. Computational Linguistics, 1996, 22(3):377–404.

[28] Lai B, Sun M, Tsou B, et al. Chinese word segmentation and POS-tagging in one step. Proceedings of of 19th Annual Meeting of the Association for Computational Linguistics, 1997. 229–236.

[29] 孙茂松, 左正平, 黄昌宁. 消解中文三字长交集型分词歧义的算法. 清华大学学报, 1999, 39(5):101–103.

[30] Chen K, Liu S. Word identification for Mandarin Chinese sentences. Proceedings of the 14th Annual Meeting of the Association for Computational Linguistics, 1992. 101–107.

[31] Sun M, Lai B, Lun S, et al. Some issues on statistical approach to Chinese word identification. Proceedings of the third International Conference on Chinese Information Processing, 1992.

[32] 黄祥喜. 书面汉语自动分词的“生成-测试”方法. 中文信息学报, 1989, 4:42–49.

[33] 徐辉, 何克抗. 书面汉语自动分词专家系统的实现. 中文信息学报, 1991, 5(003):38–47.

[34] Palmer D D. A trainable rule-based algorithm for word segmentation. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997. 321–328.

[35] 李蓉, 刘少辉, 叶世伟, 等. 基于SVM 和K-NN 结合的汉语交集型歧义切分方法. 中文信息学报, 2001, 15(6):13–18.

[36] McCallum J L A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of 18th International Conference on Machine Learning (ICML'01), 2001. 282–289.

[37] Xue N. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 2003, 8(1):29–48.

[38] Low J K, Ng H T, Guo W. A maximum entropy approach to Chinese word segmentation. Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, 2005. 161–164.

[39] Li M, Gao J, Huang C, et al. Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation. Proceedings of the First SIGHAN Workshop on Chinese Language Processing (SIGHAN'2003), 2003. 1–7.

[40] Zhao H, Kit C. An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework. the International Conference of IJCNLP'08, 2008..

[41] Gao J, Li M, Huang C N, et al. Chinese word segmentation and named entity recognition: A pragmatic approach. Computational Linguistics, 2005, 31(4):531–574.

[42] 孙茂松, 左正平, 邹嘉彦. 高频最大交集型歧义切分字段在汉语自动分词中的应用. 中文信息学报, 1999, 13(1):27–34.

[43] Qiao W, Sun M, Menzel W. Statistical properties of overlapping ambiguities in Chinese word segmentation and a strategy for their disambiguation. Proceedings of the International Conference on Text, Speech and Dialogue (TSD'08). Springer Berlin / Heidelberg. 177–186.

[44] 乔维, 孙茂松. 汉语交集型歧义切分字段关于专业领域的统计特性. 中文信息学报, 2008, 22(004):10–18.

[45] 曲维光, 吉根林, 穗志方, 等. 基于语境信息的组合型分词歧义消解方法. 计算机工程, 2006, 32(17):74–76.

[46] 肖云, 孙茂松, 邹嘉彦. 利用上下文信息解决汉语自动分词中的组合型歧义. 计算机工程与应用, 2001, 37(19):87–106.

[47] Jin H, Wong K. A Chinese dictionary construction algorithm for information retrieval. ACM Transactions on Asian Language Information Processing (TALIP), 2002, 1(4):296.

[48] Tseng H. Semantic classification of Chinese unknown words. Proceedings of the ACL'03 Student Research Workshop, Companion Volume to the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003.

[49] Chen K, Ma W. Unknown word extraction for Chinese by a corpus-based learning method. volume 3.

[50] Xue N, Converse S. Combining classifiers for Chinese word segmentation. Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing, 2002. 63–70.

[51] 刘挺, 吴岩. 串频统计和词形匹配相结合的汉语自动分词系统. 中文信息学报, 1998, 12(1):17–25.

[52] Church K, Hanks P. Word association norms, mutual information, and lexicography. Computational linguistics, 1990, 16(1):22–29.

[53] Sproat R, Shih C. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 1990, 4(4):336–351.

[54] Sun M S, Shen D, Tsou B K Y. Chinese word segmentation without using lexicon and hand-crafted training data. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, 1998. 1265–1271.

