博士学位论文

现代汉语非受限文本的实语块分析

姓　　名：　孙宏林

系　　别：　计算机科学与技术系

专　　业：　计算机科学理论

研究方向：　计算语言学

导　　师：　俞士汶 教授

二〇〇一年五月

A Content Chunk Parser for Unrestricted Chinese Text

by

Sun Honglin

Submitted to the Department of Computer Science and Technology

in partial fulfillments of the requirements for the degree of

Doctor of Science

at

Peking University

May 2001

Thesis Supervisor: Prof. Yu Shiwen

现代汉语非受限文本的实语块分析

摘　要

对非受限的自然语言文本进行自动句法分析目前仍是自然语言处理所面临的一个巨大挑战，即使对于英语这样得到充分研究的语言至今也还没有一个可以处理非受限文本的高性能的句法分析器。解决句法分析难题的途径之一是采取“分而治之”的策略，即将复杂的句法分析任务分解为若干相互独立的子任务。本文提出的实语块分析就是根据这种思想而提出的一种浅层句法分析任务，其目标是从文本中连续的实词串中分析出可能的结构。由于可以在很大程度上避开跟许多虚词相关的远距离依赖问题，因而实语块分析可以得到很高的性能和效率。实语块分析的结果可以使句子的结构得到简化，从而降低完全句法分析的歧义和复杂度。本文的研究表明，实语块分析是一个可以明确定义、相对独立的句法分析子任务，与基本名词短语分析等浅层分析任务相比，它可以得到更多的句子结构信息。

本文描述了一个完整的汉语实语块分析系统，该系统接受非受限的自然语言文本作为输入，输出包括分词、词性标注、命名实体识别和实语块分析的结果。

具体地说，本文取得了以下成果：

（1） 提出了汉语实语块分析任务，并通过实验证明实语块分析是可以明确定义、相对独立的句法分析子任务。

（2） 提出了概率上下文无关语法和概率属性相结合的汉语实语块分析模型。该模型利用跟规则相关的句法属性对上下文无关语法进行约束。这些句法属性的可学习性使模型更具有健壮性。同时，由于句法属性是带概率的，因而使模型具有更强的消歧能力。

（3） 根据汉语句法中节律对句法具有约束作用的性质提出了利用音节和长度信息对规则进行约束的思想，并使节律特征概率化。

（4） 提出了汉语并列结构的概率模型，该模型利用并列结构的对称性在若干并列项候选中选择正确的并列项。

（5） 提出了对词类标记集复杂度的定量分析方法，并在此基础上提出了基于遗传算法的词类标记集优化方法。

（6） 实现了一个完整的汉语实语块分析系统，该系统以非受限的汉语文本作为输入，以实语块分析结果作为输出。除了得到实语块之外，还可以得到分词、词性标注和命名实体识别的信息。

本文的研究成果可以应用到信息提取、信息检索、机器翻译等自然语言处理系统中。在本文描述的实语块分析系统之上，有望开发出能够处理非受限文本的汉语句法分析器。

关键词： 计算语言学　自然语言处理　浅层分析　信息提取　实语块

A Content Chunk Parser for Unrestricted Chinese Text

Abstract

Automatically parsing unrestricted natural language text is still a great challenge to natural language processing (NLP) at present. Even for the well-studied languages like English, a robust parser, which can deal with unrestricted text with high performance, is not available until now. An approach to the difficult problems like parsing is to take a divide-and-conquer strategy, i.e., dividing the whole complicated problem into several independent sub-problems, which can be solved relatively easily. Content chunk parsing, proposed in this thesis, is such a kind of subtasks in parsing, whose object is to acquire possible structures from a sequence of content words. As much long-distance dependence associated with function words, for example, the determination of the right boundary of a preposition phrase and the left boundary of a "DE structure" in Chinese, can be avoided, content chunk parsing can obtain high performance and efficiency. The result of content chunk parsing can make a sentence simplified and lead to a noticeable decrease of ambiguities and complexities in full parsing. The research in this thesis indicates that content chunk parsing is a well-defined, easy to understand and relatively independent subtask in parsing. Compared with baseNP analysis, it can acquire more structural information from sentences.

This thesis describes a whole content parser which takes unrestricted Chinese text as input and gives content chunks as output in which the information about word-segmentation, POS tagging, Named Entity recognition can be gained as well.

The thesis makes contributions as follows:

(1) It defines the task of content chunk parsing and proves that content chunk parsing is a well-defined, easy to understand and relatively independent subtask in parsing that can be dealt with independently.

(2) It proposes a model integrating probabilistic context-free grammar (PCFG) and probabilistic features for content chunk parsing. This model makes constraints on PCFG utilizing the syntactic features associated with specific rule. The learnability of the syntactic features makes the model robust and the stochastic nature of the features gives the model more power for structural disambiguation.

(3) It proposes the idea of using rhythm feature for constraining the over-generation of context-free grammar rules, based on the observation that rhythm can have an effect on syntax in linguistics.

(4) It proposes a probabilistic model for coordinate constructions based on the observation that the members of coordinate constructions tend to be symmetric in both syntactic and semantic aspects. The model can be used for making correct choices on coordinate members among two groups of candidates on both sides of a coordinate conjunction.

(5) It proposes a quantitative method for measuring the complexity of a

part-of-speech tag set and proposes a method for optimizing POS tag set using genetic algorithms.

(6) It implements a whole content chunk parser which takes unrestricted Chinese text as input and gives content chunks as output in which such information as word-segmentation, POS tagging, Named Entity recognition is available.

These results can be applied in many NLP applications, such as information extraction, machine translation and information retrieval. Based on the content chunk parser as described in this thesis, a robust full parser that can process unrestricted Chinese text with high performance can be expected.

Keywords: computational linguistics; natural language processing, shallow parsing; content chunking; information extraction

目 录

图表目录

第一章　引　言

1.1 课题的提出

自然语言文本是一个线性的字符串，人要理解其中的含义，得到一段文本所传达的信息，需要得到以下两方面的信息：

（1）字符串中的语言成分。语言成分的结构呈层级构造，从最小的语言单位到最大的语言单位逐层构造而成。最小的语言单位是语素，一个或多个语素构成词，词和词组合构成短语，短语之上是句子和篇章。

（2）语言成分之间的关系。当两个或两个以上的语言成分组合产生更大的语言成分时，这些成分之间就包含一定的关系，包括句法关系和语义关系。

人要理解自然语言，必须获得这些语言知识。对于一个理解自然语言的计算机系统来说，同样需要这些知识。句法分析任务的目标就是获得句子的结构，即获得句子中的语言成分信息。由于自然语言极其复杂，对非受限的自然语言文本进行自动句法分析目前仍是自然语言处理所面临的一个巨大挑战，即使对于英语这样得到充分研究的语言至今也还没有一个可以处理非受限文本的高性能的句法分析器。解决句法分析难题的途径之一是采取“分而治之”的策略，即将复杂的句法分析任务分解为若干相互独立的子任务。浅层分析句法分析就是在这个背景下产生的一种新的语言处理策略。

浅层句法分析(shallow parsing)，也叫部分句法分析(partial parsing)或语块分析(chunk parsing)，是近年来自然语言处理领域出现的一种新的语言处理策略。它是与完全句法分析相对的，完全句法分析要求通过一系列分析过程，最终得到句子完整的句法树。而浅层句法分析则不要求得到完整的句法分析树，它只要求识别其中的某些结构相对简单的成分，如非嵌套(non-recursive)的名词短语、动词短语等。这些识别出来的结构通常被称作语块(chunk)，语块和短语这两个概念通常可以换用。

浅层句法分析的结果并不是一棵完整的句法树，但各个语块是完整句法树的一个子图(subgraph)，只要加上语块之间的依附关系(attachment)，就可以构成完整的句法树。所以浅层句法分析将句法分析分解为两个子任务：（1）语块的识别和分析；（2）语块之间的依附关系分析。浅层句法分析的主要任务是语块的识别和分析。这样就使句法分析的任务在某种程度上得到简化，同时也利于句法分析技术在大规模真实文本处理系统中迅速得到应用。

本文提出的实语块分析就是根据这种思想而提出的一种浅层句法分析任务，其目标是从文本中连续的实词串中分析出可能的结构。本文研究的目标是实现一个完整的汉语实语块分析系统，该系统接受非受限的汉语文本作为输入，输出文本中可能的实语块。在进行实语块分析之前，需要对文本进行分词和词性标注，并进行命名实体的识别。具体来说，系统具有以下功能：

（1）　对文本进行词语切分和词性标注；

（2）　实现对文本中命名实体的识别；

（3）　实现对文本中的实语块的识别与分析。

本文的重点在于汉语实语块的识别与分析，下面将给出实语块分析的定义，并说明实语块分析的理论意义和应用价值。

1.2 什么是实语块分析

1.2.1 实语块分析的定义

实语块（content chunk）是由实词序列组成的短语。实词指表示实体概念的词，包括名词、动词（助动词和系动词除外）、形容词、区别词、状态词、时间词、处所词、实义副词。由于实语块是由实词构成的短语，所以，“实语块”和“短语”这两个概念在不发生误解的情况下可以混用，如我们把实语块按功能类型分为名词性短语、动词性短语等，而不说名词性实语块、动词性实语块等，这都是为了称说的方便。

从实词序列和实语块的对应来看，可以有三种情况：

（1）一个实词序列对应于一个完整的实语块。例如：

例 1: [[铁路/ng 建设/vg]NP [很/dd 重要/a]AP]S 。/wd

在上面这个句子中，四个词都是实词，构成了一个实词序列，这个序列构成一个主谓结构，其中，主语是名词短语“铁路/ng 建设/vg”，谓语是形容词性短语“很/dd 重要/a”。

（2）一个实词序列不构成实语块。例如：

例 2: 中国/nps 的/usd 铁路/ng 建设/vg 得/usf 不错/a 。/wd

在这个例子中，“铁路/ng 建设/vg”构成一个实词序列，但这个实词序列在句子中并不构成实语块。

（3）一个实词序列中的一部分对应于一个实语块。例如：

例3: 我国/ng 的/usd [铁路/ng 建设/vg]NP 发展/vg 得/usf [很/dd 快/a]AP 。/wd

在这个例子中，第一个实词序列是“铁路/ng 建设/vg 发展/vg”，其中“铁路/ng 建设/vg”构成一个名词短语，“发展”不是任一实语块的一部分。

实语块分析的目标是对句子中任意一个实词序列进行分析，发现其中哪些构成合法的短语，哪些不构成短语。具体来说，实语块分析的任务是：

（1） 确定实语块的边界；

（2） 确定实语块的类型；

（3） 如果实语块 A 中的一个成分 B 是实语块，则要确定 B 的边界和类型。

如果一个句子中包含的全部是实词，那么实语块分析等价于完全分析，如对于下面的输入：

例 4: 前景公司推出高级电脑排版系统

系统将给出如下的输出[1]：

[[前景/nt 公司/ng]NT [推出/vg [[高级/b [[电脑/ng 排版/vg]NP 系统]NP]NP]VP]S

在分析结果中，“前景 公司”是一个机构名短语，“推出 高级 电脑 排版 系统”是一个动词性短语，其中动词的宾语“高级 电脑 排版 系统”是一个名词短语，在这个名词短语中又嵌套了名词短语“电脑 排版 系统”，其中又嵌套了名词短语“排版 系统”。

如果句子中有虚词，则以虚词为分界符，只分析连续的实词构成的结构。如对于下面的输入：

例5: 关于李正海感人事迹的报告会，正在广西各地举行。

系统将给出如下的输出：

[1] 词类和短语类型标记的含义见附录二。

关于/pg [李正海/npc [感人/a 事迹/ng]NP]NP 的/usd 报告会/ng ,/wd 正/dr 在/pg [广西/nps 各地/s]NP 举行/vg 。/。

在这个分析结果中，给出了实语块“[李正海/npc [感人/a 事迹/ng]NP]NP”的边界和内部层次划分，介词结构“[关于/pg [李正海/npc [感人/a 事迹/ng]NP]NP]”和“的”字结构“[[关于/pg [李正海/npc [感人/a 事迹/ng]NP]NP]PP 的/usd]”以及包含“的”字结构的 NP“[[[关于/pg [李正海/npc [感人/a 事迹/ng]NP]NP]PP 的/usd]DEP 报告会/ng]NP”都不在实语块分析的范围之内。同样，给出了实语块“[广西/nps 各地/s]NP”的边界和类型，但没有给出介词结构“[在/pg [广西/nps 各地/s]NP]”以及包含介词结构的 VP“[[在/pg [广西/nps 各地/s]NP]PP 举行/vg]VP”。

尽管在特定条件下（句中不包含虚词）实语块分析等价于完全句法分析，但它在本质上还是一种浅层分析。基本名词短语（baseNP）分析也是一种浅层句法分析任务，近年在浅层句法分析领域受到了普遍的关注。下面将对实语块分析和基本名词短语分析作一简单的对比，这样能够帮助我们加深对实语块分析的认识。

1.2.2 实语块与基本名词短语

基本名词短语的概念最早由 Church(1988)提出，其实他在文章中并没有给出严格的定义，只是把这种名词短语称为“简单的非递归的名词短语”（simple non-recursive noun phrase），也没有使用 baseNP 这个术语。Ramshaw & Marcus(1995)在评析 Church 的论述中对这一概念给出了比较明确的解释，并且首次使用了 baseNP 这一术语。他们对 baseNP 分析任务的定义是“识别一个非嵌套的名词短语的开始部分直到中心词为止，包括限定词但不包括中心词后的介词短语和小句”。他们从 Penn Treebank 中抽取 baseNP 的原则是“选择内部不包含另一个 NP 的 NP”，至于 NP 中是否包含并列连词，则取决于树库中的标注，如果树库中的一个非嵌套 NP 中包含了并列连词，则 baseNP 就包含并列连词，这一点跟 Church(1988)不同，后者的非嵌套 NP 中不包含并列连词。下面是 Ramshaw & Marcus(1995)所举的一个例子：

[N The government N] has [N other agencies and instruments N] for pursuing [N these other objectives N].

赵军（1998）首次研究了汉语的基本名词短语的识别与分析，他认为 Church 定义的 baseNP 过于简单，如“自然 语言 处理”、“亚洲 金融 危机”都不是非嵌套的名词短语，所以他给出了一个包含递归的汉语基本名词短语的定义。他对基本名词短语的定义是：

baseNP → baseNP + baseNP
baseNP → baseNP + 名词 | 名动词
baseNP → 限定性定语 + baseNP
baseNP → 限定性定语 + 名词 | 名动词
限定性定语 → 形容词 | 区别词 | 动词 | 名词 | 处所词 | 西文字串
| （数词 + 量词）

从这个定义可以看出，这里定义的基本名词短语的实质是：（1）以名词或名动词结束；（2）前面的定语中不含虚词。这里定义的基本名词短语可以包含复杂的嵌套关系，只要中间不包含虚词就可以。从识别的角度看，实质上是要在一个以名词或名动词结束的实词串中

识别一个名词短语。

由上面的分析可以看出，这里定义的基本名词短语基本上属于实语块的一部分（我们定义的实词不包含数量词和西文字串[2]），但实语块指称的范围要比基本名词短语广得多，表现在：

（1） 实语块中不仅包括名词短语，而且包括动词短语、形容词短语、主谓短语等。

（2） 实语块中包含了一些基本名词短语的定义不能涵盖的名词短语，如：

（A）VP 直接作定语：

（a） VP 是状中结构。如：

违法扣车现象　　正在建房户　　依法纳税意识

今年到期国债　　合资建厂事宜　　建筑用砖

（b） VP 是述宾结构。如：

含绒量　　建房区域　　反暴利法　　养鱼专业户

无党派人士　　操纵股市意图　　建设有中国特色社会主义理论

（B）VP 直接作中心语。如：

权力再分配　　家禽优化饲养　　国债流通转让　　农产品储运加工

1.3 实语块分析的意义

为什么要提出实语块分析的任务呢？下面分别从理论和应用两个方面来讨论实语块分析的意义。

1.3.1 实语块分析的理论价值

从完全句法分析中分化出实语块分析子任务的理论意义至少有以下两个方面：

（1）**实语块分析可以减少含虚词的句法结构的分析歧义**。

由于很多虚词的远距离依赖现象比较严重，在汉语句法分析中，包含虚词的结构分析往往遇到很大困难。突出的难点有：

（A）介词短语的右边界难以确定；

（B）“的”字结构的左边界难以确定；

（C）并列结构的左边界和右边界都难以确定。

下面以一个例子来说明这些困难。

加强/vg 对/pg 企业/ng 管理/vg 人员/ng 和/c 普通/a 职工/ng 的/usd 培训/vg 和/c 教育/vg

在上面的例子中，介词“对”的宾语很难确定，可能的介词结构有：

[a1] [对/pg 企业/ng]PP

[a2] [对/pg [企业/ng 管理/vg]NP]PP

[a3] [对/pg [企业/ng 管理/vg]S]PP

[a4] [对/pg [[企业/ng 管理/vg]NP 人员/ng]NP]PP

[a5] [对/pg [企业/ng [管理/vg 人员/ng]VP]S]PP

[2] 其实，名词短语中的西文字串不能简单地理解为一个西文字串，如“ABB 公司”中的“ABB”应分析为一个公司名，是一个专有名词。“BP 机”之类的音译词则应整体视为一个复合词。

[a6] [对/pg [企业/ng [管理/vg 人员/ng]NP]NP]PP

……

可以说，从“企业”开始的所有短语都可能成为介词“对”的宾语，所以，可能的介词结构的数目随介词后词的数量的增加呈指数级增长。

另一方面，连词“和”两边的边界也很难确定，可能的并列结构有：

[b1] [人员/ng 和/c [普通/a 职工/ng]NP]NP

[b2] [[管理/vg 人员/ng]NP 和/c [普通/a 职工/ng]NP]NP

[b3] [[[企业/ng 管理/vg]NP 人员/ng]NP 和/c [普通/a 职工/ng]NP]NP

[b4] [人员/ng 和/c [[普通/a 职工/ng] 的/usd 培训/vg]NP]NP

[b5] [人员/ng 和/c [[普通/a 职工/ng 的/usd 培训/vg]NP 和/c 教育/vg]NP]NP

……

可以看出，即使我们限定并列结构两边必须是同一类型的短语，因为两边可能的名词短语数量都很大，所以，可能的并列结构的数量也是随两边词的数量的增加呈指数级增长。

类似地，结构助词“的”的左边界也很难确定，如上例中可能的“的”字结构有：

[c1] [职工/ng 的/usd]

[c2] [[普通/a 职工/ng]NP 的/usd]

[c3] [[人员/ng 和/c [普通/a 职工/ng]NP]NP 的/usd]

[c4] [[管理/vg [人员/ng 和/c [普通/a 职工/ng]NP]NP]VP 的/usd]

[c5] [[[管理/vg 人员/ng]NP 和/c [普通/a 职工/ng]NP]NP 的/usd]

……

同样，可能的“的”字结构的数量随其左边的词的数量的增加呈指数级增长。

由上面的这个例子我们可以看到，“的”字结构、介词结构和并列结构的分析的复杂度和消歧的难度都非常大。

由于这些虚词都属于最常用的词，所以这些包含虚词的句法结构的分析困难给完全句法分析带来了极大的挑战。实语块分析把虚词作为分界符，不分析跟虚词相关的句法结构，在一定程度上避免了这些困难，因而可以得到较高的分析性能和效率。当实语块分析完成之后，这些确定了边界和内部层次的实语块就像建筑中的“预制件”一样，它的内部不需打开，这样，含带虚词的句法结构的歧义就大大地减少了，因而分析的复杂度也就大大地降低了。如对于上面的例句，实语块分析可以提供下面的结果：

加强/vg 对/pg [企业/ng [管理/vg 人员/ng]NP]NP 和/c [普通/a 职工/ng]NP 的/usd 培训/vg 和/c 教育/vg

经过实语块分析之后，我们得到了三个名词短语。这样，介词“对”的宾语的歧义就大大减少了，如上面的[a1]-[a5]都不会产生。因为连词“和”两边的实语块分析出来，可能的并列结构数量也大大减少了，如上面的[b1]-[b5]都不会产生。同样，“的”字左边可能的成分也大大减少，因而“的”字结构的歧义也就减少了，如上面的[c1]、[c3]-[c5]将不会产生。

（2）实语块分析可以简化句子的结构。

由于语言结构的递归性，一部分语言结构的整体和部分在功能上具有一致性。结构主义语言学提出语言的句法结构有两种主要类型（Bloomfield 1933）：

（1） 向心(endocentric)结构，整体部分的功能一致；

（2） 离心(exocentric)结构，整体和部分的功能不一致。

向心结构中有一个词（对偏正结构或述宾结构）或几个词（并列结构中的每个词）与整

体的功能一致。这个词就是短语的中心语。利用向心结构的这一特性，我们可以用中心语来代替一个复杂的向心结构，这样就使句子的结构得到简化。例如：

输入句:　　全媒体全数字彩电日前由海尔集团研制成功

经过切分和标注变为：

全/b 媒体/ng 全/b 数字/ng 彩电/ng 日前/d 由/pg 海尔/npu 集团/ng 研制/vg 成功/a

句子的长度为 11 个词。经过实语块分析后变为：

[[[全/b 媒体/ng]B [全/b 数字/ng]B]B 彩电/ng]NP 日前/d 由/pg [海尔/npu 集团/n]NP [研制/vg 成功/a]VP

用中心语代替实语块之后，原句子紧缩为：

彩电/ng 日前/d 由/pg 集团/ng 研制/vg

句子的长度由 11 个词减少到 5 个词，紧缩之后的句子显然比原来的句子容易分析多了。当简化后的句子分析完之后，把各个实语块的分析树插入各个中心语所在的结点就可以得到原句子的完整的句法树。 如图 1-1 所示。

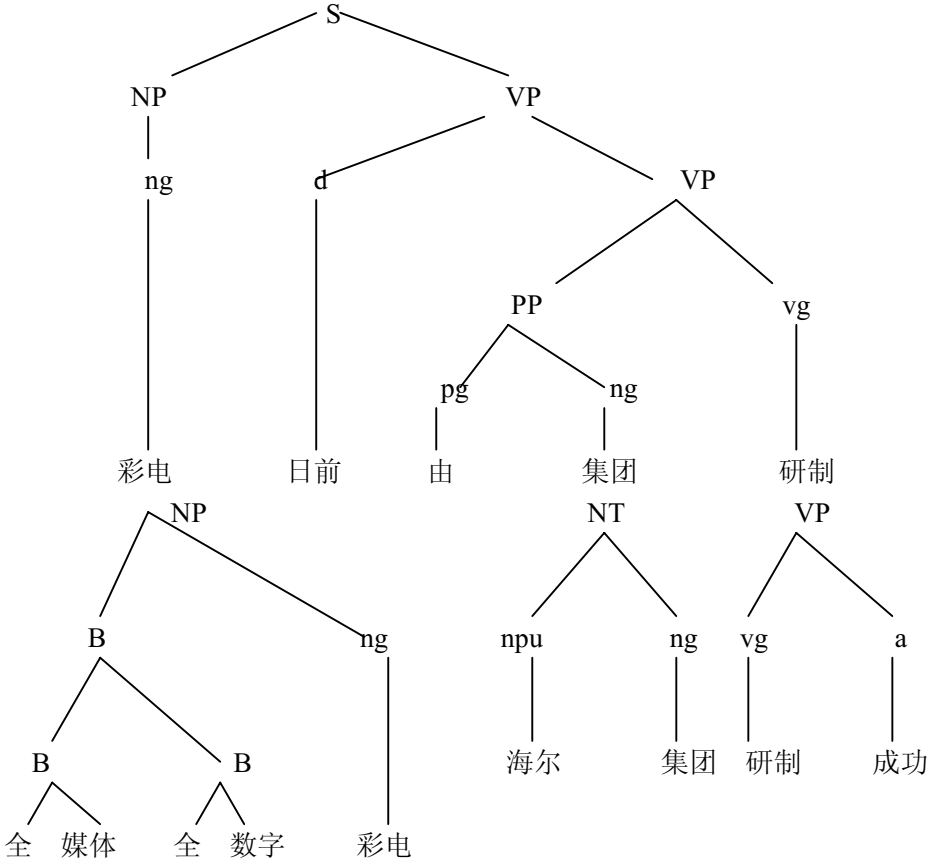


图 1-1　　实语块分析对句子结构的简化例示

1.3.2 实语块分析的应用价值

实语块分析可以应用在自然语言处理的各个方面，如信息检索、信息提取和知识获取等。下面分别说明。

1.3.2.1 信息检索

信息检索（Information Retrieval, IR）的任务是从一个文本集合中选择出与用户的查询需求相关的文本子集。评价信息检索系统的指标一般有两个：准确率(precision)和召回率(recall)，准确率衡量检出文档的相关程度，召回率衡量相关文档的检出率。在信息量激增的今天，在许多情况下（如在 Internet 上的信息检索）对准确率的要求高于对召回率的要求，而提高准确率的一条重要途径就是以短语为标引词，因为单个词的歧义程度高，难以准确地描述文本的内容。根据 Evens & Zhai(1996)的实验报告，加上短的名词短语作为标引词可以提高 IR 系统的召回率和准确率，而准确率的提高更为明显。另外，短语识别还有助于解决词义消歧问题，有利于标引词的规范化。

1.3.2.2 信息提取

信息提取(Information Extraction, IE)的任务是从自然语言文本中抽取出预先规定的信息，结果通常以关系数据库的形式贮存。现在的信息提取系统大都采用基于模板匹配的技术。对于信息提取来说，在实语块上可以进行有效的模式匹配。例如，抽取的目标信息是“产品发布”，一个最简单的模式是：

［公司］ ［发布］ ［产品］

对于输入： 前景公司推出高级电脑排版系统
实语块分析将给出如下的输出[3]：

[[前景/nt 公司/ng]NT [推出/vg [[高级/b [[电脑/ng 排版/vg]NP 系统]NP]NP]VP]S

有了这样的分析结果，自然可以得到正确的匹配。

再如，一个关于职务变动的信息提取模式是：

［人］ + “被” + ［机构］ + ［APPOINT ］ + ［职务］

在这个模式中，有五个元素，[人]、[机构]、[职务]都是语义范畴，一个词或短语的语义范畴与之相符，则匹配成功。介词“被”是被动句的标志，[APPOINT]是动词的语义范畴，它表示“任命”义，如“分配、提拔、任命、委任、推荐、选举、推举”等。对于下面的输入句：

１９９５年６月，杨丽华被中国国际航空公司任命为飞行总队副总队长。

[3] 词类和短语类型标记的含义见附录二。

实语块分析可以给出下面的分析结果，这样就可以顺利地实现匹配，达到信息提取的目的。

[１９９５年/t　６月/t]T　，/wd　杨丽华/npc　被/pbe　[中国/nps　国际/ng　航空/ng　公司/ng]NT　[[任命/vg　为/vg]V　[[飞行/vg　总队/ng]NP　[副/b　总队长/ng]NP]NP]VP

其中，关键的步骤是确定机构名“[中国/ns　国际/ng　航空/ng　公司/ng]”和实语块“[[飞行/vn　总队/n]NP　[副/b　总队长/n]NP]NP”。

1.3.2.3 语言知识获取

自然语言处理是基于知识的工程。过去，自然语言处理系统中所用的知识主要来自语言学家的内省。最近 10 年盛行的语料库语言学方法则强调从大规模语料库中获取语言知识。现在，文本语料在数量上几乎没有限制，但如果这些文本不经过句法分析，从中能够获取的有意义的语言知识是十分有限的。实语块中包含了实词组合的许多有价值的信息，在实语块的分析结果中可以获得实词之间的的搭配知识，进一步还可以获取谓词的子范畴、论元结构和语义选择限制等语言学知识。

词语搭配是关于词汇功能的重要信息，词语搭配知识在自然语言处理系统中具有十分重要的作用。过去主要利用词汇统计的方法来从大量文本中抽取词语搭配知识，由于没有引入结构信息，因而系统的准确率和召回率都不高（Smadja, 1993; 孙茂松等，1997;孙宏林，1998）。在实语块分析的基础上，就可以充分利用词语之间的结构信息，使词语搭配自动抽取的性能大大提高。

1.3 本文的贡献

本文的主要贡献有以下几点：

（7） 提出了汉语实语块分析任务，并通过实验证明实语块分析是可以明确定义、相对独立的句法分析子任务。

（8） 提出了概率上下文无关语法和概率属性相结合的汉语实语块分析模型。该模型利用跟规则相关的句法属性对上下文无关语法进行约束。由于这些句法属性是可学习的，因而使系统更具有健壮性。由于句法属性是带概率的，因而使系统具有更强的消歧能力。

（9） 根据汉语句法中节律对句法具有约束作用的性质提出了利用音节和长度信息对规则进行约束的思想，并使节律特征概率化。

（10） 提出了汉语并列结构的概率模型，该模型利用并列结构的对称性在若干并列项候选中选择正确的并列项。

（11） 提出了对词类标记集复杂度的定量分析方法，并在此基础上提出了基于遗传算法的词类标记集优化方法。

（12） 实现了一个完整的汉语实语块分析系统，该系统以非受限的汉语文本作为输入，以实语块分析结果作为输出。除了得到实语块之外，还可以得到分词、词性标注和命名实体识别的信息。

1.4 本文的组织

本文共分八章。

第一章是引言。本章给出了实语块分析的定义，并说明了实语块分析的理论意义和应用价值，概述了本文的主要贡献和全文的组织。

第二章是词类标记集的评价和优化。词性标注是句法分析的第一步，词类标记集的确定对句法分析的影响很大。本章研究了词类标记集的定量分析问题，提出了一些评价标记集复杂度的指标，并在此基础上提出了一种基于遗传算法的词类标记集优选方法，该方法可以辅助人们选择一个适合于特定任务的词类标记集。

第三章是汉语命名实体识别。本章介绍了基于词类概率模型的汉语分析和词性标注一体化的处理模型，并详细描述了专名候选词的生成算法，这些专名词语包括汉人姓名、西文译名、中国地名。

第四章是浅层句法分析方法概述。本章概述了近年来在浅层句法分析方面的研究，着重介绍了英语浅层句法分析的有关方法，同时也介绍了汉语浅层分析方面的概况。

第五章是汉语实语块分析规则与算法。本章介绍了实语块分析中的规则、实语块的语料标注和规则统计以及实语块分析的算法。

第六章是概率上下文无关语法和概率属性相结合的汉语实语块分析模型。本章提出了概率上下文语法和概率属性相结合的汉语实语块分析模型，描述了节律、上下文和词汇功能三种属性的概率估计以及这些概率属性对于句法分析的作用，然后给出了实验结果。

第七章是并列结构的概率模型。本章根据并列结构的对称性提出了并列结构的概率模型，该模型可以用来在一组并列候选项中选择正确的并列项。实验表明，利用并列结构的概率模型可以显著提高实语块分析的准确率。

第八章是总结和展望。对全文进行简单的总结，并提出下一步研究的方向。

第二章 词类标记集的评价和优化

2.1 引言

任何自然语言的词汇都是一个很大的集合，一部中等规模的词典一般也有几万个词项。这使得直接用词来描述语言的结构和建立语言模型十分困难，甚至难以实现。因而有必要对词汇进行抽象，即对词汇加以分类。分类的依据是词在句法、语义等属性上具有的共性。至于到底归为多少类，并无一定之规，因为根据抽象程度的不同可以有各种不同的结果。在过去十多年中，词性自动标注技术取得了很大的进展，统计方法的运用使得构造一个高性能的自动词性标注系统变得相对容易，现在世界主要语言几乎都有不少这样的标注系统可供使用。但是，所有的标注系统所使用的词类系统都是依靠人的知觉或经验产生的（Halteren 1999）。由于对词类体系的合理性或合适性缺乏客观的评价，我们在选择或定义一个词类标记集时就缺乏科学依据。同时，由于不同的应用系统对词性知识有不同的需求，对于特定任务到底选择什么样的词类体系也是一个问题。这些问题都属于词类标记集的评价和选择问题。

对词类体系可以从两个方面进行评价。首先，从语言学的角度看，我们可以分析对词类体系的科学性和合理性进行分析。其次，从自动标注的角度，我们可以分析一个词类体系的复杂度。评价一个词类标记集的复杂度不能简单地以标记的数量来衡量，最重要的的是它所造成的文本中词的歧义程度，歧义程度越高，标注系统消解歧义的任务产生就越重。本章首先讨论词类标记集的评价问题，然后提出一种基于遗传算法的词类标记集优选方法。

2.2 词类标记集的评价

本节首先给出标记集的几个评价指标，然后根据标注实验的结果，分析这些评级指标与标注准确率的关系。

2.2.1 三个标记集实例

为了比较不同标记集的歧义程度，我们在实验中选择了三个标记集：

（1） TS1，包含 85 个标记（不包括标点符号，下同）；

（2） TS2，包含 20 个标记；

（3） TS3，包含 15 个标记。

TS1 是我们在“现代汉语研究语料库”中所采用的标记集（孙宏林等，1996），TS2 是北大计算语言学研究所的“现代汉语语法词典”中所用的标记集（俞士汶等，1996，1998），TS3 是对外汉语教学中普遍采用的一个词类体系（胡明扬，1996）。

这三个标记集互相之间具有直接映射关系。TS2 是 TS1 的超集，即从 TS1 到 TS2 有多对一的映射关系。例如，TS1 中所有名词的子类对应于 TS2 中的一个类——名词，TS1 中所有动词的子类对应于 TS2 中的一个类——动词（见附录 2-1，2-2）。同样地，TS3 是 TS2 的

超集，同时也是 TS1 的超集。

在这三个标记集中，TS2 和 TS3 的差别较小。区别仅在于以下两点：

（1） TS2 中的时间词、处所词和方位词在 TS3 中归入名词。

（2） TS2 中的区别词和状态词在 TS3 中归入形容词。

而 TS1 则与这两个标记集相差较大。主要反映在，TS1 除了在词的分类上比 TS2 更细之外，还标注了某些词（主要是动词和形容词）在具体语境中的句法功能。例如：

公司 上市 的 检查 过程 极为 严格/a （一般形容词）
坚决 执行 财经 法规，严格/ab 财经 纪律 （形容词带宾语）
严格/ad 控制 菜田 面积 的 占用 （形容词直接作状语）
对下级要求之严格/ax 有时到了不近人情的地步 （形容词直接作 NP 中心语）

“严格”在 TS2 和 TS3 中只有一种标记，即一般形容词，在 TS1 中则带上了四种标记。再如，对动词则根据其句法功能及带宾语的不同加上不同的标记，如：

对 此 必须 认真 研究/vg （动词不带宾语）
研究/vgn 与 试点 有关 的 政策 措施 （动词带名词性宾语）
研究/vgv 深化 企业 改革 （动词带动词性宾语）
经过 研究/vgb 他们 发现 （动词作宾语）
研究/vgp 成果 返还 企业 （动词作定语）
拓展 外国 哲学 研究/vgx 的 新 领域 （动词作 NP 中心语）

在 TS2 和 TS3 中带单一标记的“研究”在 TS1 中带上了六种标记。

显然，TS1 的某些标记反映的并不光是词的静态分类，而且还带上了动态的句法功能信息。这种标记所带的信息量很大，但同时也带来歧义的增加。

2.2.2 实验语料

我们选择了“现代汉语研究语料库”的一个子集作为实验语料。该子集包含从《人民日报》中选择的 157 篇经济类文章，共有 203,499 个词例。语料用 TS1 进行了标注，并经过人工校对。因为在上述三个词类体系中存在多对一的映射关系，所以该语料库可以自动转化为用 TS2 和 TS3 标注的语料库。

2.2.3 标记集复杂度的评价指标

在一个分类体系中，一个词可能只属于一个类，也可能属于多个类。一个类使用一个标记，当一个词属于一个类时，这个标记就是这个词的可能标记之一。我们用一个词 W_i 可能具有的标记数（记为 $AMB(W_i)$）。

我们可以从静态和动态两个方面考察词的歧义程度。从静态的角度我们可以看词典中每个词型（word type）所具有的可能标记数。我们把标记集 TS 的静态歧义指数定义为词典中每个词型具有的可能标记数的均值（记为 $SAI(TS)$）。静态歧义指数可以用公式 2-1 来计算：

$$SAI(TS) = \frac{\sum_{i}^{N} AMB(W_i)}{N} \qquad (2\text{-}1)$$

这里，N 是词典中的词型总数。

从动态的角度，我们可以考察在真实文本中每个词例所具有的可能标记数。我们把标记集 TS 的动态歧义指数定义为文本中每个词例所具有的可能标记数的均值(记为 *DAI(TS)*)。动态歧义指数可以用公式 2-2 来计算：

$$DAI(TS) = \frac{\sum_{i=1}^{N} AMB(W_i) \times f(W_i)}{\sum_{i=1}^{N} f(W_i)} \qquad (2\text{-}2)$$

