大连海事大学

毕 业 论 文

装

订

线

二〇一〇年六月

中文最长功能名词短语的自动识别

专业班级：软件工程 1 班
姓　　名：裴家欢
指导教师：谢益武

信息科学技术学院

摘　要

本文依据中文最长名词短语的句法功能特征，定义了最长功能名词短语的概念。与最长名词短语相比，最长功能名词短语不仅具有不被任何名词短语所包含的特点，而且具有独立的句法功能。

最长功能名词短语的自动识别不仅有利于浅层句法分析，而且有利于自然语言处理中的其它任务，如翻译消歧、指代消解、信息检索和实体识别等。

本文首先对中文宾州树库语料进行了解析，将其处理成含有最长功能名词短语边界的链状句子，然后再将句子处理成训练和测试所用的标准格式。其次，分别使用最大熵、支持向量机和条件随机域模型对最长功能名词短语进行识别，其中最大熵识别 F 值达 64.01% ，支持向量机正向识别 F 值达 86.43%，支持向量机逆向识别 F 值达 87.97% ，条件随机域识别 F 值达 88.41% 。最后，本文将对较好的三种策略进行融合，其识别 F 值为 88.92%。另外本文设计并开发了最长功能名词短语自动识别系统。

关键词：最长功能名词短语；最大熵模型；支持向量机模型；条件随机域模型；多策略融合算法；

ABSTRACT

According to the syntactic function of the maximal-length noun phrase in Chinese, this paper proposed a new concept - the functional maximal-length noun phrase. Compared with the maximal-length noun phrase, it is defined as a noun phrase that can't be the component of any other noun phrases and has independent syntactic function.

The recognition of the functional maximal-length noun phrase not only helps the shallow phrasing, but also has significance in many natural language processing tasks, such as translation disambiguation, anaphora resolution, information retrieval, entity recognition, etc.

First, the Chinese Penn Treebank was parsed and processed into linear sentences and standard format for training and test. Then, the Maximum Entropy Models, Support Vector Machines and Conditional Random Fields for recognition were employed respectively. The F-measure of the above three models are 64.01%, 86.43% (positive SVM), 87.97% (negative SVM) and 88.41%. At last, the best three strategies were coordinated into one model and an automatic recognition system for the functional maximal-length noun phrase were developed with the F-measure reaching 88.92%.

Keywords: functional maximal-length noun phrase; Maximum Entropy Models; Support Vector Machines;Conditional Random Fields; Multi-strategy Fusion approach

目　录

中文最长功能名词短语的自动识别

第1章 绪论

1.1课题研究的背景及意义

在自然语言理解的过程中，名词短语识别对描述实体和概念至关重要，正确识别出名词短语，不仅可以更好地把握文本对象和主要内容而且可以辅助识别出自然语言的结构框架，从而提高机器学习和处理自然语言的能力。中文句法结构比较复杂，对句法结构的理解是自然语言处理中亟待解决的问题，最长名词短语识别是浅层句法分析中的重要任务，正确识别其边界可以在很大程度上解决句法分析的复杂度高的问题，辅助分析复杂句子的基本结构。本文研究的课题是中文最长功能名词短语识别，与前人考虑的最长名词短语识别问题相比，该类短语的准确识别不仅利于句法树的构建，辅助提高机器翻译的质量，而且对信息检索、实体识别以及微博情感对象识别、情感倾向分析等热门领域作用显著。

1.1.1 英文最长名词短语识别的研究现状

在英文领域中，名词短语识别的研究目前主要有三类：基本名词短语识别（Base Noun Phrase，baseNP）、最长名词短语识别（Maximal-length Noun Phrase，MNP）和功能名词短语识别（Functional Noun Phrase，funNP）。

Church[1]利用左右边界的信息构造出两个概率矩阵（名词起始矩阵和名词终止矩阵），设计了 baseNP 的抽取器；Voutilainen[2]通过 NP-肯定机制与 NP-否定机制相结合来实现 NPtool 标识 MNP；K. Chen[3]等利用基于统计的浅层句法分析与基于规则的有限状态分析相结合的方法，根据标注的句法语义信息和有限状态机的转换机制抽取出最可能的名词短语；Koehn and Knight[4]在研究英德翻译系统时，最早明确地提出了最长名词短语的定义。另外，以 Halliday[5]等为代表的一部分学者认为，语言成分的功能可以更好地解释语言结构。马建军[6]认为，baseNP 和 MNP 只根据名词短语的逻辑结构来定义而没有考虑名词短语的句法功能，不能完全消解结构歧义，因而，提出了功能名词短语的识别问题，先应用语法或语义知识来分析名词短语在某个小句中的句法功能，并根据句法功能来界定所研究的名词短语的范围，从而提高机器对句子的理解能力。从总体上看，目前英语名词短语的研究较为成熟的是 baseNP 的识别，对 MNP 和

funNP 的研究还有待提高。

1.1.2 中文最长名词短语识别的研究现状

中文名词短语识别的研究也主要分为 baseNP 识别、MNP 识别和 funNP 识别三类，对 baseNP 研究比较多[7-11]。

李文捷等人最早开始进行了 MNP 的研究，周强等[12]充分利用 NP 边界信息与内部结构特点，设计了识别 MNP 的新算法。首先对语料预处理成标注了成分边界预测信息的词语块序列，相当于组块分析（Chunking）的过程，然后根据两个概率阈值设置可能的左右边界，用 NP 栈结构进行边界识别，最后利用归约规则对 NP 内部结构进行识别，准确率为 85. 4%，召回率为 82. 3%。这种方法提供了一种 MNP 识别的新思路，但是由于预处理阶段的组块分析可能引入一些错误，如，成分组“标点分隔结构”在识别句子“周树人先生的《鲁迅全集》我已经拜读过了。”时，识别出 MNP 是【鲁迅全集】，而忽略掉了前面的修饰词。

冯冲[13]等首次把机器学习算法引入到 MNP 识别任务中，而且选择了二阶条件随机域（Conditional Random Fields，简称 CRF 或 CRFs）模型对复杂 MNP（CompleX MNP，xMNP，指所含词数大于等于 5 的 MNP）进行了识别研究，并给出了一种受限的前向-后向解码算法，识别准确率为 75.4%，召回率为 70.6%，最后将识别结果应用到真实的翻译系统 IHSMTS 中，对已经标注为 xMNP 的短语提供一些候选，从而通过简单的人工干预提高翻译质量，准确率提高 11.9%，召回率提高 9.1%。这种方法解决了长程关联（Long-range Correlation）的不足和标注偏置问题(Label Bias Problem)，提供了候选翻译的思路。

王月颖[14]应用了隐马尔可夫模型（Hidden Markov Model，HMM）和条件随机域模型（CRFs）并将 MNP 应用于指代消解任务中。使用 HMM 的方法时，先对中文宾州树库 CTB（Chinese Penn Treebank）语料进行 Chunking 解析，并映射成七类短语符号，利用统计的方法在语料中抽取特征，分别使用一阶和二阶 HMM，闭合测试识别准确率不到 55%。在使用 CRF 模型时，统计了有价值的四类特征，闭合测试准确率 99.93%（特征选取：短语+词性+汉字），然而开放测试准确率只有 65.86%，并不理想，分析其原因，一方面是受所用 Chunking 解析工具和词性标注工具性能的影响，另一方面，是否可以将短语类别作为特征输入是一个值得商榷的问题。

代翠等[15]利用统计和规则相结合的方法进行了 MNP 的识别。先用 CRFs 进行自动识别，然后引入语言知识，利用简单有限自动机和知识库（左边界信

息规则库、固定搭配表、并列词表等）对五种典型错误进行了修正，准确率达90.1%，召回率达 90.2%，都比只用统计模型高 8 个百分点。代翠[16]在后续的研究中，不仅进行了 MNP 的识别（F 值达到 90.0%），而且利用了层叠条件随机场（CRFs）对识别出的 MNP 进行了分析，并构造了浅层句法树，准确率达到 85.1%。

鉴萍，宗成庆[17]提出一种新的双向标注融合的方法识别 MNP 和最长介词短语（Maximal-length Prepositional Phrase，MPP），利用基于历史标注结果的决策模型正向和反向标注可以互为补充的特性，提出基于“分歧点（Fork Point）”的概率融合方法，并选择基于支持向量机(Support Vector Machines，SVMs)分类器的确定性标注模型，融合后，MNP 的召回率达 86.70%，F 值达 85.99%，双向标注结果的并集中所含正确短语占语料库中短语的比例达 89.36%。

Guiping Z 等[18] 使用 CRFs 方法识别 MNP，然后通过对错误实例的分析发现介词定位在 MNP 左边界的情况占 33%，通过分析介词短语（Preposition Phrase，PP）的组成，断言 PP 的左侧边界一定是介词。于是根据上下文和易于识别的左边界信息，并利用 MNP 和状语 MPP 的相互限制的特点，提出了一种基于 MPP 的 MNP 识别方法，这种方法相当于将 MNP 分为两类识别，一类是直接识别不包含在 MPP 中 MNP，另一类是将包含在状语 MPP 中的 MNP 识别转化为较为容易的状语 MPP 识别，然后通过在 MPP 内部对 MNP 识别，巧妙地回避了状语 MPP 中 MNP 识别错误的情况，最后识别准确率达 90.6627% ，召回率达 92.3313%，是目前为止比较理想的结果。

Yegang Li, Heyan Huang[19] 提出了两阶段混合的方法（2-phase Hybrid Approach）。首先确定基本组块（Base Chunk），然后进行 MNP 识别。识别的具体方法是：先利用 SVMs 模型进行双向标注，然后用带置信度的双向序列标注融合方法进行标注的融合，F1 值（F1-measure）达 90.13%。将 Base Chunk 引入作为特征可以很容易地考虑单元块信息，利于识别 MNP 边界，但是也有一个潜在问题——误差传递问题（Error Propagation Problem），具体表现是当引入自动解析的组块信息作为特征时，F1 值要比使用金标准解析（Gold-standard Chunk Parsing）时降低 1.41%

