学 校 代 码 10459
学号或申请号 201012171911
密 级

郑州大学

硕士学位论文

介词、连词用法在短语结构句法分析中的应用研究

作 者 姓 名：庞熠雅
导 师 姓 名：穆玲玲 副教授
学 科 门 类：工 学
专 业 名 称：计算机应用技术
培 养 院 系：信息工程学院
完 成 时 间：2013 年 5 月

A thesis submitted to

Zhengzhou University

for the degree of Master

Studies on the Usage of Preposition and Conjunction in Phrase Structure Syntactic Parsing

By Yiya Pang

Supervisor: Prof. Lingling Mu

Computer Application Technology

College of Information and Engineering

May 2013

摘　要

中文句法分析是自然语言处理领域中的一个重要课题。针对汉语本身的特点，本文将介词用法融入到句法分析结果中，使用介词用法属性对 Stanford Parser 进行后处理。首先，为了得到较高的介词用法自动识别结果，本文在已有的基于规则的介词用法自动标注方法的基础上，提出了基于统计的介词用法的自动标注方法，分别采用条件随机场、最大熵和支持向量机三种统计模型，以 2000 年 2 月、3 月、4 月《人民日报》分词与词性标注语料为实验语料，对常用介词进行了自动标注实验，实验结果表明基于统计的介词用法自动标注总体上优于基于规则的介词用法自动标注结果。其次，本文在 Stanford Parser 分析结果的基础上，使用由介词用法属性特征得到的边界识别结果，对已有句法分析结果进行一定的修改，从而提高中文句法分析的准确率。实验表明，融入用法属性特征的句法分析结果比之前结果有了一定的提高。最后，为了验证基于介词用法的句法分析后处理方法的适用性，本文将此方法进一步运用到了连词中，且得到了较好的实验效果。

本文主要的工作包括：

(1)根据“三位一体”广义虚词知识库，在对基于规则的介词用法自动标注结果进行人工校对所得到的正确语料的基础上，实现了基于统计的介词用法自动标注。

(2)在介词用法自动识别、基于用法的介词短语边界识别、Stanford Parser 及宾州中文树库的基础上，实现了介词用法在短语结构句法分析中的应用研究。

(3)根据介词用法在短语结构句法分析中的应用研究，在连词用法自动识别、基于用法的连词短语边界识别的基础上，实现了连词用法在短语结构句法分析中的应用研究。

最后，对本文的研究内容进行了总结，并根据研究结果对下一步工作做了展望，指出了下一步的研究方向。

关键词：自然语言处理 短语结构句法分析 介词用法自动识别 介词短语边界识别 连词短语边界识别

Abstract

Chinese syntactic analysis is an important topic in the field of natural language processing. Aiming at the characteristics of the Chinese, this paper adds the preposition's usage to the output of the parser, postprocesses the output of the Stanford Parser using the usage properties of the preposition. Firstly, in order to get a higher results of preposition' usages recognition, on the basis of the existing rule-based method of preposition' usages recognition, this paper proposes an automatically recognizing method of preposition' usages based on statistics . And three statistical models, which are CRF、 ME and SVM, are used to label some common preposition' usages on the tagged corpus of People's Daily (2000.2,3,4). And the final result shows that automatic recognition of preposition' usages based on statistics is better than the result of rule-based method of preposition' usages recognition generally. And then, on the basis of the output of the Stanford parser, this paper modifies the existing results using the boundary identification received from the property characteristics of preposition' usages, and then enhances the accuracy of Chinese parser. The experiments shows that the result that has the property characteristics of usages has been improved compared to the result that has not to some extent. Finally, in order to verify the applicability of this after-processing approach, this method is used on the conjunction, and gets a good result.

The main work of this paper includes:

(1) According to the thoughts of building the "Trinity" knowledge base of functional words, on the basis of the correct corpus received from the proofreader for the result of preposition' usages automatic annotation based on rule, this paper achieves the automatic annotation of preposition' usage based on statistics.

(2) On the basis of automatic identification of the preposition' usage, prepositional phrase boundary identification based on the usage, Stanford Parser and the Penn Chinese Treebank, this paper achieves the study on the preposition' usage in the phrase structure syntactic parsing.

(3) According to the study on the preposition' usage in the phrase structure

syntactic parsing, on the basis of the automatic identification of the conjunction' usage and conjunction phrase boundary identification based on the usage, this paper achieves the study on the conjunction' usage in phrase structure syntactic parsing.

Finally, this paper summarizes the content of this study, puts forward to the further work, and points out the future research directions.

Key Words: Natural Language Processing, Phrase Structure Parsing, Automatically Identify of Preposition' Usage, Prepositional Phrase Boundary Identification, Conjunctions Phrase Boundary Identification;

目　录

图表目录

图目录

表目录

1 引言

1.1 研究意义

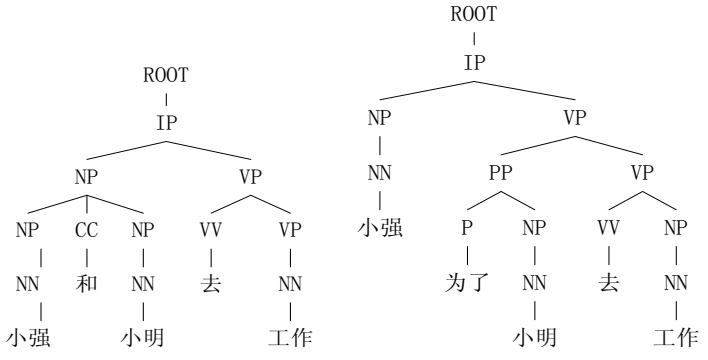
根据不同的处理深度，计算语言学中的语言技术可以分为浅层分析、深层分析两种[1]。浅层分析主要是对词汇级别的处理，一般只分析句子中的一部分，这种技术目前已基本成熟，如分词、词性标注等。深层分析是指对语言进行语法级别、语义级别甚至语用级别的处理，如句法分析等，这些分析技术的分析结果需要对句子进行整体的分析才能得到。而在深层分析技术中，句法分析一直处于十分关键的位置。

句法分析是根据一个给定的语法体系,自动的推导出句子的语法结构，分析出句子所包含的语法单元及这些语法单元之间的关系，最终将句子转化为一棵结构化的语法树[2]。汉语句子由实词和虚词组成，而虚词包括介词、连词、副词、助词、方位词和语气词，对句子的句法意义作用很大。一个汉语句子中，在相同的位置选择不同的虚词，对句子的意义就可能会有很大的影响，句法树也会有很大不同。如下面的句子：

(1)小强和小明去工作。

(2)小强为了小明去工作。

句子(1)中“和”为连词，表示并列关系，(2)中“为了”为介词，表示动作行为的目的。(1)表示小强和小明一起去工作，而(2)表示因为小明的原因，小强去工作。两个句子的句法树如图 1.1 所示。