这里，N 是词典中的词型总数。$f(W_i)$是词 W_i 在语料库中的出现频次。

上述三个标记集的静态歧义指数和动态歧义指数在表 2-1 中给出。

表 2-1　三个标记集的歧义指数

	TS1	TS2	TS3
SAI()	1.3310	1.0705	1.0671
DAI()	2.4046	1.5152	1.4913

从表 2-1 我们可以看到，标记集的动态歧义指数与静态歧义指数成正比，从 TS1 到 TS2 到 TS3，标记集越来越小，歧义程度也越来越小。但相比之下，动态歧义指数的差异程度要大于静态歧义指数。TS1 的静态歧义指数比 TS2 高 24%，而动态歧义指数则相差 59%。这是因为常用词在文本中的出现次数要高于非常用词，而越常用的词其歧义程度越高。

我们还可以用歧义词占总数的比例来衡量标记集的复杂度。这也可以从动态和静态两个角度来考察。我们把标记集 TS 的静态歧义率定义为词典中歧义词型数占总数的比例(记为 SAR(TS))，即：

$$SAR(TS) = \frac{\#\ of\ ambiguous\ word\ types}{\#\ of\ word\ types} \qquad (2\text{-}3)$$

我们把标记集 TS 的动态歧义率定义为文本中歧义词例数占总数的比例(记为 *DAR(TS)*)，即：

$$DAR(TS) = \frac{\#\ of\ ambiguous\ word\ tokens}{\#\ of\ word\ tokens} \qquad (2\text{-}4)$$

上述三个标记集的静态歧义率和动态歧义率由表 2-2 给出。

表 2-2　三个标记集的歧义率

	TS1	TS2	TS3
SAR (%)	18.5	6	5.8
DAR(%)	58.8	37	36.6

歧义率更直观地显示出词典和文本中的歧义程度。从静态的角度看，三个标记集的歧义度都不太高。根据三个标记集，词典中歧义词型所占的比例分别是 18.5%, 6% 和 5.8%。相比之下，TS2 和 TS3 相差不大，而 TS1 则与它们相差较大，这主要是因为在 TS1 中动词和形容词等标注了其在句子中的功能信息。但从动态的角度看，歧义词例所占的比重则要高得多。

即使是标记数量很小的TS2和TS3，歧义率也达到了近40%。而TS1的歧义率则达到了近60%。动态歧义率是静态歧义率的3.2至6.3倍。这也是因为常用词虽然在词典中所占的比例不大，但由于使用频率高，因而在文本词例中的比例要比在词典中的比例大得多。这和Zipf定律所反映的文本中的词汇分布规律是一致的(Zipf,1935,1949)。

根据Brown语料库的统计，其静态歧义率和动态歧义率分别是11.5%和40%（DeRose,1988)。考虑到Brown语料库所用的是一个大标记集（没有动态功能标记），英语和汉语的词类歧义情况是大体一致的。

2.2.4 标注实验

为了评价标记集对标注系统性能的影响，我们进行了一些实验。利用相同的标注算法分别采用三个标记集进行训练和标注。在实验中，我们把语料库分成两个部分：

（1） 训练集。 包括137篇文章，183，050词次。

（2） 测试集。 包括 20 篇文章， 20，449词次。

我们分别采用了两种标注算法：

（1） 算法1:对每个词选择出现频率最高的标记,即选择最可能的标记c,使得P(c|w)最大化;

（2） 算法2：使用基于HMM的统计模型，结合标记的转移概率$P(t_i|t_{i-1})$和词的生成概率$P(w_i|t_i)$，并利用动态规划算法选择最大概率路径（Church 1988; DeRose 1988)。

表2-3 三个标记集的自动标注错误率

	TS1	TS2	TS3
算法1(%)	12.74	4.50	4.42
算法2(%)	7.39	3.40	3.42

表2-3给出了标注实验的结果（不包括对未登录词的统计)。从标注的结果我们可以看到：标记集所产生的歧义程度的高低与标记的错误率成正比关系。以上描述的标记集的评价指标都与自动标注的准确率有直接的关联。但相比之下，用动态歧义指数或动态歧义率作为标记集复杂度的指标更为合理。表2-4显示了对TS1和TS2的几种复杂度指标与标注错误率的对比，表中每一列表示两个标记集的有关指标的比值，可以看到，DAI的比值(1.59)与错误率的比值(2.17)最为接近。而DAI和DAR的比值基本相同，说明这两个指标具有等价性，动态歧义指数（DAI）是衡量文本中每一个词的平均歧义程度，而动态歧义率则是衡量文本中歧义词占总词次的比例，这两个指标从不同角度反映了标记集带给标注系统的复杂度。应该说，用这两个指标来衡量标记集的复杂度都是合理的。

表2-4 评价指标与错误率的对比

SAI(TS1)/SAI(TS2)	DAI(TS1)/DAI(TS2)	SAR(TS1)/SAR(TS2)	DAR(TS1)/DAR(TS2)	ERR(TS1)/ERR(TS2)
1.24	1.59	3.08	1.56	2.17

从标注实验的结果我们还可以发现两个有意义的问题：

第一、词性标注用最简单的概率模型就可以得到不错的结果。算法1是最简单的概率模型，可以看作是自动词性标注的底线(baseline)。对于已登录词，在TS2和TS3中，利用这

一简单的算法就可以得到95%以上的准确率。我们的实验结果和有关研究的结果是基本一致的。[Charniak et al.1993]报告在Brown语料库上的实验，利用该算法可以得到91%的准确率，[Brill 1993]报告根据1000句的小规模实验，利用该算法可以达到93.6%的准确率，[白拴虎 1996]报告利用该算法在7万词次汉语语料上的封闭测试，标注准确率达到88.3%，该系统使用了一个大标记集，与TS1相近。该算法不考虑上下文影响，但可以看出，虽然算法2考虑了上下文因素，但性能提高的幅度很有限，而且，标记集越简单，提高的幅度越小。

第二、当引入上下文信息（算法 2 中的标记转移概率）之后，标记对上下文的敏感度（sensitivity）会对标注的准确率产生影响，因而导致算法2的结果与算法1的结果不尽一致。例如，在算法1中，TS2的错误率略高于TS3，但在算法2中，TS2的错误率略低于TS3。由于这两个标记集过于接近，因而差异太小。而TS1在两种算法下的结果则差别较大，上下文信息使标注的准确率提高了5个以上的百分点。标记对上下文的敏感度将在下一节讨论。

2.2.5 对单个标记的评价

与对整个标记集的评价相类似，我们可以对一个词类体系中某个类进行分析，以评价其对自动标注准确率的影响。对某个词类标记 T_i 可以从三个方面进行分析：

（1） 带 T_i 的词在词典和文本中的可能的标记数；

（2） 属于类 T_i 的词在词典和文本中的歧义率；

（3） T_i 对上下文的敏感度。

下面着重讨论标记对上下文的敏感度。

标记的敏感度是标记分布特征的一种反映。如果一个标记只在一个或几个环境中出现，那么我们则说这个标记对上下文的敏感程度高，相反，如果一个标记是随机分布的，我们则认为这个标记对上下文的敏感程度很低。举个例子来说，汉语的量词分布环境就很有限，它一般只跟在数词之后，所以当当前词是一个兼类词且其中有一个标记是量词时，如果前面是数词，那么当前词就很可能是量词而不是别的词。例如，汉语的“把”兼属介词、量词、动词，当前面出现介词时，基本上可以确定它量词而不是介词，尽管“把”作介词的概率要远大于作量词的概率。

标记的分布越集中，它对上下文的敏感度就越高，相反，如果标记的分布比较分散、均匀，则它对上下文的敏感度就不高，也就是说不容易从上下文推出它的标记来。

标记 T_i 的敏感度（记为 $SEN(T_i)$）定义如下：

$$SEN(T_i)=\frac{\sum\left|f(T_i)-\overline{f}\right|}{M\times\overline{f}} \qquad (2\text{-}5)$$

这里，M 表示包含 T_i 的标记对的总数，$\overline{f}$ 是 T_i 在每个包含 T_i 的标记对中出现次数的均值，$SEN(T_i)$实际上是由 T_i 的均方差除以一个规范化系数 $\overline{f}$ 得到。

在TS2中，敏感度值最高的标记是语气词（”y”），其分布见表2-5。

表 2-5　TS2 中语气词的分布情况

tag1	tag2	freq	tag1	tag2	Freq	tag1	tag2	freq	tag1	tag2	freq
y	v	2	n	y	65	z	y	1	u	y	5
y	m	1	t	y	1	m	y	1	y	y	2
y	d	1	f	y	3	q	y	5	k	y	2
y	y	2	v	y	240	r	y	12	I	y	8
y	x	438	a	y	90	d	y	1	x	y	10

从表 2-5 可以看到，语气词的分布集中于少数几个位置，从它的右边看，它总共出现了 444 次，其中 438 次右边都是跟一个标点符号。从左边看，它前面出现最多的是动词，其次是形容词和名词。

2.3 基于遗传算法的词类标记集的优化

2.3.1 引言

选择词类标记集可用的一个直接的简单方法是：对几个候选的词类标记集，用同一标注系统分别进行训练和标注实验，根据实验的结果从中选择一个最好的体系。但这种方法的缺点是：候选标记集只能是有限的若干个，而可能的标记集几乎是无限的。而且对每一个标记集都需要经过语料标注、训练、标注评测的过程

，其中都需要人的介入，工作量很大。

本文针对这一问题，提出了一种基于遗传算法的词类标记集优选方法，该方法可以在一个候选标记集集合中自动搜索一个最优或较优的标记集。该方法的基本思想是：首先定义一个包含一个最大标记集和最小标记集的词类层级体系，然后选择一个语料库，用最大标记集进行标注，然后对介于最大标记集和最小标记集之间的可能的标记集用一个评价函数进行度量，利用遗传算法从中选择最优或较优的标记集。标记集的评价函数综合考虑两个因素：（1）标记集的歧义度，用语料库中每个词例（token）的可能的标记数数的均值来表示；（2）标记集的信息量，用标记集的熵（entropy）来表示。这一方法的目的是帮助人们确定哪些标记应该包括在标记集中，标记应该排除在外。

2.3.2 遗传算法概述

遗传算法是根据自然进化中“适者生存”的原则提出的一种概率型优化方法（Holland 1975）。这种方法适合于具有很大搜索空间的优化问题，下面简要介绍遗传算法的主要思想，详细情况可参考（Goldberg 1989）和（Mitchell 1996）。

在遗传算法中，有一个包含个体 x_i 的群体 $P = < x_1, \ldots, x_n >$，个体 x_i 代表问题的一个解，群体就是问题的一些解的集合。某一评价函数 $F(x_i)$被用来对这些候选解进行评价，目标是优化该评价函数（搜索该函数的最大值或最小值）以解决给定的问题。这些候选解通常用位串（bit string）的形式表示，借用生物学的术语称之为染色体(chromosome)。把解表示表示为位串的过程称为编码，编码后的每个位串就表示一个个体，即问题的一个解。评价函数用以评价群体中每个个体的适应度(fitness)。在算法的每次迭代（借用

生物学术语称作一代）中，评价函数按照优化标准对每个个体 x_i 进行度量，计算其适应度 f_i，适应度最高的个体被选择允许再生，以产生新的一代。下面是算法的形式化描述：

```
{
    gen = 0;       //代号初始化
    initialize();   //初始化第一个群体
        evaluate(P(gen));   //P(gen)  是  第 gen 代的群体
    do {
        gen = gen+1;
        generate new population P(gen) from P(gen-1) through select, crossover, and
                mutation; //通过选择、交叉和变异从 P(gen-1)生成新的群体  P(gen)
        evaluate(P(gen));   //  评价 P(gen)
    } while (gen <= maxgen)
}
```

图 2-1　遗传算法概貌

遗传算法中的再生过程主要包括三个遗传算子：（1）选择；（2）交叉；（3）变异。在选择过程中，适应度高的个体被直接复制到下一代群体中。适应度越高的串，产生后代的概率就越高。在交叉过程中，两个串的部分位（称为基因）进行交换从而产生一个新串作为下一代的个体。变异用来随机地改变染色体的部分基因。交叉和变异的使用都有一定的概率，分别称为交叉概率和变异概率。

2.3.3 问题的表示

词汇的分类可以用图 2-2 所示的树型结构来表示。树中的一个结点对应于一个词类标记，树中若干个结点构成一个标记集。在遗传算法中，我们用一个位串来表示一个标记集。分类树中除根结点以外的所有结点可以用一个位来编码，每一位有两种可能的值：1 表示该结点在分类树中存在，0 表示该结点在分类树中不存在。如果最大结点数为 n，即最大标记集 M 中有 n 个标记，那么，n 个标记构成的所有可能的标记集构成 M 的幂集，其数量是 2^n。我们的目的是在一个如此大的空间中搜索一个最优或较优的标记集，即找到适应度最高的位串或染色体。

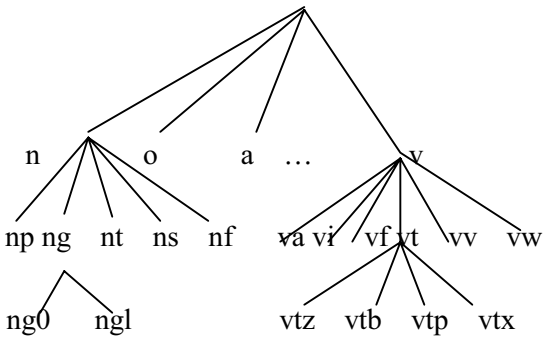


图 2-2 词分类树

由于分类树是树型结构，因此位串中各个位的的值并不是相互独立的。如果一个结点的值为 0，那么它的下位结点必为 0。所以，在样本产生过程中违反以下限制的染色体将被删

除：

如果任一子结点非空，则父结点不能为空。

例如，附录 2-1 中的标记集有 107 个标记，分为三个层级（根结点为空），如图 2-2 所示。在第一层上有 15 个结点，第二、三层上分别有 41 和 51 个结点。整个标记集可以用图 2-3 所示的位串来表示，该位串的长度为 107。

np	ng	…	il	npx	…	utg	n	…	i
0	1	…	40	41	…	91	92	…	106

图 2-3 标记集的表示

例如，在图 2-3 中，第 0 位是“np”(专有名词)，如果第 0 位的值为 0 则表示标记集中没有此标记，否则表示有此标记。

2.3.4 评价函数

对群体中的每一个个体， 要计算其适应度值作为在从前一代“进化”到下一代的过程中选择的标准。在标记集优选问题中，需要考虑两个参数：(1) 标记集对词性标注系统准确率的影响；(2) 标记集为下一步处理所提供的信息的丰富程度。词性标注追求的目标是：词性标记包含的信息越丰富越好，同时，自动标注的准确率越高越好。显然，这两个目标是有矛盾的：一方面，词类分得越粗，自动标注的准确率就越高，但词类标记提供的信息量就越少；另一方面，词类分得越细，给高一级处理提供的信息也就越丰富，但自动标注的准确率就越低。因此，我们需要在词类标记的信息量与自动标注的准确率之间找到一个最佳的平衡点。以下将分别讨论这两个参数，并在此基础上提出一个综合这两个参数的对词类标记集的评价函数。

2.3.4.1 标记集对标注准确率的影响

为了度量标记集对标注程序准确率的影响，我们可以使用直接的方法，即用不同的标记集分别进行语料的标注、标注系统的训练和标注评测。如前所述，这种方法的缺陷是：考察的标记集数量有限，而且费时费力。所以我们应该寻找间接的方法。根据我们在本章第一部分的讨论和实验结果，词类标记集的动态歧义指数（语料库中每个词可能的标记数，DAI）与标注系统的准确率成正比例关系。这一点也是比较容易理解的，因为词性标注系统的主要工作就是进行词性的消歧，语料中每个词例的歧义指数越高，那么标注系统面临的消歧困难也就越大，因而标注准确率也就越低。

在计算一个标记集的 *DAI* 之前，必须有用该标记集标注好的语料库。在遗传算法的进化过程中，需要对大量的标记集进行评价，不可能为每一个标记集准备一个标注版本。但是，如果语料库事先用最大标记集进行了标注，就可以利用最大标记集到其子集的多对一（有些标记是一对一）的映射关系自动得到用新标记集标注的语料库版本。

2.3.4.2 标记集的信息量

词的分类可粗可细。分类越粗，得到的类就越少，同时给后一步处理提供的信息也就越少。以名词的分类为例，我们分别考虑以下三种情形：

（1） 对名词不作进一步分类；

（2） 把名词分成两类：普通名词和专有名词；

（3） 在（2）的基础上，进一步把专有名词分成四个子类：人名、地名、机构名和其他专名。

对于命名实体任务（Chinchor 1998）来说，很显然以上第三种情形可以提供更多的信息，因而使命名实体任务变得容易得多。

但要想准确地估计一个标记集所包含的信息量却并非易事，一个标记所包含的信息在下一步的处理到底起什么作用，跟整个系统的理论体系有关，很难给出一致的度量。我们这里所说的信息量跟标记的具体属性含义无关，它只是一个一般性的定量的度量，而非定性的描述。

标记集所包含的标记数目可以作为一个度量标准，但它是静态的，不能反映每个标记对系统的实际贡献，因为不同的标记在文本中出现的概率不同。因此必须考虑标记的实际出现概率。从这个角度来说，熵(entropy)是对标记集信息量的一个合适的度量。

从信息论的观点看，文本中的词性标记序列可以看成一个随机过程，标记集是一个随机变量，标记序列中的特定标记就是这个随机变量的值。熵是随机变量平均不确定性的度量，不确定性越高，熵值就越大。标记的可能的取值越少，我们在猜测序列中标记的出现时得到的信息就越少；相反，标记可能的取值越多，那么我们得到的信息就越多。标记集 *TS* 的熵可以通过下面的公式来计算：

$$H(TS) = -\sum_{t \in TS} P(t) \log P(t) \qquad (2\text{-}6)$$

这里，*t* 是标记集 *TS* 中的一个标记，*P(t)*是 *t* 的出现概率，对数以 2 为底。

2.3.4.3 评价函数

在进化过程中，对每一个标记集都需要给出一个适应度的度量以指导再生过程。优化的过程就是搜索最大或最小评价值的过程。

正如在第一节中所讨论的，两个相关的参数（准确率和信息量）成反比关系。事实上，我们通常要根据特定任务的需求在这二者之间找到一个平衡点。不同的任务对词性标注有不同的要求，在某些情况下，准确率是优先考虑的因素，而在有些情况下标记的信息量则更重要。所以，应该给这两个参数赋以不同的权值，而且权值可以根据应用的要求进行调整。

我们把标记集优选的过程看成一个适应度值最大化的过程。DAI 和适应度值成反比例关系，熵和适应度值成正比例关系，所以，标记集 *TS* 的适应度可以定义如下：

$$Fitness(TS) = \frac{\lambda_1 \times (\frac{1}{D_T} - \frac{1}{D_{\max}})}{(\frac{1}{D_{\min}} - \frac{1}{D_{\max}})} + \frac{\lambda_2 \times (H_T - H_{\min})}{(H_{\max} - H_{\min})} \qquad (2\text{-}7)$$

这里，D_T 和 H_T 分别表示 *TS* 的 *DAI* 和熵，D_{max} 和 D_{min} 分别是最大标记集和最小标记集的 *DAI*，H_{max} 和 H_{min} 分别是最大标记集和最小标记集的熵值。这里，$D_{max} > D_{min}$ ，$H_{max} > H_{min}$.。λ_1 和 λ_2 是两个参数的权值，$\lambda_1 + \lambda_2 = 1$。当优先考虑标注系统的准确率时，可以给 λ_1 赋一个大于 0.5 的值，如果优先考虑标记集的信息量，则可以给 λ_2 赋一个大于 0.5 的值。当 $\lambda_1 = \lambda_2 = 0.5$ 时，两个参数的优先级相等。实际的取值可以在实验中调试。显然，λ_1 越大，结果的标记集就越小，λ_2 越大，结果的标记集就越大。

2.3.5 实验结果

2.3.5.1 实现

实验中所用的标记集是一个汉语语料库所用的标记集（孙宏林等，1996）。该分类系统是一个层级体系，分类树中共有 107 个结点，我们把这 107 个结点构成的标记集定义为最大标记集。其第一级包含 15 个标记，我们把它定义为最小标记集，这 15 个标记包含在进化过程中生成的任一标记集中。因此，共有 92 个变量。我们用长度为 92 的位串来表示这 92 个结点，变异率为 1/92，遵循这样的原则：染色体中的每一位有 1/L 的变异概率(L 为染色体中位的数量)（Bäck 1993）。交叉率为 0.8，交叉采用了简单的单点交叉。

为了评价每一个标记集，我们使用了以上语料库中的一个子集作为测试语料，其中包含 157 篇选自《人民日报》的文章，共有 203,499 个词例。这些语料用以上的最大标记集进行了标注，并经过细致的人工校对。因为最大标记集和生成的标记集之间有多对一的对应关系，该语料库可以自动地转化为用其他标记集标注的版本，所以在程序运行过程中不需要人的介入。

在评价函数中，选择 $\lambda_1 = \lambda_2 = 0.5$，给标注的准确率和标记集的信息量以相同的权值。表 2-6 给出了几个标记集的适应度，我们可以发现，在运用公式（3）中，D_{max} 和 D_{min} 的值分别是 2.42 和 1.49, H_{max} 和 H_{min} 的值分别是 4.54 和 2.83 。表 1 中的 Tagset 1 是最大标记集，它包含 107 个标记，在所有的候选标记集中 *DAI* 值和 *H* 值最大。对于最大标记集，因为 $D_T = D_{max}$ 且 $H_T = H_{max}$ ，所以其适应度值等于 λ_2。 Tagset 2 是最小标记集，它在所有候选标记集中 *DAI* 值和 *H* 值最小。因为 $D_T = D_{min}$ 且 $H_T = H_{min}$ ，所以其适应度值等于 λ_1。

表 2-6　部分标记集的适应度

Tagset	*DAI*	*H*	*Fitness(%)*	标记集说明
1	2.42	4.54	50	最大标记集，107 个标记
2	1.49	2.83	50	最小标记集，15 个标记
3	1.74	3.87	61.83	所有的第二集结点都不为空，所有的第三季结点皆为空
4	1.62	3.64	63.43	vt* and vw* 为空，其余结点皆不为空

2.3.5.2 实验结果

实验共生成了 100 代，表 2-7 给出了其中部分结果。在表 2 中，第一栏是代号，第二栏是群体中最小的适应度，第三栏是群体中最大的适应度，第四栏是群体中所有标记集适应度的均值，第五栏是已获得的最佳适应度，最后一栏是具有最佳适应度的标记集的位串表示(用 16 进制表示)。标记按照附录 2-1 中的序号编码，第 0 号标记对应于位串中的第 0 位，第 1 号标记对应于位串中的第 1 位，其余依此类推。附录 2-1 中序号为 92-106 的 15 个标记是分类树中的一级标记，如前所述，我们规定这些包括在生成的任一标记集中，所以不在位串中编码，因此位串的长度为 92。

图 2-4 显示了在前 50 代中适应度的变化情况，横轴上的 11 个点分别对应于代号 1, 5, 10, 15, … , 50。图中，横轴表示代号，纵轴表示适应度。上面的一条曲线表示最佳适应度的变

化情况，下面一条曲线表示平均适应度的变化情况。

从图 2-4 中我们可以发现，在进化过程中所生成的群体的最大适应度和平均适应度都随着代号的增加而单调增加，而且在第 23 代基本达到收敛。

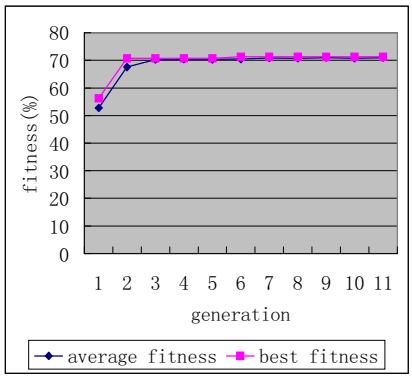


图 2-4　进化过程中适应度的变化情况

表 2-7　部分实验结果

代	最小适应度 (%)	最大适应度(%)	平均适应度 (%)	标记集的位串表示 (hex)
1	48.67	56.24	52.74	acfbc04b5f2881000574574
2	48.92	67.91	55.83	e5e3fa974ad7880005b90dc
6	53.44	70.67	67.52	e1e3ea974ad7000005790dc
16	53.44	70.67	70.21	e1e3ea974ad7000005790dc
23	54.40	71.22	70.28	b1e3ead75af4000001a9014
50	54.66	71.22	70.95	b1e3ead75af4000001a9014
100	54.66	71.22	70.89	b1e3ead75af4000001a9014

具有最高适应度的位串是 b1e3ead75af4000001a9014，在第 23 代就得到。与该位串对应的标记集所包括的标记在附录 2-1 中的序号后加了星号。该标记集包含 51 个标记，在第一、二、三级上分别有 15、25 和 11 个标记。下面以名词为例对得到的标记集进行简单的讨论。

在最大标记集中，在第一级上的名词在第二级上分为两类：普通名词和专有名词。专有名词在第三级上进一步分为五类，普通名词在第三级上又分为两类：普通名词合理和名词。其中离合名词是指一些动宾式复合词中间插入别的词语之后，原来在复合词内部只作一个构词成分的语素不得不独立成为一个词的情形，如：

如何 使 种植业 与 市场 接 上 轨

…… 向 老师们 敬 一 个 礼 ， 然后 他 深深 鞠 了 一 躬

“接轨”、“敬礼”和“鞠躬”本来都是一个复合词，由于中间插入了别的成分，使得“轨”、“礼”和“躬”都变成了独立的词。但这些词和前面的动词显然有很强的依赖关系，有的离了前面的动词之后根本就不可能存在，如“躬”。这些词和一般的名词显然不同，如果把这些词单独加上标记，对下一步的句法和语义分析显然是有益的，但这样会增加更多的兼类现象，如上例中的“礼”在上面的语境中它是一个离合名词，但在别的情况下它也可以是一般的名词。是不是需要这一类要根据任务的要求综合考虑准确率和信息量两个因素来作决定。在上面的实验中，我们设 $\lambda_1 = \lambda_2 = 0.5$，即给予两个参数相同的权重时离合名词就没有被选

中，在另外的改变参数权重的实验中，我们发现，当 λ_2 大于 0.7 时，离合名词就总是被选中。而当 λ_2 小于或等于 0.7 时，离合名词总是不被选中。

在上面的结果标记集中，我们发现，原来的专有名词分为五个小类，在结果标记集中保留了其中的四个。这主要是因为专名跟其他类兼类的现象比较少，因此尽管专名分得这样细，但引起的词性歧义并不多。不过，这并不说明专名的归类就很容易了。事实上，专名的细分类并不是一种语法分类，而是语义分类。专名的归类更多的要依靠词性标注以外的技术来解决。

显然，所得到的标记集并不能直接应用，因为它在某些方面缺乏系统性。比如，在词缀中，它选择了名词前缀（kh）、动词后缀（kv）和可能中缀（kp）,但没有选择名词后缀（kn）。从词缀的系统性考虑，这显然是不合理的，所以有必要对得到的标记集进行适当的微调。

2.4 本章小结

本章分析了一组评价词类标记集复杂度的评价指标，通过标注实验证明：用标记集的动态歧义指数或动态歧义率是比较理想的评价标记集复杂度的量化指标。

本章提出了一种利用遗传算法优选词类标记集的方法。该方法的目的是帮助人们根据任务的需求选择一个最优或较优的标记集。该方法自动生成可能的标记集，并对对每一个标记集给出一个评价函数值，在一个很大的标记集候选集中搜索评价函数值最大的标记集。该评价函数考虑了两个参数：标记集的动态歧义指数和熵。参数的权值可以根据应用进行调整。

本章给出了一个应用遗传算法进行标记集优选的实用方法。首先，在一个树型的分类体系中定义一个最大标记集和最小标记集。然后选择一个语料库，用最大标记集进行标注，并经过人工校对。在进化过程中，当一个新的标记集生成时，该语料库自动转化为用新标记集标注的版本。我们的实验表明：该方法对词类标记集的优选提供了一种有价值的辅助手段。

第三章　汉语命名实体识别

3.1 引言

英语的命名实体识别都是在词串上进行，而且英语的专有名词一般都有形式标志，即第一个字母要大写。所以英语专名识别的主要任务有两个：（1）识别专名短语；（2）对专名进行分类，确定一个专名是人名、地名、机构名或其他。与英语相比，汉语的命名实体任务则要复杂得多。首先，必须在分词过程中识别出专名词语，同时需要对这些专名词语进行分类。在分好词之后，再在词序列上识别专名短语。所以，汉语的命名实体识别工作必然要分两步：第一步是词语一级的，与分词同步进行，第二步是短语一级的，我们可以把它作为浅层句法分析的一部分。不管是专名词语的识别还是专名短语的识别，都是浅层句法分析的不可或缺的基础。

汉语词语级命名实体的主要任务是：（1）识别出汉人姓名[4]；（2）识别出外国译名，并对译名进行分类，确定译名是人名、地名或其他；（3）识别出中国地名。机构名中也有专名词语的识别问题，如：

北京益科传感器工程公司

北京市孔雀纸制品厂

其中的“益科”和“孔雀”都是公司名，按理应在分词阶段解决，即把“益科”合起来，并标注它是机构名，把“孔雀”标注为机构名而非普通名词。但由于机构名的识别往往要涉及更大的语境，例如，在上面两例中，要确定这两个词是机构名，至少要看到后面三个词，这在分词阶段处理很不方便，所以我们把机构专名词语的识别放在机构名短语的识别中进行。

与机构名相比，其他三类专名词语的识别则不要求很大的语境，一般只要看到前后一个词就足够了。所以我们把这三类专名词语放在分词阶段处理。

因为命名实体和普通词语之间存在歧义，所以有必要把词语级命名实体识别任务融入分词过程中。我们在分词阶段又采用了切分和词性标注一体化的概率模型，所以，在进行实语块分析之前，词语切分、词性标注和命名实体识别是融为一体、同步进行的。下面首先介绍切分标注的概率模型，然后讨论在这一模型下专名词语识别的策略，最后给出汉人姓名、西文译名和地名候选词的生成算法。

3.2 语言模型

3.2.1 切分标注一体化模型

我们把词语级命名实体识别任务融入基于词类的概率语言模型中。分词和词性标注一体化模型基于这样一个简单的推理：如果在一个句子中存在一个词类的最大概率路径，那么，与这条路径上的词类标记对应的词也一定在这条路径上。我们之所以选择基于词类的概率语言模型作为专名识别的语言模型，主要基于以下几方面考虑：

（1）　切分和标注一体化是目前汉语信息处理普遍采用的词法分析方法（白拴虎，1995；

[4]为了严格起见，本文使用“汉人姓名”这个称谓，而不笼统地使用“中国人名”的称谓。

Sun M. et al. 1997)。在一体化模型中，分词和标注可以互相促进（Sun & Huang, 1996）。

（2） 基于词类的概率语言模型是消解分词歧义的一种较好的方法。汉语分词中存在两大难题：一是其切分歧义的处理，一是未登录词（主要是各种专名）的处理。在处理切分歧义方面，我们做了一个小规模的实验，选择了在有关文献中经常被引用的 100 个比较“经典”的有切分歧义的句子（不包括语义歧义和语用歧义），分别用最大匹配（正向、逆向）、最大词概率和最大词类概率几种不同的算法进行切分，结果是：最大词类概率算法优于最大词概率算法，最大词概率算法优于最大匹配算法。

（3） 专名和普通词之间存在歧义，基于词类的模型有利于消解专名和普通词之间的歧义。

在基于词类的概率语言模型中，对于一个汉字串：$C_1,C_2,\ldots,C_n$，从任一个汉字 C_i（$1 \leqslant i \leqslant n$）开始有 0 到 m 个候选词，每个候选词 W_{ij}（$1 \leqslant i \leqslant n$，$1 \leqslant j \leqslant m$）有 1 到 k 个候选词性。目标是找到一个覆盖汉字串 $C_1,C_2,\ldots,C_n$ 的词串，在该词串中存在一条在所有候选词性组合中概率最大的路径。

为了避免因为概率值太小而引起数据溢出，同时也为了便于人的理解，我们在系统中采用概率的负对数表示每个结点的概率评分[5]（Sproat et al.1996），这个概率评分可以理解为该结点的耗费。概率值越大，耗费越小。因此，寻找最大概率路径问题就等价于寻找最小耗费路径问题。

3.2.2 分词歧义的并行消解策略

汉语文本中的分词歧义情况是相当复杂的，这些歧义可以归结为以下三类（Sun M. et al. 1997）：

（1） 普通词和普通词之间的歧义；

（2） 普通词和专名之间的歧义；

（3） 专名和专名之间的歧义。

在过去的许多分词系统中，对于切分歧义通常采取异步处理方式，即先消解普通词的切分歧义（主要是交集型歧义），然后再分别识别人名、地名等专名。这种异步处理方式往往顾此失彼，在解决普通词歧义的时候不能考虑专名的情况，在识别某一类专名的时候又不能考虑另一类专名的情况。而在基于词类概率的切分和标注一体化模型中，可以同时考虑所有可能的歧义，这种并行处理策略可以保证在一个概率模型中得到全局最优解，从而使以上三类歧义得到一致的、系统的解决方案。

下面分别举例说明三类歧义的消解。

例 1：汉堡地区华侨华人中人才如此之多

字串“华人中人才”形成了一个交集型歧义链，可能的双音节词有“华人、人中、中人、人才”，其中每个汉字又都可以是单音节词。上面句子中共有 15 个汉字，分别以下标 0-14 来标示，每个汉字对应一个候选词串，如下所示：

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
汉	堡	地	区	华	侨	华	人	中	人	才	如	此	之	多

0：汉堡　汉　　1：堡

2：地区　地　　3：区

4：华侨　华　　6：华人　华

[5] Church(1988)用概率的对数表示概率评分，但这样结果是零或负数，不便于人的理解。

7： 人中　人　　8： 中人　中

9： 人才　人　　10： 才

……

图 3-1 显示了字串“华人中人才”对应的状态空间和消歧过程：

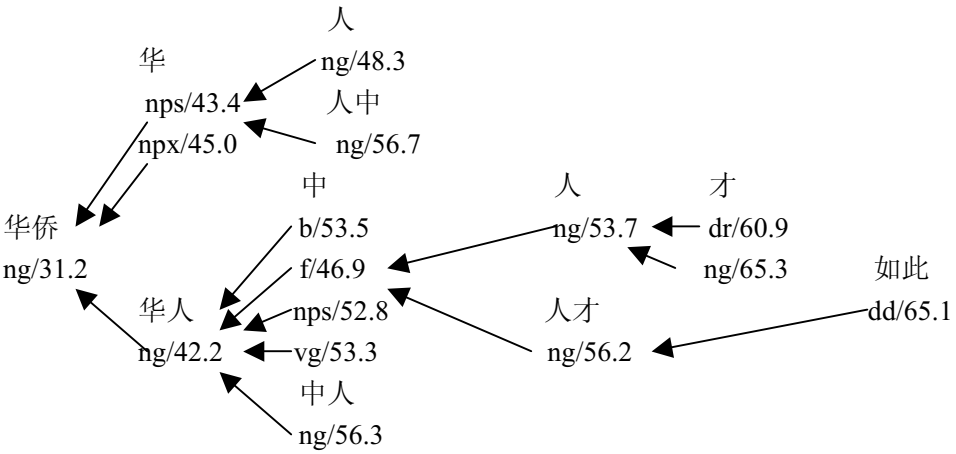


图 3-1　普通词切分歧义消解例示

在图 3-1 中，每一个候选词下列出了该词所有的可能的词性。斜杠之前是词类代码，代码的含义可参见附录二。斜杠之后是该结点的累计耗费。从“华侨”后的“华”开始有两个候选词：“华”和“华人”，“华”有两个可能的词性，一是地名（中国的简称），一是姓氏。这两个词性成为状态空间中的两个结点，这两个结点有一个共同的前驱结点，即“华侨”中的唯一词性“ng”。“华人”只有一个词性“ng”，它的前驱结点也是“华侨”中的“ng”。从第一个“人”开始有两个候选词：“人”和“人中”，它们分别只有一个词性“ng”，它们的共同最佳前驱结点是“华”中的“nps”。从“中”开始有两个候选词：“中”和“中人”，“中”有 4 个可能的词性，“中人”有一个可能的词性，这 5 个结点的共同最佳前驱是“华人”中的“ng”，在这些结点中，“中”的词性“f”（方位词）的累计耗费最小。以下依此类推。