钱小飞等对最长名词短语的研究加入了丰富的语言学知识。[20-22]对 MNP 中识别较为困难的含有“的”字的最长名词短语（deMNP）从语言学的角度进行了细致的分析研究，[23]基于对 MNP 内部语言结构的分析，提出了一种基于 baseNP 中心规约的 MNP 识别方法，首先用 CRFs 进行了 baseNP 的识别，然后用中心词（Head Word）代替 baseNP，进行规约识别 MNP，准确率达 87.58%，召回率达 88.31%。[24]提出了一种基于混合策略的中文 MNP 识别方法，该方法

包括两个部分，一是对基于语言知识评价的分类器进行了融合，二是归纳出确定性规则并对 MNP 进行进一步识别。前者使用语言规则并融合了基于 SVMs 逆向识别和基于 CRFs 归约识别的结果；后者主要针对部分连续名词边界歧义问题进行修正，准确率达 89.30%，召回率达 89.62%。

相对于英文研究，中文最长名词短语的研究更加困难。一方面中文结构比较复杂，变化比较多，仅仅添加一个“的”就可以将一些动词短语转化成名词短语。另一方面，作为中文作为一种孤立语（Isolating Language），只能为 MNP 的识别提供相对较少的语言特征，不像英语词形会根据时态、语义成分的不同有明显的变化，而且名词短语的研究受分词和词性标注结果的影响较大，不同的分词和词性标注结果识别效果相差很大。然而，正因为中文句法结构的复杂性，最长名词短语的识别是很有研究价值的。

1.1.3 中文最长名词短语识别的方法总结

对于 MNP 的识别一般分为三类方法：一是基于机器学习模型的统计方法，二是基于人工或者半自动获取规则的方法，三是统计和规则相结合的方法。

MNP 识别中常用的基于统计的方法有：边界统计方法、基于最大熵模型（Maximum Entropy Models，MaxEnt）的方法、基于隐马尔可夫模型（Hidden Markov Model，HMM）的方法、基于支持向量机（SVMs）的方法、基于条件随机域（CRFs）的方法以及多策略融合的方法等。常用的基于规则的方法有：词性串规则法、基于转换的错误驱动学习方法（Transformation-Based error-drive Learning，TBL），同时也可以由语言学家手工编写规则。

统计与规则相结合的方法一般是先用统计方法进行识别，然后再根据错误实例总结规则，最后将规则用于对统计识别结果的后处理。统计的方法属于经验主义的方法，可以自动地从训练语料中获取语言知识，有效建立学习模型，但是运行效率受统计符号类别影响比较大，也很容易出现数据稀疏的问题。规则的方法属于理性主义方法，规则主要是语言学的规则，表达清晰易理解，但是研制规则比较困难，鲁棒性较差，也很难进一步升级。所以现在更趋向于统计方法为主体，规则方法作为后处理补充和修正模型。在统计方法中，SVMs 和 CRFs 两种体现出较好的性能。

1.2本章小结

本章对英文领域和中文领域在最长名词短语识别方向的研究进行分析和讨论。最后对研究最长名词短语识别的一般方法进行了归纳和总结。

第2章 基于统计方法的中文最长功能名词短语自动识别

通过第 1 章的分析我们可以知道，统计方法中 SVMs 和 CRFs 有较好性能，而 MaxEnt 模型确是在自然语言处理应用比较广泛的模型，因此，这一章会对 MaxEnt 模型、SVMs 模型和 CRFs 模型如何处理 funMNP 的识别任务做详细的说明，并对三种模型进行比较。

2.1中文最长功能名词短语自动识别

2.1.1 中英文最长名词短语的明确定义描述

在英文研究领域，Kohen 和 Knight[4]最早提出了一种面向统计机器翻译的最长名词短语的定义：

给定一个句子 s 和它的句法分析树 t,一个名词短语是 t 的一棵子树,它至少包含一个名词而不包含动词,并且不被更大的包含名词并且不包含动词的子树包含。

在中文研究领域，周强等[12]将名词短语划分为三类：最短名词短语（mNP）、最长名词短语（MNP）和一般名词短语（GNP）。最长名词短语被定义为：被其他任何名词短语所包含的名词短语。这个定义也是目前主流认可的定义。钱小飞[23]给出的最长名词短语的定义为：不被其他任何名词短语所包含的名词短语，最长名词短语是句子级的短语单位，其上层结构即为句子根节点 S。可以看出是对主流定义的一个扩充说明。

实际上，这个定义还是有一定歧义的，一种理解是最长的连续的整体表现为名词词性的词串，另一种是考虑句法功能，在第一种理解的基础上而且要满足该词串整体在小句中有严格句法功能。有些词串虽然第一种理解的描述，但是明显是由不同句法成分的子词串构成的，而且它们之间不是严格意义上的修饰关系。下面给出一个具体的例子，“【”表示识别出的左边界，“】”表示识别出的右边界：

<原　　句>“八五”期间各项存款比“七五”末净增五十亿元，…。

<错误识别>【“八五”期间各项存款】比【“七五”】末净增五十亿元，…。

<正确识别>【“八五”期间】【各项存款】比【“七五”】末净增五十亿元，…。

在该句中，【“八五”期间】和【各项存款】虽然是邻近不间断的词串，但是【“八五”期间】明显是时间状语成分（TMP），而【各项存款】是主语成分（SBJ），把它们合并成一个 MNP 并无实际意义反而影响对句法结构的分析，

所以要分开识别。

因此，本文所研究的最长名词短语（MNP）实际上是最长功能名词短语（Functional Maximal-length Noun Phrase，funMNP），是不被其他任何名词短语所包含的具有独立句法功能的名词短语。

2.1.2 中文最长功能名词短语识别评价标准

本文对最长名词短语识别效果的评价方法选用最常用的采用的评价方法有准确率 (Precision, P)，召回率(Recall, R)和 F 值(F-1 measure, $F_{\beta=1}$)。

$$P=\frac{\text{识别出的正确funMNP数}}{\text{识别出的funMNP数}} \tag{2.1}$$

$$R=\frac{\text{识别出的正确funMNP数}}{\text{原文本中的所有funMNP数}} \tag{2.2}$$

$$F=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\times 100\% \tag{2.3}$$

其中，β 是一个权值，决定对 P 和 R 侧重视程度，这里我们取 β=1。

2.1.3 中文最长功能名词短语识别选用标记

在后续实验中，本文使用的是 IOB2 的标注方法，将最长功能名词短语的识别问题转化为其序列标注的问题。B 表示当前词是该词所在最长功能名词短语的首词，I 表示当前词属于所在短语块的内部，O 表示所在短语块的外部。选择 IOB2 标注而没有选择更细致的标注方式，是为了简化后续研究，避免细致分类导致的错误。

2.2基于最大熵模型的中文最长功能名词短语自动识别

2.2.1 最大熵模型的理论描述

最大熵模型（Maximum Entropy Models，MaxEnt）是一个指数模型，其主要思想是，对一个随机事件的概率分布的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设，在这种情况下，概率分布最均匀，预测的风险最小，因为这时概率分布的信息熵最大[25]。

对于一个有限的训练样本集合$\{(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)\}$，$x_i$ 表示当前对象上下文，y_i 表示当前位置的标记，其中 $1\leqslant i\leqslant N$。那么(x,y)的经验分布可以表示为：

$$\tilde{p}(x,y)=\frac{1}{n}\times count \tag{2.4}$$

其中，count 代表的是 (x,y) 出现在样本中的次数。

可利用样本集合的统计数据对上述大小为 N 的训练样本集合建立统计模型。引入特征函数，使模型依赖于上下文信息。假设给出 n 个特征函数 f_i，对每个特征进行条件限制，用其经验分布估计，则有：

$$p(f_i)=\tilde{p}(f_i)\quad i\in\{1,2,...,n\} \tag{2.5}$$

其中，期望值和经验值为分别为：

$$p(f)\equiv\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \tag{2.6}$$

$$\tilde{p}(f)\equiv\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{2.7}$$

为得到最优的 p(y | x) 值，则用条件熵衡量最优：

$$H(p)\equiv-\sum_{x,y}\tilde{p}(x)\,p(y|x)\log p(y|x) \tag{2.8}$$

求在限制条件下具有最大熵的模型：

$$p^*=\arg\max_{p\in C}H(p) \tag{2.9}$$

其中，C 是所有可能满足限制条件的概率分布模型的集合，约束为：

$$p(y|x)\ge 0,\forall x,\forall y \tag{2.10}$$

$$\sum_y p(y|x)=1,\forall x \tag{2.11}$$

$$\sum_{x,y}\tilde{p}(x)\,p(y|x)\,f(x,y)=\sum_{x,y}\tilde{p}(x,y)\,f(x,y),i\in\{1,2,...,n\} \tag{2.12}$$

为每个特征 f_i 引入一个拉格朗日参数 λ_i，构造条件熵的最大值拉格朗日函数得：

$$\xi(p,\Lambda,\gamma)\equiv H(p)+\sum_i\lambda_i(\tilde{p}(f)-p(f))+\gamma(\sum_x p(y|x)-1) \tag{2.13}$$

γ 和 Λ 对应 n+1 个限制，保持两个参数不变，计算(2.13)在不受限制情况下的最大值即可得到 p(y|x)的最优值 $p^*(y|x)$。

$$p^*(y|x)=Z(x)\exp(\sum_i\lambda_i f_i(x,y)) \tag{2.14}$$

其中 Z(x)是范化因子。归一化公式为：

$$Z_\lambda(x)=\sum_y\exp\sum_i\lambda_i f_i(x,y) \tag{2.15}$$

其中，f(x,y)特征函数为：

$$f(x,y)=\begin{cases}1, & \text{若 } x,y \text{ 满足某个事实；}\\ 0, & \text{否则.}\end{cases} \tag{2.16}$$