(a) (b)

图 1.1 例句 1、2 句法树

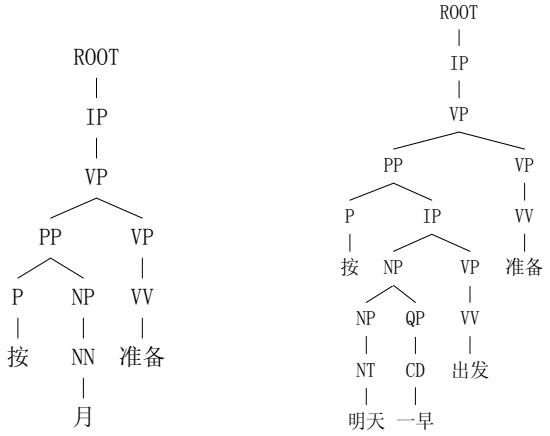
图 1.1(a)为例句(1)的句法树，由句法树可知，此句中“小强和小明”为主语，“小强”和“小明”为并列关系；图 1.1(b)为例句(2)的句法树，此句中“小强”为主语，“为了小明”为介词短语，做状语。

虚词用法的变化多种多样，同一个虚词，在不同的句子里，句法树也可以不同。如下例句：

(3)按月准备。

(4)按明天一早出发准备。

两个句子中“按”都为介词，表示遵从某种标准。而(3)中的介词短语为“按月”，“按”后名词“月”作宾语；(4)中介词短语为“按明天一早出发”，“按”后小句“明天一早出发”作宾语。两个句子的句法树如图 1.2 所示。



(a) (b)

图 1.2 例句 3、4 句法树

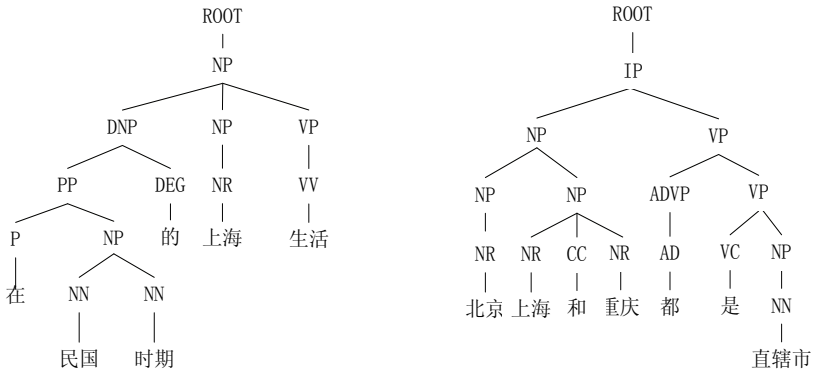
图 1.2(a)、1.2(b)分别为例句(3)、(4)的句法树，两个句子都是由介词短语和谓语组成，而图 1.2(a)中名词作介词“按”的宾语，图 1.2(b)中小句做介词“按”的宾语，句法结果就发生了很大变化。

由上面例子可知，虚词对汉语句法分析的研究是至关重要的，而现有的句法分析研究中涉及到虚词的部分通常都仅依赖于虚词本身，并没有考虑到虚词用法在句法分析中的作用。如例句(3)和(4)中，介词“按”分别介引名词和小句，而如果依靠介词“按”的这两种用法，对例句的句法分析也会变得容易。也就是说，如果在句法分析研究中引进虚词的用法，则会有助于对句法分析的研究。例如下面的例句：

(5)在民国时期的上海生活。

(6)北京、上海和重庆都是直辖市。

例句(5)中“在”为介词，与“民国时期的上海”组成介词短语；例句(6)中“和”为连词，“北京”、“上海”、“重庆”为并列关系。图 1.3 是通过没有使用虚词用法的 Stanford Parser[1]对例句(5)、(6)进行句法分析所得到的句法树。



(a) (b)

图 1.3 例句 5、6 句法树

由图 1.3(a)可看出此句的句法分析结果中，“在民国时期”为例句(5)的介词短语，即宾语指动作发生的时间，而经分析，此句中的介词短语为“在民国时期的上海”，即宾语指动作发生或事物存在的处所。在对此句进行句法分析时，如果首先考虑到此处介词“在”的用法，即首先通过介词“在”用法确定介词所引导的宾语是表示处所的，就可以识别出正确的介词短语。图 1.3(b)中，“上海”和“重庆”为并列结构，即两个词组成并列结构，而经分析，此句的并列结构中有“北京”、“上海”、“重庆”，即三个或三个以上的词组成并列结构。同样，在对此句进行句法分析时，如果首先考虑到连词“和”的用法，即首先通过连词“和”的用法确定组成并列结构词语的个数，就可以得到正确的句法树。

由此可见，在句法分析的研究中，不仅要考虑虚词本身的词性，还要考虑虚词的不同用法，从而提高句法分析的正确率。

介词和连词都是汉语虚词中重要的种类，且在文本中出现频率颇高。同时，介词句法位置多种多样，句法意义难以掌握，在意义、来源、用法、作用、分布等方面具有相当大的复杂性，在汉语句法分析中有独特的研究价值。连词是

[1] http://nlp.stanford.edu/software/stanford-parser-2010-11-30.tgz

虚词中极有个性的一类，在上下文中起到连接段落、句子、分句、词语的作用，能够表达出缜密的逻辑语义关系，在汉语句法分析中具有不可忽视的地位。

介词、连词用法在短语结构句法分析中的应用研究有非常重要的理论和应用价值，此研究将会很大程度的推进句法分析的发展，有利于信息抽取、机器翻译等自然语言处理的相关领域的研究发展。

1.2 研究背景

郑州大学自然语言处理实验室于 2006 年起开始承担“现代汉语广义虚词知识库”课题，此课题是北京大学计算语言学研究所承担的国家 973 课题“文本理解的数据基础项目”（2004CB318102）的子课题，郑州大学从事有关现代汉语广义虚词知识库建设的研究工作，课题已经于 2012 年 11 月通过河南省科学技术厅项目鉴定，完成了广义虚词知识库总体构建以及有关副词、介词、连词、助词、语气词、方位词等词类的用法知识库构建。本文所研究的第一部分，即现代汉语介词用法自动识别就是该项工作的一部分，同时也是国家自然科学基金项目“规则与统计相结合的现代汉语虚词用法自动识别研究”（60970083）的重要组成部分，并且还受模式识别国家重点实验室开放课题基金和河南省科技创新人才杰出青年基金项目（104100510026）资助。