[55] 黄萱菁, 吴立德. 基于机器学习的无需人工编制词典的切词系统. 模式识别与人工智能, 1996, 9(4):297–303.

[56] Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4. 5 learning algorithm. Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), 2000. 802–807.

[57] Chien L. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. Information Processing and Management, 1999, 35(4):501–521.

[58] Chen A, Zhou Y, Zhang A, et al. Unigram language model for Chinese word segmentation. Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, 2005. 138–141.

[59] Wu A, Jiang Z. Statistically-enhanced new word identification in a rule-based Chinese system. Proceedings of the 2nd Chinese Language Processing Workshop, volume 12, 2000. 46–51.

[60] G F. Research on statistical methods of Chinese syntactic disambiguation[D]. China: The Harbin Institute of Technology (HIT), 2001.

[61] Nie J, Hannan M, Jin W. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge. Communications of COLIPS, 1995, 5(1):47–57.

[62] Li H, Huang C, Gao J, et al. The use of SVM for Chinese new word identification. Natural Language Processing–IJCNLP 2004. 723–732.

[63] Cui S, Liu Q, Meng Y, et al. New word detection based on large-scale corpus. Jisuanji Yanjiu yu Fazhan(Computer Research and Development), 2006, 43(5):927–932.

[64] 沈达阳, 孙茂松, 黄昌宁. 局部统计在汉语未登录词辨识中应用和实现方法, 1997.

[65] 张俊盛, 刘显仲. 多语料库作法之中文姓名辨识. 中文信息学报, 1992, 6(003):7–15.

[66] 宋柔, 朱宏, 潘维桂. 基于语料库和规则库的人名识别法. 《计算语言学研究与应用》. 150–154.

[67] Luo Z, Song R. An integrated method for Chinese unknown word extraction. Proceedings of Proceedings of Third SIGHAN Workshop on Chinese Language Processing, 2004. 148–154.

[68] 孙茂松, 高海燕. 中文姓名的自动辨识. 中文信息学报, 1995, 9(002):16–27.

[69] 张华平, 刘群. 基于角色标注的中国人名自动识别研究. 计算机学报, 2004, 27(001):85–91.

[70] Wang D, Yao T. Using a semi-supervised learning mechanism to recognize Chinese names in unsegmented text..

[71] Zhang K, Liu Q, Zhang H, et al. Automatic recognition of Chinese unknown words based on roles tagging. Proceedings of Proceedings of the first SIGHAN workshop on Chinese language processing-Volume 18. Association for Computational Linguistics, 2002. 7.

[72] 孙茂松, 张维杰. 英语姓名译名的自动识别. 计算语言学研究与应用, 1993. 144–149.

[73] Gao W, Wong K. Experimental studies using statistical algorithms on transliterating phoneme sequences for English-Chinese name translation. International Journal of Computer Processing of Oriental Language, 2006, 19(1):63–88.

[74] Gao W, Wong K, Lam W. Improving transliteration with precise alignment of phoneme chunks and using contextual features. Information Retrieval Technology. 106–117.

[75] Sproat R, Tao T, Zhai C. Named entity transliteration with comparable corpora. Proceedings of Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics Morristown, NJ, USA, 2006. 73–80.

[76] Tao T, Yoon S, Fister A, et al. Unsupervised named entity transliteration using temporal and phonetic correlation. Proceedings of Proceedings of EMNLP 2006, 2006. 250–257.

[77] 沈达阳, 孙茂松, 等. 中国地名的自动辨识.《计算语言学进展与应用》, 1995. 68–74.

[78] Zhang Q, Yuan Y, Li N, et al. A New Way for Chinese Place Name Recognition. Proceedings of 2009 International Conference on Asian Language Processing. IEEE, 2009. 129–134.

[79] 张小衡, 王玲玲. 中文机构名称的识别与分析. 中文信息学报, 1997, 11(004):21–32.

[80] Chen K, Chen C. Knowledge extraction for identification of Chinese organization names. Proceedings of Proceedings of the second Chinese Language Processing Workshop, 2002. 15–21.