这样，最大熵原理就把一个有约束的优化问题转换成为一个没有约束的优化问题，通过将 λ 代入最大分布概率公式，则可解决原始问题。

2.2.2 实验语料的预处理说明

本文 funMNP 识别所用的实验语料是根据中文宾州树库 CTB 转化而来的，主要进行了以下工作：

(1) 根据 funMNP 的定义以及括号匹配原理，将 CTB5.0 中的树库转化为标记好边界的链状句子，词性标注采用 CTB 标注体系。

【SBJ 浦东/NR 开发/NN 开放/NN】是/VC【PRD 一/CD 项/M 振兴/VV 上海/NR ，/PU 建设/VV 现代化/NN 经济/NN 、/PU 贸易/NN 、/PU 金融/NN 中心/NN 的/DEC 跨世纪/JJ 工程/NN】，/PU 因此/AD 【SBJ 大量/AD 出现/VV 的/DEC】 是/VC 【PRD 以前/NT 不/AD 曾/AD 遇到/VV 过/AS 的/DEC 新/JJ 情况/NN 、/PU 新/JJ 问题/NN】 。/PU

图 2.1 链状句子实例

(2) 采用 IOB2 标记集对转化后的句子进行标记，并处理成 token 的形式。

(3) 将全部语料随机抽取 1/4 做测试语料，3/4 做训练语料。抽取 5 次，得到 5 组训练和测试语料，方便后续的交叉实验。

(4) 统计语料，总共 18782 个句子，527623 个词，87760 个 funMNP。为了便于后续工作中，窗口大小的选取，并对 funMNP 长度和数量之间关系做了统计分析，因为纵坐标数值较大，所以做了取对数的处理。

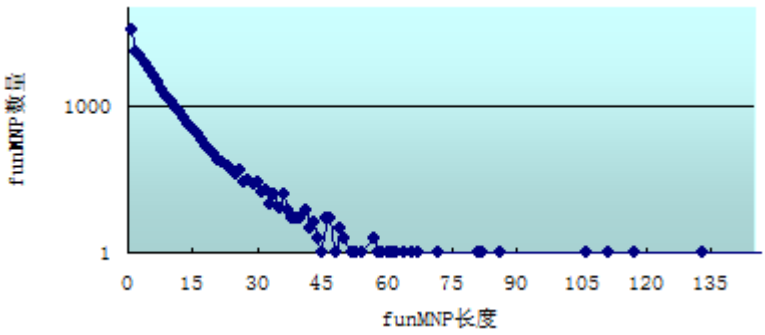

图 2.2 funMNP 长度数量关系

2.2.3 最大熵模型的特征选择

选择特征是最大熵模型中比较重要的步骤，如何采用有限的特征描述无限的语言现象是十分重要的。承认可观察的部分，对预测部分不做任何假设，在本任务中可观察到的事实可作为特征集合。

本任务中的每个特征分为两个部分：事件发生的条件和事件的最终结果。具体来讲就是，在什么样的条件下，当前词可以标注为 B（或者 I 和 O）。对于每个词分别处理时，要考虑词的上下文环境以及词本身来确定事件的特征集合。

在这部分中，我们定义特征空间为：

(1) 词（Word）信息：比如当前词，当前词的前一个词，当前词的后一个词等等。

(2) 词性（Part-of-speech, Pos）信息：比如当前词的词性，当前词的后两个词的词性等等。

(3) 组合信息：包括词与词之间的组合，词性与词性之间的组合和词与词性之间的组合。

后续需要使用 CRFs 模型，为了更好地对比两种模型处理 funMNP 问题的性能，先完成了特征模板的编写，然后编写特征抽取程序，根据特征模板抽取特征。特征模板详见 2.4 节 CRFs 识别 funMNP 部分描述。

2.2.4 最大熵模型的实验结果

本文使用东北大学张乐博士的开源 MaxEnt 工具包进行了 funMNP 的训练和测试，结果如下：

表 2.1 MaxEnt 识别 funMNP 结果

实验组别	P	R	F
Group1	57.28%	70.56%	63.23%
Group2	57.61%	71.1%	63.65%
Group3	57.65%	70.9%	63.6%
Group4	58.24%	71.06%	64.01%
Group5	56.49%	70.34%	62.66%

由表 2.1 的实验结果可知，MaxEnt 模型虽然在自然语料处理领域应用的比较广泛，但是并不适合于完成 funMNP 的识别任务，与目前的 MNP 识别效果相差很大，甚至会出现“…OI…”这种错误序列。通过分析错误实例和最大熵模型本身特点，总结原因主要有标注偏置问题(Label Bias Problem)和数据稀疏问题。

2.3基于支持向量机的中文最长功能名词短语自动识别

2.3.1 支持向量机模型的描述

支持向量机(Support Vector Machines，SVMs)作为一种监督式学习的方法，被广泛应用于统计分类以及回归分析中。SVMs 是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器。其实质就是构造一个恰当的超平面，使得两类样本能够正确分开且分类间隔尽可能大，以期任给一个新的样本，能够确定地判断是属于哪个类别。

以两类数据分类为例，给定训练样本集（x_i, y_i），i=1,2…, l, $x \in R^n$, $y \in \{\pm 1\}$，则一个线性分类的目标即是找到分类的超平面，记为：

$$w^T x + b = 0 \tag{2.17}$$

为使得分类平面正确分类且具备分类间隔，要求公式 2.18 满足约束：

$$y_i [w^T x + b] \geq 1, \ i = 1, 2, ...l \tag{2.18}$$

分类间隔（classification margin）为 2/||w||，构造最优平面问题可转换为在满足公式 2.19 的前提下求：

$$min\ \Phi(w) = \frac{1}{2} || w ||^2 = \frac{1}{2} w^T \cdot w \tag{2.19}$$

为求得 2.19 约束下的最优解，引入拉格朗日算子 a_i（a_i>0），构造函数记为：

$$L(w, b, a) = \frac{1}{2} || w || - a(y(w^T + b) - 1) \tag{2.20}$$

约束最优化问题的解由公式 2.21 的鞍点决定，且在鞍点处有：

$$\frac{\partial L(w, b, a)}{\partial w} = 0 \tag{2.21}$$

$$\frac{\partial L(w, b, a)}{\partial b} = 0 \tag{2.22}$$

最终得到最优分类函数为：

$$f(x) = sgn\{\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\} \tag{2.23}$$

如果是非线性问题，则利用非线性变换将其转化为高维空间中的线性问题，然后在变换空间求最优平面，则最优分类函数为：

$$f(x) = sgn\{\sum_{i=1}^{n} a_i^* y_i K(x_i \cdot x) + b^*\} \tag{2.24}$$

其中，K(x_i, x)为将地位空间向高维空间映射的核函数（Kernel Function），从而实现将非线性分类问题转化为线性分类问题。SVMs 基本原理如图 2.3 所

示。

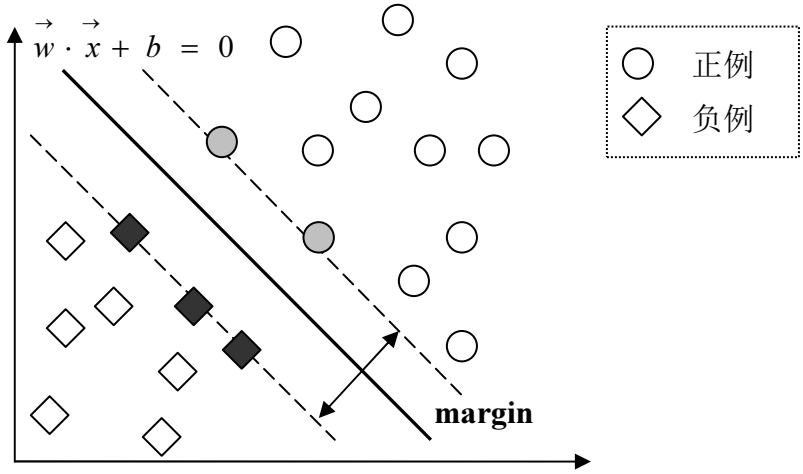

图 2.3 SVMs 基本原理

2.3.2 支持向量机模型识别实现

对于 funMNP 的识别是一个典型的序列标注问题，若要识别 funMNP 只需识别出 funMNP 的左右边界，其实也就是将连续的词串标记成“BI…I”形式即可。SVMs 是解决二类分类问题的模型，如果将语料中的每个词视为一个样本，那么要想用 SVMs 解决序列标注问题，只需对于一个词（样本）进行分类即可。由 2.3.1 的描述可知，SVMs 会将样本分成正例和负例，那么对于“B”、“I”和“O”三类分类问题，只需两两组合进行分类即可。

YamCha 是一个基于 SVMs 的通用、可定制的开源工具，它整合了 TinySVM 进行分类并打分，且将分类标签按序标出，可顺利完成序列标注任务。YamCha 的特征选取方式与 MaxEnt 的特征不同，YamCha 不仅可以指定静态特征（可以理解为上下文特征），而且可以指定动态特征，从而可以将历史标注结果加入到当前词的特征中影响分类决策结果。与 MaxEnt 模型不同，YamCha 的标注可以基于历史标注，因而可以选择正向标注（从左至右）和逆向标注（从右至左）。而且由于中文本身的特点，比如 funMNP 的右边界词是一般都是名词，对 funMNP 右边界的识别效果往往较左边界的识别效果好。因此，逆向标注的结果一般比正向标注结果好。然而，正向识别对逆向识别的修正也有一定价值，因而对双向识别进行了实验。