目前国内外在语言知识库方面的研究主要是针对实词，对虚词的研究相对较少。因此，俞士汶[3]在原有语言资源的基础之上提出了“三位一体”的思路，来实现现代汉语广义虚词知识库的构建。刘云[4]为各类虚词设计了相应的属性描述，对常用虚词进行归类总结，从而构建了现代汉语虚词词典基本框架。昝红英等[5]完成了现代汉语广义虚词知识库的构建，包括现代汉语虚词用法词典、现代汉语虚词用法规则库和现代汉语虚词语料库。刘锐等[6]初步研究了基于规则的副词用法自动识别。昝红英等[7]探讨了基于统计的副词用法自动识别。在对虚词用法研究的基础上，袁应成等[8]对现代汉语介词短语边界识别进行了研究。周丽娟等[9]对现代汉语连词结构短语自动识别进行了研究。

本文在以上研究的基础上，完善了现代汉语介词用法词典和现代汉语介词用法规则库，并对基于统计的现代汉语介词用法自动识别进行了研究[10],是现代汉语虚词知识库的重要组成部分。并且在介词用法知识库及介词短语边界识别的基础上，初步探讨了介词用法在短语结构句法分析中的应用，并进一步将其

运用到连词用法在短语结构句法分析中的应用研究中，证明了其适用性。

1.3 句法分析研究现状

1.3.1 国外研究现状

二十世纪 90 年代，自然语言处理开始从原来的小规模处理转变为大规模真实文本的处理上。此后，句法分析的研究也发生了根本的变化，即由传统的基于理论的自然语言处理方法转换为基于语料库的统计方法。

目前，形式最简单、研究最充分的统计句法分析模型为概率上下文无关文法模型(PCFG)[11]。Lari 和 Young[12]（1990）最先使用了 Inside-Outside 算法，从未标注的语料库中自动的估计 PCFG 的各种参数，通过迭代执行 Inside-Outside 算法，抛弃了大量的无用规则，而由剩余的规则组成了最终的文法，实验表明该算法使句法分析的运行效率有了很大的改善。

Collins[13]（1996）提出了一种基于词间依赖的、统计的句法分析模型，使用标准的二元概率估计方法来技术词语词之间的依赖概率。在这种方法的基础上，Collins[14]（1997）提出了一个产生式句法分析模型，此模型是目前影响最广泛的英文句法分析模型之一。

Charniak[15]（2000）采用了最大熵模型，使用了更多的特征，并且首先核心节点的词性进行预测。该方法借鉴了最大熵模型的可以方便地融合各种特征的优点，使各种特征能够均匀分布，确保了语料库中特征的充分利用，加快了句法分析的速度。

Collins[16]（2000）提出了重排序的方法，使用两个模型来进行句法处理，首先使用第一个模型对待分析句子进行处理,生成最可能的 n-best 句法分析树，再使用第二个模型对 n-best 句法分析树进行重排序，从中选择出一棵最优的句法树来作为最终分析结果。

Charniak 和 Johnson[17]（2005）提出了在重排序和产生式模型上改进的句法分析方法。在该方法中，首先输入长度小于 100 的句子，使用一个由粗到细的产生式模型，产生 50-best 候选句法树，最后使用一个基于最大熵模型的判别式重排序方法从候选句法树中选择出最优的句法分析树。

Johnson 和 Ural[18]（2010）提出了将 Brown 和 Berkeley 的句法分析器分析出的 n-best 句法树合并的方法,使用重排序器来选择合并的句法分析树的特征，

该实验证明了使用重排序器合并句法分析的有效性。

1.3.2 国内研究现状

在 90 年代之前，因为没有足够规模的统一的中文树库，中文句法分析的研究发展较为缓慢。而随着宾州中文树库（CTB）的发布，中文句法分析取得了较快发展。目前，中文句法分析已经成为了中文信息处理研究领域的一个热点问题，国内外很多学者也对中文句法分析进行了不断的研究。

清华大学的 Zhou[19]（1997）设计了基于统计的汉语句法分析器，通过预计句法分析树的构成，匹配左右括号，产生了相应的句法分析树，最后再使用消歧的方法，生发最优的句法分析树。

Zhang 等[20]2002 年提出了一种结合了 PCRG 和 GLR 算法的汉语句法分析器，该方法定义了新的 PCFG 文法，并且使用了规则和统计相结合的方法。

曹海龙[21]2006 年提出了在 CTB5.0 上的中心驱动模型结果，之后又提出了一种两级句法分析方法，该方法首先使用一个快速且有效的模型来识别较为简单的基本短语，然后在扩展中心驱动模式，来识别句子中包含有较多的递归结构的复杂短语。

Mengqiu Wang 等[22]2006 年使用移近规约决策模型对宾州树库进行了句法分析实验，比较了 Maxent、SVM、决策树等统计模型，并且利用了中文的韵律特征，提高了句法分析的准确率和解码速度。

Xiao Chen 和 Changning Huang 等[23]提出了一种新的生成句法分析树的方法，该方法采用多个句法分析器生成的 n-best 句法分析树，将其重新合并生财了新的句法分析树。

徐文智和王小捷[24]2009 年提出了一个新的基于词汇化的 PCFG 模型的句法分析模型，该方法有效的利用了汉语中的字信息，相对缓解了词汇化 PCFG 模型的数据稀疏的问题。

1.4 研究内容

本文根据已有的现代汉语虚词知识库，首先对介词用法自动识别进行了研究，其次利用介词用法自动识别及介词短语边界识别方法对短语结构句法分析进行了初步研究，最后利用介词用法在短语结构句法分析中的应用的成功思路，

将其应用在连词用法在短语结构句法分析中的相关研究上。

本文所包含的主要工作：

(1)在基于规则的现代汉语介词用法自动识别的基础上，人工校对 2000 年 2 月、3 月、4 月《人民日报》分词与词性标注语料库中的介词用法，在此正确语料的基础上，实现基于统计的介词用法自动识别，并进一步完善介词用法词典和用法规则。

(2)在基于规则和基于统计的介词边界识别及 Stanford parser 的基础上，对句法分析结果进行基于介词用法属性的后处理，从而提高句法分析的正确率。

(3)在介词用法属性在短语结构句法分析中的应用研究的思路的基础上，将其运用于连词用法在短语结构句法分析中的应用研究，验证本文所提出方法在虚词中的适用性。且因为连词用法与介词用法特征有所不同，在使用的过程中对原有方法进行调整，得到连词用法在短语结构句法分析中的应用研究方法。

1.5 论文组织框架

根据本文的研究工作，本文主要将其分为五章来阐述。各章具体安排如下：

第一章，引言。主要概述了本文的研究意义、研究背景、句法分析研究现状及本文主要研究的工作以及组织框架。

第二章，介词用法自动识别。首先介绍了基于规则的介词用法自动识别，其次讲述了 CRF、ME、SVM 三种统计模型的特征选取及基于三种统计模型的介词用法自动识别方法，分别给出了四种识别方法的实验结果，并对其进行分析。

第三章，介词用法在短语结构句法分析中的应用。首先介绍了基于介词用法的介词短语边界识别方法，且在此方法的基础上，通过获取边界识别标准库与后处理两部分详细讲述了介词用法在短语结构句法分析中的应用研究方法。最后在实验的部分给出了实验结果并且对其进行了分析。