[81] Chen H, Lee J. The identification of organization names in Chinese texts. Communication of COLIPS, 1994, 4(2):131–142.

[82] Zhang Q, Hu G, Yue L. Chinese organization entity recognition and association on web pages. Business Information Systems, 2008. 12–23.

[83] Sun J, Zhou M, Gao J. A class-based language model approach to Chinese named entity identification. Computational Linguistics and Chinese Language Processing, 2003, 8(2):1–28.

[84] Wu Y, Zhao J, Xu B. Chinese named entity recognition combining a statistical model with human knowledge. Proceedings of Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15. Association for Computational Linguistics, 2003. 65–72.

[85] Ng H, Low J. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. Proceedings of the International Conference of EMNLP'04, 2004.

[86] Zhang Y, Clark S. Joint word segmentation and POS tagging using a single perceptron. 2008..

[87] Jiang W, Huang L, Lv Y, et al. A cascaded linear model for joint chinese word segmentation and part-of-speech tagging. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008.

[88] Feng Y, Sun L, Lv Y. Chinese word segmentation and named entity recognition based on conditional random fields models. Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, 2006.

[89] Leong K, Wong F, Li Y, et al. Integration of Named Entity Information for Chinese Word Segmentation Based on Maximum Entropy. Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues. 962–969.

[90] Fuchun Peng F F, McCallum A. Chinese segmentation and new word detection using conditional random fields. Proceedings of the 20th International Conference on Computational Linguistics, 2004. 562–568.

[91] Cheng K S, Young G, Wong K. A study on word-based and integral-bit Chinese text compression algorithm. Journal of the American Society for Information Science, 1999, 50(3):218–228.

[92] Xue N. Chinese word segmentation as character tagging. Computational Linguistics, 2003, 8(1):375–393.

[93] Zhang R, Kikui G, Sumita E. Subword-based tagging by conditional random fields for chinese word segmentation. Proceedings of of Human Language Technology conference – North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL), 2006.

[94] Nakagawa T, Uchimoto K. A hybrid approach to word segmentation and pos tagging. Proceedings of the 45th International Conference on ACL'07, 2007. 217–220.

[95] Shi Y, Wang M. A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks. Proceedings of IJCAI'07, volume 7, 2007. 1707–1712.

[96] Collins M, Liang P. Semi-supervised learning for natural language. 2005..

[97] 孙茂松, 肖明, 邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词. 计算机学报, 2004, 27(006):736–742.

[98] Goldwater S, Griffiths T L, Johnson M. Contextual dependencies in unsupervised word segmentation. Proceedings of the International Conference of COLING-ACL'06, 2006. 673–680.

[99] Jin Z, Tanaka-Ishii K. Unsupervised segmentation of Chinese text by use of branching entropy. Proceedings of the International Conference of COLING/ACL'06. Association for Computational Linguistics, 2006. 435.

[100] Zhao H, Kit C. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, 2008. 106–111.

[101] 梁南元. 汉语计算机自动分词知识. 中文信息学报, 1990, 4(002):29–41.

[102] 姚天顺, 张桂平. 基于规则的汉语自动分词系统. 中文信息学报, 1990, 4(001):37–43.

[103] 韩世欣, 王开铸. 基于短语结构文法的分词研究. 中文信息学报, 1992, 6(003):48–54.

[104] Ge X, Pratt W, Smyth P. Discovering Chinese words from unsegmented text. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR'99). ACM, 1999. 272.

[105] Peng F, Schuurmans D. Self-supervised Chinese word segmentation. Advances in Intelligent Data Analysis, 2001. 238–247.

[106] Hockenmaier J, Brew C. Error driven segmentation of Chinese. Communications of COLIPS, 1998, 8(1):69–84.

[107] Florian R, Ngai G. Multidimensional transformation-based learning. Proceedings of Proceedings of the 2001 workshop on Computational Natural Language Learning. Association for Computational Linguistics, 2001. 1–8.

[108] Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics, 1995, 21(4):543–565.