实验所用到的语料是 2.2.2 中处理成 token 列的语料，每一行是一个 token，由词、词性和 IOB 标签组成并有 Tab 键隔开，静态特征选取 F:-7…7，F:-7…7，动态特征选取 T:-7…-1。在正向标注过程中，动态特征 T 的含义是参考当前词的前七个词的标注情况，而在逆向标注过程中，动态特征 T 的含义是参考当前

词的后七个词的标注情况。

2.3.3 支持向量机模型实验结果

用 Yamcha-0.33 训练好的模型文件，并进行了测试，最后对测试结果进行了统计。表 2.2 是五组语料正向标注和逆向标注的识别效果数据。

表 2.2 SVMs 识别 funMNP 结果

实验组别		P	R	F
Group1	Forward	85.49%	87.24%	86.36%
	Backward	87.19%	87.83%	87.51%
Group2	Forward	85.35%	87.35%	86.34%
	Backward	87.05%	87.49%	87.27%
Group3	Forward	85.95%	86.82%	86.38%
	Backward	88.17%	87.55%	87.86%
Group4	Forward	86.26%	85.91%	86.09%
	Backward	88.24%	86.53%	87.37%
Group5	Forward	86.13%	86.73%	86.43%
	Backward	88.36%	87.58%	87.97%

通过表 2.2 不难看出，利用 SVMs 进行 funMNP 的识别任务要优于 MaxEnt 模型，表 2.1 中第三组的识别效果最好，表 2.2 中第五组的识别效果最好，而表 2.2 中第三组的识别效果仍比 2.1 中同组的识别效果好，P 值提高 27.71%，R 值提高 14.91%。

SVMs 模型可以较好的完成 funMNP 的识别任务，训练时空间开销比较小，在 4G 内存的 Linux 服务器上，五组实验任务可以同时提交进行模型运算，然而，时间开销比较大，训练过程 BI 分类，BO 分类和 IO 分类的 CPU 运行时间和在四个小时左右，而实际提交任务得到结果的时间要更长。总体上看，SVMs 的识别性能是比较不错的，与 1.1.2 中介绍的文献不加规则时的结果很接近。

2.4基于条件随机域的中文最长功能名词短语自动识别

2.4.1 条件随机域模型描述

条件随机域（Conditional random fields，CRFs），又称条件随机场，是一个可实现统计序列标注和分类的条件概率无向图模型，其主要优点是不需要像 HMM 等模型那样的独立性假设，因为大多数序列标注时间并非独立事件。而

且解决了最大熵马尔可夫模型（MEMM）标注偏置问题，可以计算全局最优解。CRFs 模型是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，而不是在给定当前状态条件下，定义下一个状态的分布。CRFs 可以通过观察序列中非独立的、相互作用的特征及其权值很好地拟和现实数据，广泛应用于自然语言处理的各项任务中。

CRFs 的定义：需要标记的观察序列集为随机变量 X=$\{x_1,x_2,...,x_n\}$，相应地表示标记序列集为 Y=$\{y_1,y_2,\ldots,y_3\}$，其中 $Y_i \in Y$ 在有限字符集内，随机变量 X 和 Y 联合分布。设无向图 G=（V，E），Y=$\{Y_v \mid v \in V\}$，Y 与 G 中顶点一一对应。如果每个 Y_v 对 G 遵守马尔可夫属性 $p(Y_v \mid X,Y_{w,},w \neq v)=p(Y_v \mid X,Y_{w,},w \sim v)$，则（X,Y）构成一个条件随机域。

无向图 G 的结构从理论上讲是任意的，可以描述标记序列的条件独立性，而 CRFs 在构造模型时选用的是最简单且最常用的一阶链式结构，即线性链结构（Linear-chain CRFs），形式如下：

$$p(y \mid x)=\frac{1}{Z(x)} exp(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,k}\mu_k s_k(y_i,x,i)) \tag{2.25}$$

其中归一化因子 Z(x)为：

$$Z(x)=\sum_{y} exp(\sum_{i,k}\lambda_k f_k(y_{i-1},y_i,x,i)+\sum_{i,k}u_k g_k(y_i,x,i)) \tag{2.26}$$

其中转移函数 $t_k(y_{i-1},y_i,x,i)$表示的是整个观察序列和相应标记序列在 i-1 和 i 时刻的特征，状态函数 $s_k(y_i,x,i)$是在 i 时刻整个观察序列和标记的特征。λ_k 和 μ_k 可从训练语料中估计，若参数为非负数且绝对值大则优先选择该特征事件，若参数为负数且绝对值大则该特征事件发生的可能性很小。

在使用 CRFs 模型时，有三个关键问题：特征函数的选择、参数的估计和模型的推断。

CRFs 模型中特征函数 $f(y_{i-1},y_i,x,i)$是状态函数和转移函数的统一形式表示，是一个二值函数。$b(x,i)$ 代表特征函数的所有真实特征的集合，表示为：

$$b(x,i)=\begin{cases}1 & ,if \text{ certain conditions are met}\\ 0 & ,elsewise\end{cases} \tag{2.27}$$

因而，特征函数可以表示为：

$$f(yi-1,yi,x,i)=\begin{cases}b(x,i) & ,if \ certain\ conditions\ are\ met\\ 0 & ,otherwise\end{cases} \tag{2.28}$$

CRFs 模型对于参数的估计一般进行最大似然估计（Maximum Likelihood Estimation，MLE)，即运用最优化理论循环迭代，直至函数收敛或者达到给定

的迭代次数。

若有训练集 D={$(X_1,Y_1), (X_2,Y_2),\ldots,(X_\Gamma, Y_\Gamma)$}，则条件概率 $p(y|x,\lambda)$的对数似然函数形式为：

$$L(\lambda)=log\prod_{x,y}p(y\mid x,\lambda)^{\tilde{p}(x,y)}=\sum_{x,y}\tilde{p}(x,y)logp(y\mid x,\lambda) \tag{2.29}$$

已知条件概率 $p(y|x,\lambda)$的形式化公式为：

$$p(y\mid x,\lambda)=\frac{1}{Z(x)}exp(\sum_{k}\lambda_k F_k(y,x)) \tag{2.30}$$

其中，Z(x)为归一化因子，表示为：

$$Z(x)=\sum_{y}exp(\sum_{k}\lambda_k F_k(y,x)) \tag{2.31}$$

CRFs 的最大似然估计任务的实质就是从相互独立的训练数据中估计参数向量 $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_n)$，因此对数似然函数为：

$$L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\sum_{k}\lambda_k F_k(y,x)-\sum_{x}\tilde{p}(x)logZ(x) \tag{2.32}$$

令条件随机域对数似然函数的梯度公式为零，即：

$$\frac{\partial L(\lambda)}{\lambda_k}=0 \tag{2.33}$$

解公式 2.34 即可求出 λ 的值。根据最大熵原理，模型分布中特征的期望等于经验分布中的期望值。

然而，通过上述方法求解参数 λ 并不一定总是得到一个近似解，因而，通常使用迭代方法或者 L-BFGS 算法解决参数估计最优化问题。常用的迭代算法包括 GIS 算法（Generalised Iterative Scaling） 和 IIS 算法（Improved Iterative Scaling）等。GIS 算法的收敛速度有计算更新的步长决定，训练参数 c 的值与步长成反比，且 c 值越大收敛速度越慢，反之收敛速度越快。IIS 算法主要思想则是将每个对观察序列和标记序列对(x,y)起作用的特征值的和近似等于对于观察序列 x 的最大可能的观察特征的和。当然也可以通过 L-BFGS 梯度算法实现实现参数估计，本文后续用到的 CRFs 模型对参数估计使用的就是 L-BFGS 梯度算法。

2.4.2 条件随机域特征模板

CRFs 与 MaxEnt 模型不同，对于特征不需要一一列举出来，只需要给出特征模板，CRFs 就可以根据特征模板自动进行特征抽取，一方面避免了训练和

测试的语料中冗余，另一方面只要修改模板就无需重新修改代码即可生成所有的特征。

在 funMNP 的自动识别任务中，给定特征模板就可以生成相应的特征函数，而特征函数的选取对识别效果起着关键性的作用。特征函数可以描述 funMNP 的边界分布信息以及内部结构组合的信息，并供 CRFs 模型生成训练模型，这个模型中会有根据训练样本抽取的特征以及相关参数。

一般特征模板分为两种类型，一种是原子特征模板，另一种是符合特征模板。通过两种模板可以自动组合成相应的原子特征和符合特征：

(1) 原子特征：只考虑一种因素，本文主要是词（Word）和词性（POS）。在后续的模板描述时，第 *i* 个词用 W_i 表示，它的词性用 P_i 表示。

(2) 复合特征：考虑多种因素的组合情况，实际上可以理解为原子特征的组合特征。

特征选择常用频度选择法和增量选择法。频度选择法就是选择训练样本空间中出现次数大于某设定常数 N 的特征，N 相当于一个阈值，可以根据实际任务的需要设定。这种方法易于实现、效果较好，却可能产生冗余特征。增量法的思想非常容易理解，一个特征加入后如果整体的性能提高就保留，否则就删除，可以选出比频度法更精炼的特征。

本文使用的方法是增量选择法，在 CoNLL 2000 共享的 baseNP 识别特征模板基础上，不断加入新的特征，从而有目的地提高识别性能。根据 funMNP 长度的统计（见图 2.2），可看出 funMNP 长度是约是 15 的时候，因而窗口应该开 7 左右，最后经过实验证实窗口开 7 效果最佳，实验中效果最好的模板如表 2.3 所示。