第四章，连词用法在短语结构句法分析中的应用。这一章是对第三章所提方法的应用，将对介词用法的研究延伸到连词用法中，且根据连词用法与介词用法的不同，详细介绍了连词用法在短语结构句法分析中的应用研究方法。最后通过给出的实验结果来验证本文所提方法的适用性。

第五章，结论与展望。对本文所做工作进行总结，且提出了下一步所要做的研究方向。

2 介词用法自动识别

2.1 现代汉语介词用法知识库

对自然语言来说，对一个句子含义的理解是建立在对句子中所有词汇含义理解的基础上的，任何一个希望机器能够像人一样去理解自然语言句子的自然语言处理系统，首先需要具备的就是关于词汇的知识，即词汇知识库[25]。

现代汉语介词用法知识库是虚词知识库的子库，采用“三位一体”[3]构建思想，从介词用法入手，主要包括现代汉语介词用法词典、现代汉语介词用法规则库和现代汉语介词用法语料库。如图 2.1 表示现代汉语介词用法知识库。

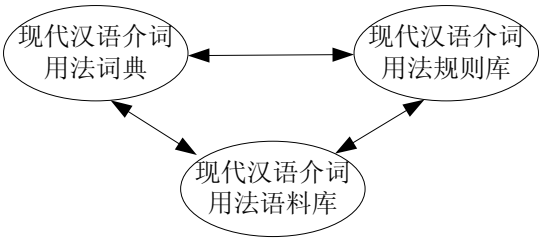


图 2.1 现代汉语介词用法知识库

2.1.1 介词用法词典

现代汉语介词用法词典是虚词知识库重要的组成部分，以“用法-属性特征”相对应的原则，将人为总结出的介词知识写入介词词典中，详细描述了每个介

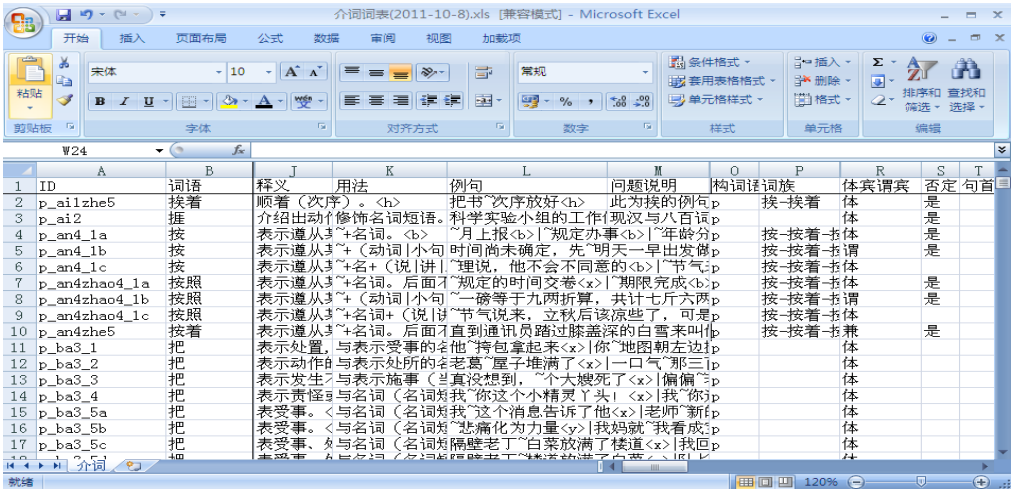


图 2.2 介词用法词典部分样例

词用法的属性特征。在介词用法词典中，共使用了 45 个属性字段来详细描述介词属性特征，其中属性字段有：释义、用法、例句、词类、体宾谓宾、否定、句首、左搭配、左紧邻、右搭配、右紧邻、句末、兼类、句法结构、省宾、否定前后、格标记、全拼音等。介词用法词典部分样例如表 2.2 所示。

介词用法词典由《现代汉语词典》[26]、《现代汉语八百词》[27]、《现代汉语虚词词典》[28]，以及《人民日报》分词与词性标注语料库整理所得。目前为止，介词用法词典中共收录了介词 139 个，涉及义项 207 个，用法 374 个。

2.1.2 介词用法规则库

为了为介词用法自动识别提供形式化依据，在介词用法词典的基础上，根据介词用法的不同特征，对现代汉语介词用法词典进行了BNF规则的形式化描述[29]，构建了介词用法规则库。下面是对介词“把”给出的用法规则描述[30][31]，其中小写字母表示词性，大写字母表示指定用法的上下文特征，如F表示句首，即词语或词性特征出现在句首；M表示左搭配，即词语或词性特征在目标介词左部搭配共现；L表示左紧邻，即词语或词性特征在目标介词左部紧邻共现；R表示右紧邻，即词语或词性特征在目标介词右部紧邻共现；N表示右搭配，即词语或词性特征在目标介词右部搭配共现；E表示句末，即词语或词性特征出现在句子末尾；汉字表示词形[32]。

$把

@<p_ba3_4>→LR ^L→我 ^R→你[这]个

@<p_ba3_5f>→E ^E→它

@<p_ba3_1>→N ^N→(n|r)*v*(p|<vq>)

@<p_ba3_5b>→N ^N→，*(当|当作|作为|比作|看作|说成|成)

@<p_ba3_5b>→N ^N→当|当作|作为|比作|看作|说成|成

@<p_ba3_6b>→LR ^L→被*(n|r) ^R→r

@<p_ba3_6a>→LR ^L→被*(n|r) ^R→n

@<p_ba3_5d>→N ^N→(<ns>|s)*v*(n|r)

@<p_ba3_5c>→N ^N→列入|纳入|写入|写进|用于|拖入

@<p_ba3_1>→RN ^R→*(n|r)*v ^N→!((n|r)*v*(n|r))

@<p_ba3_5c>→N ^N→(n|r)*v*(<ns>|s)

@<p_ba3_5a>→N ^N→(n|r)*v*(n|r)

@<p_ba3_5e>→N ^N→(n|r)*v*(n|r)

@<p_ba3_3>→LR ^L→!(n|r) ^R→(n|r)*v

@<p_ba3_2>→N ^N→(<ns>|s)*v

@<p_ba3_1>→

其中，符号“→”是定义为，符号“|”是多选一，“*”表示词或者词性之间的间隔为零次或多次，“!”表示非，()表示选组，<>用于标识词性代码中标记多于一个字母的词性，[]表示可以选择的词或词性。如规则“@<p_ba3_1>→N ^N→(n|r)*v*(p|<vq>)”为用法p_ba3_1的一条规则描述，其中N表示右搭配，此用法中的介词“把”用在动词前，与其后的名词或代词组成介宾短语，且此用法中介词“把”后只有一个宾语。规则“@<p_ba3_5b>→N ^N→当|当作|作为|比作|看作|说成|成”为用法p_ba3_5b的一条规则描述，此用法中的介词“把”与如“当”、“当作”、“作为”等的“当”类动词连用，表示受事。