例 2：水利部长钮茂生日前在江西考察工作

如果不考虑人名识别，这里存在一个交集型歧义字段“生日前”，加上人名识别以后，增加了两个候选人名“钮茂生”和“钮茂”，使交集型歧义链的链长增加。“钮茂生日前”至少存在以下三种切分可能：（1）钮茂生/日前；（2）钮茂/生日/前；（3）钮茂/生/日前。

图 3-2 显示了消歧的过程。

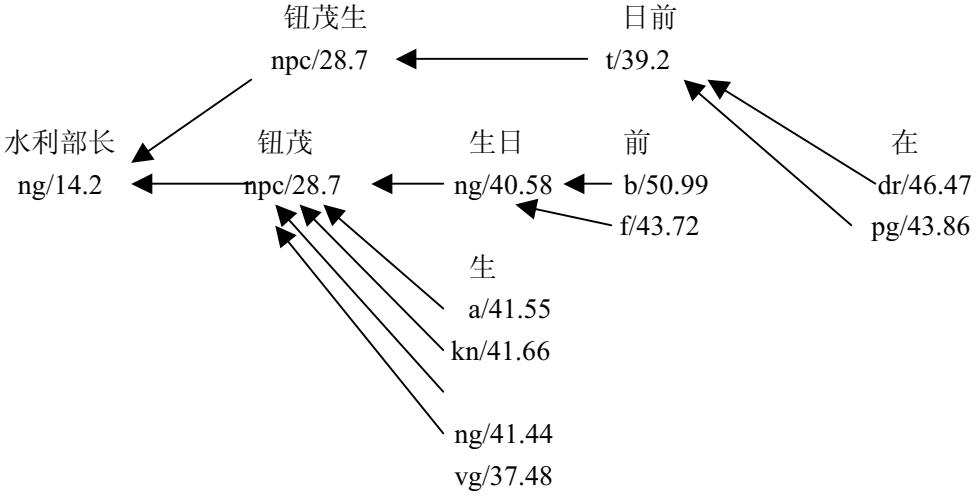


图 3-2　普通词和专名切分歧义消解例示

从图 3-2 可以看出，“水利部长/ng 钮茂生/npc 日前/t 在/pg”构成了一条最小耗费路径，即最大概率路径。两个人名“钮茂生”和“钮茂”的歧义，人名“钮茂生”和普通词“生日”的歧义，普通词“生日”和“日前”的歧义同时得到消解。

例 3：大连市甘井子区分局党委

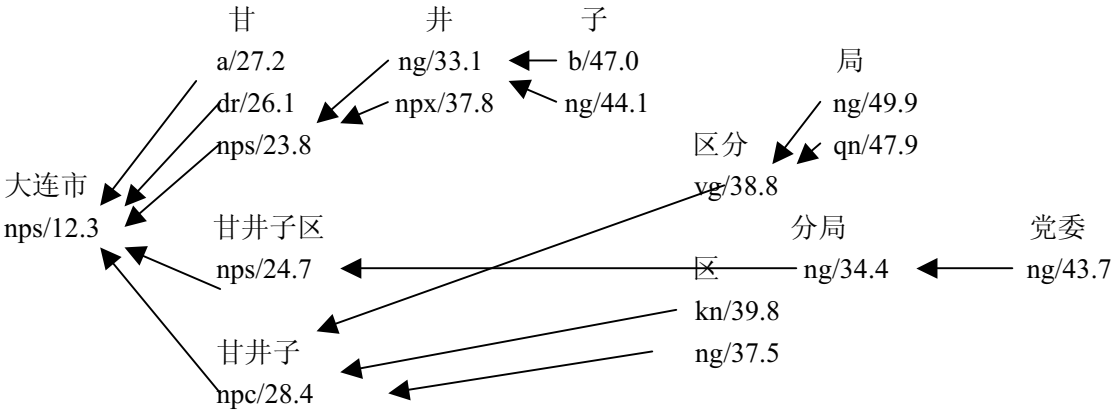


图 3-3 专名和专名及专名与普通词混合歧义消解例示

在图 3-3 中，我们可以看到，“甘井子”是一个候选的汉人姓名，“甘井子区”是一个候选的地名，“区分局”又是一个交集型歧义链。对于这样的复杂歧义，用规则方法是很难消解的。这个例子表明并行处理要优于异步处理。

3.3 生成专名词语候选的策略

在基于词类的概率模型中，汉字串首先映射到一个词的集合，词集合中每个元素又带着若干可能的词类标记，概率模型在所有可能的词类标记路径上选择一条概率最大的路径。在这条路径上的词是原字串对应的概率最大的词串。

在从汉字串映射到词集合的过程中，主要依据一部词典。与词典中任一个词对应的字串自然成为词的候选，我们把这称为候选词的静态生成，因为词典是相对封闭的，在系统运行过程中词典的大小既不会增加也不会减少。我们可以考虑把一部分常用的人名、地名等专名收入系统词典中，但这些专名是开放的，词典再大也不可能尽收，所以需要有一个生成算法从汉字串中动态生成可能的专名词语。

各类专名的性质和特点不同，这要求我们在研究专名词语候选的生成算法时要对各类专名的性质加以研究，根据它们不同的特点使用不同的算法。

汉人姓名的特点是：

（1） 度受限。汉人姓名最短的 2 个汉字，最长的 4 个汉字。

（2） 字受限。汉人的姓名分为姓和名两部分，姓氏绝大多数是单字，极少数是双字，这些姓氏用字的范围有限，常见的姓氏只有几百个汉字。

汉人姓名的这两个特点决定了汉人姓名识别不需要很大的观察范围，只需要看一个二至四字的汉字串是否符合汉人姓名的用字特征就可以了。所以生成汉人姓名候选的关键在于分析人名的用字特征。分析人名用字特征使用统计方法，利用一个大的人名库，统计其中人

名用字的特征（张俊盛等，1992；孙茂松等，1994；Chen & Lee,1996）。

西文译名的特点是：

（1）用字范围受限；

（2）长度不受限制；

（3）译名中一般不包含复合词。

虽然译名的长度不受限制，但（1）、（3）决定了生成西文译名候选主要依赖一个字串内部的特征就比较容易确定译名的边界。所以生成译名候选的关键在于确定译名的用字范围。常用的方法是使用一个译名表来统计其中的用字特征（孙茂松等，1993；Chen & Lee,1996）。

与汉人姓名和西文译名相比，中国地名的用字范围受限很小，沈达阳等（1995）统计了 17637 个中国地名，其中共使用了 2595 个汉字，所以单从用字上很难把中国地名与其他词区分开来，但中国地名在构词上的一个特点是：一般带地名后缀，而且后缀的数量是十分有限的。由于普通词也可以成为地名的一部分，如“孔雀河”、“火焰山”，而且这些地名后缀大部分属常用词，所以仅根据后缀会生成大量错误的地名候选。所以，生成地名候选除了根据地名本身的结构特点之外，还需要依靠上下文。

下面分别详细讨论汉人姓名、西文译名和中国地名候选词的生成算法。

3.4 汉人姓名候选词的生成

3.4.1 汉人姓名的构成

这里所说的汉人姓名不限于汉族人的姓名，中国境内的一部分少数民族人名的形式和汉族相似，因此汉人姓名的性质及识别方法也适用于这些少数民族。同时，韩国、朝鲜和越南人的姓名形式也与汉族的形式相同，因此这里讨论的汉人姓名也适用于这些国家的人名。

出现在文本中的汉人姓名首先可以分为两类：

（1） 完全形式。姓名的完全形式包括姓和名两部分，姓在前，名在后。

（2） 非完全形式。又可以分为五类：

a) 前缀 + 姓。 前缀有： 老、小。

如： 小李、老王。

b) 姓 + 后缀。 后缀有： 老、总、工、氏、某、某某、犯。

如： 王老、陈总、张氏、李某、王某某、张犯。

c) 有姓无名。单个姓氏可以作为人名的一种简称姓氏。

如： 王从张处得知了这一情况。

d) 有名无姓。单个名字也可以作为人名的一种简称姓氏。

e) 姓 + 称谓词。这种形式在文本中是最常见的。

如： 张妈妈、郑伯伯、黄老师、刘总经理、王校长。

3.4.2 汉人的姓氏

从历史上看，汉人的姓氏用字是比较复杂的，如台湾出版的《中国姓氏集》收集姓氏 5544 个，其中单姓 3410 个，复姓 1990 个，三字姓 144 个（刘开瑛，2000）。但这些姓氏到现代大部分已经不再被使用，现代汉人的姓氏趋于简单，用字相对集中，这也为自动识别汉人姓名提供了方便。由于我们所处理的文本主要是现代文本，所涉及的人名也基本上都是现代人名，所以对历史上汉人的姓氏情况可以不予考虑。

现代汉人的姓氏从构成来看可以分为三类：

（１）　由单个汉字构成，称为单姓。如：张、李、王、周。

（２）　由两个汉字构成，称为复姓。如：欧阳、诸葛、尉迟、端木。

（３） 双姓复合形式。如：陈方安生、范徐丽泰。

双姓复合姓氏的来源有两种：一种是所谓冠夫姓，就是妇女在结婚以后在自己原来的姓名前加上丈夫的姓。冠夫姓在我国台湾和香港地区比较流行，在中国大陆基本上没有。另一种情况是取父母两个人的姓，这种情况也比较少见。

在这三类姓氏中，单姓占绝对多数，复姓数量很少，常见的复姓不过１０几个。双姓复合姓氏由于在大陆很罕见，一般情况下可以不予考虑。

从汉人姓名的用字范围来看，姓氏的限制性最强。根据我们对一个 502,252 万个姓名的姓名库[6]的统计，单姓姓名为 501,732 个，占 99%，复姓姓名为 520 个，占 1%，双姓复合姓氏姓名没有。

在姓名库中，50 万个单姓姓名共涉及 1947 个单姓，其中 548 个单姓可覆盖全部的 99%。可见，姓氏用字的范围是十分受限的。

3.4.3 姓氏用字范围的确定

确定姓氏用字范围的方法有三种：

（1）利用一个姓名表来估计汉字用作姓氏的概率，即利用汉字在姓名表中的相对频率作为估计汉字用作姓氏的概率的依据，如果一个汉字在姓名表中没有作为姓氏出现，那么它作姓氏的概率就是零。这样就可以根据姓名表中的相对频率确定一个姓氏用字表（张俊盛等，1992；宋柔等，1993；孙茂松等，1994）。这种方法的缺陷是：汉字在姓名表中的相对频率不能准确反映一个汉字在真实文本中作姓氏的概率。这样会产生两个问题：

一、一些不常用的姓氏用字会被排除在外。例如，在我们所用的姓名库中，下面的一些姓氏由于在姓名库中出现的频次较低，就被排除在 548 个字集之外。这些字属于非常用字，在文本中出现的概率比较低，但相对作姓氏的概率比较高，有的甚至只能作姓氏。把这些字排除在外，会降低姓名识别的召回率。

翦　　苻　　庾　　弥　　钦　　郧　　萨　　邾　　尕　　郓

同时，一些作姓氏概率很高的非常用字，因为在姓名库中的频次比较低，虽然被包括进姓氏字集中，但概率值很低。如“逄”（音 pang2），它在姓名库中共作姓氏 20 次，相对频率很低，只有 0.0004，但从文本中来看，它只能作姓氏，所以它作姓氏的概率应为 1。类似的还有以下一些字：

恽　　仉　　蒯　　扈　　濮　　贝　　宦　　谭　　钭　　铙

二、会使一些很少作姓氏的高频字进入姓氏字集。这些常用字在文本中出现的频率很高，但相对作姓氏的概率很低。但由于在姓名表中的相对频次较高而进入姓氏字集。这些字进入姓氏字集虽有利于提高识别的召回率，但会大大降低识别的准确率，所得大于所失，因而是不可取的。例如，在上述的 548 个高频姓氏用字中（最低频次是 18）就包括了以下一些常用字：

天	广	言	望	竹	加	须	火	刀	暴	力	从	羊
买	多	鱼	催	巨	幸	运	畅	寻	员	山	生	青
庆	岩	税	冶	德	扎	延	续	依	植	怀	种	玉

[6] 汉人姓名库和外国人名译名库由清华大学计算机系孙茂松教授和周强博士提供，谨此致谢。

(2) 利用语料库中实际出现的姓名进行统计确定姓氏用字的范围（郑家恒等，1993）。这种方法虽然能准确地估计出每个汉字用作姓氏的概率，但需要一个非常大的经过正确切分的语料库。根据我们的统计，在一个 150 万词的综合语料库中，出现了 3179 个不同的完全姓名，共出现了 12,054 次，占文本的总词次的 0.8%。如果按这个比例，要得到 10 万个不同的姓名，就需要 4700 万词次的语料库，而要得到 50 万个不同的姓名，则需要 2.36 亿词次的语料。这显然在目前是做不到的。

（3）综合使用姓名库和语料库。这种方法是假设姓名表是从一个大的语料库中得到的，这样就不需要对语料库进行分词，只需要统计语料库中每个字的频次就可以了。用一个字在姓名表中作姓氏的频次除以它在语料库中的频次就可以估计出这个字用作姓氏的概率。(Chen & Lee 1996) 使用了这种方法，但所用的语料库太小。其所用的姓名表共有 219,738 个姓名，但语料库只有 400 多万字，不能准确地估计出汉字用作姓氏的概率。

我们确定姓氏用字依据两个标准：

一、汉字 C 用作姓氏的概率，定义如下：

$$P(C = Surname \mid C) = \frac{Count(C = Surname)}{Count(C)}$$

我们用上述第三种方法来估计汉字用作姓氏的概率，Count(C=Surname)是汉字 C 在姓名库中作姓氏的频次，Count(C)是 C 在一个大的真实文本语料库中的频次。我们使用的姓名库如上所述，包括了 50 多万不同的姓名。我们使用的文本语料库包括约 2.5 亿汉字。

二、汉字 C 用作姓氏的相对频次。即：

$$Count(C = Surname)$$

仅根据 P(C=Surname|C)来确定姓氏用字的范围，会使一些作姓氏概率低、但作姓氏绝对词数多的字被排除在姓氏用字表中。如：

关	文	华	成	应	项	时	解	车	司
和	全	谈	师	海	计	宣	门	劳	南
花	查	明	席	商	区	国	平	农	巴

如果按 P(C=Surname|C)来排序，上面这些字的位次都在 500 以后，但这些字在姓名表中作姓氏的频次都在 80 以上。凭直觉我们会发现这些字虽不属常见姓氏，但也不是十分罕见的。如果把这些字排除在外，将会降低识别的的召回率。因此需要综合使用以上两个标准。

具体说来，我们确定姓氏用字集合的步骤是：

一、取文本姓氏概率最高的前 250 个姓氏。这些姓氏覆盖姓名表的 78.8%。这 250 个姓氏中，在姓名表中用作姓氏频次小于 80 的共 86 个，其中 43 个频次小于 20。如果仅按标准 2（即作姓氏的相对频次）来选择的话，这些字将会排除在姓氏字集之外。如：

仝	竺	亢	晁	钮	郦	嵇	郗	逯
亓	仉	闫	咎	逄	蹇	笪	茆	茆

这些字虽然不常用，但作姓氏的概率很高，有的只能作姓氏。把这些字收入姓氏表中，对于提高人名识别的准确率和召回率都是有利的。

二、剩下的姓氏用字中，取频次大于 80 的，共 171 个，覆盖姓名表的 19%。这其中包括一些在文本中作姓氏概率比较低的，但绝对数量比较多，把这些字包括进来，对于提高召

回率是有利的，但会降低准确率。如上面所举的“关、文、华”等。

两部分合起来共选择姓氏用字 421 个，覆盖姓名表的 97.8%.

3.4.4 汉人姓名候选词的生成算法

输入：汉字串 $S = C_1, C_2, \cdots, C_n$

资源：单字姓氏集 SN, 双字姓氏集 DN, 姓名中间字集 MN，姓名末字集 TN。SN,MN,TN 中储存每个字作姓氏的概率 Ps(C), 作姓名中间字的概率 Pm(C) 和作姓名末字的概率 Pt(C).

常量：θ_1为单名概率阈值，θ_2为双名概率阈值。

过程：

1） 在 S 上扫描，如果发现双字组合 $C_{i-1}C_i \in DN$，则生成候选姓名：

NameCandidate1 = $C_{i-1}C_iC_{i+1}$ // 双姓单名

NameCandidate2 = $C_{i-1}C_iC_{i+1}C_{i+2}$ // 双姓双名

并把它们加入候选词集中。转 1）.

2）如果发现 $C_i \in SN$,

如果 $C_{i+1} \in TN$, 则生成候选姓名：

NameCandidate1 = C_iC_{i+1} // 单姓单名

如果 $Ps(C_i) * Pt(C_{i+1}) > \theta_1$则将 NameCandidate1 加入候选词集中。

如果 $C_{i+1} \in MN$ 且 $C_{i+2} \in TN$，则生成候选姓名：

NameCandidate1 = $C_iC_{i+1}C_{i+2}$

如果 $Ps(C_i) * Pm(C_{i+1}) * Pt(C_{i+1}) > \theta_2$则将 NameCandidate1 加入候选词集中。

3） 如果到达串尾则退出，否则转 1)。

3.4.5 汉人姓名识别结果

我们选择了《人民日报》1998 年 7 月 30 日全部 179K（80,508 字）语料作为测试语料（下同)。其中出现汉人姓名 640 词次，系统识别出 672 个，其中 606 个正确，漏识 34 个。

正确率 = 606/672 = 90.2%

召回率 = 606/640 = 94.7%

F-measure = 2* 0.902*0.947/(0.902+0.947) = 92.4%

3.4.6 错误分析

系统的召回率比较高，但错误率较高，主要错误类型有：

（1） 把中国地名误作人名。

如： 郭东（大堤） 石光梁（上） （抵达）黄州 黎钦线 黎湛线

吴中（第一名胜） 周家咀（险段） （北有）秦陵

主要原因在于：

一、一些地名后缀没有收入地名后缀表中，如“大堤”、“陵”、“咀”。

二、语境词没有收入语境词表中。如“抵达”。

（2） 把普通词误作人名。

如： 白雪、房源、高丽、洪峰、林立、楼群、魏晋、宁为玉（碎）、
夏玉米、战洪峰

主要原因在于：

一、由于姓名库中双字人名太少（只有 2 万多，与 50 万的总数显然不成比例），因而不能准确地估计双字人名的概率，这一部分大部分错误来自双字人名。

二、有的词没有收入词典。如：“房源”、“魏晋”、“宁为玉碎”。

（３） 把外国译名误作汉人姓名。

如：马吉德、谢里夫、卓娅、夏庇若

（４） 把机构名误作人名。

如： 储金会 、乐华

在 34 个漏识的错误中，主要有以下几种类型：

（1）罕见姓氏。如“况钟”、“展红”。

（2）笔名。如：“冰心”、“乘舟”。

（3）人名中包含高频单字词，如“于右任”。

（4）把三字人名误识为双字人名。

3.5 西文译名候选词的生成

3.5.1 西文译名的特征

印欧语言中的专名翻译成中文主要用音译的方式。音译的译名有两个重要特征：

（１） 译名的用字受到很大限制。汉语的一个音节往往有许多汉字，但在译名中往往只用其中少数的几个。如“玛丽”，不会写成“马力”、“码例”等。根据《英语姓名译名手册》中近 4 万个译名的统计，总共用字不到 500 个（孙茂松等，1993）。这一特征为译名识别提供了很好的条件。

孙茂松等（1993）曾尝试仅根据译名用字的限制在字串上识别译名，结果发现虽然可以得到较高的召回率，但准确率不高（据其报告，召回率为 98%，准确率为 63%）。准确率低的原因是，译名用字虽然受限，但其中包含许多很常用的单音节词。仅仅依据用字特征就会把非译名的单音节词误识为译名的一部分。如：

克林顿总统和夫人希拉里来到上海图书馆参观。

由于“来”是译名用字，所以仅根据译名用字会把“希拉里来”误识为译名。因此，仅根据译名用字来识别译名是不够的。

（２） 组成译名的字串在汉语中无意义。这一性质有两个含义：一、译名中没有有意义的词语组合；二、译名中没有普通的复合词。这一性质使我们在阅读时很容易把译名与其他汉语词区别开来而不至于产生误解。这也是目前计算机译名识别系统所用策略的依据（Chen& Lee 1996; Sun M. et al. 1997）。该方法的基本思想是：在最大匹配切分后的单字词序列中寻找一个属于译名用字集的最大字串。这种方法就是基于这样一个假设：译名不跨越汉语的复合词。这种方法可以避免在一般字串上匹配所产生的边界错误，如上例中“来到”是一个复合词，因而避免将“来”误作译名用字，这样可以提高识别的准确率。

在单字词串上识别译名的方法仍然会遇到以下两个问题：

a) 一些普通的单字词串由于符合译名用字特征而被误识为译名。如：

不 负 港人 所托

这里，由于“所”和“托”都是译名用字，所以被误识。

b) 译名中的部分字串是词典中的复合词，因而会被漏识。如：

不久前 马克 斯·维勒 在 北京 的 画展 非常 成功
原 内阁 部长 留任 的 有 外交部长 西亚 松

这里，由于“马克”和“西亚”都是词典中的复合词，所以不在一个单字词序列中。

第一类问题等同于组合型歧义，不太容易解决，第二类问题则可以通过一个词表来解决。针对第二类问题，我们对《英语译名手册》中所有译名的内部结构进行了分析，看其中包含了哪些复合词。《英语译名手册》共有译名 38,862 条，排重以后有 29,510 条（因为有一些不同的英语人名翻译成相同的汉译名）。在这些译名中共得到字的二元组合(bigram)13,226 条，三元组合(trigram) 27,392 条。

在二元组合中，与词典中复合词重合的有 170 个，其中 119 个是地名（含外国地名），如：

安多 安吉 伊宁 安西 安图 安泽 达卡 彭泽 巴西 巴林
瑞安 西安 西林 西沙 悉尼 南沙 利辛 廷布 罗平 莱西

51 个是普通复合词，其中不少是译音词，如：

坦克 克隆 吉普 纳米 雷达 马达 马克 杜马 拉丁 夸克 里拉

有些则纯属巧合，如：

福利 利索 奇特 阿拉 吉利 皮科 斯文 普法 多达 圣地 内宾
谢恩 劳顿 白金 金文 科班 甘为 西皮 多加 马扎 恩德 恩泽

在三元组合中，与复合词重合的只有三条，都是译音词：

埃米尔 华尔兹 迪斯科

我们把这些词收入一个特殊词表 FrnAmbList，在识别译名的时候，这些词等同于单字词串。

3.5.2 西文译名候选词生成算法

输入：最大匹配分词后的词串 S = W1, W2, …, Wn.
资源：译名首字表，译名中间字表，译名末字表，由译名用字组成的复合词表 FrnAmbList.
过程：

1) 在 S 中找到一个最长单音节词串 MonoWordString（FrnAmbList 中的复合词等同于单音节词串）。
2) 如果 MonoWordString 包含两个或两个以上的汉字，则在其中寻找一个子串 S_i ，其满足以下条件：
   a) 至少包括两个汉字；
   b) 串的首字在译名首字表中；
   c) 串的尾字在译名末字表中；
   d) 串中除首字和尾字以外的其他字在译名中间字表中；
   e) S_i 不在词表 FrnAmbList 中。
3) 如果找到符合条件的 S_i，则把 S_i 加到候选词集中。
4) 如果到达 S 尾，则退出，否则转 1）。

3.5.3 西文译名识别结果

语料中共出现译名 183 词次，系统识别出 191 个，其中正确的 174 个，漏识 9 个。

正确率 = 174/191 = 91.1%
召回率 = 174/183 = 95.1%
F-measure = 2 * 0.911*0.951/(0.911+0.951) = 93.1%

错误的主要原因是：

（1） 由于译名样本的不足，有些译名用字没有包括进来（这里也有文本中译名用字不规范的问题）。这类错误占全部错误的 40%。

如： 阿姆斯特郎 （人名，“郎”非译名尾字）
阿瓦铎 （地名，“铎”非译名尾字）
艾卜拉莱篙 （地名，“篙”非译名尾字）
奉辛比克 （组织名，“奉”非译名首字）
旁遮普邦 （地名，“遮”非译名中间字）
哥美士安 （人名，“美”非译名中间字）
拉吉夫·甘地 （人名，“地”非译名尾字）

（2） 单音节词串误作译名。这类错误占全部错误的 30%。

如： （扎下）根来 切莫 （此去）彼来
塔西因（挨着山崖而无须开掘） 因多宗（暴力罪案被香港判刑）

（3） 译名误作汉人姓名，如前所述。

（4） 中国少数民族人名错误。中国少数民族人名的汉译与西文译名有许多相似之处，但用字上有不同，因此西文译名生成算法能够正确生成一部分少数民族人名，如“达赖”，但也会产生错误，如“巴音朝鲁”，因为“朝”非译名中间字，因而把“巴音”误识为一个译名。

（5） 缩略语引起的错误。几个错误都是由铁路线名引起的。

如：达万（线） 内昆（线） 西康（线）

3.6 地名候选词的产生

3.6.1 地名的特征及识别策略

与汉人姓名和译名相比，地名相对数量要少，因为世界上的地名总是有限的。地名有两个特点：

一、一部分特别重要的地名已经进入了基本词汇。这些地名由于特别重要，因而在日常生活中出现的频率很高。这些地名已经进入基本词汇，成为社会上一般人知识的一部分。比如，一些大国的国名，如“美国、日本、法国”等；世界上一些著名城市的名称，如“柏林、巴黎、伦敦、芝加哥”等；中国的省级行政区名，如“北京、上海、安徽、江西”等；省会城市名和重要的省辖市名，如“广州、杭州、深圳、厦门”等；名山大川等重要的地理名称和著名旅游风景区名，如“黄山、长江、太湖、西湖、故宫、颐和园”等。由于这些地名词语已经进入了基本词汇，所以在文本中出现的时候，一般不需要加上地名的标志或加上特定的限定成分，比如，在中国如果说“深圳”的话，不需要在“深圳”加上表示地名属性的标志词，即不需要说“深圳市”，当我们听到或看到“深圳”这个词的时候，自然会根据我们关于这个地名的知识补充属性信息，即它是一个城市。同时，我们也不必说“中国深圳”或

“广东深圳”，因为绝大多数中国人都知道深圳的地理位置。既然这些地名已经进入了基本词汇，所以有必要在系统词典中收入这些常见的地名。

二、非常用地名在使用时经常要加标志词或限定成分。与上面所说的著名的地名相比，当我们在使用一个不太知名的地名的时候，往往需要加上限定成分或表示地名属性的标志词，如“个旧”，很多人可能不知道这个地名，所以要加上地名标志词“市”，表示它是一个城市，或在前面加上表示范围限定的词语，如“云南个旧”，表示它的地理位置。这些信息帮助人在文本中识别一个未知的地名。这些非常用地名数量十分庞大，不可能都收入词典，但这些标志为动态识别文本中的地名提供了线索。

根据地名的这两个特征，我们制定了地名的识别策略。按区域可以将地名分为两类：一是中国地名，一是外国地名。

对外国地名的处理策略是：将常见地名收入词典，这些地名包括所有的国名（全称和简称）、首都名和世界知名城市名。目前收入词典的有所有国名的全称、简称加上首都名，共612 个。同时也在词典中收入了 100 个左右著名的自然地理名称和世界著名城市名。对于非常用地名，地名的主体识别在西文译名识别阶段完成，西文译名的默认类是人名，但当译名有地名的上下文特征（下面将要讨论）时，则将译名的类别改为地名。

对中国地名的处理策略与外国地名的处理策略类似，首先将常用地名收入词典。目前我们在词典中收入了中国县以上行政地名。根据新版《辞海》的附录，中国县和县级以上行政地名共 2584 个，其中包括了不少极少见到的县名。同时收入了 100 多个常见自然地理名称和著名旅游风景区名。对于县以下的行政地名和一般的自然地理名称一律不收入词典中，需要在文本中动态地识别。

为了了解实际文本中地名的使用情况，我们统计了《人民日报》1998 年 7 月份语料中地名的分布情况。语料共包括 62,404 个词型、1,115,232 词次（不含标点符号和数字），其中地名词语出现了 3,809 个词型和 35,277 词次，分别占总数的 6.1%和 3.2%。其中在词典中查到的词型和词次分别是 1,849 和 3,1040，也就是说，文本中 88%的地名都可以在一个包括 3000 多个地名的地名表中查到。而且这些地名在文本中的重复率很高，这和基本词汇的特征是一致的。在词典中收入 3000 多个常用地名，虽然要占一定的存储空间，但跟它所得到的性能相比是值得的。

3.6.2 中国地名候选词的生成算法

动态生成地名候选词以及确定西文译名是否地名的关键是找到地名的构词特征和上下文特征。

从内部结构来看，地名的一般构造方式是：专名+通名。如“凤凰山”，“凤凰”是某一个山的专有名称，“山”是通名，所有的山名后面都可以带上这个专名。我们把通名称为地名后缀，地名后缀是十分有限的一个词语集合。（李如龙，1998）把地名后缀按地名的类型分为以下几类：

（１） 自然地理实体名通名。
　　如：山、岭、峰、冈、江、河、湖、泊、岛、港、沙漠、盆地。
（２） 聚落通名。
　　如：乡、村、坊、里、屯、庄、镇、巷、街、道、集、路、园。
（３） 人工建筑地物的通名。
　　如：城、楼、阁、亭、门、桥、塔、庙、院、渡、渠、寨、坝。
（４） 行政区划通名。
　　如：州、县、市、区、乡、村。

从语法特征上看，地名后缀可以分为两类：

（1）不成词后缀。如：“岭、峰、冈、阁、集、屯、湾、庄、矶”等。

（2）成词后缀。如：“江、河、山、冰川、城堡、盆地、山庄、胡同、大街”。

在地名后缀中，大部分都是成词后缀，不成词后缀所占的比例很小。

正如我们在前面分析的那样，非常用地名在文本中出现时，一般要出现地名后缀，所以地名后缀将成为生成地名候选词的触发条件。

地名出现的上下文环境也是识别地名的重要线索。我们通过分析语料中地名的上下文发现地名出现的典型的上下文环境有：

（1）前面是一个地名。这种情况一般是：前面是一个更大范围的地名，后面是这个范围内的一个地名。前面一个地名一般是行政地名，后面一个可能是行政地名，也可能是非行政地名。前者如“哈尔滨市　太平区”，后者如“昌平 小汤山”。由于我们在词典中收入了中国县以上的所有行政地名，这些行政地名为新地名的识别提供了重要线索。

（2）前面出现带处所宾语的动词和介词。地名是处所词的一种，因此能带处所宾语的动词和介词都可以出现在地名之前。这些词如：

在 到 至 向 往 从 位于 经 经过 由 去 经由 通过 家住 住 坐镇 奔 抵达

例如：

洪峰 2 7 日下午 5 时已通过城陵矶

家住建设街的孤寡老人王淑芝大娘

（3）几个地名并列。如：

组织家属去兴隆山、刘家峡、鲁沟等名胜参观。

（4）前面是“的”，“的”前是一个地名或处所，表示所属关系。如：

大连金石滩国家旅游度假区的神月湾畔

根据以上的分析，我们制定了以下的地名生成算法。该算法的基本思想是：以地名后缀为触发条件，在遇到地名后缀时，提出一个地名候选词。然后检查候选词的上下文，如果发现上下文复合地名上下文的特征，则将该地名候选词加入候选词集中。具体算法描述如下：

输入： 一个句子经过切分后的候选词集，每个 token 对应一个候选词串。候选词集用数组 wordCandidateSet[0..n]表示，其中每个元素是指向一个链表的指针，链表中存储从同一个汉字开始的所有的候选词。

资源： 地名后缀表 locSuffixList，带处所宾语的介词和动词表 locVerbPrepList。

过程：

1）i = 0;

2) 检查 wordCandidateSet[i]中每个候选词 wi， 如果 wi ∈ locSuffixList

如果在 wi 之前找到一个 j，满足：

a) idx-tempidx) <= 4 //地名主体最多 4 个汉字

b) wordCandidateSet[j]中有一个候选词 wj，满足：

wj 为地名

或 wj ∈ locVerbPrepList

或 wj 是顿号且 wordCandidateSet[j-1]中有候选词是地名

或 wj 是结构助词“的”且 wordCandidateSet[j-1]中有地名或处所词

则将 locCandidate 设为从 j+1 到 i 的汉字组成的字串, 并将 locCandidate 加入候选词集中。

3) 如果 i< n-1 则转 1），否则退出。

从上面的算法可以看出，前面生成的地名候选词会成为后面地名候选词的语境，例如：

组织家属去兴隆山、刘家峡、鲁沟等名胜参观

这里，三个地名并列在一起，第一个地名候选词“兴隆山”因为前面有动词“去”，所以被加入候选词集中，第二个地名候选词“刘家峡”前一个词是顿号，顿号的前面有一个地名候选词，这个地名候选词就是刚才加进去的“兴隆山”，这样就把“刘家峡”加入候选词集中。同样，“刘家峡”又成为下一个候选地名“鲁沟”的上下文条件。

3.6.3 中国地名识别结果

语料中共出现中国地名词 1447 词次，系统识别出 1463 词次，其中正确的 1396 词次。

召回率 = 1396 / 1447 = 96.5%

正确率 = 1396 / 1463 = 95.4%

F-measure = 2*0.965*0.954/(0.965+0.954) = 96%

词典中查到的 1315 词次，占总数的近 91%。把这部分去掉，真正动态识别的准确率和召回率分别是 55%和 61%，提高的潜力还很大。

主要错误原因是：

（1）缺少地名语境。如：

十几位细山村农民在雨中上船搬运碎石

担任中支村党支部书记兼吕巷管理区农技员

（2）地名后缀没有收入后缀表中。如：

记者随省委主要领导冒雨来到安庆市广成圩大堤、同马大堤

沉埋在印山２５００年的越王陵气势恢弘地呈现在越国子孙的眼前

（3）没有地名后缀。如：

房山石经成为轰动幽燕十六州的盛事

3.7 本章小结

本章描述了实语块分析前的切分、标注和命名实体识别。介绍了切分标注一体化的概率模型，并介绍了在这种概率模型中的命名识别策略。然后，详细描述了汉人姓名、西文译名和地名候选词语的生成算法，并给出了命名实体识别的实验结果。

第四章　浅层句法分析方法综述

4.1 引言

90 年代以来，浅层句法分析研究受到普遍关注，英语的有关研究很多，产生了一些新的方法，国内也有一些学者采用英语中的方法探索汉语的浅层句法分析。本章主要就在英语浅层句法分析中所应用的一些技术进行简要的介绍，并简单介绍汉语的有关研究。其中有些方法虽然是面向完全句法分析的，但由于其对完全句法分析的任务进行了分解，所以其技术也可以归入浅层分析的范畴。概括起来，句法分析的方法基本上可以分成两类：基于统计的方法和基于规则的方法。当然也可以采用规则和统计相结合的混合方法。下面第 2 节介绍基于统计的方法，第 3 节介绍基于规则的方法，第 4 节简要介绍汉语的有关研究。

4.2 基于统计的方法

随着语料库技术的发展，近 10 年来许多统计方法被用在短语识别和分析方面。这些方法的理论主要来自概率统计和信息论。以下将介绍其中具有代表性的几种方法：（1）基于隐马尔科夫模型的方法；（2）互信息方法；（3）ϕ^2统计方法；（4）基于中心词依存概率的方法。

4.2.1 基于隐马尔科夫模型(HMM)的方法

隐马尔科夫模型(Hidden Markov Models,HMMs)是从语音识别中发展出来的一种统计技术(Rabiner,1989)，它提供了一种基于训练数据提供的概率自动构造识别系统的技术。一个隐马尔科夫模型包含两层：一个可观察层和一个隐藏层，这个隐藏层是一个马尔科夫过程，即是一个有限状态机，其中每个状态转移都带有转移概率。在语音识别中，可观察层是声音片段的序列，隐藏层是用音素序列表示的词的发音的有限状态模型。用口语录音片段及其转写(transcription)作为训练数据训练HMM，它就可以用作识别器，用于识别未训练过的声音片段，从而生成口语的转写形式。