表 2.3 CRFs 特征模板

编号	特征	编号	特征
1	W_i,-7≤i≤7,i ∈Z	8	$P_{i-1}/P_i/P_{i+1}$, -2≤i≤2, i ∈Z
2	W_i/W_0, -5≤i≤5, i ∈Z	9	$P_{-1}/P_0/W_i$, -2≤i≤2, i ∈Z
3	W_{i-1}/W_i, i=-2,-1, 2, 3	10	$P_0/P_1/W_i$, -2≤i≤2, i ∈Z
4	P_i,-7≤i≤7, i ∈Z	11	$P_{-3}/P_{-2}/P_{-1}/P_0$
5	P_{i-1}/P_i, -1≤i≤3, i ∈Z	12	$P_0/P_1/P_2/P_3$
6	W_{-1}/P_0	13	$P_{-2}/P_{-1}/P_0/P_1/P_2$
7	W_1/P_0		

其中，当 i=0 时，代表是当前词所在的位置，所以 W_0 代表当前的词，P_0 代表当前词的词性；当 i<0 时，代表相当于当前词往前数 i 个位置，所以 W_i 代表当前词往前数 i 个位置的词，P_i 代表当前词往前数 i 个位置的词性；

2.4.3 分词粒度的对比实验

在 2.2.2 语料预处理部分已经提到，本文的实验使用 CTB 本身带有的词性标注，这一节将对实验过程中发现的分词粒度对 funMNP 识别效果的影响问题进行讨论。

下面这组对比实验是对于五组实验语料分别用 CTB 和 Nihao 分词与词性标注体系进行标注，因为只是想要比较分词粒度对 funMNP 识别效果的影响，因而只是对两个实验使用同样的特征模板并没有特意选择 2.4.1 中给出的最好效果的模板。实验结果如表所示：

表 2.4 五组粒度对比实验

组别/类别		P	R	F
Group1	CTB	85.60%	86.96%	86.27%
	Nihao	75.56%	75.03%	75.30%
Group2	CTB	85.15%	86.71%	85.92%
	Nihao	76.52%	76.16%	76.34%
Group3	CTB	86.42%	86.64%	86.53%
	Nihao	76.87%	75.31%	76.08%
Group4	CTB	86.31%	85.71%	86.01%
	Nihao	76.41%	74.42%	75.40%
Group5	CTB	86.47%	86.82%	86.64%
	Nihao	75.12%	74.12%	74.61%

由表 2.4 可以明显看出，使用 CTB 标注要比使用 Nihao 标注的 P、R、F 三项指标均高 10%左右，而且每组对比试实验的数据都比较稳定而并非偶然现象。分析其主要原因是 CTB 标注体系的分词粒度要比 Nihao 标注体系的分词粒度大。前人的论文中普遍承认名词短语（或者句子）的长度对识别效果的影响较大，不妨称之为绝对长度（字数）。实际上，其绝对长度并非是真正的影响因子，相对长度（token 数）也可以说是分词的粒度才是真正的影响因子。在自然语言处理的任务中，最小的语言理解单位应该是词而并非单个字，对于同一个段字序列，分词粒度大，则词数少，那么识别就比较容易。因而，本文选择了 CTB 的标注体系进行标注，但并不意味着对于分词系统而言，分词粒度越大越

好，因为，粒度大相对包含的语言信息就会减少，可能会不利于自然语言处理的其他任务的进行。因而，较为理想的方案是分层次的标注，在标注体系的最高层是较为较大的类别，最底层是细致的分类，而在不同任务中选用不同层次的标注。

2.4.4 条件随机域实验结果

训练语料和测试语料与 YamCha 模型(SVMs 模型)所用的语料相同，都是 token 形式的，条件随机域模型的训练和测试使用了开源序列标注工具 CRF++0.58。表为使用表 2.3 中的特征模板的 funMNP 自动识别结果。

表 2.5 条件随机域模型实验结果

实验组别	P	R	F
Group1	87.67%	88.78%	88.22%
Group2	87.10%	88.43%	87.76%
Group3	88.33%	88.48%	88.41%
Group4	88.33%	87.36%	87.84%
Group5	88.23%	88.29%	88.26%

由表 2.5 的结果不难看出，CRFs 对于 funMNP 的识别性能比 MaxEnt 模型和 SVMs 模型识别的效果要好，CRFs 模型的第五组实验 F 值比 SVMs 模型最好的第五组反向标注实验的 F 值（见表 2.2）高 0.29%，而且 CRFs 模型最好的第三组实验的 F 值比 SVMs 最好的第五组反向标注实验的 F 值高 0.44%。

如表 2.6 所示，Template1~Template6 是实验的一些特征模板，1~6 模板特征递增，有表中数据可知，模板中的特征并非越多越好，模板 5、6 再增加特征时，识别效果反而下降，分析原因主要是冗余特征会成为噪声影响识别效果。

表 2.6 Group3 用不同特征模板的识别效果

模板	P	R	F
Template1	86.42%	86.64%	86.53%
Template2	86.72%	86.74%	86.73%
Template3	87.22%	87.05%	87.14%
Template4	88.33%	88.48%	88.41%
Template5	88.30%	88.40%	88.35%
Template6	87.74%	86.82%	87.28%

然后我们考虑一下训练语料和测试语料的大小对识别效果的影响。如表 2.7 所示，是实验中五组实验的训练和测试语料的大小。对比表 2.5 的 funMNP 识别结果，Group3 的识别效果是最好的，但是该组训练语料和测试语料既不是最大的，也不是最小的，因而识别效果与语料的大小并不是完全线性相关的。一方面，我们希望有足够的训练语料得到包含更多有用特征，另一方面我们有不希望过多的语料带来噪声问题，使得模型性能降低，因而选择合适规模而且没有噪声的语料也是一个非常重要的环节。

表 2.7 条件随机域模型所用语料大小

	Group1	Group2	Group3	Group4	Group5
Train	3.66 MB	3.71 MB	3.58 MB	***3.57 MB***	3.46 MB
Test	0.83 MB	***0.78 MB***	0.91 KB	0.92 MB	1.03 MB

2.5本章小结

第二章主要用最大熵模型（MaxEnt）、支持向量机模型（SVMs）和条件随机域模型（CRFs）对最长功能名词短语（funMNP）进行识别，MaxEnt 的识别性能最差而且与 SVMs、CRFs 模型的识别性能相差至少 20%以上，因而认为 MaxEnt 不适合 funMNP 的识别。

SVMs 模型和 CRFs 模型相比，SVMs 相对较差但是差别不大，SVMs 模型可以基于历史标注，考虑之前标记好的状态标签，而且正向和逆向标注相比，逆向标注的结果要好于正向标注的结果，分析其原因主要是中文最长功能名词短语的右边界比较容易识别，funMNP 内最右边界词往往是名词，所以进行新的标注时之前类别标记正确的概率要大于正向标注时的之前类别标记正确的概率，因而也有一定可取性。

综上，后续的多模型融合算法中 MaxEnt 模型不参与决策，只考虑 SVMs 和 CRFs 之间的融合。

第3章 多策略融合的中文最长功能名词自动短语识别

3.1多策略融合识别框架

上一章已经利用 MaxEnt、SVMs 和 CRFs 三种模型进行了中文最长功能名词短语（funMNP）的识别，受文献[16]和[17]和的启发，本章使用多策略融合的方法，对 funMNP 进行识别。方法整体框架如图所示。