2.1.3 介词用法语料库

现代汉语介词用法语料库共有介词用法词典例句语料和《人民日报》介词用法标注语料两种形式。例句语料即介词用法词典中词条“例句”属性中所包含的例句集，如下例句为介词“把”的例句及例句语料。

把悲痛化为力量。

把/p<p_ba3_5b> 悲痛/n 化为/v 力量/n 。/w

《人民日报》介词用法标注语料共包含了1998年1月、2000年1月到6月的《人民日报》语料。如下例句为介词“把”的《人民日报》分词与词性标注语料样例和样例语料。

天津/ns 汽车/n 不仅/c 引进/v 了/ul 合作方/n 先进/a 的/ud 技术/n ，/wd 也/d **把**/p 丰田/nz 的/ud 管理/vn 方法/n 用/v 在/p 了/ul 生产/v 之中/f ，/wd

天津/ns 汽车/n 不仅/c 引进/v 了/ul 合作方/n 先进/a 的/ud 技术/n ，/wd 也/d **把**/p<p_ba3_1> 丰田/nz 的/ud 管理/vn 方法/n 用/v 在/p 了/ul 生产/v 之中/f ，/wd

2.2 基于规则的介词用法自动识别

2.2.1 基于规则的介词用法自动识别方法

介词用法自动标注是由基于介词规则的标注软件自动完成[33]，标注流程图如图2.3所示。

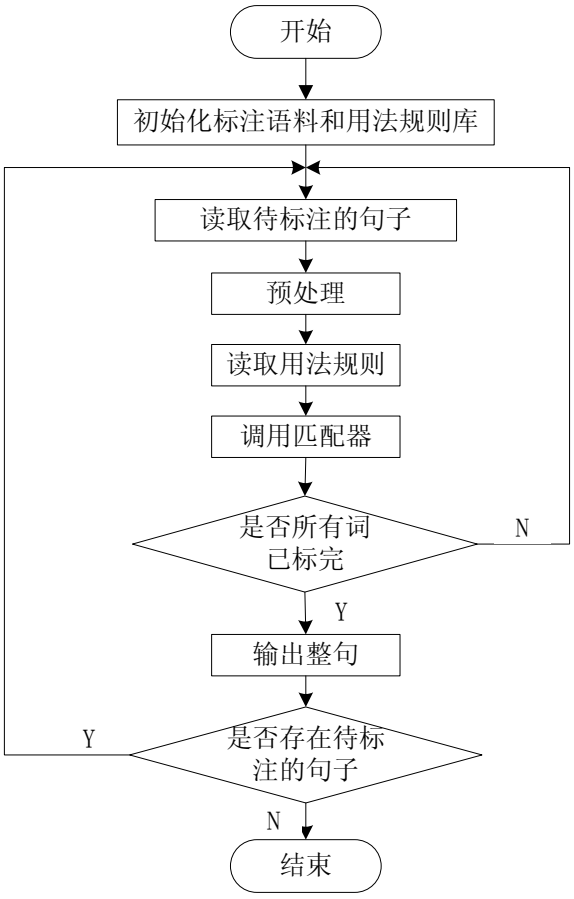


图 2.3 介词用法自动识别标注系统流程图

标注系统的具体步骤如下：

(1)初始化待标注语料和介词用法规则库。语料读取时系统将语料内容以句号、问号、叹号和换行符为截句标志切分为若干整句，语料和规则分别以动态数组和哈希表的形式写入内存。

(2)读取待标注整句，对句子进行预处理，找出句子中需要标注的介词及与此介词对应的规则，调用匹配器对句子与规则进行匹配，匹配成功则在句子中标注相应的介词用法，反之标注为<?FALL>。标注完成后查看句子中是否所有介

词已标注完成，若未完成则再次读取相应介词规则进行匹配，反之输出整句，进入步骤3。

(3)查看是否所有整句已标注完成，若完成则标注系统结束，反之进入步骤2，读取下一个整句。

2.2.2 实验评价方法

为了对介词用法自动识别的性能进行评价，本文的实验中使用了准确率（P）作为评价指标，具体定义如下：

$$P=\frac{\text{机器标注和人工标注相同的个数}}{\text{机器标注总个数}}\times 100\% \qquad (2.1)$$

2.2.3 实验结果与分析

本文选用2000年2月、3月、4月《人民日报》分词与词性标注语料作为实验语料，其语料形式如下所示:

“20000301-01-002-007/m 许多/m 委员/n 在/p 审议/vn 过程/n 中/f 认为/v ，/wd 常委会/n 在/p 人大/jn 工作/vn 中/f 始终/d 自觉/a 地/ui 接受/v 党/n 的/ud 领导/vn ，/wd **把/p** 对/p “/wyz 一府两院/j ”/wyy 的/ud 监督/vn 和/c 支持/vn 有机/d 地/ui 结合/v 起来/vq 。/wd”

“希望/v 全体/n 常委/n 和/c 委员/n ，/wd 聚精会神/iv ，/wd 群策群力/iv ，/wd **把**/p 大会/n 开/v 成/v 一个/mq 民主/a 、/wu 求实/vi 、/wu 团结/a 、/wu 鼓劲/vi 的/ud 大会/n 。/wd”

上面是实验语料中的一个小节的语料，每个词语的词性在“/”后标注，其中“p”表示该词语的词性为介词[34]。

将实验语料使用介词用法自动标注系统进行介词用法自动识别，介词用法识别后语料如下所示：

“20000301-01-002-007/m 许多/m 委员/n 在/p 审议/vn 过程/n 中/f 认为/v ，/wd 常委会/n 在/p 人大/jn 工作/vn 中/f 始终/d 自觉/a 地/ui 接受/v 党/n 的/ud 领导/vn ，/wd **把/p<p_ba3_1>** 对/p “/wyz 一府两院/j ”/wyy 的/ud 监督/vn 和/c 支持/vn 有机/d 地/ui 结合/v 起来/vq 。/wd”

“希望/v 全体/n 常委/n 和/c 委员/n ，/wd 聚精会神/iv ，/wd 群策群力/iv ，/wd **把/p<p_ba3_5b>** 大会/n 开/v 成/v 一个/mq 民主/a 、/wu 求实/vi 、/wu 团结/a 、/wu 鼓劲/vi 的/ud 大会/n 。/wd”

基于规则的用法识别方法具有直观、简单、针对性强等优点，但是其覆盖程度低，且难于进一步优化。如表2.1为基于规则的16种常用介词用法自动识别实验结果，基于规则的介词用法自动识别结果准确率平均值保持在70%左右，如介词“把”在语料中共出现了10268次，而基于规则标注正确的只有6978个，平均准确率为67.96%。