计算语言学家最早把HMM技术应用于英语的词性标注并取得了极大的成功，仅依靠简单的统计方法就可以达到 95%左右的正确率。在词性标注中，可观察层是词的序列，隐藏层是词类标记的序列，训练数据是标注好词性的文本语料，经过训练的HMM就成为自动标注器，它可以给只包含词序列的文本中的每个词标注上词类标记。

Church(1988)进一步把HMM用于识别英语中简单的非递归的名词短语，他把短语边界识别化为一个在词类标记对之间插入NP的左边界（"["]和NP的右边界(")"）的问题。如果不考虑空短语（即"[]）和短语的嵌套（如"[["，"]]"，"）[["等]，那么在一对词类标记之间只有四种情况：(1) [；(2)] ；(3)][；(4) 空（即无NP边界）。进一步可以把最后一种分为两种情况：(a)无NP边界但在NP之内（I）；(b)无NP边界但在NP之外（O）。这样任意一对词类标记之间就只存在 5 种可能的状态：(1) [；(2)] ；(3)][；(4) I；(5) O。Church的方法是：首先，在标注词性的语料中人工或半自动标注NP边界，以此作为训练数据，然后统计出任意一对词类标记之间出现以上 5 种状态的概率。统计得到的概率就成为短语边界标注的根据。这实际上把短语边界的识别变成了一个与词性标注类似的问题。如：

```
输入：      $  the  procecutor  said   in  closing  that          （词序列）
               DT      NN        VB    IN    NN      CS          （词性序列）
输出：  <$,DT>   <DT,NN>  <NN,VB> <VB,IN> <IN,NN><NN,CS>  （词性标记对）
           |         |        |       |       |       |
           [         I        ]       O       [       ]     （NP 边界）
```

4.2.2 互信息方法

互信息(mutual information)是信息论中的一个概念(Fano,1961)，它用来度量一个消息中两个信号之间的相互依赖程度。二元互信息是两个事件的概率的函数：

$$MI(X,Y) = \log_2 \frac{P(X,Y)}{P(X) \times P(Y)} \tag{1}$$

我们可以把词类序列看成随机事件，这样就可以计算一对词类标记之间的互信息。如果 X 和 Y 在一起出现的机会多于它们随机出现的机会，则 $P(X,Y)>>P(X) \times P(Y)$，即 $MI(X,Y)>>0$；如果 X 和 Y 是随机分布的，则 $P(X,Y) \approx P(X) \times P(Y)$，即 $MI(X,Y) \approx 0$；如果 X 和 Y 是互补分布的，则 $P(X,Y)<<P(X) \times P(Y)$，即 $MI(X,Y)<<0$。互信息值越高，X 和 Y 组成短语的可能性越大，互信息值越低，X 和 Y 之间存在短语边界的可能性越大。

为了确定句子中短语的边界，不能局限于 bigram（两个符号的组合）内部的互信息，需要看更多的上下文，即把二元互信息扩展为 n-gram（n 个符号的组合）内部的互信息。Magerman & Marcus(1990)提出了广义互信息(generalized mutual information) 的概念，它根据两个相邻的词类标记的上下文（在一个观察窗口内）来决定它们之间是否是一个短语边界所在。在下面的公式中，MI 表示二元互信息，MI_n 是一个向量，它表示 n-gram($x_1 \cdots x_n$)中任意两个部分之间的互信息，MI_n^k 表示这个向量中的第 k 个分量（1≤k<n），它表示 $x_1 \cdots x_k$ 和 $x_{k+1} \cdots x_n$ 之间的二元互信息。

一个 n-gram($x_1 \cdots x_n$)内部有 n-1 个二分切分点，每一个切分点的二元互信息为：

$$MI_n^k(x_1 \cdots x_n) = MI（x_1 \cdots x_k , x_{k+1} \cdots x_n） \tag{2}$$

$$= \log \frac{P(x_1 \cdots x_n)}{P(x_1 \cdots x_k)P(x_{k+1} \cdots x_n)} \tag{3}$$

在公式（3）中，对于每个 MI_n^k(k=1,2, …, n-1)，分子都相同，当分母最大时，互信息值最小。

基于互信息的短语边界划分的理论基础是：在 n-gram 中，局部广义互信息值最小的一对标记之间就是短语边界所在的位置。理论推导参见 Mamerman & Marcus (1990)。

在 n-gram($x_1 \cdots x_i, y_1 \cdots y_j$) (1≤i<n, 1<j<n, i+j=n) 内部，以两个相邻的词类标记 x_i 和 y_1 之间为界，共有 i×j 个二元组合(bigram)，要计算其间的互信息，应当综合考虑每一个 bigram 之间的二元互信息，因此产生了广义互信息的概念。广义互信息的计算公式是：

$$GMI_{n(i+j)}(x_1 \cdots x_i, y_1 \cdots y_j) = \sum_{\substack{X以x_i结束 \\ Y以y_1开始}} \frac{1}{\sigma XY} MI(X,Y) \tag{4}$$

这里，$GMI_{n(i+j)}(x_1 \cdots x_i, y_1 \cdots y_j)$表示在一个 n-gram 中两个相邻的元素 x_i 和 y_1 之间的广义互信息，X 表示 n-gram 中以 x_i 结束的词类标记串，Y 表示 n-gram 中以 y_1 开始的词类标记串。σXY 是 XY 中 $MI_{|XY|}^k$（$|XY|$ = i+j，0<k< i+j）的标准差。

4.2.3 ϕ^2统计方法

Gale & Church(1991)用ϕ^2统计方法来度量两个词之间的关联度。Chen & Lee(1995)用这种方法来确定短语的边界。

对于两个词 w_1 和 w_2，可以建立如下的联立表(contingency table)：

	W_2	$!W_2$	Σ
W_1	a	b	a+b
$!W_1$	c	d	c+d
Σ	a+c	b+d	a+b+c+d

在上表中，a 表示串 w_1 w_2 出现的次数，b 表示不在 w_1 w_2 中的 w_1 的出现次数，c 表示不在 w_1 w_2 中的 w_2 的出现次数，d 表示既不是 w_1 又不是 w_2 的词的次数。a+b 是 w_1 的出现次数，a+c 是 w_2 的出现次数，c+d 是非 w_1 的总词次，b+d 是非 w_2 的总词次，N=a+b+c+d 表示语料库中的总词次。根据上面的联立表，ϕ^2统计量定义如下：

$$\phi^2 = \frac{(a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \qquad (5)$$

当 a=0 时，ϕ^2近于 0，即当 w_1 和 w_2 从不共现时，ϕ^2取极小值。当 b=c=0 时，ϕ^2=1，即当 w_1 和 w_2 总是共现时，ϕ^2取极大值。ϕ^2值越大，说明 w_1 和 w_2 共现的机会越多，相反，ϕ^2值越小，则说明 w_1 和 w_2 共现的机会越少。

如果把上面的两个词换成两个词类标记，则可以进行标记对之间的ϕ^2统计。进一步推广则可以在一个词类序列的两个子序列之间进行ϕ^2统计。

4.2.4 基于中心词依存概率的方法

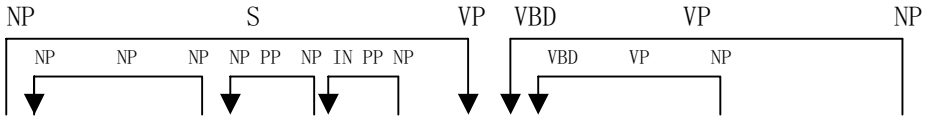
Collins(1996)提出了一种基于分析树中中心词之间依存概率的统计分析算法，该方法的要点是：把分析树归结为一个非递归的基本名词短语（base noun phrase, 简称 base NP）集合及依存关系的集合。在这些依存关系中，base NP 中除了中心词其他词都被忽略，所以依存关系就是 base NP 的中心词和其他词之间的依存关系，依存概率可以通过树库中的统计得到。分析算法是一个自底向上的线图分析器，利用动态规划来查找训练数据中所有的依存关系空间。

例如，由（a）中句子的分析树(b)可以得到 base NP 的集合 B 及中心词之间的依存关系集合 D(c)。

（a）John/NNP Smith/NNP, the/DT president/NN of/IN IBM/NNP,announced/VBD his/PRP$ resignation/NN yesterday/NN .

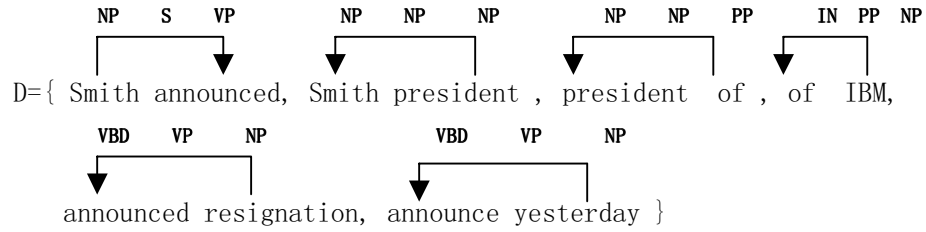
（b） (S (NP (NP John/NNP Smith/NNP),(NP (NP the/DT president/NN)
(PP of/IN IBM/NNP)) ,)
(VP announced/VBD (NP his/PRP$ resignation/NN) yesterday/NN) .)

(c)



[John **Smith**] [the **president**] of [**IBM**] announced [his **resignation**] [**yesterday**]

B={[John Smith],[the president],[IBM],[his resignation], [yesterday]}



基于以上模型的分析器的分析过程可以描述如下：对于一个输入句，首先分析出其中的 base NP，然后根据训练数据中得到的依存概率计算各个依存关系集合的可能性，可能性最大的依存关系集合就成为句子成分结构的最佳估计。由于依存关系表示为一个三元组，因此依存关系集合和 base NP 集合就可以映射为句子的短语结构树。

Collins 的算法是以大规模树库为基础的，而且以完全句法分析为目标。但这种方法以基本名词短语识别为前提，而且其利用具体词之间依存概率的思想和把短语归结为其中心词的思想对于部分句法分析也是很有借鉴意义的。

4.3 基于规则的方法

规则方法就是根据人工书写的或（半）自动获取的语法规则标注出短语的边界和短语的类型。根据标注策略的不同可以把规则方法分为两种：（1）增加句法标记法(incremental/constructive approach)，即在词串中插入短语边界和短语类型等句法标记。（2）删除句法标记法(reductionist approach)，即从多个候选的句法标记中删除不合法的标记。

4.3.1 增加句法标记法

增加句法标记的句法分析包括一个状态转换器(transducers)序列，转换器由正则式(regular expression)构成，即语法规则是有限状态语法的形式。大部分的规则系统都采用这种方法，如 Abney 的语块分析系统 CASS(Abney,1991,1996b)， Mokhtar & Chanod(1997)等。下面以 Abney 的有限状态层叠(finite-state cascades)方法为例对此加以说明。

有限状态层叠包括多个层级，分析逐层进行。每一级上短语的建立都只能在前一级的基础之上进行，没有递归，即任何一个短语都不包含同一级的短语或高一级的短语。

分析过程包括一系列状态转换，用 T_i 表示。通常的状态转换操作的结果是在词串中插入句法标记，而有限状态层叠则在每一级转换上用单个的元素来替换输入串中的一个元素序列，就跟传统的自底向上的句法分析一样。

每一个转换定义为一个模式的集合。每一个模式包括一个范畴符号和一个正则式。正则式转换为有限状态自动机，模式自动机合在一起就产生一个单一的、确定性的有限状态层级识别器(level recognizer)T_i，它以 L_{i-1} 级的输出为输入，并产生 Li 作为输出。在模式匹配过程中，如遇到冲突（即两个或两个以上的模式都可以运用），则按最长匹配原则选择合适的模式。如果输入中的一个元素找不到相应的匹配模式，则把它直接输出，继续下一个元素的匹配。

例如，给定下面的规则：

T_1: { NP --> (D) A* N^+
VP --> V-tns | Aux V-ing
NP --> Pron }

T_2: { PP --> P NP }

T_3: { S -->PP* NP PP* VP PP* }

对输入句" the woman in the lab coat thought you were sleeping."的分析过程如下所示：

L_3 ----------S ----------S (T_3)
L_2 ---------NP ----------PP VP NP ----------VP (T_2)
L_1 ---------NP P ----------NP VP NP ----------VP (T_1)
L_0 D N P D N N V-tns Pron Aux V-ing
the woman in the lab coat thought you were sleeping
0 1 2 3 4 5 6 7 8 9

识别器 T_1 从第 0 个词开始，在 L_0 级上进行匹配，在到达位置 2 时，得到一个与 NP 模式相匹配的状态序列，于是在 L_1 级上，从位置 0 到位置 2 输出一个 NP。然后从位置 2 重新开始，因为没有与之匹配的模式，所以把 P 直接输出。然后又从位置 3 开始，在位置 5 和位置 6 上分别有一个与 NP 模式相匹配的模式，这时采用最长匹配，于是在 L_1 级上从位置 3 到 6 输出一个 NP，然后又从位置 6 继续匹配。

4.3.2 删除句法标记法

这种方法的思想来自词性标注。在词性标注中，首先从词典中查出每个词可能具有的所有词性，然后根据上下文来消歧，从中选择一个正确的词性。这种思想用到句法标注上就是首先标注出每个词可能的句法功能，然后根据上下文来消歧，从中选择出一个正确的句法功能标记。也就是说，句法分析包括两个主要步骤：

（1）给出输入词可能的句法功能标记（与上下文无关，可能有多个候选）；

（2）删去在上下文中不可接受的句法标记，或从几个候选中选出一个最合理的句法标记（即同时排除其他标记）。

这样，句法分析实际上成了一个删除在上下文中不合法的句法标记过程。下面举例加以说明。

输入句： others moved away from traditional jazz practice.

经过词性标注后，加上可能的句法标记：

```
"<others>"          "other"         PRON        @>N     @NH
"<moved>"           "move"          V           @V
"<away>"            "away"          ADV         @>A     @AH
"<from>"            "from"          PREP        @DUMMY
"<traditional>"     "traditional"   A           @>N     @N<     @NH
"<jazz>"            "jazz"          N           @>N     @NH
"<practice>"        "practice"      N           @>N     @NH
```

标记注释：@>N (前定语) @N< (后定语) @NH (NP 核心) @>A (前状语)
@A< (后状语) @AH (副词短语核心) @V (动词和助动词) @DUMMY (介词)

在上面的例子中，第 1 列是词语，第 2 列是词的原形，第 3 列是词性标记，第四列是句法标记。所有的句法标记都以@开头。如果一个词有两个或两个以上的句法标记，则说明它在句法上是有歧义的。在句法分析过程中根据语法规则进行消歧。如果一个词只有一个标记，则不运用规则，如果一个词虽有歧义的标记，但没有与之匹配的规则则保留歧义。即分析结束后不保证每个词都只有一个句法标记。

规则采用所谓限制语法(Constraint Grammar)的形式。如：

```
REMOVE (@>N) (*1C <<< OR (@V) OR (@CS) BARRIER (@NH))
```

这条规则的含义是：如果上下文满足下面的条件则从有歧义的词中删去前定语标记@>N：其右边某个词（*1，*表示一个或多个，1 表示右边）是非歧义的（C），这个词是句子边界(<<<)、动词(@V)或主从连词(@CS)，并且该词和当前词之间没有哪个词有@NH 标记。

这些规则主要是人工书写的(Voutilainen,1993;Voutilainen&Padro,1997)，但完全靠人工总结这些上下文限制规则十分费时费力。因此有必要研究从语料库中自动获取这些语法规则的方法(Brill,1995; Samuellson et al.,1996)。

4.3.3 语法规则的自动学习

在基于规则的方法中，主要的困难在于语法规则的获取以及语法规则之间的优先顺序排列。Eric Brill (1995)提出了一种基于转换的错误驱动的学习方法，这种方法首先被用于词性标注，得到的结果可以和统计方法相媲美。Ramshaw & Marcus(1995)把这种自学习方法用于识别英语中的基本名词短语(base NP)。这种方法通过学习得到一组有序的识别基本名词短语的规则。另一种语法规则自动获取的方法是采用机器学习中基于实例的方法（instance-based learning）或基于记忆的方法(memory-based learning)，如 Cardie & Pierce(1998)和 Argamon et al.(1998)。下面首先介绍基于转换的学习方法，然后介绍基于实例的方法。

4.3.3.1 基于转换的规则学习方法

如图 4-1 所示，基于转换的学习方法以下列三部分资源为基础：（1）带标注的训练语料库。对于 base NP 识别任务来说，训练语料要标注出其中所有正确的 base NP（在此之前当然要先标注词性）。（2）规则模板集合。规则模板集合用于确定可能的转换规则空间。（3）一个初始标注程序。

基于转换的错误驱动的学习算法是：

（1）初始标注。把训练语料中所有的 base NP 标记去掉，用一个简单的初始标注程序标注出训练集中可能的 base NP。把这个结果作为系统的底线(baseline)。

（2）生成候选规则集。在每个初始标注错误的地方，规则模板便用来生成候选规则，规则的条件就是词的上下文环境，动作就是改正错误标记所要做的动作。

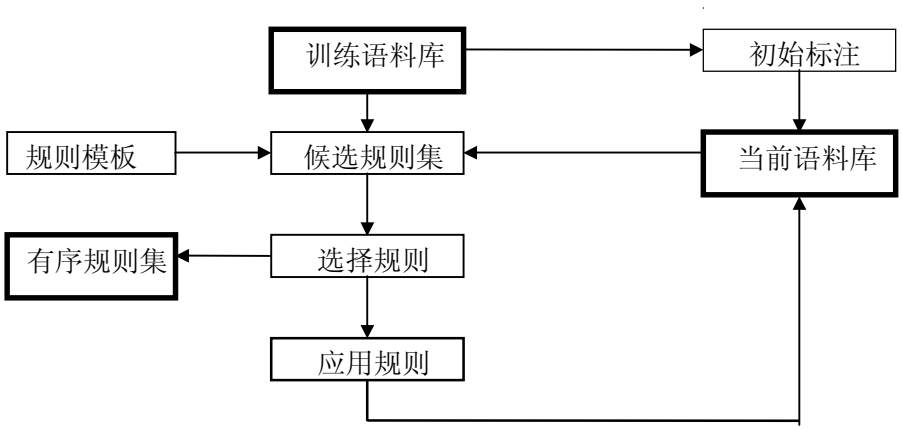


图 4-1　基于转换的错误驱动的学习过程

（3）获取规则。把候选规则集中的每条规则分别运用于初始标注的结果，选出得分最高的规则（得分为正确的修改数减去错误的修改数得到的结果）。把这条规则运用于初始标注的结果作为下一轮循环的基础，并把这条规则作为规则序列中的第一条规则输出。重复以下过程直到得分最高的规则的得分为 0 或低于某个阈值为止：获取候选规则集，给其中每条规则打分，选择得分最高的规则输出到规则集中，并把这条规则作用于当前语料库。

通过以上的自动学习过程就可以得到一个有序的规则集。base NP 识别的过程是：首先运用初始标注程序标注出输入句中可能的 base NP，然后顺序运用规则集中的规则对初始标注的结果进行转换操作。

4.3.3.2 基于实例的规则学习方法

前面所介绍的基于转换的学习方法在学习过程之后得到的是识别短语的规则，这样的的规则描述在什么条件下一个序列是一个基本名词短语，在什么条件下不是一个基本名词短语。而基于实例的学习方法是通过学习得到一组短语的组成模式，分析的时候利用这样的模式去和文本中的词类序列进行匹配。

Cardie & Pierce(1998)把标注好短语信息的语料库分为两个部分，一部分用于训练，另一部分用于剪枝。首先从训练的语料中得到一组名词短语的组成模式规则，然后把得到的这些规则应用到剪枝的语料中，对这些规则进行打分。比如，如果一个规则识别出一个正确的短语得 1 分，识别出一个错误的短语得-1 分，这样根据每条规则的总的得分情况对规则进行删减，去掉那些得分低的规则。最后得到的一组规则能保证得到较高的正确率。应用这些规则来识别文本中的名词短语的方法很简单，就是简单的模式匹配方法，在遇到规则冲突时，采用最长匹配原则。

Argamon et al.(1998)并不是通过学习过程显性地得到一组短语的组成模式（词类序列及其上下文环境），这些模式隐含在标注好短语的实例中。在训练阶段，把标注好词性和短语边界的语料用一种可以快捷检索的数据结构存储起来，在识别阶段，拿文本中的词类序列和训练语料中的实例进行匹配：把句子中的每个子串作为候选，对于每个候选，通过查找实例库计算它的概率分数，对分数高于某个阈值的候选予以保留。这一技术的关键在于候选子串和实例的匹配，因为子串可能是若干词的序列，而且还要考虑上下文，如果拿整个子串去和实例匹配的话就有严重的数据稀疏问题，这是基于事例的方法中普遍存在的一个问题。他们提出的一种覆盖(cover)技术较好地解决了这一问题。覆盖技术的基本思想是对待分析的串（包括上下文）进行分解，把它分解成若干更小的子串，利用这些子串去匹配。最后找到一个覆盖原串的子串集合，这些子串的总的概率分数最高。这种基于实例的方法把语法规则隐含在标注好的实例之中，跟前两种学习方法相比，它并没有一套显性的用于识别的语法规则，所以这种方法似乎更像基于统计的方法。

4.4 汉语的有关研究

近年来，中国学者也开始借鉴国外的方法进行汉语浅层句法分析的探索。李文捷等（1995）用短语边界与词性标记对共现概率的方法研究汉语中最长名词短语的识别。首先在训练集中统计 NP 起始和 NP 终止两个概率矩阵，然后根据这些概率信息在输入句的词性标记对之间插入 NP 起始标记和 NP 终止标记，然后对标记进行匹配处理。张国煊等（1995）用简单的互信息方法划分短语边界，郭志立等（1996）用互信息方法确定汉语“的”字短语的边界。这些研究都是基于统计方法的。

孙宏林（1997）用规则方法识别汉语的 VO（动宾）结构，刘长征（1998）采用规则方

法识别由名词序列构成的 NP。这两项研究中的语法规则是根据语料库中的统计通过人工归纳得到。

赵军（1998）系统地研究了汉语基本名词短语的识别和分析。在识别方面，从预先定义的句法模板（组成 base NP 的词类序列）出发，探讨了两种 base NP 识别方法：一种是统计的 N 元模型，该模型利用了 base NP 组成成分的词性信息、音节信息及上下文信息，研究表明这种模型比单纯的基于词类序列的模型要好。另一种是规则方法，其规则通过基于转换的学习算法从训练语料中自动获取（赵军、黄昌宁,1999）。

周强（1996）的研究目标虽是句子的完整句法分析，但其第一步是短语边界的初界定，利用从树库中得到的统计信息确定一个词前后的边界类型。穗志方（1998）的目标也是完全分析，这种方法的基本思想是：首先确定谓语中心词，然后围绕谓语中心词进行自底向上的组块分析，以确定谓语中心词的支配成分。其中的组块分析如果独立出来，就是一个部分句法分析器。在这个组块分析过程中主要利用了词语之间的依存关系。

4.5 本章小结

本章对近年来在浅层句法分析领域出现的一些有代表性的方法进行了总结，由此我们可以看到该领域近年来的基本动向。九十年代以来，在自然语言处理的技术发展中出现了两个十分突出的现象：一是统计方法得到了广泛的应用，二是有限状态方法进一步扩大了其应用领域。这两种现象都在浅层句法分析领域得到了充分的反映。另外值得注意的是，采用机器学习方法进行语言知识的自动获取研究成为一个十分明显的趋势。从大规模树库中获取语法规则以及规则的统计信息，使规则方法和统计方法相互融合成为新的研究动向。这都为我们的汉语实语块分析提供了有益的启示。

第五章　汉语实语块分析规则与算法

5.1 实语块分析规则

5.1.1 实语块类型

实语块按功能分为以下几类：（1）名词性短语（NP）；（2）简单动词性短语（V）；（3）复杂动词性短语（VP）；（4）形容词性短语（AP）；（5）区别词性短语（B）；（6）副词性短语（DP）；（7）主谓结构（S）；（8）非结构（NO）。

这里的短语分类与北大“现代汉语短语结构规则库”（以下简称“规则库”，详见[詹卫东，2000]）中的分类标准是一致的，即都以短语的功能为标准。两个分类体系的对应关系见表 5-1（“规则库”中涉及虚词的短语，如介词结构，因与实语块分析无关，故忽略不计）。

表 5-1　两种短语类型体系对照表

实语块短语类型	“规则库”短语类型
名词性短语	名词性短语
简单动词性短语	动词性短语
复杂动词性短语	动词性短语
形容词性短语	形容词性短语
区别词性短语	形容词性短语
主谓短语	单句型短语
副词性短语	副词性短语

从表 5-1 中可以看出，实语块分析中的短语分类体系与“规则库”中的相比，有三点小的差异：

（1）“主谓短语”与“单句型短语”只是名称不同，实际含义相同。

（2）“规则库”中的动词性短语在我们的体系中被分化为两类：简单动词性短语和复杂动词性短语。

（3）“规则库”中的形容词性短语在我们的体系中被分化为两类：区别词性短语和形容词性短语。

简单动词性短语是指由两个词组成的动词性短语（述宾结构除外）。之所以要把这些动词性短语分离出来单立一类，主要是因为这类动词性短语在功能上与单个动词更接近。

简单动词性短语的主要类型有：

(a) 述补式。如：

走上市场　　迈出更大步伐　　买到粮食　　评为优秀教师

打好基础　　摸清家底　　看准目标　　炒高股价　　打牢基础

发展迅猛　　掌握不当　　认识不足　　把握好时机　用足政策

（b）并列式。如：

改革开放　审议批准　解释说明　保值增值　参政议政

库存积压　培养提高　推广应用　信任支持　暴涨暴跌

（c）状中式。如：

正式挂牌　迅速回落　稳定发展　公开发行　积极参与　严厉打击
婉言谢绝　巍然兴起　更为加剧　高度重视　最为关心　大大降低
当时公布　今年到期　现场指挥　整体推进　法制化　　具体化

这样做的目的是为了减少规则的歧义。例如，对于下面一条规则：

VP → VP NP

该规则表示一个动词性短语加上一个名词性短语可以构成一个动词性短语。对于“无党派人士”就有可能产生下面的分析结果：

[[无 党派]VP 人士]VP

当动词性短语区分为简单动词性短语（V）和复杂动词性短语（VP）两类之后，上面的规则就改写为：

VP → V NP

它表示一个简单动词性短语加上一个 NP 可以构成一个复杂动词性短语。而规则“VP → VP NP”是不成立的，即一个复杂的动词短语加上一个 NP 不能构成一个更大的动词性短语。这样，对于“无党派人士”就不会产生 VP 的分析结果。

从形容词性短语中分化出区别词性短语也是基于同样的考虑。形容词性短语的典型功能有：（1）作述语；（2）作补语；（3）作定语；（4）作状语（詹卫东，2000）。但是，区别词性短语只有功能（3），即它只能作定语，而不能作其他成分。区别词性短语的例子如：

指令性计划　地方性法规　群众性组织　速度型发展路子
消费性资金　混合型经济　封闭式运营　开发性扶贫　综合性大学
非常设机构　非高档商品　非基础性产业　非正常性亏损
假冒伪劣商品　定量定性分析　稳产高产农田　跨世纪工程

以上这些例子中划线部分都是区别词性短语。区分出这一类型之后，就可以减少跟形容词性短语相关的规则的歧义。

另外，值得一提的是，在我们的规则体系中，引入了一个特殊的非终结符“非结构”。下面的小节将解释为什么要引入这样一个特殊的非终结符。

5.1.2 非终结符 NO

在一般的上下文无关语法（CFG）中，规则 A → β 表示一个串 β 能够被一个非终结符 A 替换，而且这种替换是无条件的，但事实上在自然语言中这种替换往往是有条件的，也就是说，β 并不是在任何情况下都能被 A 替换。这有两种情况：

（一）β 可以被另一非终结符 B 替换，这就是通常所说的句法歧义。如一个“ng + vg ”（名词+动词）序列既可以被 S（主谓结构）替换，也可以被 NP（名词性短语）替换。

（二）β 不能被任一非终结符替换，如一个“ng + vg ”序列在特定的上下文中有可能不是一个结构（详细讨论参见[马真、陆俭明，1996]）。

事实上，这两种情况的性质是相同的，都是句法歧义。但一般的 CFG 只能表示前一种情况，而不能表示后一种情况。也就是说，如果一个串 β 出现在规则的右部，则 CFG 假定它永远是成结构的，而不能说明在什么情况下 β 是不成结构的。例如，假定给定下面的规则：

NP → ng vg
VP → ng vg
S → ng vg

对于输入句“中国/nps 的/usd 铁路/ng 建设/vg 得/usf 不错/a”，其中的“铁路/ng 建设/vg”会生成三条边，它们的类型分别是 NP,VP 和 S。这三条边对输入句来说都是错误的。

基于约束的 CFG 则利用一些属性约束来处理非结构的问题，即在规则的约束中加入一些条件，说明在何种条件下 β 能够归结为 A，在何种条件下不能归结为 A。条件约束可以分为两类：一类是对一个结构内部组成成分的约束，也就是关于 β 的条件约束，一类是关于上下文的约束，也就是对 β 出现的上下文的条件约束。

例如，对于下面的规则：

NP → VP NP

加上下面的约束条件（詹卫东，2000）：

（1） VP 能够作定语；

（2） NP 能够作中心语；

（3） NP 是一个词；

（4） NP 前能够受动词修饰。

以上任何一个条件不满足，上面的规则都不能成立。

对于一个串 β，如果规则集中的所有规则 A → β 都不能成立，则说明 β 归结为一个隐性的非终结符，这就是“非结构”。我们引入非终结符 NO 就是要把这个隐性的范畴显性化。

引入一个特殊的非终结符来处理非结构的问题，把非结构看成一种特殊的结构，使得对非结构和结构的处理一致化，这样就使非结构处理问题得到系统的解决。

对于完全句法分析来说，不一定需要引入非终结符 NO。以前面所举的例子“中国/nps 的/usd 铁路/ng 建设/vg 得/usf 不错/a”来说，虽然从局部来看，“铁路/ng 建设/vg”不成结构的情况不能通过语法规则反映出来，但从全局来看，我们可以得到下面的分析结果：

[中国/nps 的/usd 铁路/ng]NP [建设/vg 得/usf 不错/a]VP

在这个分析结果中，就隐含了“铁路/ng 建设/vg”是非结构。也就是说，局部的非结构可以在更大的结构中反映出来。

本文的目标是进行实语块分析。由于它并不是完全分析，它只能在句子中的一个实词串中进行分析。正如我们在第一章指出的那样，一个实词串不一定能构成一个短语。如：

学习/vg 知识/ng 和/c 技能/ng

这里“ 知识/ng 和/c 技能/ng”是一个结构，但“学习/vg 知识/ng”并不是一个结构。

由于没有全局的结构分析，如果仅用成结构的规则来分析的话，就很难得到非结构的结果。如对上面的例子，仅就实词序列“学习/vg 知识/ng ”来分析，可能的结果有：

0: [学习/vg 知识/ng]VP

1: [学习/vg 知识/ng]NP

2: [学习/vg 知识/ng]B

这是用概率上下文无关语法(PCFG)分析得到的结果，按概率的大小排序，第 0 个结果的概率最大。下面的消歧也就只能在这些候选中进行选择，但是这些结果对输入句来说都是错误的。所以，在这种情况下，有必要把隐性的“非结构”显性化，以增加候选的空间。下面就是引入非终结符 NO 之后的 PCFG 分析结果，最右边表示边的概率。

0: [学习/vg 知识/ng]VP 0.325161

1: [学习/vg 知识/ng]NP 0.13523

2: [学习/vg 知识/ng]NO 0.0782199

3: [学习/vg 知识/ng]B 0.015

引入非终结符 NO 之后，增加了 1 条边，这样就真正穷尽了实词序列“学习/vg 知识/ng”在结构上的所有可能性。

从上面的例子可以看出，引入非终结符 NO 之后，可以处理各种非结构的情形。否则我们需要在应用每一条规则的时候都要加上一大堆条件，说明在什么条件下，这条规则不能成立。这些条件约束是很难描述的，在两类条件约束中，关于上下文的约束要比关于内部组成成分的约束更复杂，这是人的内省的知识难以胜任的。而很多消歧则需要依靠上下文。如在“学习知识和技能”中，离开上下文，“学习知识”无论如何都是合法的结构。

在实语块分析中，对一个实词序列β，我们首先要回答的问题是：它是否成结构？如果是，接着第二个问题是：它是什么类型？我们的实语块分析利用了概率上下文无关语法和概率属性相结合的概率语言模型（下一章将详细介绍）。在这样一个概率语言模型中，把非结构作为一种特殊的结构来处理，这样就可以同时回答这两个问题。如上面所举的例子，“[学习/vg 知识]NO ”和其他三条成结构的边一起竞争。在没有引入上下文信息的情况下，在三个候选边中，“[学习/vg 知识/ng]VP”的概率最大，但当引入上下文属性时，就可以得到下面的结果：

```
0: [ 学习/vg 知识/ng ]NO        0.0186106
1: [ 学习/vg 知识/ng ]VP        0.0159797
2: [ 学习/vg 知识/ng ]NP        0.00771012
3: [ 学习/vg 知识/ng ]B         1.5e-016
```

这时，“[学习/vg 知识/ng]NO ”就成为概率最大的边，这样就达到了消解歧义的目的。在概率消歧的过程中，我们没有为非结构的处理使用任何规则，也没有使用任何特别的处理手段，但可以得到很好的消歧效果。

5.1.3 规则形式

在我们的规则形式中，采用了二分的规则形式，即一个规则的右部有且只有两个符号。这意味着：

（1） 词类标记不直接上升为非终结符，即不采用下面的规则形式：

NP → ng

VP → vg

B → b

这样做的结果会导致规则数量的增加，但好处是可以减少规则的歧义。对于下面一条规则：

NP → ng vg (a)

如果采用上面的上升式规则，就会成为：

NP → NP VP (b)

规则(b)引起的歧义要比规则(a)大得多。例如：对于下面的句子：

新/a 技术/ng 呼唤/vg 新型/b 现代/t 企业/ng 组织/ng

如果用规则（a）的话，仅会生成一条边：

[技术/ng 呼唤/vg]NP

但如果用规则（b）的话，就会产生大量的NP边。假定跟规则右部NP匹配的只有两条边：

[技术/ng]NP

[新/a 技术/ng]NP

从“呼唤”开始的所有VP都可以和这两条NP边组合，如：

[[新/a 技术/ng]NP [呼唤/vg [新型/b [现代/t [企业/ng 组织/ng]NP]NP]NP]VP]NP

[[新/a 技术/ng]NP [呼唤/vg [[新型/b [现代/t 企业/ng]NP]NP 组织/ng]NP]VP]NP

[[新/a 技术/ng]NP [呼唤/vg [新型/b [[现代/t 企业/ng]NP 组织/ng]NP]NP]VP]NP
[[新/a 技术/ng]NP [呼唤/vg [新型/b [现代/t 企业/ng]NP]NP]VP]NP
[[新/a 技术/ng]NP [呼唤/vg 新型/b]VP]NP
[技术/ng [呼唤/vg [新型/b [现代/t [企业/ng 组织/ng]NP]NP]NP]VP]NP
[技术/ng [呼唤/vg [[新型/b [现代/t 企业/ng]NP]NP 组织/ng]NP]VP]NP
[技术/ng [呼唤/vg [新型/b [[现代/t 企业/ng]NP 组织/ng]NP]NP]VP]NP
[技术/ng [呼唤/vg [新型/b [现代/t 企业/ng]NP]NP]VP]NP
[技术/ng [呼唤/vg 新型/b]VP]NP

从规则数量来说，不使用上升式规则，并不会使规则增加多少。在我们的实语块分析规则中，共涉及到 19 个词类标记，8 个短语标记，采用二分规则形式，可能的规则数量是 27×27×8 = 5832，但我们在 20 万词的语料库中统计得到的规则数量只有 699 条（见 5.2 节），因为有很多不合法的组合。

（2） 对多项并列结构强制进行二分处理。

在实语块中，两个成分组合成一个更大的成分，采用二分的规则形式是十分自然的。唯一需要多分的是多项并列结构。对此，我们做出了下面的规定，三项并列结构切分为：

[[A B] C]