[109] Ponte J, Croft W. Useg: A retargetable word segmentation procedure for information retrieval. Proceedings of the 5th International Conference: Symposium on Document Analysis and Information Retrieval, 1996.

[110] 沈达阳, 孙茂松. 汉语分词系统中的信息集成和最佳路径搜索方法. 中文信息学报, 1997, 11(002):34–47.

[111] Gao J, Li M, Huang C. Improved source-channel models for Chinese word segmentation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL'03), 2003. 272–279.

[112] Vapnik V, Cortes C. Support vector networks. Machine Learning, 1995, 20(3):273–297.

[113] Taskar B, Guestrin C, Koller D. Max-margin Markov networks. Advances in neural information processing systems, 2004, 16:25–32.

[114] Zhao H, Kit C. A simple and efficient model pruning method for conditional random fields. Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy. 145–155.

[115] Zhang H, Yu H, Xiong D, et al. HHMM-based Chinese lexical analyzer ICTCLAS. Proceedings of the 2th SIGHAN workshop on Chinese language processing. ACL, 2003. 187.

[116] Liu E S. Frequency Dictionary of Chinese Words. USA: Stanford University, 1975.

[117] Dai X L. Chinese Morphology and its Interface with the Syntax[D]. USA: Ohio State University, 1992.

[118] Chen G L. On Chinese Morphology. Shanghai, China: Xuelin Publisher, 1994.

[119] Zhu D X. Lectures on Grammar. Beijing, China: The Commercial Press, 1982.

[120] Tang T C. Chinese Mrophology and Syntax. Taipei: Taiwan Student Publisher, 1992.

[121] Nagata M. A self-organizing Japanese word segmenter using heuristic word identification and re-estimation. Proceedings of the 5th Workshop on Very Large Corpora, 1997. 203–215.

[122] Sun M S, Zhang Z, Tsou B K Y, et al. Word frequency approximation for Chinese without using manually annotated corpus. Proceedings of 7th International Conference on Intelligent Text Processing and Computational Linguistics, 2006. 105–116.

[123] 付国宏, 王晓龙. 汉语词语边界自动划分的模型与算法。. 计算机研究与发展, 1999, 36(9):1142–1147.

[124] Holland J H. Adaptation in Natural and Artificial Systems. Ann Arbor, USA: MI: University of Michigan Press, 1975.

[125] 黄昌宁. 中文信息处理中的分词问题. 语言文字应用, 1997, 17(1):72–78.

[126] 李国杰. 智能机研究动态. 第五届全国汉字识别、语言识别与合成系统及自然语言处理系统评测结果. 43–46.

[127] 刘开瑛. 现代汉语自动分词评测技术研究. 语言文字应用, 1997, 21(1):101–106.

[128] 郑家恒, 刘开瑛. 中文文本歧义切分技术研究. 语言工程. 北京: 清华大学出版社, 1997..

[129] Bing S, Yu S. A graded approach for the efficient resolution of Chinese word segmentation ambiguities. Proceedings of the 5th Natural Language Processing Pacific Rim Symposium, 1999. 19–24.

[130] 侯敏, 陈琼璜, 初田天, 等. 汉语自动分词中的上下文相关歧义字段(CSAS)研究. 北京: 清华大学出版社, 2005: 214–220.

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

[132] 俞士汶. 现代汉语语法信息词典详解(第二版). 北京: 清华大学出版社, 2003.

[133] 《第25次中国互联网络发展状况统计报告》. Technical report, 中国互联网络信息中心, January 15, 2010. http://www.cnnic.net.cn/uploadfiles/pdf/2010/1/15/101600.pdf.

[134] Wang X, Qin Y, Liu W. A search-based Chinese word segmentation method. Proceedings of the 16th International Conference on World Wide Web (WWW'07). ACM, 2007. 1129–1130.

[135] 李月伦, 常宝宝. 基于最大间隔马尔可夫网模型的汉语分词方法. 中国计算机语言学研究前沿进展(2007-2009), 2009..

[136] 王睿, 张洁, 张由仪, et al. 基于混合模型的中文命名实体抽取系统. 清华大学学报(自然科学版), 2005, 45(9).