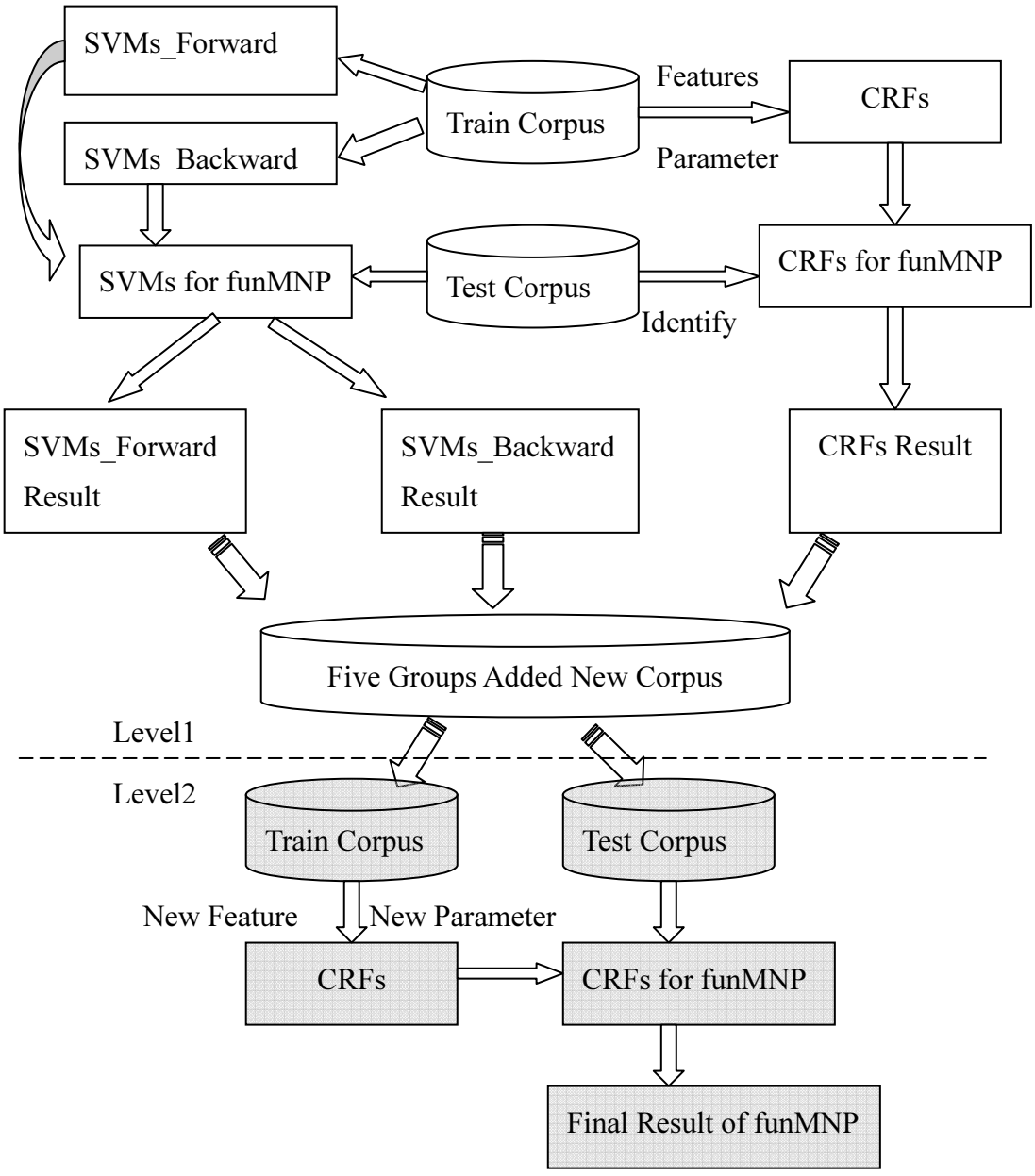

图 3.1 多策略识别 funMNP 框架

多策略的核心思想是在第二层 CRFs 识别时，将第一层的多个分类结果作为第二层的特征输入，以期有更多的可供学习特征，从而强化对 funMNP 的识别性能。第二层识别选择 CRFs 出于两方面考虑：

(1) CRFs 模型是目前性能最好的模型。

(2) CRFs 模型的训练和测试时间比 SVMs 模型短。

第一层 SVMs 模型正向和逆向识别的两个结果以及 CRFs 模型识别的结果语料格式可表示为 $W_i/P_i/T_i/AT_i$ ，其中 i 表示当前处理到的 token 的位置，也可以理解为当前词的位置，W_i 是当前词，P_i 是当前词的词性，T_i 是当前词的正确标注，AT_i 是当前词的自动标注。

第二层所用语料预处理的具体方法是，将第一层的三个结果文件选择性进行合并，生成格式为 $W_i/P_i/$ YamCha_B_AT_i / YamCha_F_AT_i / CRFs_AT_i /T_i ，然后，为了将第二层识别的结果与第一层进行比较，需要对应第一层的语料，将第二层所需语料分成训练语料和测试语料，分预料过程执行五次，得到与第一层对应的五组实验语料。

在第二层识别时也需要用到特征模板，在第一层所用特征模板的基础上，增加三列原子特征：

(1) YamCha_B_AT_i, $-7 \leqslant i \leqslant 7, i \in Z$

(2) YamCha_F_AT_i, $-7 \leqslant i \leqslant 7, i \in Z$

(3) CRFs_AT_i, $-7 \leqslant i \leqslant 7, i \in Z$

在准备好第二层所需语料和特征模板后进行 funMNP 的自动识别。在多策略多层框架中，识别策略本文只选用了 SVMs 正向、SVMs 逆向和 CRFs 模型进行识别有两个原因：

第一，三者的性能比较好，而且相差不大。第二，三者迭代训练模型的过程比较慢，但是测试的速度都比较快，这样就可以在计算性能较强的 Linux 服务器下完成模型的训练，在客户端进行测试，有可应用的前景。

在多层框架中，本文只使用两层框架，也有两个主要原因，一是考虑到预处理的复杂性，二是防止层数过多提高的性能受限，复杂的处理代价可能并不会换来更好的性能。

3.2多策略融合实验结果

如表 3.1 所示，是 3-策略 2-层融合的实验结果，从表中数据不难看出，多策略融合的方法使得识别性能稳步提高，而且第三组的识别性能最好。

多策略融合的方法最好一组的 F 值比一层 SVMs 正向标注最好一组的 F 值

提高 1.86%，比一层 SVMs 逆向标注最好一组的 F 值提高 0.95%，比一层 CRFs 标注的最好一组的 F 值提高 0.51%，更值得关注的是，对于它们识别最好的效果而言，一层的 SVMs 模型和一层 CRFs 模型的召回率比较高，正确率相对低，其本质可以理解为“找到的多，但找对的少”，当然这里的“多”和“少”都是相对的，而 3-策略 2-层融合的结果，是正确率明显提高，也就是识别正确的样本增多了，由此可见，多策略融合的方法对修正机器学习模型对 funMNP 识别是有效的。

表 3.1 多策略融合实验结果

实验组别	P	R	F
Group1	88.43%	89.21%	88.82%
Group2	88.11%	88.80%	88.45%
Group3	89.03%	88.82%	88.92%
Group4	88.82%	87.55%	88.18%
Group5	88.89%	88.71%	88.80%

3.3本章小结

本章主要介绍了多策略多层融合的方法来识别 funMNP，并取得了一定提高。在选择每个策略时应当遵循两个原则：

(1) 分策略择优原则：每个策略的性能不能太差，比如，MaxEnt 模型并没有加入到融合策略中来。尽量选择本身性能就比较好的策略。

(2) 分策略差异原则：分策略的模型应该尽可能具备差异性，这样的决策才能有说服力。

(3) 高层选最优原则：对于上层的策略要选择性能最好的策略。

在层数选择上，本文选择两层，但也可以根据实际需要选择更高的层数。总体上看，多策略多层融合的方法是有效可行的。

第4章 单策略识别与多策略融合的识别演示系统实现

4.1系统框架

基于第1~3章的理论研究，本章将实现单策略以及多策略融合识别funMNP的系统，该系统目标功能是对输入的新语料进行funMNP的识别。开发环境为Visual Studio 2010，系统框架如图4.1所示：

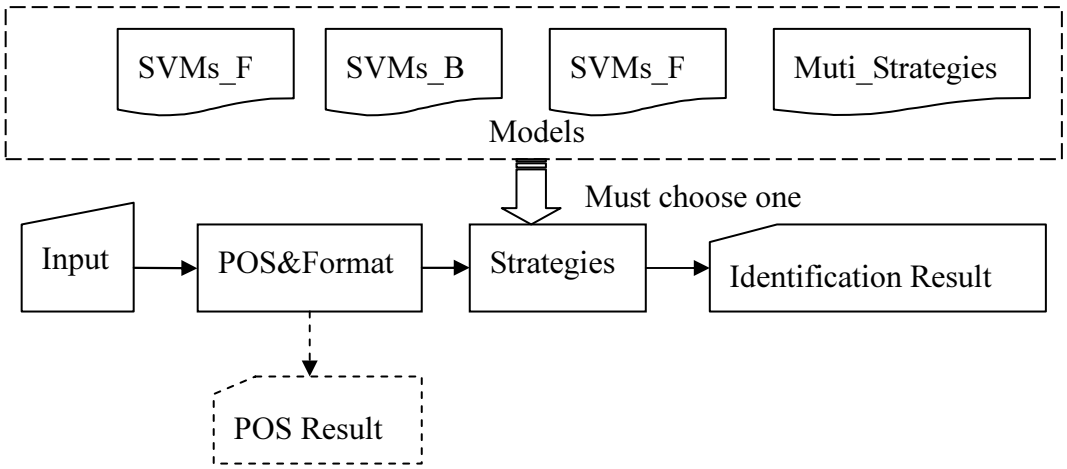

图 4.1 系统框架图

该系统主要分为三个部分：

(1) 语料预处理：对输入的句子进行分词、词性标注并处理成所需的标准格式。本文中所使用的分词工具是宾州树库标注体系的 ctbparser_0.12，并使用 VS2010 对 ctbparser_0.12 源码进行了修改和编译。标注输入格式是指每类特征为一列，每列之间用Tab符号隔开，每一行是一个token的形式。

(2) 选择策略自动识别：先要选择模型，然后进行自动识别。这个过程实质上是根据模型进行测试的过程。选择多策略模型的前提是三种单策略的识别结果文件已经生成，这样规定是为了分步骤观察识别效果。