如：

[[新型/b 高效/b]B 低毒/b]B 的/usd 治虫药/ng
[[高产/b 优质/b]B 高效/b]B 农业/ng

四项并列结构切分为：

[[A B] [C D]]

如：

[[交通/ng 运输/ng]NP [邮电/ng 通信/ng]NP]NP
[[山峰/ng 奇石/ng]NP [海滩/ng 岛礁/ng]NP]NP

从数量上看，中间不含标点和连词多项并列结构在文本中并不多见，在我们统计的规则实例中，多项并列结构只占总数的 1 ‰。

从理论上看，对多项并列结构进行二分处理也是合理的。如把由三个区别词组成的并列结构“新型/b 高效/b 低毒/b”分析成一个区别词性短语加上一个区别词，从功能的角度看，区别词性短语和区别词的功能是一致的，所以它们并列在一起仍然产生一个区别词性短语。

采用二分的规则形式会带来一些计算上的好处。例如，我们在计算上下文属性的概率时，会利用下面的条件概率：

$$P(\beta \rightarrow A \mid \beta, Context = c)$$

它表示在给定 β 的条件下和上下文的条件下，β 归结为 A 的概率。假设 β 为“ng vg”（“名词 + 动词”序列），上下文 Context 为 β 左边一个词的词性，c 等于“usd”（结构助词“的”），A 等于 S（主谓结构），这个条件概率表示：一个“名词 + 动词”序列，当它左边的词为结构助词“的”的时候，这个序列归结为主谓结构的概率。采用最大似然估计，上下文属性的概率估计为：

$$P(\beta \rightarrow A \mid \beta, Context = c) = \frac{Count(\beta \rightarrow A, Context = c)}{Count(\beta, Context = c)}$$

就我们上面所举的实例，计算 Count(β，Context=c)实际上就是进行 trigram 的统计，因为 β 是一个由两个词类组成的序列（“ng vg”），Context 又是这个序列左边的一个词类，所以 Count(β，Context=c)就是计算序列“usd ng vg”的频次。显然，在这种情况下，β 越短，所需要的训练数据就越少。我们把限制 β 的长度为 2，则会给这种统计带来很大的方便。

5.2 实语块的语料标注和规则统计

5.2.1 语料标注

为了获取短语规则，我们对 20 万词（约 30 万汉字）的语料进行了实语块的标注。这些语料全部选自《人民日报》1993-1994 年经济类文章，共有 157 篇，203,499 个词例。为了使标注的结果具有一致性，我们制订了一个短语标注的规范（见附录）。对一些比较特殊的结构作了规定，如：

（1） 兼语结构“V 兼 NP VP”的层次划分为：[[V 兼 NP]VP VP]VP。例如：
[[授权 全国人大常委会]VP [审议 批准]V]VP

（2） 对于因为数量结构后置而产生的“vg + ng + 数量结构”，“vg + ng”标注 VP。例如：
这家企业 [生产/vg 味精/ng]VP 5 万吨，[完成/vg 产值/ng] VP 8 亿元

真实语料中有许多复杂的现象，有的层次划分不好确定，有的短语类型不好确定。我们对这些问题的处理还没有十分的把握，只好凭自己的感觉来决定。下面就层次划分和短语类型的确定分别各举一例说明。

（一） [NP_1 + vg + NP_2]NP 的内部层次

这类 NP 有两种可能的层次划分，如图 5-1 所示：

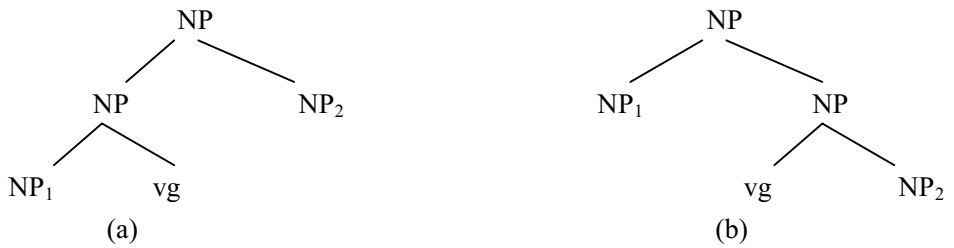


图 5-1 [NP_1 + vg + NP_2]NP 的两种层次划分

邢福义（1994）提出按照 NP_1 和 NP_2 音节的不同分别采取两种不同的层次切分，例如认为当 NP_1 和 NP_2 都是双音节时，应该按(a) 进行划分[7]。例如：

[[果树/ng 栽培/vg]NP 技术/ng]NP （a）

但似乎没有充足的理由说明（b）就不对。因为[NP_1 + vg]和[vg + NP_2]都成结构，而且都

[7]邢福义（1994）认为这里的[NP_1 + vg]是“受事性主谓结构”，但这有两个问题：

（1） NP_1 不一定是受事。如当动词是不及物动词时就谈不上受事，如：

外商投资企业 儿童入学年龄 汇率浮动幅度

有的动词虽是及物动词，但和 NP_1 不构成语义上的支配关系，如：

工农业生产计划 前期准备工作

（2）从功能上看，当 NP_1 是动词的受事时，[NP_1 + vg]只具有体词性，不具有谓词性。它不能作谓语，也不能独立成句，所以，我们认为是名词性短语更合适。但当 NP_1 是动词的施事时，就有可能被理解为主谓结构。

可以在其他语法位置上出现，例如：

NP 作宾语：　　提高 [栽培/vg 技术/ng]NP

NP 作主语：　　[栽培/vg 技术/ng]NP 有待　提高

在“的”前：　　[栽培/vg 技术/ng]NP 的 提高

（二）[NP　VP]短语类型的确定

[NP　VP]可以是名词性短语，也可以是主谓结构，这看起来好像很清楚，但在某些语法位置上，当主谓结构和名词短语都可以出现时，短语的类型就不好确定了。这些语法位置有：

（1）　在主语位置。如(括号中是作者的判断)：

外资进入势如潮涌　(S)

山东改革起步较早　(NP)

（2）在结构助词“的”前。如：

在国家计划执行的过程中　(NP)

在列车运行的过程中　(S)

是社会主义市场经济发展的前提条件　(NP)

（3）在方位词前。如：

在钢材涨价中发了一笔财　(S)

在经济改革中大显身手　(NP)

（4）在宾语位置。如果动词只能带名词性宾语，则不会产生问题，如：

冲破计划经济束缚

但如果动词既能带名词性宾语，又能带主谓结构作宾语的话，则会产生歧义，如：

促进两岸经贸发展　　支持部队训练

推动创建活动健康发展　　不反映价格变化

造成交通堵塞　　保证养老金足额发放

（5）在定语位置，如：

外商投资企业　　外汇保值服务

电力发展步伐　　汇率浮动幅度

5.2.2 规则统计

对上一节所述语料库中实语块标注的统计结果见表 5-2。

表 5-2　实语块规则统计结果

规则类型	规则数目	出现频次
AP	13	1130
B	28	295
DP	11	42
NP	117	21707
V	20	4839
VP	74	8701
S	70	2906
NO	366	54343
合计	699	93963

关于规则的统计，有一点需要说明：对于非结构 NO，实际上在语料中并没有显性地标出，我们只是根据下面这条规则抽取出来的：

对于两个相邻的边（包括词汇边）e_1，e_2，如果找不到一个边 e_3 覆盖 e_1，e_2，则生成一个 NO 边。（如果 e_3 的起点等于 e_1 的起点且 e_3 的终点等于 e_2 的终点，则 e_3 覆盖 e_1，e_2。）如：

[解决/vg [居民/ng [[吃/vg 菜/ng]VP 问题/ng]NP]NP]VP

在“解决”和“居民”之间，左边只有一条边：

(a) 词汇边：解决/vg

右边有两条边：

(b) 词汇边： 居民/ng

(c) 非终结符边：[居民/ng [[吃/vg 菜/ng]VP 问题/ng]NP]NP

(a)和(c)相邻，但有另一条边“[解决/vg [居民/ng [[吃/vg 菜/ng]VP 问题/ng]NP]NP]VP”覆盖了(a)和(c)，所以不产生一条 NO 边。

(a) 和(b)也相邻，但没有一条边可以覆盖(a)和(c)，所以生成一条 NO 边，即：

[解决/vg 居民/ng]NO

这样就产生了一个规则 “NO → vg ng ”的实例。

类似地，在“居民”和“吃”之间，左边只有一条边：

(d) 词汇边：居民/ng

右边有三条相邻的边：

(e) 词汇边：吃/vg

(f) 非终结符边： [[吃/vg 菜/ng]VP 问题/ng]NP]

(g) 非终结符边： [吃/vg 菜/ng]VP

（d）和（e）生成一条 NO 边，即： [居民/ng 吃/vg]NO

（d）和（g）生成一条 NO 边，即： [居民/ng [吃/vg 菜/ng]VP]NO

因为有一条边“[居民/ng [[吃/vg 菜/ng]VP 问题/ng]NP]NP ”覆盖了（d）和（f），所以（d）和（f）不产生一条 NO 边。

5.3 实语块分析算法

实语块分析系统采用线图分析（chart parsing）算法，在一个实词序列中根据语法规则自底向上生成所有可能的边。在生成边的过程中，利用概率上下文无关语法和概率属性相结合的概率模型（详见第六章）对每一条边给出概率评分，概率为零或极小的边将被删除。在一个实词序列中选择长度最长、概率最大的一条边作为输出。

关于线图分析算法，详细的介绍见[Winograd, 1983]。

5.4 本章小结

本章首先介绍了实语块分析中所用的规则。在规则部分，我们提出了一个特殊的表示非结构的非终结符 NO，用来处理实语块分析中的非结构问题。本章对非终结符 NO 的含义、意义及其在语料库中的统计进行了解释。然后，说明了在实语块语料中标注中所遇到的一些问题。最后，简要介绍了实语块分析所用的线图分析算法。

第六章　概率上下文无关语法和概率属性相结合的汉语实语块分析模型

6.1 概率上下文无关语法

6.1.1 概率上下文无关语法的定义

概率上下文无关语法（Probabilistic Context Free Grammar，简称 PCFG；或 Stochastic Context Free Grammar，简称 SCFG）是上下文无关语法的扩展，它在上下文无关规则中的每一条规则上加上一个条件概率。

PCFG 是一个五元组，G = (N, Σ, P, S, D)，这里，D 是一个函数，它给 P 中的每一条规则赋一个概率值。这个函数表示在给定非终结符 A 的条件下，A 改写为序列 β 的概率 p，它可以表示为：

$$P(A \rightarrow \beta) \qquad (6\text{-}1)$$

或：

$$P(A \rightarrow \beta \mid A) \qquad (6\text{-}2)$$

一个分析树 T 的概率定义为用来扩展 T 中每个结点 n 的规则 r 的概率的乘积。即：

$$P(T,S) = \prod_{n \in T} P(r(n)) \qquad (6\text{-}3)$$

6.1.2 概率上下文无关语法的局限性

概率上下文无关语法是上下文无关语法的扩展，它增加了对分析结果进行概率评分的功能，因而有助于对句法歧义的消解。上下文无关语法的致命缺陷是生成能力过强，因此在分析自然语言时会产生大量的歧义。概率上下文无关语法并不能减少歧义，而且对歧义的消解能力也有限，因为歧义产生的原因是多方面的，光靠规则的使用概率只能解决一部分歧义。具体说来，CFG 和 PCFG 的共同缺陷在于：

（1）不能反映规则的内部约束。规则的内部约束包括多个方面，如：成分（constituent）中的中心词；短语及其中心词的句法属性、语义属性等。

（2）不能反映规则的外部约束。规则的外部约束即上下文的约束，它反映规则在什么样的上下文环境中可以使用，在什么样的上下文环境中不能使用。

6.2 概率属性

6.2.1 什么是属性概率（probabilistic features）

针对 CFG 的缺陷，语言学家提出了许多基于约束的(constraint_based)语法模型，其中最著名的有词汇功能语法（Lexical Functional Grammar, LFG）（Bresnan, 1982）、中心词驱动的

短语结构语法（Head-driven Phrase Structure Grammar，HPSG）（Pollad & Sag, 1994; Sag & Wasow, 1999）、广义短语结构语法（Generalize Phrase Structure Grammar，GPSG）（Gazdar et al. 1985; Gazdar et al. 1988）。这些语法模型的缺陷是：（1）属性的描述非常困难；（2）在属性描述上是“说一不二”的，要么具有某个属性值，要么不具有某个属性值。它虽然能消除一些绝对不可能的组合，但对若干“合理”的结果，却没有更有效的手段来消解歧义。

针对这些语法模型的不足，一些计算语言学家提出了概率属性语法（probabilistic feature grammar, PFG）的思想。Brew(1995)提出了概率 HPSG 的设想，Goodman(1997, 1998)进一步提出了概率属性语法的思想。Abney(1997)讨论了概率属性值语法中的参数估计问题。概率属性语法用一个属性描述的集合来代替 PCFG 中的一个终结符或非终结符，其中，结点的范畴（PCFG 中的终结符或非终结符）也可以看成一个属性，因此 PFG 可以视作 PCFG 的超集。PFG 固然是一个理想的语法模型，但属性相关性造成的严重的数据稀疏问题使其难以实用。例如，对于下面的 CFG 规则：

NP → DET N

其 PCFG 表示是：

P (NP → DET N)

用 PFG 则表示为：

P (A1 = NP, A2 = singular, A3 = man → B1 = DET, B2 = singular,
B3 = the, C1 = N, C2 = singular, C3 = man)

图 6-1 显示了 PFG 和 PCFG 的差异。

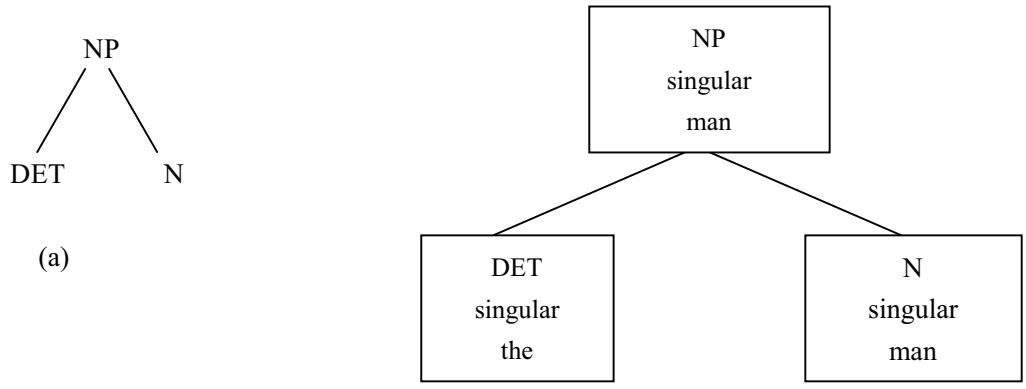


图 6-1 PFG 示意图

概率属性语法的思想是用一个属性描述的集合来代替 PCFG 中的单一标记。如图 6-1 所示，(a)是 PCFG 的规则形式，(b)是 PFG 的表示形式，(a)中的结点 NP 对应于(b)中的一个属性描述的集合，其中范畴属性的值是“NP”，性范畴属性的值是“singular”（单数），中心词属性的值是“man”。下面是 PFG 的生成过程：

P(B1 = DET | A1 = NP, A2 = singular, A3 = man)
≈ P(B1 = DET | A1 = NP, A2 = singular)

P(B2 = singular | A1 = NP, A2 = singular, A3 = man, B1 = DET)
≈ P(B2 = singular | A2 = singular, B1 = DET)

P(B3 = the | A1 = NP, A2 = singular, …, A3 = man, B2 = singular)
≈ P(B3 = the | B1 = DET, A3 = man)

…

P(C3 = man | A1 = NP, A2 = singular, …, C2 = singular)

≈ P(C3 = man | A3 = man, A1 = NP)

从上面的例子可以看出，概率属性语法存在严重的数据稀疏问题，尽管可以通过各种独立性假设减少属性之间的相关性，但属性相关带来的数据稀疏问题仍然相当严重。

本文的思想是：以 PCFG 为基础，在 PCFG 上加上概率属性约束。这里，各个属性是相对独立的，这样就避免了属性概率估计中的数据稀疏问题。

总的来说，属性包括两个方面：

（1） 上下文无关属性。在 A → β 中，β 中的部分或全部结点所具有的属性。如成分中心词的句法属性、语义属性等。

（2） 上下文有关属性。β 所在的上下文的属性。如前一个词的词性，后一个词的词性，前面第二个词的词性，后面第二个词的词性等。

在我们的实语块分析实验中，使用了三种属性：（1）结构的节律属性；（2）词的词汇功能属性；（3）上下文属性。其中，（1）、（2）是上下文无关属性，（3）是上下文有关属性。

6.2.2. 属性概率的估计

设 F 是与规则 A → β 相关的一个属性，F 的取值范围为 $f_1, f_2, \cdots, f_n$，规则 A → β 在语料库中的实例集为 E，那么 E 根据 F 可以划分为 n 个子集 $E_1, E_2, \cdots, E_n$，分别与 $f_1, f_2, \cdots, f_n$ 相对应，根据最大似然估计有：

$$= \quad _i \qquad \rightarrow \ \beta \ ) = \frac{|E_i|}{|E|} \qquad (6\text{-}4)$$

由此可以看出，属性 F 对 CFG 规则 A → β 的约束表现在：将 A → β 的状态空间 Ω 划分为 n 个互不相交的子空间 ω_1， $\omega_2, \cdots$， ω_n，我们把在 A → β 给定的情况下属性 F 取值为 f_i 的情形看作一个随机事件，并度量每个随机事件的概率。

6.2.3 PCFG + PF 的概率语言模型

设跟规则相关的属性集合为 FS，那么在句子的推导过程中，规则 r 的概率为：

$$P(A \rightarrow \beta \mid A) \times P \ \ FS \quad A \rightarrow \beta) \qquad (6\text{-}5)$$

这里，$P(A \rightarrow \beta \mid A)$ 表示 PCFG 模型，$P(FS \mid A \rightarrow \beta)$ 表示概率属性模型。其中

$$P(FS \mid A \rightarrow \beta) = \prod_{F \in FS} P(F = fi \mid A \rightarrow \beta) \qquad (6\text{-}6)$$

这里，f_i 为属性 F 在当前实例中的取值。

6.3 结构的节律属性

6.3.1 节律对句法的作用

节律指音节的长短。结构的节律属性（rhythm feature，简称 RF）是指一个结构在音

节上的组合特征。语法研究表明，韵律特征（包括重音、节律、语调等）对句法有一定的制约作用。例如，Quirk et al.(1985)指出节律对语序排列有一定作用，如有的并列结构的两个并列项不能调换位置，其中影响的因素之一是节律，如：

man and woman	* woman and man
ladies and gentlemen	* gentlemen and ladies

汉语中节律对句法的制约作用更为明显，如“种”和“种植”都是动词，而且意义相同，但只能说“种树”，不能说“种植树”,但是如果宾语是“果树、杨树、桃树”等，两个动词都能用，如：

种 树	* 种植 树
种 花	* 种植 花
种 草	* 种植 草
种 果树	种植 果树
种 棉花	种植 棉花

这些语言现象说明，汉语的述宾结构不大能接受“头重脚轻”的节律格局，也就是说，当动词的音节长度大于宾语的音节长度时，述宾结构一般不能成立(详细讨论见冯胜利，2000)。节律对句法结构是否成立的制约作用，不仅表现在述宾结构中，在偏正结构中也有类似的情况，如：在由“名词+动词”构成的名词性短语中，动词一般要求是双音节的，而且名词不能是单音节的，如：

棉花 种植	* 棉花 种	* 花 种植	* 花 种
果树 栽培	* 果树 栽	* 树 种植	* 树 栽

这种节律的制约往往并不是绝对的,如前面说述宾结构中动词的音节长度不能大于宾语的长度，但这并不是绝对的，例如，“害怕 光”,“整理 书”等就能够成立。这说明节律对句法有限制作用，但这种限制又往往不是绝对的，它往往表现出一种倾向性，很难用绝对的“是”或“否”来下断语。用概率可以比较恰当地描述这些现象，比如，假如我们这样来描述：在述宾结构中，动词的长度大于宾语长度的概率为 5%，动词与宾语长度相等的概率为40%,动词的长度小于宾语长度的概率为 55%，这样既能解释上面的规律，又不至于排斥小概率的现象。

本文基于节律对句法结构具有制约作用的性质，提出了利用概率化的节律属性对 FCFG 进行约束的思想。下面首先分析简单短语中的节律属性,然后再分析复杂短语中的节律属性。所谓简单短语是指其中的成分都是词，而复杂短语中至少一个成分是短语。

6.3.2 简单短语中的节律属性

一个词的音节属性取值有三种：

（1） 单音节；

（2） 双音节；

（3） 多音节。

由于我们采用了二分的规则体系，β 中只有两个结点。所以，在同时考虑两个词的音节属性时，共有 3×3 = 9 种可能的取值。规则 A →β 的节律属性的属性值集合为：

$$F = \{ (0, 0), (0,1), (0,2), (1,0), (1,1), (1,2), (2, 0), (2, 1), (2, 2) \}$$

这里，0 表示单音节，1 表示双音节，2 表示多音节。

根据(6-4)，对于规则 A →β，节律属性 RF 取值为 rf(rf ∈ F)的概率可以用(6-7)

来估计：

$$P(RF = rf \mid A \to \beta) = \frac{Count(RF = rf \mid A \to \beta)}{Count(A \to \beta)} \qquad (6\text{-}7)$$

表 6-1 显示了规则 NP → ng vg 和 NP → vg ng 的实例中词的音节组合分布。在 NP → ng vg 中，“双音节 + 双音节”组合所占的比例为：1275/1371 = 93%，在 NP → vg ng 中“双音节 + 双音节”组合所占的比例为：2328/2889 = 81%。也就是说，无论是动词直接作 NP 的修饰语还是动词直接作 NP 的中心语，短语的音节组合模式都主要是“双音节+双音节”模式。

表 6-1　短语中的音节组合分布（一）

规则	总次数	0,0	0,1	0,2	1,0	1,1	1,2	2,0	2,1	2,2
NP → ng vg	1371	0	4	0	0	1275	4	0	88	0
NP → vg ng	2889	13	10	0	401	2328	91	0	44	2

表 6-2 显示了歧义序列“vg + ng”在 NP 和 VP 中的音节组合分布。

表 6-2　短语中的音节组合分布（二）

规则	总次数	0,0	0,1	0,2	1,0	1,1	1,2	2,0	2,1	2,2
VP → vg ng	2777	826	640	49	80	1121	121	0	11	1
NP → vg ng	2889	13	10	0	401	2328	91	0	44	2

从表 6-2 可以看出，同样的词类序列，在不同类型的短语中，音节组合模式有很大差异。对于“vg + ng”序列，组成名词短语和动词短语的数量大致相当，但在音节组合模式的分布上存在很大差异。在名词短语中，分布主要集中于动词为双音节的，动词为双音节的动词短语总数为：401+2328+91 = 2820，占总数的 98%，动词为单音节的只有 13+10+0 = 23 个，占总数的 0.8%。也就是说，双音节动词进入这一格式的机会是单音节动词的 122 倍！在相应的动词短语中，动词为双音节的短语总数为：80+1121+121 = 1322，占总数的 48%，动词为单音节的短语总数为：826+640+49 = 1515，占总数的 55%。显然，在动词短语中，音节分布的差异并不明显。通过这个对比可以清楚地看出，利用结构的节律属性可以有效地帮助消解“vg + ng”序列的句法歧义。

表 6-3 显示了歧义序列“ng + vg”在 NP、NO 和 S 中的音节组合分布。

表 6-3　短语中的音节组合分布（三）

规则	总次数	0,0	0,1	0,2	1,0	1,1	1,2	2,0	2,1	2,2
NP → ng vg	1371	0	4	0	0	1275	4	0	88	0
NO → ng vg	6536	384	578	42	1131	3718	143	90	435	15
S → ng vg	470	28	1	2	17	347	22	2	43	8

从表 6-3 可以看出，在规则 NP → ng vg 中，vg 绝对不能是单音节，即概率为 0。“双音节 + 双音节”模式总数为 1275，占总数的 93%。在规则 S → ng vg 中，“双音节+双音节”模式仍占绝对优势，占 74%，但其他类型也有少量分布。其中 vg 为单音节的占 10%。在 NO → ng vg（即“名词+动词”不成结构）中，“双音节+双音节”模式虽仍占大多数，占

57%，但在其他音节类型上也有相当数量的分布，如动词为单音节的总数为 384+1131+90 = 1605，占总数的 25%。从统计可以发现，在“ng + vg”序列中，假设 ng + vg 只构成 NP,NO 和 S 三种类型，如果 vg 为单音节，那么它成为 NP 的概率为 0，成为 S 的概率 = 47/(1605 + 47) = 3%，它不成结构的概率 = 1605/(1605 + 47) = 97%。

表 6-4 显示了歧义序列“ng + ng”在 NP 和 NO 中的音节组合分布。

表 6-4　短语中的音节组合分布（四）

规则	总次数	0,0	0,1	0,2	1,0	1,1	1,2	2,0	2,1	2,2
NO → ng ng	2930	105	480	60	147	1646	149	27	298	18
NP → ng ng	5769	275	70	8	374	4464	260	23	263	32

“ng + ng”只有两种可能：（1）成结构，为名词短语；（2）不成结构。从上表可以看出，在“单音节+双音节”组合模式中，不成结构的概率要远大于成结构的概率。不成结构的概率为 480/（480+70）= 87%，成结构的概率为 70/（480+70） = 13%。

6.3.3 复杂短语中的节律属性

跟两个词之间的组合相类似，在一个短语和一个词组合或一个短语和另一个短语组合的过程中，节律同样具有制约作用。例如，对于下面的例子：

跨/vg　进/vg　三峡/nps　工程/ng　大门/ng

我们在用 PCFG 进行分析实验的时候，得到了这样的结果：

[跨　[[进　[三峡　工程]NP]VP　大门]NP]VP

这里的错误是：动词短语“[进　[三峡　工程]NP]VP” 修饰名词“大门”构成一个 NP，因为存在规则 NP → VP ng （如“种 粮 大户”)。我们经过统计发现，规则 NP → VP + ng 在语料库中共使用了 216 次，其中 VP 长度为 2 的 168 次，VP 长度为 3 的 30 次，VP 长度大于 3 的 18 次。由此可以看出，VP 修饰名词主要限于由两个词构成的述宾结构，如“种粮大户”、“学雷锋标兵”等，而由三个词组成的动词短语直接修饰名词的概率比较小。从这个例子可以看出引入短语的长度信息有助于句法消歧。

我们用所包含的词的数目来度量短语的长度。据此短语的长度属性可以有三种取值：0 表示词数为 2，1 表示词数为 3， 2 表示词数大于 3。这样，不管是词和词的组合还是词和短语的组合或短语和短语的组合，结构的节律属性都可以一致地用 3×3 的矩阵来描述。

6.3.4 节律属性在分析中的作用

（1） 节律属性可以帮助消解短语类型歧义。

例：该/r 校/ng 有/vg 9 0 0/mx 学子/ng 为/pg 国/ng 捐躯/vg 。/wd

对于该句中的实词序列“国/ng 捐躯/vg”，表 6-5 和表 6-6 分别给出了用 PCFG 模型和 PCFG+RF 模型的分析结果。

表 6-5　PCFG 模型分析例示（一）

实词序列	规则(A → β)	P(A → β ｜ A)
国/ng 捐躯/vg	S → ng vg	0.161679
国/ng 捐躯/vg	NO → ng vg	0.120273
国/ng 捐躯/vg	NP → ng vg	0.063159
国/ng 捐躯/vg	V → ng vg	0.011573

表 6-6　PCFG + RF 属性模型分析例示

实词序列	RF	规则(A → β)	P(RF= (0,1) | A → β)	P(A → β | A)* P(RF = (0,1))
国/ng 捐躯/vg	(0,1)	NO → ng vg	0.08843	0.010636
国/ng 捐躯/vg	(0,1)	S → ng vg	0.00292	0.000344
国/ng 捐躯/vg	(0,1)	NP → ng vg	0.00213	0.000184
国/ng 捐躯/vg	(0,1)	V → ng vg	0.0	0.0

从表 6-5 中可以看到，在 PCFG 模型中，它分析为 S 的概率最大。从在 6-6 中可以看到，它分析为 NO 的概率最大。表 6-6 中第二列表示实词序列的节律模式，（0，1）表示“单音节+双音节”，第四列表示节律属性概率，第五列表示 PCFG 概率和节律属性概率的乘积。

（2） 音节属性可以帮助剪枝，避免不必要的搜索。

利用音节属性，可以帮助剪枝，使不可能产生的边尽早删除，避免不必要的搜索，从而提高系统的效率。

例： 解决/vg 居民/ng 吃/vg 菜/ng 问题/ng 十分/dd 困难/a 。/wd

这个句子虽然只有 7 个词，但结构相当复杂，应用一般的 PCFG 后产生的边数为 1238，利用音节属性之后，使 chart 中产生的边数由 1238 减少为 348，减少了 73%。

由表 6-1 可见，在 NP → ng vg 中，P(RF=(1,0)) = 0，在 NP → vg ng 中，P(RF=(0,1))=0.003,)。利用节律属性概率之后，边“ [居民/ng 吃/vg]NP ”被删除，因为它的概率为 0。含“ [吃/vg 菜/ng]NP ”的边也被删除，因为它的概率很低，使包含它的边的概率变得极小。

表 6-7 给出了三种模型的实验结果，第一种模型是 PCFG 模型，第二种则是在 PCFG 上加上了简单短语中的节律属性，第三种模型是在 PCFG 上加上所有短语中的节律属性。从这三个模型的实验结果可以看出，加上 RF(1)比单纯的 PCFG 模型的性能有显著的提高，带标记的准确率和召回率分别提高了 4.5 和 5.1 个百分点。利用所有短语中的节律属性之后，带标记的准确率和召回率又进一步提高，分别比 PCFG 模型提高 8.2 和 8.8 个百分点。

表 6-7　音节属性对分析的作用

模型	带标记			不带标记		
	P	R	F	P	R	F
PCFG	51.3	64.7	57.2	66.5	83.8	74.1
PCFG +RF（1）	55.8	69.8	62.1	66.1	82.7	73.5
PCFG +RF（2）	59.5	73.5	65.8	67.9	83.9	75.0

6.4 上下文属性

上下文属性是指在应用规则 A →β 时，在β前后出现的成分的特征。为了减少数据稀疏问题，本文把上下文限定为β左边第一个词的词性和右边第一个词的词性，把左边第一个词的词性称为前文，记为 LT，把右边第一个词的词性称为后文，记为 RT。下面首先介绍上下文属性的两种概率估计，然后通过实验对这两种概率估计进行对比。

6.4.1 上下文属性的两种概率估计

上下文属性的概率可以有两种估计形式。第一种形式是：

$$P(A \rightarrow \beta \mid A, Context = c) \qquad (6\text{-}8)$$

这里，Context 表示上下文属性，c 是上下文属性的取值。例如，当 Context 为 LT 时，LT 的可能取值是词类标记集中的某个标记。当 Context 为 RT 时，RT 的可能取值也是词类标记集中的某个标记。(6-8) 中的条件概率表示在给定上下文条件下，A 被改写为β的概率。我们把它称为自顶向下的上下文属性概率。

上下文属性概率的第二种估计形式是：

$$P(\beta \rightarrow A \mid \beta, Context = c) \qquad (6\text{-}9)$$

(6-9)中的条件概率表示在给定上下文条件下，一个串β归结为非终结符 A 的概率。我们把它称为自底向上的上下文属性概率。

在 (6-8) 和 (6-9) 中，Context 分别可取 LT 和 RT，这样共有四种可能的上下文概率，分别记为 $P_{TD}(LT)$，$P_{TD}(RT)$，$P_{BU}(LT)$，$P_{BU}(RT)$ ， P_{TD}表示自顶向下形式的概率，P_{BU}表示自底向上形式的概率。表 6-8 表示了这四种上下文属性概率的差异。

表 6-8　四种上下文属性概率

推导方向 \ 前后文	前文	后文
自顶向下	$P_{TD}(LT)$	$P_{TD}(RT)$
自底向上	$P_{BU}(LT)$	$P_{BU}(RT)$

按 (6-8)，则用下面的公式来估计上下文概率：

$$P(A \rightarrow \beta \mid A, Context = c) = \frac{Count(A \rightarrow \beta, Context = c)}{Count(A, Context = c)} \qquad (6\text{-}10)$$

这里，$Count(A \rightarrow \beta, Context = c)$ 表示在上下文 Context 的取值为 c 的条件下，非终结符 A 被串β替换的次数，Count(A,Context = c)表示在上下文 Context 的取值为 c 的条件下

出现非终结符 A 的次数。

按（6-9），则用下面的公式来估计上下文概率：

$$P(\beta \to A \mid \beta, Context = c) = \frac{Count(\beta \to A, Context = c)}{Count(\beta, Context = c)} \qquad (6\text{-}11)$$

这里，$Count(\beta \to A, Context = c)$ 表示在上下文 Context 的取值为 c 的条件下串 β 归结为 A 的次数，Count(β,Context = c)表示在上下文 Context 的取值为 c 的条件下串 β 出现的次数。

6.4.2 自顶向下的上下文属性概率的应用

6.4.2.1 前文属性

下面以一个具体例子来说明自顶向下的前文属性概率的作用。

例： 知道/vg 了/ut 这/r 个/qn 名字/ng 代表/vg 的/usd 内涵/ng 。/wd

对于该句中的实词序列“名字/ng 代表/vg”，表 6-9 和表 6-10 分别给出了用 PCFG 模型和 PCFG+ $P_{TD}(LT)$ 模型分析的结果。

表 6-9 PCFG 模型分析例示（二）

实词序列	规则(A → β)	P(A → β ｜ A)
名字/ng 代表/vg	S → ng vg	0.161679
名字/ng 代表/vg	NO → ng vg	0.120273
名字/ng 代表/vg	NP → ng vg	0.063159
名字/ng 代表/vg	V → ng vg	0.011573

表 6-10 PCFG+ $P_{TD}(LT)$ 模型分析例示

实词序列	LT	规则(A → β)	$P_{TD}(LT)$	P(A → β ｜ A) * $P_{TD}(LT)$
名字/ng 代表/vg	qn	NO → ng vg	0.452680	0.054445
名字/ng 代表/vg	qn	S → ng vg	0.124384	0.033370
名字/ng 代表/vg	qn	NP → ng vg	0.206398	0.007856
名字/ng 代表/vg	qn	V → ng vg	0.020665	0.000239

从表 6-9 中可以看出，应用 PCFG 模型，这个序列分析为 S 的概率最大，在表 6-10 中可以看到，应用 PCFG+ $P_{TD}(LT)$ 模型，这个序列分析为 NO 的概率最大，这里的 $P_{TD}(LT)$ 表示 P(A→β｜ LT = “qn”, A ）， 最后一栏表示 PCFG 模型的概率和 $P_{TD}(LT)$ 模型概率的乘积。

6.4.2.2 后文属性

例：中国/nps 的/usd 铁路/ng 建设/vg 得/usf 不错/a 。/wd

对于该句中的实词序列“铁路/ng 建设/vg”，表 6-11 和表 6-12 分别给出了用 PCFG 模型和 PCFG+P_{TD}(RT)模型的分析结果。

表 6-11　PCFG 模型分析例示（三）

实词序列	规则(A → β)	P(A → β ｜ A)
铁路/ng 建设/vg	S → ng vg	0.161679
铁路/ng 建设/vg	NO → ng vg	0.120273
铁路/ng 建设/vg	NP → ng vg	0.063159
铁路/ng 建设/vg	V → ng vg	0.011573

表 6-12　PCFG+ P_{TD}(RT)模型分析例示

实词序列	RT	规则(A → β)	P_{TD}(RT)	P(A → β ｜ A)* P_{TD}(RT)
铁路/ng 建设/vg	usf	NO → ng vg	0.0460041	0.00553305
铁路/ng 建设/vg	usf	S → ng vg	0.0000001	1.61679e-008
铁路/ng 建设/vg	usf	NP → ng vg	0.0000001	6.31594e-009
铁路/ng 建设/vg	usf	V → ng vg	0.0000001	1.15726e-009

从表 6-11 中可以看出，应用 PCFG 模型，序列“铁路/ng 建设/vg”分析为 S 的概率最大，在表 6-12 中可以看到，应用 PCFG+ P_{TD}(RT)模型，这个序列分析为 NO 的概率最大，这里的 P_{TD}(RT)表示 P(A→β｜ RT = “usf”, A)， 最后一栏表示 PCFG 模型的概率和 P_{TD}(RT)模型概率的乘积。