表 2.1 基于规则的常用介词用法自动识别结果分析

介词	出现次数	准确率(%)	介词	出现次数	准确率(%)
在	76207	64.90	由	7362	55.21
对	27684	67.23	据	6147	99.40
以	12820	88.64	将	4694	99.21
从	13263	58.05	通过	4692	49.54
把	10268	67.96	给	4091	65.52
向	8012	93.58	到	3255	89.38
被	7771	65.19	为了	2902	77.38
于	7526	72.60	比	2814	88.57

如果要提高基于规则的介词用法自动识别方法的准确率，必须不断修改和调整介词的用法规则。如2.1.2节中举例的用法p_ba3_5b的规则，对于句中存在“当类动词”，规则用“当|当作|作为|比作|看作|说成|成”来描述，这样的规则是不完备的。如“把悲痛化为力量”、“把这些青年培养成为有用的人才”等句中“化为”、“培养成”从语义上分析也应属于用法p_ba3_5b，而现有规则却无法进行正确的识别。此外，还存在一些其他问题，如规则难区分、规则难书写等问题。针对这些问题，本文提出了基于统计的介词用法的自动识别方法，进一步提高了介词用法自动识别的准确率。

2.3 基于统计的介词用法自动识别

2.3.1 统计模型介绍

根据对介词用法及其所在上下文语料的考察，本文选择了条件随机场(Conditional Random Fields, CRF)、最大熵(Maximum Entropy, ME)以及支持向量机(Support Vector Machine, SVM)三种机器学习统计模型。

2.3.1.1 条件随机场模型

条件随机场的概念自2001年被提出[35]，在图像处理和自然语言处理等领域就得到了广泛应用。近几年，CRF先后被应用于语义角色标注[36]、汉语词义消歧[37]、词性标注[38]等领域，并且已经取得了一定的成功。在本文的研究中，介词用法识别的问题可以被看作为序列标注任务。也就是说，将一个中文句子看作一个有序的词序列，通过判断该介词所在的不同上下文语境，进一步来判断介词应该标注的用法。

条件随机场模型是一个无向图模型，在给定了输入节点的条件下计算出输出节点的条件概率，它用来考察给定的输入序列所对应的标注序列的条件概率，其训练目标是使条件概率最大化。近年来，该模型广泛应用于歧义消解、中文分词等自然语言处理任务中。

在本文的研究中，首先分别假设X, Y表示待标记的观察序列和它所对应的标记序列的联合分布随机变量，而条件随机场(X, Y)就是一个以待标记的观测序列X为全局条件的无向图模型。

通常，我们定义$G=(V, E)$是一个无向图，定义$Y=\{Y_v | v \in V\}$是以节点v为索引的随机变量Y_v所构成的集合，其中，V指节点集合，E指无向边的集合，即V中的每个节点对应着一个随机变量Y_v。如果以待标注的观察序列X为条件，则每个随机变量Y_v都满足如下的马尔科夫特性：

$$p(Y_v \mid X, Y_w, w \neq v) \quad p_v(Y \mid X_w Y), \tag{2.2}$$

其中，$w \sim v$指两个节点在图G中是邻近的节点，那么，(X, Y)是一个条件随机场。

为了给出条件随机场的形式化描述，对于第j种特征，分别定义局部特征函数$f_j(y_{i-1}, y_i, X, i)$和与该特征对应的函数权重λ，$f_j(y_{i-1}, y_i, X, i)$表示状态特征

s(y_i, X, i)以及转移函数t(y_{i-1}, y_i, X, i),其中y_{i-1}，y_i为标识序列，X为输入序列，i为输入序列X的第i个位置。对于输入序列X和标记序列Y，可定义CRF的特征函数为：

$$F(Y,X)=\sum_{i=1}^{n} f_i(y_{i-1},y_i,i) \tag{2.3}$$

由此，CRF定义的条件概率可以由以下式子给出：

$$p(Y\mid X,\lambda)=\frac{1}{Z(X)}\exp(\lambda\cdot F(Y,X)) \tag{2.4}$$

分母Z(X)为归一化因子，表示为：

$$Z(X)=\sum_{Y}\exp(\lambda\cdot F(Y,X)) \tag{2.5}$$

2.3.1.2 最大熵模型

ME模型是在1957年被E.T.Jaynes提出的统计模型，其基本原理是，对于一个给定事实的集合，选择一个模型，该模型与所有已知的事实相一致，而对未知的可能发生的事实，模型赋予其分布均匀的概率[39]。在语言处理方面，基于ME所建立的语言模型不依赖于领域知识，且独立于特定的任务，在情感分类[40]、命名实体识别[41]、词义消歧[42]、组块分析[43]等自然语言处理领域取得了较好效果。

将最大熵的方法应用于介词用法自动识别任务中，满足最大熵条件的$p(y/x)$的形式如下：

$$p(y\mid x)=\frac{1}{Z(x)}\exp\sum_{i=1}^{n}\lambda_i f_i(x,y) \tag{2.6}$$

其中，分母$Z(X)$为归一化因子,f_i是第i个特征函数，λ_i是待求解的参数，也是特征函数f_i的权重，可以通过训练集学习得出。归一化因子定义如下：

$$Z(X)=\sum_{y}\exp\sum_{i=1}^{n}\lambda_i f_i(x,y) \tag{2.7}$$

2.3.1.3 支持向量机模型

SVM模型是由Vapnik等人提出的机器学习方法，该模型通过使用一些策略，将具有不同特征数据的中间界限最大化，使用数据的特征来判断该数据具体属于哪个类别[44]。在自然语言处理中，SVM广泛应用于词义消歧[45]和文本自动分类[46]等方面。

SVM是由线性可分情况下的最优分类面发展而来的，基本思想可用图2.4的二维平面来说明。

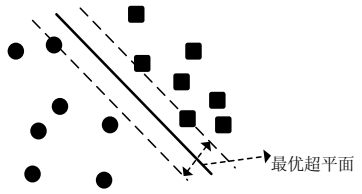


图 2.4 两类线性划分的最优超平面

图中，方块点和圆点分别代表两类样本，中间的粗实线指分类器，其附近的两虚线分别为各类中距分类线最近的样本并且平行于分类线的直线，两虚线之间的距离指的就是分类间隔。最优分类线就是分类线不仅能将两类正确的分开，而且要使分类间隔最大。如果说训练数据可以无误差地被划分开，那么，以最大的间隔分开数据的超平面就称为最优超平面。

而SVM的基本原理就是寻找此最优超平面。此处可以定义二值分类样本为：$(x_1, y_1), \cdots, (x_n, y_n)$，其中$x_i \in R_n$, $y_i \in \{-1, +1\}$，SVM的决策函数为：

$$g(x) = sign\left(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\right) \quad (2.8)$$