[137] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the 7th Conference on Natural Language Learning (CoNLL'03), 2003. 188–191.

[138] Wu C W, Tsai R, Hsu W L. Semi-joint Labeling for Chinese Named Entity Recognition. Information Retrieval Technology. 107–116.

# 致　谢

衷心感谢我的导师孙茂松教授多年来在科研上对本人的悉心指导，在生活上对本人的关怀和帮助。他为我创造了良好宽松的研究环境，提供了优越的科研设施和宝贵的国际交流机会。他的为人为学之道深深地影响了我，将使我受益终身。

感谢本人在美国伊利诺伊大学香槟分校联合培养期间的合作导师Richard Sproat教授以及在中德CINACS项目中的联合指导老师Wolfgang Menzel教授在国际交流中给予本人的帮助，在研究工作中给予的指点和建议。

感谢人工智能实验室的全体老师和自然语言处理组的同学们在本人博士期间的学习和生活上给予的支持和帮助。

本课题承蒙国家863项目及自然科学基金资助，特此致谢。

感谢我挚爱的父亲母亲，他们伟大而无私的爱永远是我前进的强大动力。

# 声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

# 个人简历、在学期间发表的学术论文与研究成果

## 个人简历

1983年 8 月 10 日出生于河南省洛阳市。

2000年 9 月免试进入西安电子科技大学计算机科学与技术系，2004年 7 月本科毕业并获得工学学士学位。

2004年 9 月免试进入清华大学计算机科学与技术系，直接攻读博士学位至今。

2008年 10 月至 2009年5月赴美国伊利诺伊大学香槟分校（UIUC）进行联合培养，联合指导老师为Richard Sproat教授。

## 发表的学术论文

[1] **Wei Qiao**, Sun Maosong and Wolfgang Menzel. Chinese Word Frequency Approximation Using Multiple-type Corpora. International Journal of Quantitative Linguistics. 2010, 17(2):142-166. (SSCI 源刊)

[2] **乔维**，孙茂松. 基于$M^3N$的中文分词与命名实体识别一体化方法.《清华大学学报自然科学版（中文版）》, 2009,（已录用，EI 源刊）

[3] **Wei Qiao**, Sun Maosong and Wolfgang Menzel. Overlapping Ambiguity in Chinese Word Segmentation and a Strategy for Their Disambiguation. Proceedings of 11th International Conference on Text Speech and Dialogue(TSD ’08), 2008, 177-186. (EI 收录, 检索号:20084111630493.)

[4] **Wei Qiao**, Sun Maosong. Word Frequency Approximation for Chinese Using Raw, MM-Segmented and Manually Segmented Corpora. Proceedings of the 21st International Conference of Computer Processing of Oriental Languages (ICCPOL ’06), 2006, 256-267. (ISTP 收录)

[5] **Wei Qiao**, Sun Maosong. Incorporate Web Search Technology to Solve Out-of-Vocabulary Words in Chinese Word Segmentation. Proceedings of 11th Pacific

Asia Conference on Language, Information and Computation (PACLIC), 2009, 454-463. (ISTP 收录)

[6] **乔维**, 孙茂松. 中文交集型切分歧义关于专业领域的统计性质. 中文信息学报, 2008, 22(4):10-18. (中文核心期刊)

[7] **乔维**, 孙茂松. 基于生语料、最大匹配切分语料以及熟语料的中文词频估计方法. 第三届学生计算语言学研讨会论文集（SWCL ’06）, 2006, 261-268.

## 参与的科研项目

[1] 国家自然科学基金项目：超对等语义搜索引擎。项目编号：60520130299。立项部门：国家自然科学基金委。时间：2005-2007。

[2] 国家863计划项目：大规模网络图文数据的语义分类及适度理解。项目编号：2007AA01Z148。立项部门：中华人民共和国科技部。时间：2007-2009。

[3] 搜狐-清华联合实验室项目。时间：2008.1-2008.10。

[4] 中德清华-汉堡CINACS联合培养项目。时间：2006.9-2010.3。