6.4.3 自底向上的上下文属性概率的应用

我们在前面一节看到，自顶向下的上下文概率模型确实比单纯的 PCFG 模型的消歧能力强，但是，我们发现，自顶向下的上下文概率的模型不能充分地反映上下文对结构的制约作用。例如，在 6.4.2.1 所举的例子中，β 等于“ng vg”，LT 等于“qn”（名量词），四条可用规则的属性概率如下表所示：

规则(A → β)	P_{TD}(LT)
NO → ng vg	0.452680
S → ng vg	0.124384
NP → ng vg	0.206398
V → ng vg	0.020665

从表中可以看出，在各个规则中上下文属性概率的差异并不大。但我们从统计中发现，在语料中“qn + ng + vg ”序列 （即“名量词 + 普通名词 + 一般动词”）共出现了 280 次，其中“ng + vg ”不成结构的有 246 次，占总数的 88%，NP 为 27 次，占总数的 10%，S 为 6 次，占总数的 2%，V 为 1 次。如果用 P_{BU}(LT)来表示上下文属性概率的话，当 LT= “qn”时，“ng + vg”成为 NO 和 S 的概率分别为：

P(NO → ng vg) × P(ng vg → NO | NO, LT = ‘qn’) = 0.120273 × 246/280 = 0.1057

P(S → ng vg) × P(ng vg → S | S, LT = ‘qn’) = 0.161679 × 6/280 = 0.0034

也就是说，“ng + vg”在“qn”后成为非结构的概率是成为主谓结构的 30.5 倍。但在用 P_{TD}(LT)的情况下，它成为 NO 的概率只是成为 S 的概率的 1.6 倍（见表 6-10）。显然，P_{BU} (LT)

更能反映语料中表现的上下文对结构的作用。

再比如，语料库中“NP + VP”前加“usd”（结构助词“的”）的实例为 80 个，其中归结为 NO 的 79 次，归结为 S 的 1 次。简单地说，结构助词“的”后的“NP + VP”不成结构的概率为 79/80，成主谓结构的概率为 1/80，差异非常显著。但是，如果用 $P_{TD}(LT)$ 来表示上下文属性概率的话就会得到下面的结果：

P(NO → NP VP | NO) = 0.01192603

P(NO → NP VP | NO, LT = ‘usd’) = 0.10854332

P(S → NP VP | S) = 0.0718599

P(S → NP VP | S, LT = ‘usd’) = 0.03019324

P(NO → NP VP | NO) × P(NO → NP VP | NO, LT = ‘usd’) = 0.00129449

P(S → NP VP | S) × P(S → NP VP | S, LT = ‘usd’) = 0.00216968

由此可见，如果用 $P_{TD}(LT)$ 表示前文概率的话，当 LT = “usd”时，“NP + VP”归结为 S 的概率要大于归结为 NO 的概率，显然这不能正确反映结构助词“的”后“NP + VP”归结的统计规律。但如果用 $P_{BU}(LT)$ 来代替 $P_{TD}(LT)$ 的话就可以避免这一问题。下面是应用 $P_{BU}(LT)$ 的结果：

P (NO → NP VP | NP VP, LT = ‘usd’) = 79/80

P (S → NP VP | NP VP, LT = ‘usd’) = 1/80

P(NO → NP VP | NO) × P (NO → NP VP | NP VP, LT = ‘usd’) = 0.011777

P(S → NP VP | S) ×P (S → NP VP | NP VP, LT = ‘usd’) = 0.000898

由此可见，它成为非结构的概率是成为主谓结构的概率的 13 倍。

表 6-13 给出了在一些规则当 LT 为结构助词“的”或介词的情况下的统计。

表 6-13 在给定上下文条件下规则的统计举例

LT	规则右部	频次	规则左部及其频次
usd	NP vg	398	NO – 367; NP – 28; S – 3
usd	NP V	87	NO – 86; S – 1
usd	NP VP	80	NO – 79; S – 1
usd	ng vg	781	NO – 572; NP – 188; S – 14; V – 7
usd	ng V	129	NO – 123; NP – 3; S – 3
usd	ng VP	128	NO – 125; NP – 2; VP – 1
pg	NP vg	744	NO – 638; NP – 53; S – 53
pg	NP V	112	NO – 100; S – 12
pg	NP VP	219	NO – 208; S – 11
pg	ng vg	833	NO – 616; NP – 155; S – 59; V – 3
pg	ng V	107	NO – 80; NP – 2; S – 20
pg	ng VP	179	NO – 157; S – 22

表 6-13 反映了这样一个语言事实：一个名词或名词短语后面加上一个动词或动词短，当前一个词是结构助词“的”或介词时，它不成结构的概率远远大于成结构的概率，特别是成为主谓结构的概率很低。可是通过下面的表我们清楚地看到：用 $P_{TD}(LT)$ 不能正确地反映这一语言事实，而 $P_{BU}(LT)$ 可以很好地反映这一统计事实。表 6-14 列出了两种上下文概率模型对这些结构的分析结果，从中可以看到 $P_{BU}(LT)$ 模型更能反映上下文的制约作用。

表 6-14 两种上下文概率模型的对比

左部	右部	频次	概率	LT	F(LT)	PCFG+ $P_{TD}(LT)$	PCFG+ $P_{BU}(LT)$
NO	NP VP	868	0.011926	usd	79	0.001294	0.011777
S	NP VP	238	0.071860	usd	1	0.002169	0.000898
NO	NP vg	3749	0.051510	usd	367	0.025974	0.047498
NP	NP vg	386	0.014414	usd	28	0.001511	0.001014
S	NP vg	375	0.111715	usd	3	0.010119	0.000842
NO	NP V	635	0.008725	usd	86	0.001031	0.008624
S	NP V	155	0.046800	usd	1	0.001413	0.000538
NO	ng vg	8040	0.110467	usd	572	0.086817	0.080905
NP	ng vg	1644	0.061550	usd	188	0.043322	0.014816
S	ng vg	537	0.162138	usd	14	0.068536	0.002906
V	ng vg	74	0.010795	usd	7	0.001102	0.000077
NO	NP vg	3749	0.051510	pg	638	0.045153	0.044171
NP	NP vg	386	0.014414	pg	53	0.002860	0.001027
S	NP vg	375	0.111715	pg	53	0.175398	0.007958
NO	NP V	635	0.008725	pg	100	0.001199	0.007789
S	NP V	155	0.046800	pg	12	0.016956	0.005014
NP	NP VP	868	0.011926	pg	208	0.003408	0.011327
S	NP VP	238	0.071860	pg	11	0.023867	0.003609
NO	ng vg	8040	0.110467	pg	616	0.093495	0.081690
NP	ng vg	1644	0.061550	pg	155	0.035718	0.013891
S	ng vg	537	0.162138	pg	59	0.288832	0.002725
V	ng vg	74	0.010795	pg	3	0.000472	0.000090
NO	ng V	1363	0.018727	pg	85	0.002187	0.014877
S	ng V	133	0.040157	pg	20	0.024249	0.007506
NP	ng V	31	0.001161	pg	2	0.000009	0.000022
NO	ng VP	1861	0.025583	pg	157	0.005519	0.022439
S	ng VP	275	0.083031	pg	22	0.055154	0.010205

表 6-15 两种上文概率属性模型的实验结果对比

模型	带标记			不带标记		
	准确率	召回率	F 度量	准确率	召回率	F 度量
PCFG + $P_{TD}(LT)$	57.3	69.8	62.9	67.8	82.7	74.5
PCFG + $P_{BU}(LT)$	67.3	74.1	70.5	75.0	82.7	78.7

从上表中的实验结果可以看出，PCFG + $P_{BU}(LT)$ 模型显然比 PCFG + $P_{TD}(LT)$ 好得多，它对于提高短语分析的准确率的作用尤其明显。在带标记的结果中，第二种模型比第一种模型的准确率提高了 10 个百分点，在短语边界的识别中，第二种模型比第一种模型的准确率提高了 7.2 个百分点。实验结果证明：应用自底向上的上下文概率属性可以更好地反映上下文对结构消歧的作用。

6.5 词汇功能属性

6.5.1 词汇功能属性的定义

CFG 生成能力过强的根本原因在于：在规则的推导过程中，语言中的词汇项被范畴符号代替，而范畴并不能反映词汇项的个性特征。所以有必要利用词汇的功能、语义属性对规则加以约束。传统的词汇属性描述是定性描述，它说明词有或没有某个属性，这种是/否型的描述过于简单，不能反映词汇属性上的概率差异。例如，俞士汶（1998，2000）提出了概率属性描述的思想，其基本想法是用属性的概率值代替是/否型的属性描述。本文将以动词的功能属性为例，尝试实现词汇的概率属性描述，并将概率属性信息应用于句法分析。下面首先给出有关的定义。

词汇功能属性

定义： 给定一个词 w，它属于范畴 t。设 t 的分布环境为 $D_1,D_2,\ldots,D_n$（其中 Di 也是一个集合，集合中的元素是词的具体用例），分别对应范畴 t 的词汇动能属性 $F_1,F_2,\ldots,F_n$。若 w 能在分布环境 D_i (1≤ i ≤ n) 中出现，则 w 具有功能属性 F_i，否则 w 不具有功能属性 F_i。

词汇功能属性 F_i 可以看作一个函数，它将范畴 t 映射到一个分布环境集合 D_i 中。

例如，动词“存在”有下面两个用例：

（1）现代资本主义社会之所以存在失业，就是由于这些因素交相作用而造成的有效需求不足。

（2）物质的各种存在形式和运动形式之间普遍存在着联系。

例（1）中，“存在”的宾语是“失业”，例（2）中“存在”的宾语是“联系”，“失业”和“联系”都是动词，所以，我们可以把这两个用例概括为“带加动词宾语”。与此相类似，我们可以根据“存在”的其他用例得到“存在”的其他特征，如“带名词性宾语”、“作 NP 修饰语”、“作 NP 中心语”等，在表 6-16 中可以看到关于动词“存在”的功能分布的统计，统计的语料包括 120 万词。

表 6-16　　词汇功能分布统计举例

词	不带宾语	带名词性宾语	带动词宾语	作 NP 修饰语	作 NP 中心语	合计
存在	260	175	2	11	4	459
流动	46	0	0	21	7	74

概率词汇功能属性

定义：给定一个词 w，它属于范畴 t。设 t 的分布环境为 $D_1,D_2,\ldots,D_n$，分别对应范畴 t 的词汇动能属性 $F_1,F_2,\ldots,F_n$。w 具有功能属性 F_i 的概率为：

$$P(F_i \mid w,t) = \frac{Count(D_i,w,t)}{Count(w,t)} \qquad (6\text{-}12)$$

这里，Count（D_i,w,t）表示属于类 t 的词 w 在分布环境 D_i 中的出现次数，Count(w,t)表示词 w 作为 t 出现的次数。例如，如表 6-16 所示，动词“存在”在分布环境“带动词宾语”中出现了 2 次，它作为动词总共出现了 459 次，那么动词“存在”具有“带动词宾语”这一属

性的概率为 2/459 = 0.00436。

6.5.2 词汇动能属性对消歧的作用

词类 t 在规则 A → α t β 中有确定的功能属性值 F_i。例如，在规则“ VP → vg ng ”中，vg 功能属性值为“带名词性宾语”。在规则“ NP → vg ng ”中，vg 的功能属性值为“作 NP 修饰语”。下面例句中有一个实词序列“流动/vg 人口/ng”，表 6-17 和表 6-18 分别显示了 PCFG 模型和 PCFG+概率词汇功能属性模型的分析结果。

例：广东/nps 还/dr 有/vg 大量/b 的/usd 流动/vg 人口/ng ，/wd 都/dr 要/va 吃饭/vg 。/wd

表 6-17　　PCFG 模型分析例示（四）

实词序列	规则(A → β)	P(A → β | A)
流动/vg 人口/ng	VP → vg ng	0.325161
流动/vg 人口/ng	NP → vg ng	0.13523
流动/vg 人口/ng	NO → vg ng	0.0782199
流动/vg 人口/ng	B → vg ng	0.015

表 6-18　　PCFG+概率词汇功能属性模型分析例示

实词序列	规则(A → β)	vg 属性	P(F= “vgn” |w, t)	P(A → β | A)*P(F= “vgn” |w, t)
流动/vg 人口/ng	VP → vg ng	vgn	0.0	0.0
流动/vg 人口/ng	NP → vg ng	vgp	0.2838	0.0383762
流动/vg 人口/ng	NO → vg ng	vgi	0.6216	0.0486232
流动/vg 人口/ng	B → vg ng	vgn	0.0	0.0

从表 6-17 中，我们可以看出，按照 PCFG 来分析，“流动/vg 人口/ng”成为 VP 的概率最大，加上概率词汇功能属性，因为“流动”不能带名词性宾语，所以 VP 和 B 的概率均为 0。从表 6-18 可以看到 NO 的概率要略大于 NP 的概率，但应用上下文属性概率之后，NP 的概率就要大于 NO 的概率。词汇功能属性在这里删除了两条不合法的边。

6.6 实验结果及分析

6.6.1 实验语料

为了测试各种模型的性能，我们从标注的 20 万字语料中抽取了 100 句构成测试集。为了保证测试语料准确反映实际语料的情况，我们采取了随机抽取的方法。语料共有 7611 句，平均每句 26.6 词。我们抽取的 100 句，共 2533 词，平均每句长度为 25.3 词，最短的为 3 个词，最长的 80 词，具体长度分布如下表所示：

表 6-19　实验语料句长分布

句长	3-10	11-20	21-30	31-50	51-80	平均
数量	13	32	27	20	8	25.3

测试集中共有普通短语 507 个（不包括机构名短语 17 个和地名短语 6 个）。

6.6.2 评价指标

在我们的实验中，采取精确匹配的标准来评价分析系统的性能。短语匹配分成带标记的匹配和不带标记的匹配两种情况，前一种是指两个短语的边界相同且标记相同，后一种是指短语的边界相同但标记未必相同。

设系统分析结果中边界和标记完全正确的短语数为 ExactMatchNum，结果中边界正确但标记不正确的短语数为 BoundaryOnlyMatchNum，答案中有而分析结果中没有的短语数为 MissedNum，结果中有而答案中没有的短语数为 OverGeneratedNum。利用这几个统计量就可以定义系统的评价指标。

$$Labeled\ \mathrm{Pr}ecision = \frac{ExactMatchNum}{ExactMatchNum + BoundaryOnlyMatchNum + OverGeneratedNum}$$

$$Labeled\ \ \mathrm{Re}call\ = \frac{ExactMatch\ \ Num}{ExactMatch\ \ Num\ + BoundaryOn\ \ lyMatchNum\ \ + MissedNum}$$

$$LabeledF-measure = \frac{Labeled\ \mathrm{Pr}ecision \times Labeled\ \mathrm{Re}call \times 2}{Labeled\ \mathrm{Pr}ecision + Labed\ \mathrm{Re}call}$$

$$UnLabeled\ \mathrm{Pr}ecision = \frac{ExactMatch\,Num + BoundaryOn\,lyMatchNum}{ExactMatch\,Num + BoundaryOn\,lyMatchNum\ + OverGenera\,tedNum}$$

$$UnLabeled\,\mathrm{Re}call = \frac{ExactMatchNum + BoundaryOnlyMatchNum}{ExactMatchNum + BoundaryOnlyMatchNum + MissedNum}$$

$$UnLabeledF-measure = \frac{UnLabeled\ \mathrm{Pr}ecision \times UnLabeled\ \mathrm{Re}call \times 2}{UnLabeled\ \mathrm{Pr}ecision + UnLabed\ \mathrm{Re}call}$$

6.6.3 实验结果

下面的表给出了在各种模型下的实验结果，表中 P，R，F 分别表示准确率、召回率和综合指标 F-measure。

表 6.20 各种模型实验结果对比

模型	带标记			不带标记		
	P	R	F	P	R	F
PCFG	51.3	64.7	57.2	66.5	83.8	74.1
PCFG + RF（1）	55.8	69.8	62.1	66.1	82.7	73.5
PCFG + RF（2）	59.5	73.5	65.8	67.9	83.9	75.0
PCFG + P_{TD}(LT)	57.3	69.8	62.9	67.8	82.7	74.5
PCFG + P_{BU}(LT)	67.3	74.1	70.5	75.0	82.7	78.7

PCFG + P_{TD}(RT)	61.2	72.1	66.2	71.1	83.8	76.9
PCFG + P_{BU}(RT)	63.7	71.7	67.5	72.1	81.3	76.4
PCFG + LF	57.9	71.3	63.9	67.7	83.4	74.8
PCFG+FS(1)	68.4	77.2	72.5	73.7	83.2	78.2
PCFG+FS(2)	78.4	83.2	80.8	81.2	86.2	83.6

在表 6-20 中，PCFG 表示概率上下文无关语法模型，RF 表示节律属性模型，RF(1)表示仅使用简单短语中的节律属性，RF(2)表示使用所有短语中的节律属性。P_{TD}(LT)表示自顶向下的前文概率模型，P_{BU}(LT)表示自底向上的前问概率模型，P_{TD}(RT)表示自顶向下的后文概率模型，P_{BU}(RT)表示自底向上的后文概率模型。LF 表示词汇功能属性，FS(1)表示 PCFG + RF(1) + P_{TD}(LT) + P_{TD}(RT) + LF，FS(2)表示 PCFG + RF(2) + P_{BU}(LT) + P_{BU}(RT) + LF。

6.6.4 实验结果分析

从实验的结果来看，在 PCFG 之上加上任何一个概率属性，系统的性能都比 PCFG 好，各个属性加起来的性能又比单个属性的性能要好，这说明各个属性之间有互相补充的作用。

从单个属性对系统的贡献率来看，上下文属性的作用最大，其中前文属性比后文属性作用更大。

对于音节属性和上下文属性，我们分别测试了两种模型，实验结果证明考虑所有短语中的节律属性要比只考虑简单短语中的节律属性要好。对上下文来说，自底向上的上下文属性概率模型要比自顶向下的上下文概率模型要好。

6.6.5 错误分析

下面的错误分析以最后一个综合模型的结果为依据。错误的原因大致可以分为以下几类：

（一）由长距离依赖现象产生的错误。具体说来，主要有以下两种情况：

A. 动词的宾语是复杂的“的”字结构。

这类结构的主要形式是： V + X + 的 + Y

正确的分析结果是：V + [X + 的 + Y]

系统错误地分析为：[V + X] + 的 + Y

这是汉语中一种典型的歧义结构。如“咬死猎人的狗”就允许两种不同的分析，分别表示两种不同的意义：

[咬死 [猎人 的 狗]

[[[咬死 猎人] 的] 狗]

在具体的上下文中，并不存在这种歧义。消解歧义的关键在于确定 X 的归属。在我们的实语块分析中把“的”作为分界符，不考虑位于“的”两边的 X 和 Y 之间的关系，对于 X 的归属我们不是采用规则方法，而是通过非结构的 NO 符号来解决。如果[V+X]构成非结构，则说明 X 先和“的”结合。这可以解决一部分歧义，但不能解决全部的问题，特别是当 X 比较长的时候。下面是一些错误的实例。

（1）就/dr 能/va [[正确/a 利用/vg]V [资本主义/ng 社会/ng]NP]VP 创造/vg 的/usd [文明/ng 成果/ng]NP

（2）[[[种/vg 菇/ng]VP 户/kn]NP [占/vg [[全/b 村/ng]NP [总/b 户数/ng]NP]NP]VP]S 的/usd ８０%/mx 以上/mab 。/wd
（3）要/va 尽快/dr [改变/vg 我国/ng]VP 的/usd [落后/a 面貌/ng]NP ，/wd
（4）[提高/vg 企业/ng]VP 的/usd [管理/vg 水平/ng]NP 和/c [企业/ng 劳动者/ng]NP 的/usd 素质/ng ，/wd 从而/c [提高/vg [[企业/ng 劳动者/ng]NP [具体/a 劳动/ng]NP]NP]VP 的/usd 生产率/ng ；/wd
（5）[受到/vg [市场/ng 机制/ng]NP]VP 的/usd 调节/vg 。/wd
（6）[[精心/a 安置/vg]V [移民/ng 们/kn]NP]VP 的/usd 生产/vg 和/c 生活/ng 。/wd
（7）[增强/vg 企业/ng]VP 的/usd 活力/ng ，/wd [建立/vg 企业/ng]VP 的/usd 激励/vg 和/c 约束/vg 机制/ng
（8）[作/vg 相应/vg]V 的/usd 调整/vg 。/wd
（9）[保证/vg [国有/b 资产/ng]NP]VP 的/usd [保值/vg 增值/vg]V
（10）[加强/vg 政府/ng]VP 对/pg [市场/ng 物价/ng]NP 的/usd [调控/vg 管理/vg]V
（11）这里/r 只/dr [有/vg 抓假/vg]V 之/usd 功/ng 而/c [无/vg 灭假/vg]V 之/usd 力/ng 。/wd
（12）[促进/vg [社会主义/ng [社会/ng 生产力/ng]NP]NP]VP 的/usd 发展/vg 。/wd

B. 动词的宾语是复杂的句子。如：
（1）他/r 总是/dr [主张/vg [制定/vg [经济/ng [发展/vg 计划/ng]NP]NP]VP]VP 既/dr 要/va 积极/a ，/wd 又/dr 要/va 留有余地/vg ，/wd 力争/vg “/wb1 适度/a 的/usd 发展/vg ”/wb2 。/wd

（二）跟并列结构相关的错误。下一章将详细讨论，参见 7.4.2 节的分析。
（三）复杂 NP 内部的层次错误。如：
（1）[国有/b [资产/ng 所有权/ng]NP]NP
（2）[建立/vg [[社会主义/ng 市场/ng]NP [经济/ng 体制/ng]NP]NP]VP
（3）[从事/vg [社会/ng [实践/ng 活动/ng]NP]NP]VP
（4）[促进/vg [社会主义/ng [社会/ng 生产力/ng]NP]NP]VP 的/usd 发展/vg 。/wd
（5）[[新/a 建成/vg]V [[国家/ng 重点/ng]NP 实验室/ng]NP]VP ４/mx 个/qn 。/wd
（6）[[前景/npu 公司/ng]NT [推出/vg [[高级/b 电脑/ng]NP [排版/vg 系统/ng]NP]NP]VP]S
（7）[[职工/ng 文化/ng]NP 生活/ng]NP
（8）[传授/vg [农业/ng [技术/ng 知识/ng]NP]NP]VP
（9）[[跨/vg 进/vg]V [三峡/nps [工程/ng 大门/ng]NP]NP]VP
（10）[国有/b [[大中型/b 企业/ng]NP 负责人/ng]NP]NP
（11）[大中专/ng [学校/ng 学生/ng]NP]NP
（四）误将相邻的两个动词识别为并列结构。如：
（1）靠/pg [信贷/ng 规模/ng]NP [[膨胀/vg 拉动/vg]V [经济/ng 发展/vg]NP]VP
（2）继续/vg 把/pba 企业/ng [改革/vg 推/vg]V 向/pg 前进/vg
（3）[知道/vg 购黄/vg]V 是/vi 不/dr 光彩/a 的/usd 事/ng
（五）连谓结构的边界错误。如：
（1）它/r [[使/vg [[外出/vg 务工/vg]V [经商/vg 者/kn]NP]NP]VP [无/vg 后顾之忧/ng]VP]VP
（2）我/r 受/vg [[现在/t 挂职/vg]V 工作/vg]VP 的/usd [湖南/nps 张家界/nps 市

委/ng]NT 委托/vg
（3）相反/c 还/dr [有/vg 余额/ng]VP 倒贴/vg 。/wd
（4）[[防止/vg 农业/ng]VP 萎缩/vg]VP
（六）含动词的复杂NP的边界错误。如：
（1）确定/vg 了/ut [干部/ng 安全/ng]NP 管理/vg 的/usd [量化/vg 标准/ng]NP
（2）在/pg [充分/a 肯定/vg]V 前/f 一/mx 段/qn 物价/ng [[大/a 检查/vg]V 成绩/ng]VP 的/usd 同时/t
（3）兑付/vg 时/f 按/pg 当时/t 公布/vg 的/usd 保值/vg 贴补/vg 率/kn [给予/vf 保值/vg]VP 。/wd
（4）[河南省/nps 电力/ng]NP 建设/vg
（5）[基础/ng 设施/ng]NP 建设/vg 。/wd
（6）[纸张/ng [印刷/vg [器材/ng 博览会/ng]NP]VP]S 将/dr 举行/vg
（7）在/pg [广泛/a [[发展/vg 劳动/vg]V [[密集/a 型/kn]NP 产业/ng]NP]VP]VP 的/usd 同时/t
（七）把VP误作V。如：
（1）[喜欢/vg 自嘲/vg]V
（2）[深化/vg 改革/vg]V ，/wd [扩大/vg 开放/vg]V
（3）[负责/vg 解释/vg]V
（八）多切分。这类两种切分都允许，应不算错误。
（1）[中国/nps 政局/ng]NP 稳定/a]S
（2）[经济/ng [发展/vg 计划/ng]NP]NP]
（3）[社会主义/ng [社会/ng 生产力/ng]NP]NP

6.7 本章小结

本章提出了概率上下文无关语法和概率属性相结合的汉语实语块分析模型，并提出了属性概率的估计方法。根据节律对句法有制约作用的性质提出了利用结构的节律属性对概率上下文无关语法进行约束的思想，并进行了两个实验利用节律属性的实验：一是仅用简单短语中的节律属性，一是利用所有短语（包括复杂短语）的节律属性，这两个试验的结果证明，后者比前者有大幅度的提高，说明在复杂短语的构造中，节律也起着重要的制约作用。其次，本章探讨了上下文属性概率的估计，探讨了自顶向下和自底向上两种上下文概率属性估计方法，通过对比发现自底向上的上下文概率属性更能反映上下文对结构的约束作用。同时，我们也简单地探讨了以下词汇功能属性对消歧地作用。最后，我们给出了 10 个不同模型的实语块分析结果，并对最后的综合模型的错误进行了分析。

本章的研究表明，用概率上下文无关语法和概率属性相结合的分析模型对非受限的汉语文本进行实语块分析在不需要太多资源的条件下达到了令人满意的效果。从错误分析中我们可以看到，在目前的分析模型中，系统性能提高的潜力还很大。下一章讨论的并列结构的概率模型就是提高性能的途径之一。

第七章　并列结构的概率模型

7.1 引言

并列结构在本文中出现的频率很高，而且情况相当复杂，因为几乎各种类型的成分都可以进入并列结构中。并列结构是句法分析的难点之一，主要困难在于确定并列结构的边界。我们可以根据用不用关联词语把并列结构分为无标记并列结构和有标记并列结构两种。不用关联词语的是无标记的并列结构，用关联词语的是有标记的并列结构。本文只讨论有标记的并列结构。

并列连词包括连词“和”、“与”、“及”、“或”、“同”和一个特殊的标点符号——顿号。在我们标注的 20 万词语料中，这些连词共出现了 5670 次（见表 7-1），虽然只占总词次的 3%，但有 37%的句子中都包括并列连词。所以复杂并列结构的正确处理对句法分析是十分重要的。

表 7-1　　并列连词频率统计

连词	顿号	和	与	及	或
频次	3471	1810	149	143	120

从表 7-1 可以看出，在几个并列连词中，顿号和“和”的出现频率最高。“同”则很少出现，在语料中没见到一个用例。

并列连词连接的并列项可以是词，也可以是短语。词可以是各种实词，短语也包括各种功能类型。由于在我们的实语块分析中，把连词作为实语块的分隔符，所以一般不考虑并列连词两边成分之间的关系，这样自然会产生错误。例如，下面的例子就是前面系统的一个输出实例。

（1）非但/c [党务/ng 工作/ng]NP ，/wd 她/r 还/dr 得/va [协助/vg 董事长/ng]VP 、/wm [总经理/ng [做/vg [人事/ng 工作/ng]NP]VP]S 、/wm [[思想/ng 教育/vg]NP 工作/ng]NP 、/wm [[职工/ng 文化/ng]NP 生活/ng]NP 等等/ur 。/wd

在上面这个例子中，如果不考虑并列成分之间的关系，分别看顿号两边成分的分析，那么分析的结果都是最合理的。如“协助　董事长”是一个VP，“总经理 做 人事 工作”是一个主谓结构。在实语块分析中不考虑连词两边成分之间的关系是基于这样一个假设：并列项之间是相互独立的。这样一个独立性假设有时是对的，如：

（2）要/va　表彰/vg　一/mx　批/qnu　[[市场/ng　管理/vg]NP　严格/a]S　、/wm　[[执行/vg　政策/ng]VP　坚决/a]S　的/usde　典型/ng

在（2）中，连接词连接的是两个主谓结构，这两个并列结构并列之后加上“的”构成“的”字结构，所以在实语块分析阶段可以不考虑并列结构连词两端的实词之间的关系。

在实语块分析中，并列成分独立性的条件有两个：
（1）并列结构不和右边相邻的实词或实语块组成直接成分；
（2）并列结构不和左边相邻的实词或实语块组成直接成分。
违背以上任何一个条件，独立性都不能成立。例如：

（3）要/va [层层/dr 负责/vg]V 向/pg 农民/ng [做/vg 好/a]V [宣传/vg 教育/vg]V 和/c 动员/vg 工作/ng 。/。

在例（3）中，并列结构“宣传教育和动员”作“工作”的定语，违背了上面的条件（1），所以不能将连词后的“动员工作”合起来。

（4）建立/vg [[情报/ng 、/、 信息/ng] 网络/ng]NP ，/， [[筛选/vg 、/、 整理/vg]V 出/vg]V 1 1 7 2/mx 条/qn [实用/a 信息/ng]NP

在例（4）中“情报、信息”和右边的“网络”构成直接成分，“筛选、整理”和右边的“出”构成直接成分，都违反了条件（2）。

在例（1）中，并列结构“董事长、总经理”作动词“协助”的宾语，并列结构“人事工作、思想教育工作 、职工文化生活”作动词“做”的宾语，违背了条件（2），所以引起错误。再看下面的例子：

（5）国有/b 粮食/ng 系统/ng 要/va 建立/vg 起/vg 从/pg 粮食/ng [收购/vg 、/、 加工/vg 、/、 批发/vg] ，/， 一直/dr 到/pg 零售/vg 的/usd [网络/ng 式/kn]B 的/usd [主/b 渠道/ng]NP 。/。

在例（5）中，并列结构“收购、加工、批发”和左边的“粮食”构成直接成分，违反了条件（1），所以“粮食收购”不能先捆。

对于并列结构的处理，关键在于并列项的确定，也就是确定连接词两边并列成分的边界在什么地方。如在例（3）中，如果我们确定了“和”连接的并列项是“宣传 教育”和“动员”，那么就会避免出现下面的分析结果：

[[做/vg 好/a]V [宣传/vg 教育/vg]V]VP
[动员/vg 工作/ng]NP

因为在这两个短语中，并列项分别和其左右的相邻成分构成了直接成分，没有机会在一起构成一个并列结构。

具体说来，确定并列项的任务是确定左边并列项的左边界和右边并列项的右边界，因为往往并列项淹没在一串实词中，会有多个候选项。本文提出了一种并列结构的概率模型，在众多候选项中选择可能性最大的并列项。

7.2 并列结构的对称性

前面讨论了并列结构的复杂性，但是，并列结构并非没有规律可循。并列结构有一个非常显著的特点，这就是：在并列结构中，各个并列项之间存在着对称性。对称性表现在以下方面：
（1）功能类型；

（2）结构关系；
（3）长度；
（4）语义范畴。

所谓对称是指相同或相似。下面分别讨论这几种对称关系。

7.2.1 功能类型的对称性

两个成分具有相同或相似的功能，则这两个成分在功能上具有对称性。如果两个成分的功能类型绝对相同，则自然是对称的。如名词对名词，动词对动词，NP 对 NP，VP 对 VP 等。这是比较容易确定的，问题在于对于两个功能类型不相同的成分如何确定它们之间的对称性。我们知道，名词和 NP 具有相似的功能，动词和 VP 具有相似的功能。那么名词和动词之间有没有相似的功能呢？名词和 VP 之间、动词和 NP、NP 和 VP 之间是否有相似的功能呢？这些问题恐怕不好一下子回答出来。名词和名词性短语可归入体词，动词和动词性短语可归入谓词（朱德熙，1984）。一般来说，体词和谓词存在着对立，但这种对立又不是绝对的。在并列结构中功能类型的对称是比较严格的要求，一般来说，体词性成分和谓词性成分不能并列。通常谓词要进入一个体词化短语中才可以和体词并列。如：

（6）重点是化肥、农业机械、农膜、饲料以及农药、畜禽用疫苗和药品的生产 。
（7）包括农业气象预测预报，农村技术咨询，农村金融、农业生产资料、农产品购销和农产品储运加工，农村信息服务等。

但体词性成分和谓词性成分也不是绝对不能并列，如：

（8） 形式主义、搞花架子甚至弄虚作假
（9） 就业、住房、子女入学等方面的困难
（10） 贡献自己的聪明和才智

（8）中“形式主义”是名词，“摆花架子”是 VP。（9）中“就业”是动词，“住房”是名词。（10）中“聪明”是形容词，“才智”是名词。
下面这些例子都是并列项不太相似的例子。

（11）经历了刚刚过去的粮食涨价、抢购，人们对国有粮店撤并更是担忧。
（12）成为浙赣线上吞吐量最大、最繁忙的大站之一
（14）鼓励国有、集体、个体、私营、外资一齐上
（15）假冒、伪劣商品

这些例子说明，成分在功能上的对称性并不是绝对的，往往存在程度上的差异，不好用简单的“是/否”来描述。但不可否认，并列结构在功能类型上一般是要求对称的。

7.2.2 结构关系的对称性

结构关系和功能类型是紧密相关的。在汉语中，短语的类型确定之后，就可以部分地确定短语内部的结构关系。如：名词性短语内部只能是偏正或并列关系，不可能是述宾关系。主谓结构内部只能是主谓关系。动词性内部关系要复杂一些，可以有述宾、述补、并列、状

中四种。

7.2.3 长度的对称性

对称的本质是均衡，并列结构有一种强烈的倾向，即要求并列成分在长度上尽量相同或接近。表 7-2 是关于并列成分长度的一个小规模的统计（只涉及短语和短语的并列，不包括一个词和短语的并列）。

表 7-2 并列短语的长度统计

短语类型	长度差为 0	长度差为 1	长度差为 2	合计
NP	592	190	32	814
VP	177	38	11	226
V	73			73
S	87	27	3	117

从这个统计可以看出，大部分并列短语的长度是相同的，完全相同的，最低是 73%（NP），最高的是 100%（V），长度相差 2 的不到 5%，长度差大于 2 的没有。

7.2.4 语义的对称性

语义的对称性体现在并列项一般位于同一语义场中，具有相同的义类特征。例如：

（15）农副产品、日用工业品等消费品；建材机电、电子元配件、有色金属等生产资料市场；金融、劳动力、房地产、技术等要素市场；仓储、运输、通讯、旅游、文化等服务市场，还有动产拍卖市场、产权转让市场、期货市场等，多已初具规模。

在例（15）中的主语是一个复杂的并列结构，第一层是各种市场的划分：消费品市场、生产资料市场、要素市场、服务市场等。在每个市场中又对交易对象进行进一步分类，如消费品包括农副产品、日用工业品等。语义分类构成了一个树型结构，如下图所示。

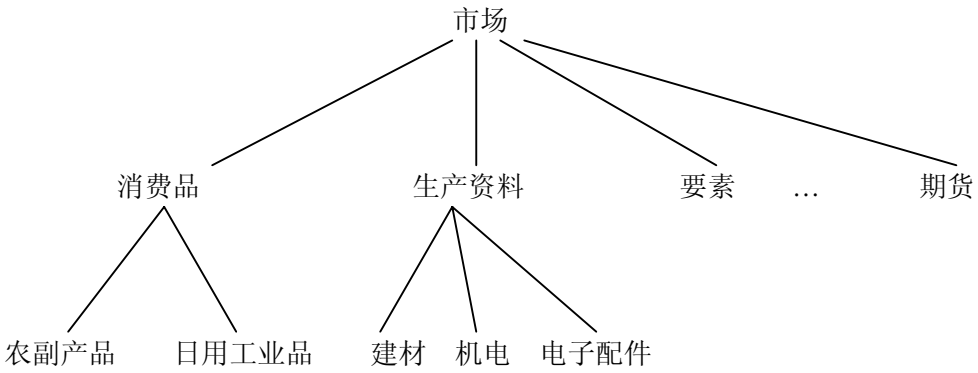


图 7-1　语义分类例示

并列项在语义上的对称性体现在：并列项不能跨越语义分类树中的层次。尽管两个成分具有共同的上位，但如果跨越层次则不能并列，如在上图中，“农副产品”和“生产资料”就不能并列。

7.3 基于对称性原则的并列结构概率模型

7.3.1 基本思想

并列结构处理的实质性问题是：当并列连词两边都存在一组可能的并列项时，如何从中选择一对正确的并列项？下面以一个简单的例子来说明：

可以/va　到/vg　上述/b　机构/ng　和/c　网点/ng　购买/vg

在连词“和”的左边有三个实词，可能的并列项有：

[到/vg　[上述/b　机构/ng]NP]VP
[到/vg　[上述/b　机构/ng]NP]NO
[[到/vg　上述/b] VP　机构/ng]NO
[[到/vg　上述/b] VP　机构/ng]NP
[上述/b　机构/ng]NP
[上述/b　机构/ng]NO

在连词的右边有两个实词，可能的并列项候选有：

[网点/ng 购买]NO
[网点/ng 购买]S
[网点/ng 购买]V

词候选项已隐含在类型为 NO 的短语中，如并列候选项“机构”隐含在“[上述/b　机构/ng]NO”中，“网点”隐含在“[网点/ng 购买]NO”中。这样，我们就可以只考虑短语候选项，而不必考虑词和短语的组合。第一组有 6 个候选，第二组有 3 个候选，两两组合共有 6×3 = 18 种可能性。

我们的基本思想是：对于每一种组合，根据对称性原则对组合进行概率评分，然后再以短语原来的概率分别乘以并列结构的概率评分，从中选择一个组合使得两边成分的概率都最大化。

7.3.2 对称性概率评价

由于功能类型可以大致控制结构关系，而语义范畴又需要庞大的语义分类词典，所以在此我们只考虑功能类型和长度的对称性。

设并列连词 c 是句中第 i 个词，给定 c 左边的词性串 L 和右边的词性串 R，左右两边候选并列项的短语类型分别为 T_1,T_2，则 T_1,T_2 的对称性评价可定义如下：

$$Sym(T_1,T_2) = P(T_1,T_2 \mid L,c,R) \qquad (7\text{-}1)$$

在评价短语功能类型概率的时候，当然希望知道的上下文越多越好，即 L 和 R 越长越好，但这将导致严重的数据稀疏问题，所以需要进行独立性假设。为了缩小统计数据的规模，我们假设 L 为 c 左边第一个词的词性 t_1，R 为右边第一个词的词性 t_2，那么上式可简化为：

$$Sym(T_1,T_2) = P(T_1,T_2 \mid t_1,c,t_2) \qquad (7\text{-}2)$$

根据最大似然估计，有：

$$Sym(T_1,T_2)=\frac{Count(T_1,T_2,t_1,c,t_2)}{Count(t_1,c,t_2)} \qquad (7\text{-}3)$$

如果两个候选短语的长度分别为 L1,L2,则 L1,L2 的对称性评价可定义如下：

$$Sym(L_1,L_2)=P(L_1,L_2\mid T_1,T_2,L,c,R) \qquad (7\text{-}4)$$

我们假设：短语的长度和短语的类型无关，同时设 L 为 c 左边第一个词的词性 t1，R 为右边第一个词的词性 t2，那么上式可简化为：

$$Sym(L_1,L_2)=P(L_1,L_2\mid t_1,c,t_2) \qquad (7\text{-}5)$$

根据最大似然估计，有：

$$Sym(L_1,L_2)=\frac{Count(L_1,L_2,t_1,c,t_2)}{Count(t_1,c,t_2)} \qquad (7\text{-}6)$$

短语的长度定义为其中词的数量，如果短语的类型为 NO,则 c 左边的短语定义为其右子结点的长度，右边的短语定义为其左子结点的长度。

7.3.3 算法描述

Let c be the i-th word in the sentence, t1 be the POS tag for the i-1th word, t2 be the POS tag for the i+1-th word.