其中，$K(x_i, x)$为核函数。

2.3.2 特征抽取

在基于统计的介词用法自动识别系统中，特征的选择直接影响到系统的性能，词语的上下文语境对识别研究起到了至关重要的作用。

2.3.2.1 CRF模型特征选取

在介词的用法识别中，不管是介词所在上下文中的词，还是所在上下文中词的词性对其的用法自动识别都有重要意义。因此，本文的实验中主要选择介词所在上下文中的词性信息、词语信息、以及词性与词语的组合信息作为CRF模型的特征，通过设定上下文窗口大小来选择不同的特征进行实验。

模型的训练和测试使用了工具包CRF++0.53[2]，首先将测试文件和训练文件转换成可识别的文件格式，依据CRF++对数据格式的要求，对于介词的用法识别任务，可将每个待识别的介词所在的中文句子转换成一个块。如2.2.3实验语料中

[2] http://crfpp.sourceforge.net/

的两个例子，要识别介词“把”的用法，上下文窗口大小设定为2，即分别选取介词“把”左右两边词语及词性，则介词“把”所在的中文句子可转换为如表2.2所示：

表 2.2 介词“把”的训练数据示例

0	1	2	3	4	5	6	7	8	9	10
把	p	领导	vn	，	wd	对	p	“	wyz	〈p_ba3_1〉
把	p	群策群力	Iv	，	wd	大会	n	开	v	〈p_ba3_5b〉

表2.2中第1列为待标注的介词“把”，最后一列为该词的用法编码，其他列的数据为介词“把”所在上下文窗口为2范围内的词语及词性特征。

2.3.2.2 ME模型特征选取

在此实验中，模型的训练和测试使用了最大熵工具包maxent[3]。同CRF一样，主要选择介词所在上下文中的词性信息、词语信息、以及词性与词语的组合信息作为ME模型的特征，通过设定上下文窗口大小来选择不同的特征进行实验。使用此工具包之前同样需要将测试文件和训练文件转换成工具可识别的文件格式，如2.2.3实验语料中的两个例子，当上下文窗口为2时，可转换为如下所示：

“<p_ba3_1> w0=把把 p0=p w-2=领导 p-2=vn wp-2= 领 导 vn w-1=，p-1=wd wp-1=，wd w+1=对 p+1=p wp+1=对p w+2= “ p+2=wyz wp+2=“wyz”

“<p_ba3_5b> w0=把把 p0=p w-2=群策群力 p-2=iv wp-2=群策群力iv w-1=，p-1=wd wp-1=，wd w+1=大会 p+1=n wp+1=大会n w+2= 开 p+2=v wp+2=开v”

上面的例子中，w为词语，p表示词性，w0和p0分别为待标注的介词及其词性，±1、±2表示相应的特征序号，wp为词语与词性的组合特征，<>中为待标注的介词的用法编码，如<p_ba3_1>为介词“把”在例句中的用法编码。

2.3.2.3 SVM模型特征选取

SVM的特征是数值型的。对于句子中出现的介词，通过选定该介词所在的上下文的窗口，然后计算窗口内的特征词语与该介词用法的互信息，以此来作为特征向量。如介词w的用法u与特征词t之间的互信息可计算如下：

[3] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

$$I_u = \log \frac{p}{p_1 * p_2} \quad (2.9)$$

P_1：词w的用法u在语料中出现的频率

P_2：特征t在语料中出现的频率

P：词w的用法u与特征t共同出现的频率

在基于支持向量机的介词“把”用法标注试验中，模型的训练和测试使用了LibSVM[4]工具包。这里需将训练文件和测试文件转换成原语料的数据格式，如2.2.3实验语料中的两个例子，当上下文窗口为2时，可转换为如下所示：

1 1:0.23501013521331898 2:0.14190073082183763 3:-0.4430617698993186 4:-0.09158939939794092

6 1:2.1724998638800206 2:-0.11157108697146872 3:1.0025748624377087 4:0.9172428086379463

其中1和6分别代表介词“把”的用法p_ba3_1和p_ba3_5b，其后分别为互信息特征和特征编码。

2.3.3 实验结果与分析

实验选取如2.2.3节所示的2000年2月、3月、4月的《人民日报》介词用法标注语料作为实验语料，首先利用在3.1节中介绍的基于规则的方法对实验语料中的介词进行标注，对基于规则的介词自动识别的结果进行了多人交叉人工校对，形成介词用法标注的标准语料。

实验中将含有介词的句子按用法类别基本均匀地散列为4份数据集，采用4折交叉验证试验。标注系统的性能很大程度上取决于训练和测试模型所使用的特征，根据CRF+ +、max ent及LibSVM 的训练数据格式以及介词用法的语境特点，实验中特征模板选取介词前后n个词语的词形与词性。为了评价统计方法识别介词用法的性能，且与基于规则的自动识别方法进行比较，本文在实验中使用了2.2.2节中的准确率P来作为评价指标。

为了考察上下文窗口大小对使用结果的影响，分别对常用介词在三种统计模型下进行了实验，图2.5是表2.1中使用的16种常用介词的用法自动识别的准确率分别在三种统计模型下的平均实验结果，其中横坐标2~9指上下文窗口的大小，纵坐标指几个常用介词的准确率宏平均值。

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm

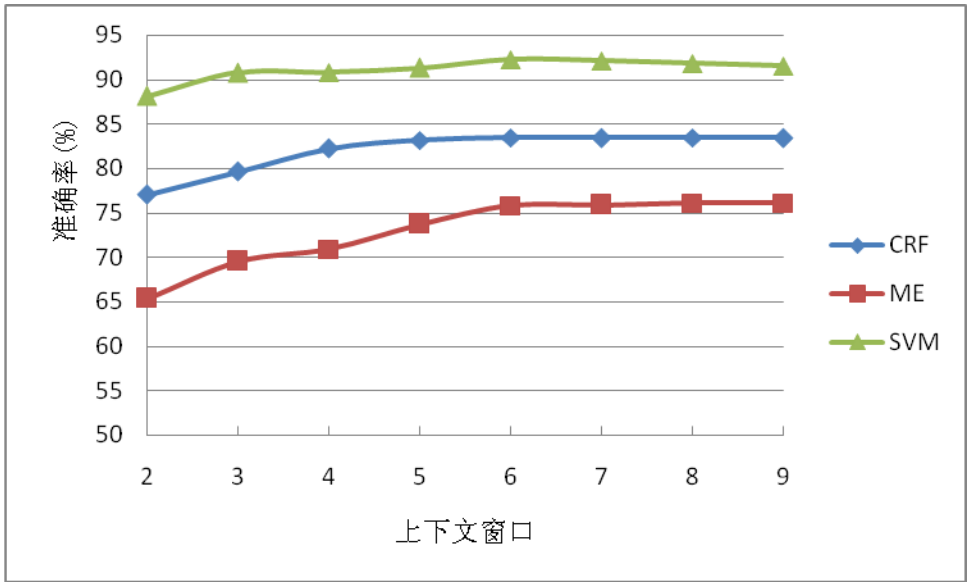


图 2.5 在不同上下文窗口中三种模型的识别效果

从图2.5的识别效果看，对于三种模型来说，识别效果并不是随着上下文有效范围增大而增大的，对CRF和ME，当窗口为6时达到最大值，之后趋于稳定；对于SVM来说，随着窗口的增大，识别效果却在慢慢减小。因此，在当前语料库的规模下，介词的用法自动识别并不是上下文窗口越大越好，随着窗口的增大可能会给识别带来更多的噪音。