(3) 显示格式的后处理：将带有 IOB 标注的 token 形式的语料转化成将funMNP用“【】”括起来的句子形式，并显示在Output区域。

4.2 系统流程

如图 4.2 所示是本文实现的识别系统的流程图。

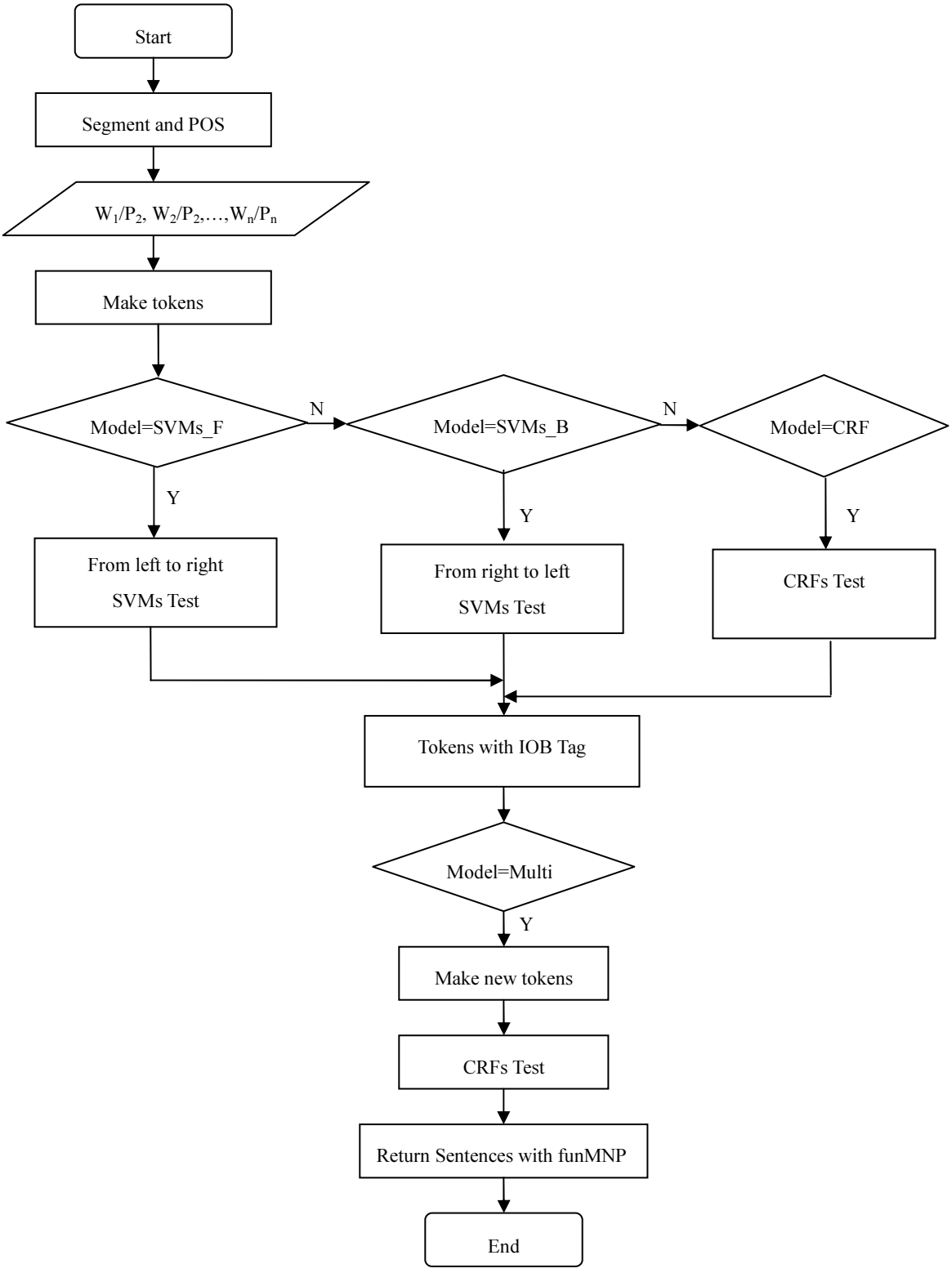

图 4.2 funMNP 识别系统流程图

输入的文本要求是生语料，即不经过任何处理的文本。例如：

“习近平在主持学习时发表了讲话。他指出，党的十八届三中全会提出，经济体制改革是全面深化改革的重点，核心问题是处理好政府和市场的关系，使市场在资源配置中起决定性作用，更好发挥政府作用。”

以该句为例，系统流程如下：

(1) 在 Input 框内输入生语料，点击 Seg_Pos 按钮，进行分词和词性标注。结果如图所示：

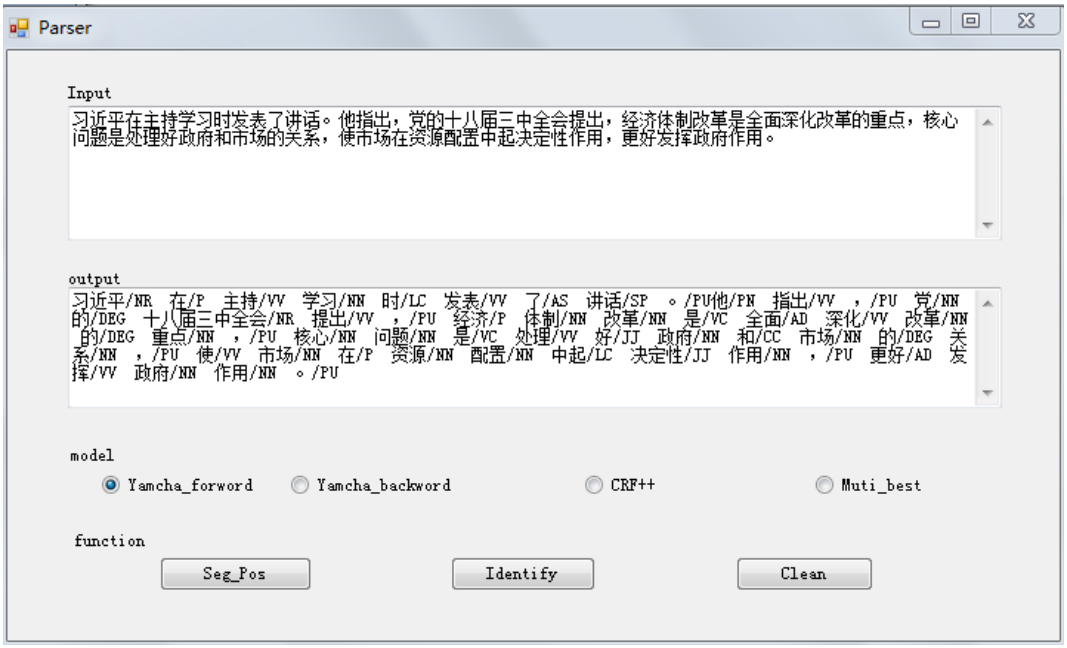

图 4.3 分词结果图

(2) 选择 Yamcha_forword 模型，点击 Identify 按钮。生成的格式化的测文件如图 4.3 所示。识别结果如图 4.4 所示。识别结果文件如图 4.5 所示。

```
1  习近平→NR
2  在→P
3  主持——→VV
4  学习——→NN
5  时→LC
6  发表——→VV
7  了→AS
8  讲话——→SP
9  。→PU
10
11 他→PN
12 指出——→VV
13 ，→PU
14 党→NN
15 的→DEG
16 十八届三中全会→NR
17 提出——→VV
18 ，→PU
```

图 4.4 一层模型标准输入格式

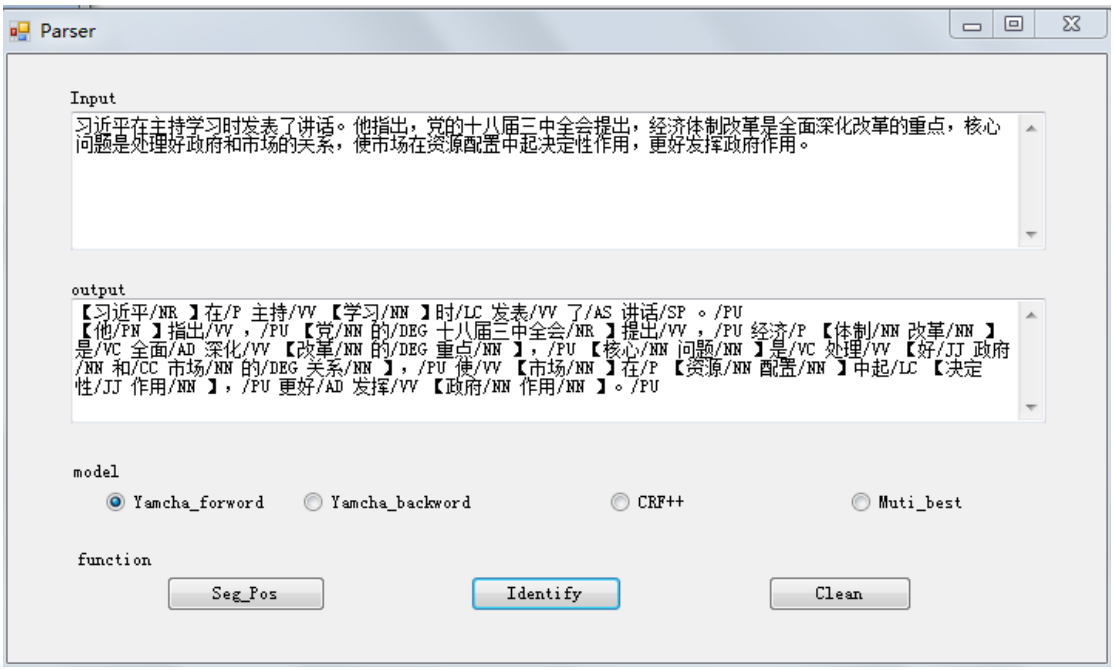

图 4.5 Yamcha 正向标注结果

```
1  习近平→NR→B
2  在→P→O
3  主持→VV→O
4  学习→NN→B
5  时→LC→O
6  发表→VV→O
7  了→AS→O
8  讲话→SP→O
9  。→PU→O
10
11 他→PN→B
12 指出→VV→O
13 ，→PU→O
14 党→NN→B
15 的→DEG→I
16 十八届三中全会→NR→I
17 提出→VV→O
18 ，→PU→O
```

图 4.6 识别结果文件标准输出格式

(3) 同理依次选择 Yamcha_backword 模型和 CRF++模型。三个模型对该段文本识别都很好，因而 Yamcha_backword 模型和 CRF++模型的识别结果与 Yamcha_forward 识别结果是相同的。

(4) 在确保三个单策略模型识别完成，并生成相应的结果文件时，可以选择 Multi_best 模型，是多策略的最优模型。多策略模型的标准输入格式如图 4.6 所示。由于三个单策略的模型都可以很好地识别该段文本，因而识别结果相同。

```
1  习近平→NR→B→B→B→B
2  在→P→O→O→O→O
3  主持→VV→O→O→O→O
4  学习→NN→B→B→B→B
5  时→LC→O→O→O→O
6  发表→VV→O→O→O→O
7  了→AS→O→O→O→O
8  讲话→SP→O→O→O→O
9  。→PU→O→O→O→O
10
11 他→PN→B→B→B→B
12 指出→VV→O→O→O→O
13 ，→PU→O→O→O→O
14 党→NN→B→B→B→B
15 的→DEG→I→I→I→I
16 十八届三中全会→NR→I→I→I→I
17 提出→VV→O→O→O→O
18 ，→PU→O→O→O→O
```