(1) Find all the maximum edges extending to i-1 to form set EF, and
Find all the maximum edges starting from i+1 to form set ER.

(2) maxEdgeNum1 = maxEdgeNum2 = 0; maxProb1 = maxProb2 = 0

(3)
```
for(i =0; i< EF.size(); i++)
  for(j = 0; j < ER.size(); j++)
  {
    P1 = P(EF[i]) * Sym(Type(EF[i]),Type(EF[j])) * Sym(Len(EF[i]),Len(EF[j]))
    P2 = P(EF[j]) * Sym(Type(EF[i]),Type(EF[j])) * Sym(Len(EF[i]),Len(EF[j]))
    if(P1 > maxProb1 && P2 > maxProb2)
    {   maxProb1 = P1; maxProb2 = P2;
        maxEdgeNum1 = i; maxEdgeNum2 = j;
    }
  }
```

7.4 实验结果及分析

7.4.1 实验结果

引入并列结构处理模型之后，系统的输出结果和原来的结果相比，共有 28 处改动，其中修改短语类型 2 处，1 处是改错为错，1 处正确。新产生短语 3 个，全部正确。删除短语 23 处，19 处正确，4 处错误。表 7-3 给出了引入并列结构概率模型前后实语块分析系统的性能对比。

表 7-3 引入并列结构概率模型前后的系统性能对比

模型	带标记			不带标记		
	准确率	召回率	F 度量	准确率	召回率	F 度量
引入并列结构概率模型前	78.4	83.2	80.8	81.2	86.2	83.6
引入并列结构概率模型后	81.5	83.2	82.3	84.2	86.0	85.1

从表 7-3 中我们可以发现，引入并列结构概率模型之后，由于删除了一些跟并列结构有关的错误的边，因而使系统的正确率有较大的提高。带标记和不带标记的正确率分别提高了 3 个百分点，分别达到了 81.5%和 84.2%。同时，召回率基本没有变化。

7.4.2 实验结果分析

为了仔细分析并列结构统计模型的得失，也为了给进一步研究并列结构的处理提供一些资料，下面将列出前面所说的 28 处改动，并加以简单的分析。下面所列的例句是引入并列结构统计模型之前的输出，在分析中将说明新的输出与之相比的差异。

（一）边界一致，但标记不同。2 例。

(1) [促进/vg 发展/vg]V 和/c [保持/vg 稳定/a]S

(2) 同/pg [白裤瑶/npr [移民/ng 们/kn]NP]NP 一道/dr [睡/vg 草棚/ng]VP 、/wm [啃/vg 干粮/ng]VP 、/wm [吃/vg 野菜/ng]VP 、/wm [喝/vg 山泉/ng]VP 、/wm [搞/vg 开发/vg]V

（1）中短语“[保持/vg 稳定/a]”的标记由 S 改为 V，尽管“和”前后两个短语的标记都应该是 VP，但两个类型相同的结果比原来的结果显然更合理。（2）中的短语“[搞/vg 开发/vg]”的标记由 V 改为 VP。这两例都体现了并列结构中功能类型对称性的作用。

（二）产生新的短语。3 例。

(3) 建立/vg [产权/ng 明晰/a]S 、/wm [责任/ng 明确/a]S 、/wm 政企/ng 分开/vg 、/wm [管理/ng 科学/a]S 的/usd [现代/t [企业/ng 制度/ng]NP]NP

(4) 这/r 是/vi 百/mw 万/mw [海内外/s [[长乐/nps 人/ng]NP 政治/ng]NP]NP 、/wm [经济/ng 生活/ng]NP 中/f 的/usd 一/mx 件/qn [大/a 喜事/ng]NP

(4’) 这/r 是/vi 百/mw 万/mw [海内外/s [长乐/nps 人/ng]NP]NP 政治/ng 、/wm 经济/ng 生活/ng 中/f 的/usd 一/mx 件/qn [大/a 喜事/ng]NP

(5) [李鹏/npc 总理/ng]NP 和/c 国务委员/ng 陈俊生/npc 由/pg [[[中共/npu 河南/nps 省委/ng]NT 书记/ng]NP 李长春/npc]NP 、/wm [省长/ng 马忠臣/npc]NP 陪同/vg

上面（3）中的“政企/ng 分开/vg”原来分析为非结构，新标注为“[政企/ng 分开]S”，这是类型对称的作用。（4）中前后差异较大，所以在（4’）中列出新的版本，以便于说明。新产生的短语是“[海内外/s [长乐/nps 人/ng]NP]NP”。这是长度对称的作用，因为“政治”和“经济”的长度一致，它们并列的概率最大。（5）中新产生的短语是“[国务委员/ng 陈俊生/npc]NP”，这也是类型对称和长度对称双重作用的结果。

（三）原输出中有，新输出中没有的短语，共 23 例。其中正确的为 19 例，错误的为 4 例。

(6) 市里/ng 在/pg 吃/vg 、/wm 住/vg 、/wm [行/vg 方面/ng]VP 提供/vg 了/ut 诸多/mm

方便/a 。/wd

“行 方面”在新的输出中为非结构，因为“行”和前面的“住”在功能和长度上都对称。

(7) 从/pg[国有/b[资产/ng 所有权/ng]NP]NP 中/f[[分离/vg 出/vg]V[国家/ng[终极/b 所有权/ng]NP]NP]VP 和/c[企业/ng[法人/ng 财产权/ng]NP]NP

“分离 出 国家 终极 所有权”在新的输出中为非结构，因为“[国家/ng [终极/b 所有权/ng]NP]NP”和 “[企业/ng [法人/ng 财产权/ng]NP]NP”在功能和长度上都对称。

(8)[查处/vg[价格/ng[违法/vg 案件/ng]NP]NP]VP 的/usd 难度/ng 和/c[阻力/ng[比较/dd 大/a]AP]S 。/wd

“[阻力/ng [比较/dd 大/a]AP]S”在新的输出中为非结构，因为“难度”和“阻力”作为并列项的概率最大。

(9) 靠/pg 企业/ng 自己/r [赚/vg 钱/ng]VP 、/wm [[攒/vg 钱/ng]VP 发展/vg]VP

“[[攒/vg 钱/ng]VP 发展/vg]VP”在新的输出中为非结构，虽然“[赚/vg 钱/ng]VP”和“[[攒/vg 钱/ng]VP 发展/vg]VP”在功能上一致，但在长度上“[赚/vg 钱/ng]VP”和“[攒/vg 钱/ng]VP”更一致。

(10)[达到/vg 临界点/ng]VP 的/usd ，/wd 不/dr [发/vg 工资/ng]VP 和/c 奖金/ng

“发工资”被判为非结构。因为“发工资”和“奖金”在功能和长度上都不一致。

(11) 车间/ng 和/c 单位/ng 把/pba 提取/vg 的/usd 2 5 %/mx [作为/vg [职工/ng 工资/ng]NP]VP 和/c 奖金/ng 。/wd

在新的输出中，“[作为/vg 职工/ng 工资/ng]VP” 和“[职工/ng 工资/ng]NP”都是非结构，因为，前一个短语的类型是VP，和名词在功能上不对称，后一个短语和后面的并列项虽然在功能上对称，但在长度上不完全对称。所以“工资”和“奖金”并列的概率最大。

(12) 可以/va [到/vg [上述/b 机构/ng]NP]VP 和/c 网点/ng 购买/vg 。/wd

在新的输出中，“[到/vg 上述/b 机构/ng]VP” 和“[上述/b 机构/ng]NP”都是非结构，原因与前一句相同。

(13) 是/vi 今年/t 全党/ng 和/c [全国/ng 工作/ng]NP 的/usd 大局/ng 。/wd

“[全国/ng 工作/ng]NP”在新的输出中为非结构，“全党”和“全国”在长度上更对称。

(14) 已经/dr [争取/vg 到/vg]V [世界/ng 银行/ng]NT 、/wm [西班牙/nps 政府/ng]NT 和/c [香港/nps 投资/ng]NP 等/ur 贷款/ng 共/dr 8． 2/mx 亿/mw 美元/qn 。/wd

“[香港/nps 投资/ng]NP”被误判为非结构，这是因为，“西班牙 政府”作为一个机构名在我们的系统中被当作一个词，所以，“香港”和“西班牙 政府”在长度上一致。

(15) [加快/vg 仓储/ng]VP 、/wm 运输/ng 、/wm [通讯/ng 信息/ng]NP 等/ur [基础/ng 设施/ng]NP 建设/vg 。/wd

“[加快/vg 仓储/ng]VP ”和“[通讯/ng 信息/ng]NP”在新的输出中都是非结构。后一例为误判。如果引入语义信息或搭配信息，就有可能判定“通讯信息”为并列结构，以避免这一错误。

(16) 非但/c [党务/ng 工作/ng]NP ，/wd 她/r 还/dr 得/va [协助/vg 董事长/ng]VP 、/wm [总经理/ng [做/vg [人事/ng 工作/ng]NP]VP]S 、/wm [[思想/ng 教育/vg]NP 工作/ng]NP 、/wm [[职工/ng 文化/ng]NP 生活/ng]NP 等等/ur 。/wd

(16') 非但/c [党务/ng 工作/ng]NP ，/wd 她/r 还/dr 得/va 协助/vg 董事长/ng 、/wm 总经理/ng 做/vg [人事/ng 工作/ng]NP 、/wm [[思想/ng 教育/vg]NP 工作/ng]NP 、/wm [[职工/ng 文化/ng]NP 生活/ng]NP 等等/ur 。/wd

该例共有三个短语被识别为非结构，全部正确。

(17) [演讲/vg 稿/ng]NP 被/pbe 印发/vg 给/vg 成千上万/mg 的/usd 官员/ng 和/c [企

业/ng [管理/vg 人员/ng]NP]NP 。/wd

该例中的“[企业/ng 管理/vg 人员/ng]NP”被误判为非结构，因为“官员”和“企业”在功能和长度上的对称度最高。因为没有引入语义因素，所以不知道“官员”和“人员”的对称度更高。

（18）为了/pg [进一步/dr 解放/vg]V 和/c 发展/vg 生产力/ng

“[进一步/dr 解放/vg]V”在新的输出中为非结构。

（19）他/r [占有/vg 土地/ng]VP 和/c 一部分/mm [生产/vg 工具/ng]NP

“[占有/vg 土地/ng]VP”在新的输出中为非结构。

（20）[[国家计委/npu 主任/ng]NP 陈锦华/npc]NP 、/wm [[[中国/nps 人民/ng 银行/ng]NT [副/b 行长/ng]NP]NP 周正庆/npc]NP 在/pg 会/ng 上/f 发/vg 了/ut 言/ng 。/wd

该例中被识别的后一个并列项是“[[中国/nps 人民/ng 银行/ng]NT [副/b 行长/ng]NP”，因为它跟前一个并列项的长度一致，都是三个词（“中国人民银行”是机构名，被当作一个词）。后面的人名“周正庆”没有被包括进来，如果能加上短语核心词的对称评价，应能避免这一错误。

7.5 本章小结

本章根据并列结构的对称性原理提出了并列结构的概率模型，该模型可以很好地描述并列结构在功能类型和节律上的对称性。本章给出了功能和长度对称性的概率评价，并描述了利用并列结构的概率模型识别并列项的算法。把这一模型作用于实语块分析系统的输出可以纠正很多因为并列连词而产生的错误。实验表明，利用并列结构的概率模型可以利用有限的代价得到并列项的高效识别。

第八章　总结与展望

8.1 全文总结

本文的主要工作是对汉语非受限文本进行实语块分析。实语块分析不是完全分析，所以它应该归入浅层句法分析或部分句法分析的范畴。但本文提出的实语块分析和一般的浅层分析（如 Church 提出的基本名词短语和 Abney 提出的非嵌套的语块）有很大差别。一般的浅层分析只涉及短语边界的界定，而不涉及短语的层次划分。其实质是在词间的空格处找到隐藏的短语边界。从所用的方法来看，一般都是采用一些“线性”的方法（如 HMM 之类），力图避开句子中的层次结构。我们认为，层次性是自然语言结构的本质特征，离开句子的层次构造去寻找短语的边界，其结果只能得到一些“简单的非递归的”短语。要识别实语块这样可能有复杂层次构造的短语，用一般的线性的处理办法难以奏效。这也正是我们采用 CFG 作为实语块分析的基本手段的原因所在。

CFG 的最大优点在于它可以很方便地描述自然语言句子中的层次构造，但是 CFG(包括 PCFG)的缺点在于生成能力过强，分析时会产生大量的歧义，分析效率很差。其原因在于 CFG 过于概括，它在把丰富的自然语言句子抽象为一个简洁的形式系统的过程中丢掉了很多关于词汇的、关于具体结构的丰富的信息。所以，我们应该设法弥补 CFG 的缺点，同时充分发挥它的优点。本文提出概率上下文无关语法和概率属性相结合的实语块分析模型的动机正在于此。CFG 的缺点在于过于概括，把很多性质很不相同的对象混为一谈。我们要弥补 CFG 的缺陷，就要设法将 CFG 的规则细化，各种基于合一或基于限制的语法模型都因此而提出。但是这些复杂而繁琐的限制要靠语言学家一一给出是十分费时费力的，而且很难保证知识的完备性和正确性。

我们认为，应该区分语言知识和语言学知识这两个概念。所谓语言知识是一个人关于某种自然语言的知识，只要一个人能够听懂并能说某种语言，那么他就具备关于这种语言的语言知识。语言学知识是语言学家从大量语言现象或言语中抽象出来的关于语言的理性认识，这些知识跟语言学家对语言的研究有关，可以说永无止境。即使是一位很有成就的语言学家，她（他）也只掌握这些知识中的一小部分，因为有很多语言学知识至今还没有被发现。基于限制的语法理论中的“限制”就涉及很多至今还没发现的知识。比如，我们说“吃”的宾语是“食物”，但对“吃了一记闷棍”、“吃一回苦头”怎么描述？对“研究”的宾语的限制怎么描述？凡此种种，不胜枚举。但是基于语料库的方法并不依赖这么高级的知识，只要一个人具有某种语言的语言知识，加上一些并不很多的语言学知识，她（他）就可以进行语料的标注工作。比如，我们给语料进行实语块标注，就需要具有汉语语法学中关于词、词类和短语结构的基本知识，只要具备这些知识（当然还需要责任心）就可以按照某种规范对语料进行标注。标注的这些知识是语言学知识，不过这些语言学知识并不是抽象的，它反映语言学家对具体语言现象的认识，比如，给“研究”标上动词，给“研究 问题”标上动词短语，就反映了语言学家的认识。当标注的语料达到一定规模之后，我们就可以通过统计手段得到语料库中所包含的语言学知识。例如，从标注了实语块的语料库中我们就可以得到一个描述实语块结构的规则集，这个规则集有几百条规则，如果靠人来写这些规则是十分困难的。这说明了基于语料库的方法的优势。它对于解决自然语言处理中的知识瓶颈问题是至关重要的。

基于语料库的方法并不排斥规则。事实上，在我们的工作中就很好地体现了规则和统计的统一。给语料库标注什么样的知识，以及要从语料库中获取什么样的知识，这些都离不开我们对语言和语言处理的认识。比如，我们利用结构的节律属性对规则进行约束就受到了语法研究中关于节律对句法具有制约作用的理论的影响。结构的消歧到底跟哪些因素相关，这些因素之间又是如何作用的，这些问题都值得深入研究，这些研究成果必将对自然语言的句法分析产生重要的影响。

8.2 下一步的研究

目前的汉语实语块分析系统已经具有比较好的基础，但这里仍有很多工作要做。例如，对专名短语还需要作进一步的研究，对因为“的”字结构引起的错误（比例最高）需要找到有效的解决办法等。

除了对现有的系统进一步完善之外，还可以在目前工作的基础上进一步拓展，主要的工作方向有：

1. 对概率属性问题进行进一步研究。其中一个方向是探讨跟具体词相关的属性。本文虽对词汇功能属性进行了一些探讨，但由于缺乏充足的语料，很多工作难以深入。词汇信息是最具体的信息，它对句法分析的消歧作用十分重要。近年英语有关句法分析研究的一个动向就是在句法分析过程中引入词汇化的信息。

2. 在实语块分析的基础上进行完全句法分析的研究。本文提出的实语块分析模型完全可以用在完全句法分析上。我们希望下一步能够应用这一模型进行完全句法分析的实验。这一工作的难点在于，目前还没有较大规模的汉语树库[8]。开发大规模高质量的汉语树库工作任重而道远。

3. 将实语块分析技术应用到信息提取系统中。本文所描述的命名实体识别就是信息提取的一个重要内容。下一步将尝试把实语块分析系统和信息提取中的模板匹配结合起来。在实语块分析的基础上实现模板的多层次匹配是我们努力的一个方向。

[8] 就笔者所知，目前宾州大学开发的的汉语树库是规模最大的汉语树库，但也只有 10 万词。

附录一　词类标记集优化实验中所用的三个标记集

附录 1-1　实验中使用的最大标记集及获得的优化标记集

序号	标记	说明	序号	标记	说明	序号	标记	说明
0*	np	专有名词	36*	kp	可能中缀	72*	mam	中助数词
1	ng	普通名词	37	in	名词性成语	73	mab	后助数词
2*	nt	时间词	38*	iv	动词性成语	74*	qns	个体量词
3*	ns	处所词	39	id	副词性成语	75	qnu	集合量词
4	nf	方位词	40*	il	连接语	76*	qnk	种类量词
5	ag	一般形容词	41*	npx	汉人姓氏	77	qng	量词“个”
6	az	状态词	42*	npm	人名	78	qnm	度量词
7*	ab	区别词	43*	npu	机构名	79*	qnc	不定量词
8*	va	助动词	44	nps	地名	80	qnt	临时量词
9*	vi	系动词	45*	npr	其他专名	81	qv0	动量词
10*	vf	形式动词	46	ng0	普通名词	82	qvt	临时动量词
11	vv	"来/去" + VP	47	ngl	离合名词	83	usd	助词“的”
12	vt	动词用作体词	48	ag0	普通形容词	84	usz	助词“之”
13	vw	动词用作谓词	49	agz	形容词作状语	85	usy	助词“似的”
14*	mg	一般数词	50	agb	形容词带宾语	86	usi	助词“地”
15*	ma	助数词	51	ags	形容词作主宾语	87*	usf	助词“得”
16*	qn	名量词	52	agx	形容词作 NP 中心语	88	uss	助词“所”
17*	qv	动量词	53	vtz	动词作主语	89*	utl	助词“了”
18*	qt	时间量词	54	vtb	动词作宾语	90	utz	助词“着”
19	ra	代词作定语	55	vtp	动词作定语	91	utg	助词“过”
20*	rs	代词作主宾语	56	vtx	动词作 NP 中心语	92*	n	名词
21	rp	代词作谓语	57	vw0	动词不带宾语	93*	a	形容词
22*	rd	代词作状语	58	vwn	动词带 NP 宾语	94*	v	动词
23	pg	一般介词	59	vwv	动词带 VP 宾语	95*	m	数词
24*	pa	介词“把”	60	vwa	动词带形容词宾语	96*	q	量词
25*	pe	介词“被”	61	vws	动词带小句宾语	97*	r	代词
26	pz	介词“在”	62	vwd	动词带双宾语	98*	p	介词
27*	db	否定前副词	63	vwj	动词带兼语	99*	d	副词
28	dd	程度副词	64	vwc	动词作补语	100*	c	连词
29*	dr	其他副词	65	mgx	基数词	101*	u	助词
30*	us	结构助词	66	mgw	位数词	102*	y	语气词
31*	ut	时态助词	67	mgg	概数词	103*	o	拟声词
32	ur	其他助词	68	mgm	数量词	104*	e	叹词
33*	kh	前缀	69	mgh	数词“半”	105*	k	词缀
34	kn	名词后缀	70	mgo	数词“零”	106*	i	成语
35*	kv	动词后缀	71*	maf	前助数词			

附录 1-2　标记集 TS2

标记	注释	标记	注释	标记	注释
n	名词	v	动词	u	助词
t	时间词	m	数词	y	语气词
s	处所词	q	量词	o	拟声词
f	方位词	r	代词	e	叹词
a	形容词	p	介词	k	词缀
b	区别词	d	副词	i	成语
z	状态词	c	连词		

附录 1-3　标记集 TS3

标记	注释	标记	注释	标记	注释
n	名词	r	代词	y	语气词
a	形容词	p	介词	o	拟声词
v	动词	d	副词	e	叹词
m	数词	c	连词	k	词缀
q	量词	u	助词	i	成语

附录二　实语块分析中所用的词类标记集和短语标记集

附录 2-1　词类标记集

标记	说明	标记	说明	标记	说明
ng	普通名词	dd	程度副词	usi	结构助词“地”
npc	汉人姓名	dr	其他副词	usf	结构助词“得”
npf	外国人名	maf	前助数词	uss	结构助词“所”
npm	人名	mam	中助数词	ut	时间助词
npx	汉人姓氏	mab	后助数词	ur	其他助词
nps	地名	mx	系数词	e	叹词
npu	机构名	mw	位数词	o	拟声词
npr	其他专名	mg	概数词	y	语气词
t	时间词	mm	数量词	l	连接语
s	处所词	ms	序列词	kh	名词前缀
f	方位词	qn	名量词	kn	名词后缀
va	助动词	qv	动量词	kp	可能中缀
vi	系动词	qt	时量词	kv	动词后缀
vf	形式动词	qu	不定量词	wb1	左标号
vg	一般动词	qc	复合量词	wb2	右标号
a	形容词	pba	介词“把”	wm	中点号
b	区别词	pbe	介词“被”	wd	点号
z	状态词	pg	其他介词	xe	英文
r	代词	c	连词	xm	数字
db	否定前副词	usd	结构助词“的”	xg	其他符号

附录 2-2 实语块分析中的短语标记集

标记	注释	标记	注释
NP	名词性短语	AP	形容词性短语
V	简单动词性短语	B	区别词性短语
VP	复杂动词性短语	DP	副词性短语
S	主谓结构	NO	非结构

附录三　实语块短语标注规范

一、　目标

在句子中的一个实词序列中标注出最大的短语及其内部结构。
实词包括：名词（含专名）、动词（助动词、系动词除外）、形容词、区别词、状态词、时间词、处所词、实义副词。

二、　短语类型

1.　主谓短语

由主语和谓语两部分组成的短语。标记为 S。

（1）名词或名词性短语作主语。如：

农业丰收　太阳落山　婚姻破裂　股价上升　财政包干　价格失控
农民使用　国家规定　单位持有　军队开办　老百姓关心　（+ “的”）
人多　客满　质量好　档次高　名气大　规模小
天气晴朗　气候恶劣　价格低廉　条件艰苦　道路狭窄　方法简单

（2）动词或动词性短语作主语。

影响重大　开工不足　分工明细　创收难　调节有力　亏损严重
发行结束　立法滞后　发展进入较高阶段　偿还有保障
卖粮难　存在政治分歧在所难免　购买国债安全可靠

（3）主谓短语作主语。

花样翻新快　乡镇企业嫁接外资发展迅速　外资进入胶东半岛势如潮涌

（4）主谓短语作谓语。如：

管理人员素质低下　北京市场食油断档　个别零配件质量欠佳
高档服装价格高扬　城乡居民收入增加　村民们喜悦心情溢于言表
台湾股市违法脱序现象普遍　解决居民吃菜难问题难度很大

2．名词性短语

功能相当于名词的短语。标记为 NP。

（1）　偏正式。

名词 + 名词：

市场经济　边防部队　业务工作　边疆地区　人力机械　公有制模式

动词 + 名词：

供给量　投资人　调查权　竞争力　接待费　变质煤
上市股票　涉及金额　投资规模　发展速度　改革难点　协作伙伴

名词 + 动词：

经济犯罪　商业竞争　服装定价　股份合作　市场定位　业务往来
批量生产　后勤服务　产权转让　信息交流　工序分解　款式设计

形容词 + 名词：

高价　大厂　鲜鱼　新技术　大公司　小山村　旧体制　硬道理

贫困村　精细菜　走俏商品　落后观念　科学论断　强大动力　严峻事实

区别词 + 名词：

主渠道　新型梁　合资企业　高额利润　违禁药品　人为障碍

便民措施　一贯立场　切身利益　闲散资金　国有企业　超低高度

名词 + 动词：

农业生产　市场竞争　价格违法　技术创新　行政管理　基础建设

国际分工　民意测验　设备挖潜　收入安排　钢材生产　食品出口

形容词 + 动词：

高增长　大检查　新变化　小调整　新发现　正当竞争

重大改革　正确领导　突出表现　有益补充　不良影响　科学总结

动词 + 动词：

生产需要　廉政建设　抽样调查　咨询服务　改革试验　投资管理

VP 直接作定语：

（c） VP 是状中结构。如：

违法扣车事件　正在建房户　依法纳税意识

今年到期国债　合资建厂事宜　建筑用砖

（d） VP 是述宾结构。如：

含绒量　建房区域　反暴利法　养鱼专业户

无党派人士　操纵股市意图　建设有中国特色社会主义理论

VP 作中心语。如：

权力再分配　家禽优化饲养　国债流通转让　农产品储运加工

（2） 并列式。

客货　军民　省区　县乡　粮棉油　鱼禽猪　贸工农　农工商

戈壁草原　边贸旅游　广播电视　卫生保健

（3） 附加式。

动词 + 后缀：

受害者　购买者　普及率　转化率　操作员　宣传员　投资额　成交额

加工业　制造业　衰退期　生长期　发展观　开放度　发酵法　劣质品

名词 + 后缀：

股份制　公司制　成衣率　利润率　物价员　季节性　风险金　同事们

饮食业　旅游业　间歇期　速度观　水产热　房地产热　实业家　策略家

形容词 + 后缀：

一般性　隐蔽性　知名度　透明度　神秘感　成熟期　充足率　饥渴症

NP + 后缀：

农产品加工业　社会服务业　股份合作制　浮动汇率制　质量监督员

行政指令性　合同外资额

VP + 后缀：

有消息来源者　产酸率　透光率　罚款额　试生产期　负债建设期

主谓结构 + 后缀：

外商投资热　厂长负责制

3．动词性短语

功能相当于动词的短语。我们把动词性短语分为两类：

（1） 简单动词性短语，标记为 V，包括由两个具体词组成的动词性短语（述宾结构除外）。之所以要把这些动词性短语分离出来单立一类，主要是因为这类动词性短语与单个动词更接近。

（2） 复杂动词短语，标记为 VP，简单动词性短语之外的动词性短语。

3.1 简单动词性短语

主要类型有：

（2） 述补式。

(a) 补语是动词。如：

走上市场　迈出更大步伐　买到粮食　评为优秀教师
列入重点课题　上交给学校　增至 20 万　改建成宿舍

(b) 补语是形容词。如：

打好基础　摸清家底　看准目标　炒高股价　打牢基础
发展迅猛　掌握不当　认识不足　把握好时机　用足政策

（3） 并列式。如：

改革开放　审议批准　解释说明　保值增值　参政议政
库存积压　培养提高　推广应用　信任支持　暴涨暴跌

（4） 状中式。

(a) 状语是形容词。如：

正式挂牌　迅速回落　稳定发展　公开发行　积极参与　严厉打击
顺利完成　及时改进　正常使用　多挑剔　少买　妥善解决

(b) 状语是副词。如：

相互协调　持续发展　有所增长　长期徘徊　逐步解决　大幅度增加
婉言谢绝　巍然兴起　更为加剧　高度重视　最为关心　大大降低

(c) 状语是时间词。如：

当时公布　今年到期

(d) 状语是处所词。如：

庞国增现场指挥

(e) 状语是普通名词。如：

在改革整体推进的攻坚阶段
司机出身的粗人　高中毕业后　与松下技术合作
工业用沙　机车用油　电用瓷瓶　建筑用砖

(f) 状语是动词。如：

配套改革　联合发起　放心购买　抓紧干　合伙持股
搭车涨价　违约交割　到期偿还　胜利完成　负债经营

(4) 附加式。后缀只有“化”。

(a) 名词 + 后缀。 如：

法制化　制度化　社会化　国际化　时装化

(b) 形容词 + 后缀。如：

具体化　系统化　准时化　合理化

(c) 区别词 + 后缀。如：

间接化　　外向化

3.2 复杂动词短语

（1） 述宾式。

(A) 述语和宾语都是单个的词。

(a) 宾语是名词。如：

含泪　长庄稼　负责任　静候佳音　转换机制　嫁接外资

(b) 宾语是动词。如：

有保证　受损害　值得重视　深化改革　实行封锁　构成犯罪

(c) 宾语是形容词。如：

获得成功　保持稳定　摆脱贫穷　趋向健康　有奥妙

（B）述语和宾语中至少有一个不是单个的词。

(a) 宾语是名词性的。如：

狠抓基层建设　建立健全规章制度　拉动经济发展　跑遍大小沙原

（b）宾语是动词性的。如：

开始攒钱 着手制定计划 实行敞开销售　获得迅猛发展　坚持改革开放

（c）宾语是形容词性的。如：

显得更为必要　充满诡异多变　促进普遍繁荣

（2） 状中式。如：

（a）副词作状语。如：

进一步发展经济　难以平抑物价　大力调整经营战略

四处寻找合作伙伴　不断写信　大量收集证据材料

（b）形容词作状语。如：

科学养鱼　明显无效益　多储备原料　充分利用外资

直接经营企业　密切联系群众　盲目追求速度　努力提高经济效益

（3） 并列式。如：

缺粮缺肉　有职有责　爱说爱笑　还本付息　给职给权

缺窗少门　找矿探宝　引凤筑巢　求高档逐名牌

(4) 连谓式。如：

深入实际进行教改　去煤矿运煤　想办法提高质量

支持政府搞城建　引导农民进入市场　使产权进入企业

4．形容词性短语

以形容词为核心的谓词性短语。标记为 AP。

主要结构类型有：

（1）并列式：　新老（用户）　正确有效　广泛深入持久

（2）状中式：　最大　更突出　极度短缺　尤为迅速　（程度副词 + 形容词）

日趋明显　着实可观　长期稳定　空前巨大 （其他副词 + 形容词）

（3）补充式：　好极了　热闹起来

5．区别词性短语

功能相当于区别词的短语，即只能作定语的短语。标记为 B。

主要结构类型有：

（1） 附加式。

X + 后缀， X 可以是名词、动词、区别词、形容词、副词或主谓结构，其中前两类较为常见。如：

名词 + 后缀： 指令性计划　地方性法规　群众性组织　速度型发展路子

动词 + 后缀： 消费性资金　混合型经济　封闭式运营　开发性扶贫

区别词+后缀： 综合性大学

副词 + 后缀： 暂时性因素

形容词+后缀： 紧密型经济联合

主谓结构 + 后缀： 劳动密集型产业

前缀 +X， 前缀主要是“非”，X 可以是名词、区别词或区别词性短语。如：

前缀 + 名词： 非公有制企业　非名牌时装

前缀 + 区别词：非常设机构　非高档商品

前缀 + 区别词性短语： 非基础性产业　非正常性亏损

（2） 并列式，如：

假冒伪劣商品　定量定性分析　稳产高产农田

优质低价商品　高产优质高效农业

（3） 述宾式。主要由“跨 +NP”构成，这类述宾结构往往既可以作定语，又可以作状语，作定语时是区别词性的，作状语时是副词性的。如：

跨世纪工程

具有跨领域、跨学科的特点

组建跨地区、跨行业的大型企业集团

6．副词性短语

功能相当于副词的短语，即只能作状语的短语。标记为 DP。

主要类型有：

（1） 定中式。这类短语一般兼属区别词性短语。如：

多渠道筹集资金

多层次、全方位对外开放

（2） 并列式。如：

全面超额完成任务

保持经济持续快速健康发展

（3） 述宾式。如：

跨地区流动

跨乡跨县搞横向联合

读完一个专业，跨学科继续学习者也不计其数

附录四　实语块分析系统部分输出结果

@/wd [[前景/npu　公司/ng]NT [推出/vg [[高级/b 电脑/ng]NP [排版/vg 系统/ng]NP]NP]VP]S

绝大部分/mm 干部/ng 连/pg 星期天/t 、/wm 节假日/t ，/wd 甚至/c 春节/t 也/dr [难得/vg 休息/vg]VP 。/wd

非但/c [党务/ng 工作/ng]NP ，/wd 她/r 还/dr 得/va 协助/vg 董事长/ng 、/wm 总经理/ng 做/vg [人事/ng 工作/ng]NP 、/wm [[思想/ng 教育/vg]NP 工作/ng]NP 、/wm [[职工/ng 文化/ng]NP 生活/ng]NP 等等/ur 。/wd

@/wd [[[广西/nps　玉林/nps　地委/ng]NT 书记/ng]NP 李新明/npc]NP 对/pg 我们/r 说/vg ：/wd “/wb1 ‘/wb1 玉柴/npu ’/wb2 是/vi [玉林/nps　地区/ng]NS ９００/mx 万/mw 人民/ng [引以为/vg 骄傲/a]VP 的/usd 一/mx 颗/qn 明珠/ng ，/wd 是/vi [玉林/nps　地区/ng]NS 的/usd 希望/ng 所/usu 在/vg 。/wd

其次/c ，/wd 在/pg 发展/vg 的/usd 层次/ng 上/f ，/wd 要/va 破除/vg “/wb1 [落后/a 地区/ng]NP 只/dr 能/va [起点/ng 低/a]S ”/wb2 的/usd [片面/a 观点/ng]NP ，/wd 在/pg [广泛/a [[发展/vg 劳动/vg]V [[密集/a 型/kn]NP 产业/ng]NP]VP]VP 的/usd 同时/t ，/wd [力争/vg 发展/vg]VP 一/mx 批/qn [[技术/ng 起点/ng]NP 高/a]S 、/wm [[关联/vg 效应/ng]NP 强/a]S 、/wm [[市场/ng 前景/ng]NP 好/a]S 的/usd [高科技/ng 项目/ng]NP 。/wd

拥有/vg 一/mx 个/qn [博士后/ng [流动/vg 站/ng]NP]NP 、/wm 七/mx 个/qn [博士/ng 点/ng]NP ，/wd [在读/b 学生/ng]NP 达/vg　９００/mx 人/ng 的/usd [法律/ng 系/ng]NP ，/wd 把/pba 学科/ng 建设/vg　与/pg 开辟/vg 新/a　的/usd [学术/ng 阵地/ng]NP [结合/vg 起来/vg]V　，/wd 在/pg 国内/s 率先/dr [设立/vg 机构/ng]VP 对/pg [知识/ng [产权/ng 法/ng]NP]NP 、/wm [物证/ng 技术/ng]NP 、/wm [金融/ng 法/ng]NP 、/wm [国际/ng 刑法/ng]NP 等/ur [开展/vg 研究/vg]V ；/wd

@/wd 我/r 就/dr 举/vg 一些/mm [农村/ng 常见/a]S 的/usd 例子/ng ，/wd 如/pg [[农业/ng 技术/ng]NP 人员/ng]NP 通过/pg 广播/ng 、/wm 电视/ng 向/pg 农民/ng [传授/vg [农业/ng [技术/ng 知识/ng]NP]NP]VP　，/wd 可/c 许多/mm 农民/ng 不/dr [去/vg 听讲/vg]V 。/wd

@/wd 本/r 规定/ng 由/pg 财政部/ng [负责/vg 解释/vg]V　。/wd

[[中国/nps 政局/ng]NP 稳定/a]S ，/wd 对/pg 外资/ng [实行/vg [优惠/a 政策/ng]NP]VP　，/wd [[基础/ng 设施/ng]NP 完善/a]S ，/wd [资源/ng 丰富/a]S ，/wd [[劳动力/ng 工值/ng]NP 低/a]S ，/wd 是/vi [世界/ng 少有/a]S 的/usd 良好/a 的/usd [投资/vg 市场/ng]NP 。/wd

三/mx 、/wm [加强/vg 政府/ng]VP 对/pg [市场/ng 物价/ng]NP 的/usd [调控/vg 管理/vg]V ，/wd [包括/vg [建立/vg 粮食/ng]NP]VP 和/c [[副食品/ng 价格/ng]NP [风险/ng 基金/ng]NP]NP ，/wd 加强/vg 对/pg ２０/mx 种/qn [居民/ng [基本/a [生活/ng 必需品/ng]NP]NP]NP 和/c 服务/vg 价格/ng 的/usd 监审/vg ，/wd 对/pg 商品/ng 和/c [[服务/vg 实行/vg]V 明码标价/vg]VP ，/wd 在/pg 全

国/ng [[开展/vg 物价/ng]VP [大/a 检查/vg]V]VP ； /wd

后来/t 我/r 想/vg ， /wd 这/r 巨响/ng 也许/dr 是/vi 这/r 座/qn [文化/ng 名城/ng]NP ， /wd 在/pg [遭遇/vg [文化/ng 破坏/vg]NP]VP 时/f [发/vg 出/vg]V 的/usd 一/mx 声/qv 无奈/a 的/usd 叹息/vg 。/wd

@/wd 第/maf 一/mx ， /wd [定向/dr 招生/vg]V ， /wd [定向/dr 代培/vg]V ， /wd [利用/vg [高等/b 院校/ng]NP]VP 等/ur [[社会/ng 教学/ng]NP 力量/ng]NP 为/pg 企业/ng [培养/vg [技术/ng 骨干/ng]NP]VP 。/wd

各/r [成员/ng 国/ng]NP 的/usd 利率/ng 去年/t 底/f 以来/f 几/mg 次/qv 下调/vg 。/wd

据/pg 统计/vg ， /wd 山西省/nps [调整/vg 班子/ng]VP ８０/mx 个/qn ， /wd [占/vg [[亏损/vg 企业/ng]NP 数/ng]NP]VP 的/usd ３５/mx ·/wm ７%/mx ； /wd 天津市/nps 调整/vg ４２/mx 个/qn ， /wd 占/vg ４３/mx ·/wm ３%/mx ； /wd 沈阳市/nps 调整/vg １３６/mx 个/qn ， /wd 占/vg ４３/mx ·/wm ８%/mx 。/wd

遗憾/a 的/usd 是/vi ， /wd 种/qn 种/qn 原因/ng ， /wd 十/mw 余/mab 年/qt 来/f ， /wd 未/dr 能/va 与/pg 华西/nps 谋面/vg ；/wd 直到/pg 今年/t ９/mx 月/qt ，/wd 我/r 受/vg [[现在/t 挂职/vg]V 工作/vg]VP 的/usd [湖南/nps 张家界/nps 市委/ng]NT 委托/vg ， /wd 送/vg ４０/mx 名/qn [乡镇/ng 干部/ng]NP “/wb1 求学/vg ”/wb2 华西/nps ， /wd 才/dr [踏/vg 上/vg]V 苏南/nps 这/r 块/qn 小小/z 的/usd 土地/ng 、/wm 又/dr 是/vi 全国/ng 赫赫有名/vg 的/usd 村庄/ng 。/wd

[齐铁/npu 人/ng]NP 在/pg [总结/vg [历史/ng 教训/ng]NP]VP 中/f 确定/vg 了/ut [干部/ng 安全/ng]NP 管理/vg 的/usd [量化/vg 标准/ng]NP ， /wd 并/c 用/vg “/wb1 五定/ng ”/wb2 的/usd 形式/ng [固定/vg 下来/vg]V 。/wd

我/r 后来/t 听说/vg ， /wd 我/r 的/usd [演讲/vg 稿/ng]NP 被/pbe 印发/vg 给/vg 成千上万/mg 的/usd 官员/ng 和/c 企业/ng [管理/vg 人员/ng]NP 。/wd

@/wd 长乐/nps ， /wd 正/dr [跃/vg 上/vg]V 新/a 的/usd [历史/ng 制高点/ng]NP 。/wd

这/r 是/vi 百/mw 万/mw [海内外/s [长乐/nps 人/ng]NP]NP 政治/ng 、/wm 经济/ng 生活/ng 中/f 的/usd 一/mx 件/qn [大/a 喜事/ng]NP ， /wd 是/vi [长乐/nps [发展/vg 史/ng]NP]NP 上/f 的/usd [新/a 里程碑/ng]NP ， /wd 标志/vg 着/ut [建/vg 县/ng]VP １３７１/mx 年/qt 的/usd 长乐/nps ， /wd 其/r 经济/ng 、/wm 社会/ng 发展/vg 将/dr 进入/vg 一/mx 个/qn 崭新/b 的/usd 阶段/ng 。/wd

消息/ng 还/dr 未/dr [正式/a 公布/vg]V ，/， 先期/dr 获悉/vg 此/r 情/ng 的/usd 人们/ng 便/dr 纷纷/dr [来到/vg 国家科委/npu]VP ， /wd 要求/vg 在/pg 本/r 地区/ng 、/wm 本/r 行业/ng 、/wm 本/r 企业/ng 推广/vg 。/wd

@/wd 他/r 说/vg ， /wd 为了/pg 进一步/dr 解放/vg 和/c 发展/vg 生产力/ng ， /wd [增强/vg 企业/ng]VP 的/usd 活力/ng ， /wd [建立/vg 企业/ng]VP 的/usd 激励/vg 和/c 约束/vg 机制/ng ， /wd 必须/dr 按照/pg [建立/vg [[社会主义/ng 市场/ng]NP [经济/ng 体制/ng]NP]NP]VP 的/usd 要求/ng ，

/wd 在/pg [巩固/vg [公有制/ng [主体/ng 地位/ng]NP]NP]VP 的/usd 前提/ng 下/f ，/wd 继续/vg 把/pba 企业/ng [改革/vg 推/vg]V 向/pg 前进/vg ，/wd [建立/vg [现代/t [企业/ng 制度/ng]NP]NP]VP 。/wd

要/va 在/pg 这/r 个/qn 前提/ng 下/f ，/wd [明确/vg [[国有/b 资产/ng]NP [投资/vg 主体/ng]NP]NP]VP ，/wd 即/vi [国有/b 股/ng]NP 的/usd [[持股/vg 机构/ng]NP 问题/ng]NP 。/wd

企业/ng 先/dr [富/a 起来/vg]V 后/f 怎么办/r ？/wd

@/wd [亚细亚/npu 大/a 酒店/ng]NT 并/db 不/dr [追求/vg [虚/a 名/ng]NP]VP 。/wd

同时/c ，/wd 它/r [[使/vg [[外出/vg 务工/vg]V [经商/vg 者/kn]NP]NP]VP [无/vg 后顾之忧/ng]VP]VP ，/wd [避免/vg [粗放/b 经营/vg]NP]VP ，/wd [[防止/vg 农业/ng]VP 萎缩/vg]VP 。/wd

@/wd １９９３/mx 年/qt １０/mx 月/qt ２４/mx 日/qt ，/wd [长委会/npu 宜昌/nps 监理/vg 中心/ng]NT 第/maf 一/mx 次/qv [[跨/vg 进/vg]V [三峡/nps [工程/ng 大门/ng]NP]NP]VP 。/wd

@/wd 同/pg 白裤瑶/npr 一块/dr “/wb1 迁移/vg ”/wb2 的/usd 还/dr 有/vg 由/pg [县委/ng [主要/b 领导/ng]NP]NP 挂帅/vg 的/usd [移民/ng [领导/vg 小组/ng]NP]NP 的/usd [干部/ng 们/kn]NP ，/wd 整整/dr 一/mx 年/qt 多/mab 的/usd 时间/ng 里/f ，/wd 他们/r [常驻/vg [移民/ng 点/ng]NP]VP ，/wd 同/pg [白裤瑶/npr [移民/ng 们/kn]NP]NP 一道/dr [睡/vg 草棚/ng]VP 、/wm [啃/vg 干粮/ng]VP 、/wm [吃/vg 野菜/ng]VP 、/wm [喝/vg 山泉/ng]VP 、/wm [搞/vg 开发/vg]VP ，/wd [[精心/a 安置/vg]V [移民/ng 们/kn]NP]VP 的/usd 生产/vg 和/c 生活/ng 。/wd

@/wd 廖朝元/npc 、/wm 蓝元清/npc 、/wm 何光贵/npc 、/wm 何小立/npc 等/ur 户主/ng 向/pg 记者/ng 谈/vg 了/ut 感受/ng ，/wd [[得意/a 神情/ng]NP 溢于言表/vg]S 。/wd

正/dr 是/vi 带/vg 着/ut 这/r 个/qn 引进/vg “/wb1 外资/ng ”/wb2 再展宏图/vg 的/usd 设想/ng ，/wd 他/r [走/vg 进/vg]V 了/ut [西山/nps 矿务局/ng]NT [装潢/vg 考究/a]S 的/usd 会议室/ng 。/wd

[回到/vg 邵州/nps]VP ，/wd 已/dr 是/vi 半夜/t １０/mx 点/qt 半/mx 。/wd

@/wd 党/ng 的/usd 十/mw 四/mx 届/qn 三中全会/npr 通过/vg 的/usd《/wb1 决定/ng 》/wb2 ，/wd 明确/a 地/usi 将/pba [劳动力/ng 市场/ng]NP 写/vg 在/pg [中央/ng 文件/ng]NP 上/f ，/wd 将/pba [劳动力/ng 市场/ng]NP [作为/vg [[培育/vg 市场/ng]VP 体系/ng]NP]VP 的/usd 重点/ng 之一/r ，/wd 并/c 用/vg 一/mx 段/qn 文字/ng 对/pg [改革/vg [劳动/vg 制度/ng]NP]VP ，/wd 逐步/dr [[形成/vg [劳动力/ng 市场/ng]NP]VP [加以/vf 阐述/vg]VP]VP ，/wd 这/r 是/vi [改革/vg 开放/vg]V １５/mx 年/qt 来/f 的/usd [新/a 认识/ng]NP 、/wm [新/a 思想/ng]NP 、/wm [新/a 观点/ng]NP ，/wd 对于/pg [建立/vg [[社会主义/ng 市场/ng]NP [经济/ng 体制/ng]NP]NP]VP [具有/vg 重要/a]V 的/usd 意义/ng 。/wd

职工/ng 不/dr 能/va 从/pg 需要/va 调整/vg 的/usd [生产/vg 部门/ng]NP [转移/vg 到/vg]V 需要/va 发展/vg 的/usd [生产/vg 部门/ng]NP ，/wd 与/pg 此/r 相关/vg 的/usd 固定资产/ng 和/c 流动资

金/ng 也/dr 不/dr 能/va [作/vg 相应/vg]V 的/usd 调整/vg 。/wd

相反/vg 的/usd 情况/ng 是/vi ，/wd 劳动者/ng 并非/vi “/wb1 一无所有/vg ”/wb2 ，/wd 他/r 占有/vg 土地/ng 和/c 一部分/mm [生产/vg 工具/ng]NP ，/wd 为了/pg 追求/vg [更/dd 多/a]AP 的/usd 收入/ng ，/wd 还/dr 要/va [[去/vg 出卖/vg]V 劳动力/ng]VP 。/wd

例如/l ，/wd 为了/pg [保证/vg [国有/b 资产/ng]NP]VP 的/usd [保值/vg 增值/vg]V ，/wd [济南/nps 局/ng]NP 规定/vg 的/usd [考核/vg 内容/ng]NP 有/vg 五/mx 点/ng 。/wd

@/wd [书面/b 通知/ng]NP 和/c [书面/b 回复/ng]NP 的/usd [具体/a 形式/ng]NP 由/pg 公司/ng 在/pg [公司/ng 章程/ng]NP 中/f [作出/vf 规定/vg]VP 。/wd

@/wd 九/mx 月/qt 十/mw 日/qt 至/pg 十/mw 二/mx 日/qt ，/wd [李鹏/npc 总理/ng]NP 和/c [国务委员/ng 陈俊生/npc]NP 由/pg [[[中共/npu 河南/nps 省委/ng]NT 书记/ng]NP 李长春/npc]NP 、/wm [省长/ng 马忠臣/npc]NP 陪同/vg ，/wd 在/pg 洛阳/nps 考察/vg 了/ut [[黄河/nps 小浪底/nps]NS [[水利/ng 枢纽/ng]NP 工程/ng]NP]NP ，/wd 并/c 参加/vg 了/ut [工程/ng [开工/vg 典礼/ng]NP]NP ；/wd 参观/vg 了/ut [洛阳/nps 棉纺织厂/ng]NT 和/c [洛阳/nps 浮法玻璃/ng 集团/ng 公司/ng]NT ，/wd 并/c 同/pg 一些/mm [国有/b [[大中型/b 企业/ng]NP 负责人/ng]NP]NP 座谈/vg ；/wd 还/dr 走访/vg 了/ut 粮店/ng 和/c 农户/ng 。/wd

由于/c [制假/vg 地/ng]NP 的/usd 保护/vg ，/wd 这里/r 只/dr [有/vg 抓假/vg]V 之/usd 功/ng 而/c [无/vg 灭假/vg]V 之/usd 力/ng 。/wd

各/r 级/qn 政府/ng 要/va [精心/a 组织/vg]V ，/wd [严格/a 管理/vg]V ，/wd [全面/a 完成/vg]V 今年/t 的/usd [棉花/ng [收购/vg 任务/ng]NP]NP 。/wd

近/a 2 0/mx 多/mab 年/qt 来/f ，/wd 法国/nps 的/usd [农业/ng [合作/vg 组织/ng]NP]NP 发生/vg 了/ut [[很/dd 大/a]AP 变化/vg]NP 。/wd

@/@ [开发/vg [旅游/vg 资源/ng]NP]VP ，/wd [面临/vg [资金/ng 难题/ng]NP]VP 。/wd

各/r 级/qn [税务/ng 部门/ng]NP 要/va 加强/vg 对/pg [个人/ng 所得税/ng]NP 的/usd 征管/vg 和/c 检查/vg 工作/ng 。/wd

同时/c ，/wd 要/va [主动/a 关心/vg]V [有/vg 困难/ng]VP 的/usd 企业/ng 和/c 职工/ng ，/wd 以及/c 一些/mm [大中专/ng [学校/ng 学生/ng]NP]NP ，/wd 帮助/vg 他们/r 解决/vg 在/pg 工作/ng 中/f 和/c 生活/ng 上/f 的/usd [实际/b 困难/ng]NP 。/wd

@/wd [[国家计委/npu 主任/ng]NP 陈锦华/npc]NP 、/wm [[中国/nps 人民/ng 银行/ng]NT [副/b 行长/ng]NP]NP 周正庆/npc 在/pg 会/ng 上/f 发/vg 了/ut 言/ng 。/wd

@/wd [[[长江/nps 三峡/nps]NS 工程/ng]NP [正式/a 开工/vg]V]S

参考文献

[1] Abney, S. 1991. Parsing by chunks. In *Principle_Based Parsing*. Berwick, Abney and Tenny (eds.),. Kluwer Academic Publishers.

[2] Abney, S. 1996a. Part-of-speech tagging and partial parsing. In *Corpus- Based Methods in Language and Speech*. Church, Young and Bloothooft (eds.),Kluwer Academic Publishers. pp. 119-136.

[3] Abney, S. 1996b. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI'96 Robust Parsing Workshop*.

[4] Abney, S. 1997. Stochastic Attribute-Value Grammars, *Computational Linguistics*, 23(4): 597-618.

[5] Argamon, S. , Dagon I., and Krymolowsky Y., 1998. A memory-based approach to learning shallow natural language patterns. In *Proceedings of COLING-ACL'98*. pp. 67-73.

[6] Bäck, T. (1993). *Optimal mutation rates in genetic search*. In "Proceedings of the 5th International Conference on Genetic Algorithms" (ICGA′93), Morgan Kaufmann, pp. 2-9.

[7] Bloomfield,L. 1933. Language. 中译本《语言论》（商务印书馆，1980）

[8] Bresnan, J. and Kaplan, R.M. , 1982. Introduction: Grammars as mental representations of language. In Bresnan, J.(Ed.), The Mental Representation of Grammatical Relations. MIT Press.

[9] Brew, C. 1995. Stochastic HPSG, In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pp.83-89.

[10] Brill,E. 1993. A Corpus_Based Approach to Language Learning, Ph.D. dissertation, University of Pennsylvania.

[11] Brill, E. 1995a. Transformation_Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Computational Linguistics*, Vol. 21(4), pp. 543-565.

[12]Brill, E. 1995b. Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the 3rd Workshop on Very Large Corpora*. pp. 1-13.

[13] Cardie,Claire and Pierce,David, 1998. Error-driven pruning of treebank grammars for base noun phrase identification. In *Proceedings of COLING-ACL'98*. pp. 218-224.

[14] Charniak, E., 1993. Statistical Language Learning, The MIT Press.

[15] Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M.,1993. Equations for part-of-speech tagging. In Proceedings of the Eleventh National Conference on Artificial Intelligence. AAAI/MIT Press. pp.784-789.

[16] Chen H., Bian,G. 1998. White Pate Construction from Web Pages for Finding People in Internet, In *International Journal of Computational Linguistics and Chinese Language Processing*, 3 (1):75-100.

[17] Chen H., Ding Y., Tsai S. and Bian G. 1998. Description of the NTU System used for MET-2, In Proceedings of the Seventh Message Understanding Conference (MUC-7).

[18] Chen, H., Lee, Y.,1995. Development of a partially bracketed corpus with part-of-speech information only. In *Proceedings of the 3rd Workshop on Very Large Corpora.*. pp.162-172.

[19] Chen, H., Lee, J.C. 1996. Identification and Classification of Proper Nouns in Chinese Texts, In Proceeding of the 16th COLING, Copenhagen, .pp. 222-229.

[20] Chen, K., Chen, C. 2000. Knowledge Extraction for Identification of Chinese Organization Names,

In *Proceedings of the Second Workshop on Chinese Language Processing*, pp. 15-21.

[21] Chen, K., Chen, H.1994. Extracting Noun Phrases from Large-Scale Texts: A hybrid approach and its automatic evaluation. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* pp. 234-241.

[22] Chinchor, N. 1998. *MUC-7 Named Entity task definition*. In "Proceedings of the Seventh Message Understanding Conference" (MUC-7). http:// www.muc.saic.com/proceedings/muc_7_toc.html.

[23] Church,,K.,1988.A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp136-143.

[24] Collins,M., 1996.A new statistical parser based on bigram lexical dependencies, In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.184-191.

[25] DeRose S. 1988. Grammatical Category Disambiguation by Statistical Optimization, *Computational Linguistics*, Vol. 14 (1).

[26] Evans, D., Zhai, C. 1996. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.pp.17-24.

[27] Fano, R. M., 1961. Transmission of Information , A Statistical Theory of Communication. MIT Press.

[28] Gale,W.A. & Church, K.W.1991. Identifying word correspondences in parallel texts, in *Proceedings of DARPA Speech and Natural Language Workshop*, pp.152-157.

[29] Gazdar, G., Klein, E., Pullum, G. K. 1985. Generalized Phrase Structure Grammar. Basil Blackwell.

[30] Gazdar, G., Pullum, G.K., et al. 1988. Category structures. Computational Linguistics, 14(1):1-19.

[31] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

[32] Goodman, J. 1997. Probabilistic Feature Grammars, In *Proceedings of the International Workshop on Parsing Technologies*, September 1997

[33] Goodman, J. 1998. Parsing Inside-Out, PhD thesis, Harvard University.

[34] Grishman, R, Sundheim, B. 1996. Message Understanding Conference-6: A Brief History, In Proceedings of the 16th COLING, Copenhagen, pp.466-471.

[35] Halteren, H. 1999. *Syntactic Wordclass Tagging*, edited by Hans van Halteren, Kluwer Academic Publishers.

[36] Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press. (Second edition: MIT Press, 1992).

[37] Magerman, D. & Marcus, M. 1990, Parsing a natural language using mutual information statistics, in *Proceedings of AAAI '90*. pp.984-989.

[38] Marcus, M. 1980. A Theory of Syntactic Recognition for Natural Language, MIT Press.

[39] Marcus,M., Santorini, B., and Marcinkiewicz, M. 1993. Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistic*s, 19(2): 313-330.

[40] Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. MIT Press.

[41] Mokhtar S.,Chanod J.,1997. Incremental finite-state parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. pp.72-79.

[42] MUC-3, 1991. Proceedings of the Third Message Understanding Conference, Morgan Kaufmann.

[43] MUC-4, 1992. Proceedings of the Four thMessage Understanding Conference, Morgan Kaufmann.

[44] MUC-5, 1993. Proceedings of the Fifth Message Understanding Conference, Morgan Kaufmann.

[45] MUC-6, 1995. Proceedings of the Sixth Message Understanding Conference, Morgan Kaufmann.

[46] MUC-7, 1998. Proceedings of the Seventh Message Understanding Conference, Morgan Kaufmann.

[47] Pollard, C. and Sag, I.A. 1994. Head-driven Phrase Structure Grammar. University of Chicago Press.

[48] Quirk,R, Greenbaum, S., Leech, G, Svartvik, J. 1985. A Comprehensive Grammar of the English Language, Longman.

[49] Rabiner,Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, reprinted in *Readings in Speech Recognition* , Waibel and Lee(eds.), Morgan Kaufmann ,1990, pp. 267-96.

[50] Ramshaw L., and Marcus M. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora.*

[51] Sag, I.A. and Wasow,T.(Eds.), 1999. Syntactic Theory: A Formal Introduction. CSLI Publications, Stanford.

[52] Samuelsson C., Tapanainen P. and Voutilainen A.,1996. Inducing constraint grammars. In *Grammatical Inference: Learning Syntaxfrom Sentences,* Springer-Verlag.

[53] Skut, W. and Brants,T.,1998. A maximum-entropy partial parser for unrestricted text. In *Proceedings of the 6th Workshop on Very Large Corpora*, Montreal,Quebec,1998.Also available at cmp-lg/9807006.

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

[55] Sproat, R., Shih, C., Gale, W., Chang,N. 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, *Computational Linguistics*, 22(3):377-404.

[56] Sun, H., Yu, S. and Lu, Q. 1999. *Evaluations on Part-of-speech Tagset.* In "Proceedings of the 5th Natural Language Processing Pacific Rim Symposium" (NLPRS′99), Tsinghua University Press, pp. 25-31.

[57] Sun M., Huang,C. 1996. Word Segmentation and Part-of-speech Tagging for Unrestricted Chinese Texts, A Tutorial on the International Conference on Chinese Computing'96, Singapore.

[58] Sun, M., Shen, D. and Huang, C. 1997. CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts, In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp.119-126.

[59] Voutilainen A.,1993. Nptool, a detector of English noun phrases. In *Proceedings of the First Workshop on Very Large Corpora*.

[60] Voutilainen A. and Padro L.,1997. Developing a hybrid NP parser. In *Proceedings of the 5th Conference on Applied NaturalLanguag*e . pp.80-87.

[61] Winograd, T. 1983. Language as a Cognitive Process. Addison-Wesley.

[62] Yu S.H., Bai S.H., Wu P. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. In Proceedings of the Seventh Message Understanding Conference (MUC-7).

[63] Zhang, Y., Zhou, J. 2000. A trainable method for extracting Chinese entity names and their relateions, In *Proceedings of the Second Workshop on Chinese Language Processing*, pp. 66-72.

[64] Zipf,G. 1935. The Psycho-Biology of Language. Houghton Mifflin.

[65] Zipf,G.1949. Human Behavior and the Principle of Least Effort. Hafner, New York.

[66] 白拴虎（1995）， 汉语词切分及词性标注一体化方法，《计算语言学进展与应用》，清华大学出版社，56-61 页。

[67] 白拴虎（1996），基于统计的汉语语料库词性自动标注的研究与实现，载《语言信息处理专论》，黄昌宁、夏莹主编，清华大学出版社， 37-77 页。
[68] 陈信希，李振昌（1994） 中文文本组织名之辨认，Communications of COLIPS, 4(2):131-142.
[69] 冯胜利（2000），汉语韵律句法学，上海教育出版社。
[70] 郭志立，苑春法，黄昌宁(1996)，用统计方法研究“的”字短语的结构与边界，载《计算机时代的汉语和汉字研究》（罗振声、袁毓林主编），清华大学出版社。174-183 页。
[71] 胡明扬（1996），词类问题考察，北京语言文化大学出版社。
[72] 李如龙（1998） 汉语地名学论稿，上海教育出版社。
[73] 李文捷，周明等(1995)，基于语料库的中文最长名词短语的自动抽取，载《计算语言学进展与应用》（陈力为、袁琦主编），清华大学出版社。119-124 页。
[74] 刘长征（1998），基于词性标注语料库中普通名词序列的捆绑研究，载《1998 中文信息处理国际会议论文集》（黄昌宁主编），清华大学出版社。244-250 页。
[75] 刘开瑛 (2000). 中文文本自动分词和标注，商务印书馆。
[76] 马真，陆俭明（1996），“名词+动词”词语串研究，《中国语文》4 期。
[77] 沈达阳，孙茂松，黄昌宁（1995），中国地名的自动辨识，《计算语言学进展与应用》，清华大学出版社，68-74 页。
[78] 宋柔，邱超杰，欧阳龙根，徐绿兵，王鑫（1996）， 二元接续关系及其在汉语分词和校对中的应用，Proceedings of International Conference on Chinese Computing’96, Singapore, pp.428-433.
[79] 宋柔，朱宏，潘维桂，尹振海 (1993)， 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》（陈力为主编），北京语言学院出版社，150-154 页。
[80] 穗志方（1998），语句相似度研究中的骨架依存分析法及其应用，北京大学计算机系博士论文。
[81] 孙茂松，黄昌宁，方捷(1997), 汉语搭配定量分析初探，《中国语文》第 1 期，29-38 页。
[82] 孙茂松，黄昌宁，高海燕，方捷 (1994), 中文姓名的自动识别，Communications of COLIPS, 4(2): 113-122.
[83] 孙茂松，张维杰 (1993). 英语姓名译名的自动辨识，《计算语言学研究与应用》（陈力为主编），北京语言学院出版社，134-149 页。
[84] 孙宏林，黄建平，孙德金，李德钧，邢红兵（1996），“现代汉语研究语料库系统”概述，载《计算机时代的汉语和汉字研究》，罗振声、袁毓林主编，清华大学出版社。
[85] 孙宏林（1997），从标注语料库中归纳语法规则：“V+N”序列实验分析，载《语言工程》（陈力为、袁琦主编），清华大学出版社。157-163 页。
[86] 孙宏林(1998)，词语搭配在文本中的分布特征，《1998 中文信息处理国际会议论文集》（黄昌宁主编），清华大学出版社，230-236 页。
[87] 邢福义（1994），NVN 造名结构及其 NV|VN 简省形式，《语言研究》第 2 期。
[88] 俞士汶，朱学锋，王惠，张芸芸（1996）， 《现代汉语语法信息词典》规格说明书，《中文信息学报》，10(2):1-22.
[89] 俞士汶，朱学锋，王惠，张芸芸（1998），现代汉语语法信息词典详解，清华大学出版社。
[90] 俞士汶、朱学锋、李峰（1998），现代汉语词语的语法属性描述，《汉语计量与计算研究》（邹嘉彦主编），香港城市大学，353-373 页。
[91] 俞士汶、朱学锋（2000），“现代汉语词语的语法属性描述研究”的目标与进展，《语言文字应用》1 期。
[92] 詹卫东，2000. 面向中文信息处理的现代汉语短语结构规则研究，清华大学出版社。
[93] 张国煊，郁梅，王小华（1995），基于语料库的汉语边界划分的研究，载《计算语言学进展

与应用》（陈力为、袁琦主编），清华大学出版社。94-99 页。
[94] 张小衡，王玲玲 1997. 中文机构名称的识别与分析，《中文信息学报》Vol. 11, No. 4, 21-31.
[95] 赵军（1998），汉语基本名词短语识别及结构分析，清华大学计算机系博士论文。
[96] 赵军、黄昌宁（1999），基于转换的汉语基本名词短语识别模型，《中文信息学报》Vol.13,No.2。
[97] 郑家恒，刘开瑛 (1993). 自动分词系统中姓氏人名处理策略探讨，《计算语言学研究与应用》（陈力为主编），北京语言学院出版社，139-143 页。
[98] 周强（1996），一个短语自动定界模型，《软件学报》第 7 卷增刊，315-322 页。
[99] 朱德熙（1984），语法讲义，商务印书馆。

作者在攻读博士学位期间发表的论文

[1] 孙宏林、陆勤、俞士汶（2001）. 利用遗传算法实现词类标记集的优化，《中文信息学报》，15 卷 1 期。

[2] 孙宏林，俞士汶（2000）. 浅层句法分析方法概述，《当代语言学》2000 年第 2 期。

[3] 孙宏林（1998）. 词语搭配在文本中的分布特征，《1998 中文信息处理国际会议论文集》，黄昌宁主编，67-72 页，1998，清华大学出版社。

[4] 孙宏林，段慧明（1998）. 面向自然语言处理的汉语短语信息库，《术语标准化和信息技术》1998 年第 2 期，26-31 页。

[5] 孙宏林，黄建平，孙德金，李德钧，邢红兵（1997），“现代汉语研究语料库系统”概述，《第五届世界汉语教学讨论会论文选》，胡明扬主编，北京大学出版社，459-466 页。

[6] 孙宏林，1997a. 浅谈汉语分词的标准，《语言文字应用》1997 年第 4 期。

[7] 孙宏林，1997b. 从标注语料库中归纳语法规则：“V+N”序列实验分析，《语言工程》（第 4 届全国计算语言学联合学术会议论文集），陈力为、袁琦主编，清华大学出版社，157-163 页。

[8] Sun Honglin, Lu Qin & Yu Shiwen 2000. *Using Genetic Algorithms for Optimizing POS Tagset*, In Proceedings of the International Conference on Language Information Processing, Aug. 16-20, 2000, Urumuqi, China,pp.176-173.

[9] Sun Honglin, Yu Shiwen and Lu Qin,1999. *Evaluations on Part-of-speech Tagset*. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium*, Nov. 5-7, Beijing, China. Published by Tsinghua University Press. pp. 25-31.

[10] Sun Honglin, Lu Qin and Yu Shiwen, 1999. *Two-level Shallow Parser for Unrestricted Chinese Text*. In *Papers on Computational Linguistics (Proceedings of the 5th China National Joint Conference on Computational Linguistics*), edited by Huang Changning and Dong Zhendong, Tsinghua University Press. pp.280-286.

[11] Sun Maosong, Sun Honglin et al. 2000. *Hua Yu: A Word-segmented and Part-Of-Speech Tagged Chinese Corpus*, In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), May 20-22, Athens, Greece.

[12] Sun Maosong, Zhou Qiang & Sun Honglin et al. 2000. *Constructing Word-segmented & POS-tagged Chinese Corpus and a Chinese Tree bank*, 2000 International Conference on Chinese Language Computing, July 8 - 9 , 2000, Chicago, USA.

致 谢

在论文完成之际，我希望感谢所有在过去帮助和支持过我的人们。

首先，我要感谢我的导师俞士汶教授。我在攻读博士学位的五年期间，得到了俞老师的悉心指导和热情帮助。论文从选题到研究的每个阶段，老师都提出了许多重要的意见和建议。感谢老师对我学业上的严格要求，使我不断地认识到自己的不足，从而找到前进的方向。老师严谨求实的治学态度、勤奋踏实的工作作风为学生树立了一个榜样，将成为我今后努力的目标。

感谢北京语言文化大学语言信息处理研究所所长张普教授，在我们共事的十多年里，一直得到张老师的支持、鼓励和帮助。如果没有他的全力支持，我是很难完成攻读博士学位阶段的学业的。

感谢指导小组的陆俭明教授、冯志伟教授对我的指导和帮助。

感谢北大计算语言学研究所的所有老师和同学们，我从他们那里学到了很多。特别要感谢朱学锋老师、段慧明老师、李保利同学为我提供词典和语料方面的支持，感谢詹卫东博士、孙斌博士、于江生博士、刘群副研究员等，我在与他们的多次讨论中获益良多。

感谢清华大学计算机系的黄昌宁教授、孙茂松教授和周强博士，在 1995 至 1998 年期间，我有幸参加了黄昌宁教授主持的国家自然科学基金重点项目“语料库语言学的理论、方法和工具”。在这期间，我从他们那里学到了不少关于统计自然语言处理的思想和方法。孙茂松教授为我提供了许多重要的文献资料和专名方面的资源，我在论文选题和写作过程中和周强博士进行过多次有益的讨论。

感谢香港理工大学电子计算学系的陆勤博士，陆博士为我提供机会，使我能够在香港理工大学进行十个月的访问研究。在这期间，我接触了不少在北京难以看到的文献资料，使我的眼界得到了开阔。感谢陆博士对我学业和生活上提供的帮助。

感谢北京语言文化大学研究生廉竹钧同学帮助我校对了部分语料，她的细致工作使语料标注的一致性和准确性得到了保障。

最后，我要感谢我的妻子沈建华和女儿孙烨，感谢她们对我的理解和支持。