图 4.7 多策略融合二层输入文件格式

4.3本章小结

本章实现了最长功能名词短语识别演示系统的开发，将对 funMNP 识别的研究成果应用到真实的系统中。首先，从总体上对系统框架进行了分析和说明，然后通过一段生语料的处理实例，描述了系统从标注、形式化处理到识别出 funMNP 的整个流程，得到了预期效果。

结论

最长功能名词短语是指不被任何其它名词短语包含的，具有独立句法功能的名词短语。对最长功能名词短语的实验研究，不仅有利于浅层句法分析，而且对自然语言处理的其它任务也有很大价值，如实体识别、指代消解、翻译消歧、信息检索等。

本文在对现有识别技术充分了解的前提下，主要使用机器学习的方法，进

行最长功能名词短语的识别。一般而言，统计方法与规则结合会达到较高 F 值，但是出于两点原因本文并没有加入规则后处理进行提高，一是对规则的总结需要较深的语言功底，否则并不会带来很好的效益，比如在研究期间曾尝试加入规则，结果并没有提高，甚至会出现下降的情况。二是规则对语料有依赖性，一旦更换语料，规则可能会不适用。然而这并不意味着本文否定规则方法的作用，只是在当前研究范围内，只是将统计方法作为研究重点。

在研究过程中，发现了几点需要说明的问题。一是在以往的最长名词短语研究中并没有提出功能的概念，而且对于最长名词短语的定义也是有歧义，因而在本文的研究中，对最长功能名词短语的概念进行了定义。二是发现句子的相对长度（token 数）对最长功能名词短语的识别有很大的影响，在本文的研究中只是使用了粒度较大的分词、词性标注系统，就比粒度小的情况下的性能高10%以上。

在单层识别模型中，MaxEnt 模型识别性能较差，最好的一组 F 值只有64.01%，不适于最长功能名词短语的识别任务。CRFs 模型的识别性能最好，最好的一组 F 值可达 88.41%，与现在最好的研究结果相近，训练和测试的时间成本也在可控范围内，一个较大的缺点是内存消耗很大，在实验过程中，一旦内存分配不错就会出现错误，无法继续进行训练和测试。SVMs 模型的识别性能也不错，正向识别最好的一组 F 值可达 86.43%，逆向识别最好的一组 F 值可达 87.97%，与 CRFs 识别效果相差不大。

在第三章，本文实现了多策略融合的方法对最长功能名词短语进行了识别，该方法的主要思想是通过多策略决策修正单策略决策的错误，以期得到更好的识别效果。虽然理论上对于策略数量和多策略框架的层数没有要求，但是出于对预处理过程复杂性和增加层数的性能提高幅度两方面考虑，本文实现的是 3-策略 2-层的方法，得到了较为理想的识别效果，最佳的一组 F 值达 88.92%，比正向 SVMs 单策略高 1.86%，比逆向 SVMs 单策略高 0.95%，比 CRFs 单策略提高 0.51%，而且多策略识别的 P 值大于 R 值，说明多策略方法可以修正单策略标注的错误，得到更多识别正确的最长名词短语。

在第四章，本文实现了最长功能名词短语的自动识别系统，将理论研究应于与实际系统。

今后的研究工作主要分为两个方面：

(1) 对单策略识别的提高，包括挖掘更好的特征、选择适当的训练语料、去除语料中的噪音、调节参数值等方法。
(2) 发现更多高效的单策略模型进行参加决策。
(3) 对第二层策略的改进，包括采用比 CRFs 识别性能更好的模型作为第二

层的识别模型，改进第二层的特征模板，调整参数等方法。

(4) 将识别结果应用于真实的翻译系统中，提高该系统的翻译性能。

参 考 文 献

[1] CHURCH K.A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text[C]//In Proceedings of the second Conference on Applied Natural Language Processing,1988:136-143.

[2] VOUTILAMEN A.NPTool, A Detector of English Noun Phrases[C]//In Proceedings of the Workshop on Very Large Corpora:Academic and Industrial Perspectives ,1993:48-57.

[3] K. Chen and H. Chen.Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation[C]//Proceedings of 32nd Annual Meeting of Association of Computational Linguistics,1994:234-241.

[4] KOEHN P,KNIGHT K.Feature-Rich Statistical Translation of Noun Phrases[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,2003:311-318.

[5] HALLIDAY M A K.An Introduction to Functional Grammar (third edition)[M].Beijing:Language Teaching and Research Press,2008.

[6] 马建军.面向机器翻译的英语功能名词短语识别研究[D].大连:大连理工大学,2012.

[7] 马艳军,刘颖.基于隐马尔可夫模型和候选排序的汉语基本名词短语识别[C]//清华大学出版社,2005:107-112.

[8] 张惠春.基于最大熵模型的中文名词短语识别[J].电脑知识与技术,2009(8):146-148.

[9] 梁颖红,赵铁军,姚建民,等.基于混合策略的英语基本名词短语识别——边界统计和词性串规则校正相结合的策略[J].计算机工程与应用,2004(35):4-6+124.

[10] 孟迎,冯丽辉,赵铁军.基于决策树的汉语基本名词短语识别[J].黑龙江工程学院学报,2004(2):3-6.

[11] 徐昉,宗成庆,王霞.中文 Base NP 识别:错误驱动的组合分类器方法[J].中文信息学报,2007(1):117-121.

[12] 周强,孙茂松,黄昌宁.汉语最长名词短语的自动识别[J].软件学报,2000(2):176-181.

[13] 冯冲,陈肇雄,黄河燕,等.基于条件随机域的复杂最长名词短语识别[J].小型微型计算机系统,2006(6):176-181.

[14] 王月颖.中文最长名词短语识别研究[D].哈尔滨:哈尔滨工业大学,2007.

[15] 代翠,周俏丽,蔡东风,等.统计和规则相结合的汉语最长名词短语自动识别_代翠[J].中文信息学报,2008(6):110-115.

[16] 代翠.汉语最长名词短语的自动识别与分析[D].沈阳:沈阳航空工业学院,2009.

[17] 鉴萍,宗成庆.基于双向标注融合的汉语最长短语识别方法[J].智能系统学

报,2009(5):34-41.

[18] Zhang G,Lang W,Zhou Q .et al.Identification of Maximal Length Noun Phrases Based on Maximal Length Preposition Phrases in Chinese[C]//Asian Language Processing (IALP), 2010 International Conference on. IEEE,2010:65-68.

[19] Yegang Li,Heyan Huang.Automatic identifying of maximal length noun phrase[C]//Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on ,2012:1445-1448.

[20] 钱小飞.以“的”字结构为核心的最长名词短语识别研究[J].计算机工程与应用,2010,46(18):138-141.

[21] 钱小飞,陈小荷.含“的”字偏正结构的最长名词短语的自动识别[C]//内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集,2007:96-101.

[22] 钱小飞.含“的”最长名词短语的自动识别[D].南京:南京师范大学,2007.

[23] 钱小飞.最长名词短语识别研究[J].现代语文(语言研究版),2009(7):124-126.

[24] 钱小飞,侯敏.基于混合策略的汉语最长名词短语识别[J].中文信息学报,2013,27(6):16-22.

[25] 吴军.数学之美[M].北京:人民邮电出版社,2012.

致　　谢

首先感谢海事大学为我提供良好的学习环境，给我一个发展和提升的舞台；感谢软件工程专业的各位老师在四年里对我的辛苦培育，不仅教会了我专业知识与技能，而且给予了我无私的关怀和帮助；感谢我的同学和朋友，在生活和学习上给予我宽容和照顾。

特别感谢我的指导老师谢益武老师。课堂上，谢老师讲课清晰、幽默又有耐心，不仅教会我们专业知识，也常常给我们讲人生道理，让我终身受益，不畏困难，知难而上。在此，我谢老师表示崇高的敬意和衷心的感谢！

附录 1

Part-Of-Speech tags: 33 tags

标记	英语解释	中文解释
AD	adverbs	副词
AS	Aspect marker	体态词，体标记（例如：了，在，着，过）
BA	把 in ba-const	“把”，“将”的词性标记
CC	Coordinating conjunction	并列连词，“和”
CD	Cardinal numbers	数字，“一百”
CS	Subordinating conj	从属连词 （例子：若，如果，如…）
DEC	的 for relative-clause etc	“的”词性标记
DEG	Associative 的	联结词“的”
DER	得 in V-de construction, and V-de-R	“得”
DEV	地 before VP	地
DT	Determiner	限定词，“这”
ETC	等，等等 in coordination phrase	等，等等
FW	Foreign words	例子：ISO
IJ	interjection	感叹词
JJ	Noun-modifier other than nouns	
LB	被 in long bei-construction	例子：被，给
LC	Localizer	定位词，例子：“里”
M	Measure word (including classifiers)	量词，例子：“个”
MSP	Some particles	例子：“所”
NN	Common nouns	普通名词
NR	Proper nouns	专有名词
NT	Temporal nouns	时序词，表示时间的名词
OD	Ordinal numbers	序数词，“第一”
ON	Onomatopoeia	拟声词，“哈哈”
P	Prepositions (excluding 把 and 被)	介词
PN	pronouns	代词
PU	Punctuations	标点
SB	被 in long bei-construction	例子：“被，给”
SP	Sentence-final particle	句尾小品词，“吗”
VA	Predicative adjective	表语形容词，“红”
VC	Copula 是	系动词，“是”
VE	有 as the main verb	“有”
VV	Other verbs	其他动词