根据上面对上下文的考察，在下面的实验中设定上下文窗口为6。为了比较三种统计模型的实验效果，本文进行了第二个实验，当上下文窗口为6的情况下，分别采用基于规则的方法、CRF方法、ME方法、SVM统计方法进行实验，实验对比结果如表2.3所示。

从表2.3的实验结果也可看出，统计模型在介词的用法自动识别方面具有较好的适应性，相对于基于规则的用法自动标注的准确率，基于统计的用法自动标注在总体上取得较高的准确率，尤其是SVM模型得到了良好的结果。虽然基于统计的方法得到了良好的效果，但是也可以看出，在介词“据”中，基于规则的准确率为99.40%，而基于统计的方法中准确率最高的为93.06%。因此，我们并不能断言，关于介词的用法自动识别，统计的方法一定优于规则的方法。针对这些问题，我们优先考虑采用规则方法进行标注，而对于规则识别效果差，统计识别效果好的用法，采用统计方法进行标注，这样便可实现简单的规则与统计结果的方法对介词进行用法标注，这也是今后介词用法自动识别研究的主

要课题。

表 2.3 基于统计的常用介词用法自动识别结果

介词	CRF准确率(%)	ME准确率(%)	SVM准确率(%)	介词	CRF准确率(%)	ME准确率(%)	SVM准确率(%)
在	78.02	75.54	86.32	由	84.88	76.32	85.08
对	65.62	67.86	73.54	据	67.74	90.17	93.06
以	91.82	88.06	90.02	将	100.00	99.33	100.00
从	72.89	77.22	85.39	通过	70.85	60.39	66.63
把	83.46	75.83	92.23	给	82.83	89.16	98.19
向	98.49	98.15	100	到	90.49	90.07	98.47
被	70.17	62.26	71.08	为了	90.09	86.40	96.80
于	87.40	73.67	57.99	比	63.61	87.90	91.13

2.4 本章小结

本章介绍了基于统计的介词用法自动识别。该方法针对规则方法在介词的用法自动识别中存在的不足，利用条件随机场、最大熵和支持向量机三种模型对介词的用法自动识别进行了实验，从实验结果可以看出，基于统计的介词用法自动识别比基于规则的介词用法自动识别有更好的识别效果，能够很好的对未知情况进行预测。

3 介词用法在短语结构句法分析中的应用

介词在句法结构中的功能主要为介引某些句法成分，并与其一同组成介词短语，作谓语动词的修饰语[47]。因而，从句法层次上看，介词在句法结构中的最主要的功能为“介引”，即对某些词语起到介引的作用，本文着重于研究介词所起介引作用中的词语。如句子“在民国时期的上海生活”中，由Stanford Parser分析得到的介词短语为“在民国时期”，而此句正确的介词短语却为“在民国时期的上海”。针对这个问题，如果首先考虑到介词“在”的用法，就可以很容易的得到解决。因此，本文主要研究介词在句法分析中的边界，即使用介词的边界识别结果来修改句法分析中介词的边界，以此来考察介词用法特征在句法分析中的影响。

3.1 介词短语边界识别

在本文对句法分析的研究中，主要使用了基于规则和基于CRF统计的介词边界识别[8]方法。

3.1.1 基于规则的介词短语边界识别

在介词用法规则库中，我们用形式化编码和符号系统来表示语言知识，其目的就是利用这些知识使对自然语言任务的处理能力增强。本文从介词用法规则库中提取出对介词短语边界识别有用的信息，用其对介词短语边界识别进行实验。在介词用法规则库中含有很多介词短语后边界的界定信息，如下介词“在”的用法规则：

$在

@<p_zai4_5>→N ^N→看来|来说|而言|说来|来看|来讲

@<p_zai4_3b>→L ^L→控制|限制|保持|维持|稳定|表现|体现

@<p_zai4_3a>→N ^N→方面|问题上|实践中|生活中|领域|工作中

@<p_zai4_1c>→N ^N→过程|活动|会议|比赛|斗争|接触

@<p_zai4_4>→N ^N→(v|<vn>)<下/f>|(条件|前提|情况|情形|形势|背景|原则|努力)下|基础上

@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之间|之中|之际|夜晚|t)*v

@<p_zai4_1b>→LN ^L→v ^N→年|月|日|号|天|星期|世纪|期间|初|时|秒|之后|之前|之间|之中|之际|夜晚|t

@<p_zai4_2a>→N ^N→(ns|s)*v

@<p_zai4_2b>→LN ^L→v ^N→n|f

@<p_zai4_2a>→N ^N→(ns|s)*, *v

@<p_zai4_1a>→N ^N→(年|月|日|天|号|星期|世纪|期间|初|时|秒|之后|之前|之际|之间|之中|夜晚|t)*, *v

@<p_zai4_3a>→R ^R→a|v|n

在介词“在”规则中，可以看到很多规则都含有介词“在”短语边界的界定信息，如第一条规则中“看来”、“来说”，第五条规则的方位词“下”、“上”，第六条规则的时间词等。

例句3.1: **在**/p 国际/n 形势/n 发生/v 深刻/a 变化/vn 的/ud 今天/t ，/wd 作为/p 世界/n 上{shang5}/f 两/m 个/qe 最/d 大/a 的/ud 发展中国家/l ，/wd 我们/rr 都/d 肩负/v 着/uz 发展/v 经济/n 、/wu 提高/v 人民/n 生活/n 水平/n 的/ud 重任/n 。/wj

上面例句中“在”的后界为“今天/t”的后面，从第六条规则中我们可以看出时间词可以做“在”的后边界。所以可以从介词规则库中提取出这些对介词短语边界识别有利的规则，形成相应的介词短语边界规则文件。如下是介词“在”的短语边界规则：

$在

@<p_1>→N ^N→<时/X>|方面|日子|同时

@<p_2>→N ^N→f

@<p_3>→N ^N→t

@<p_4>→N ^N→s|<ns>

@<p_5>→N ^N→n</PP>v

@<p_6>→N ^N→</PP>的

@<p_7>→N ^N→</PP>,

其中第五条规则中1~7分别表示介词的不同用法ID，“n</PP>v”表示边界确定在动词v之前，下面两条分别表示边界确定在“的”前和“，”前。

在对句子进行短语边界识别时，首先要对待识别句子进行用法自动识别，再利用用法识别后的句子进行短语边界识别，如下是对例句3.1进行介词短语边界识别后的结果：

<PP> 在/p<p_zai4_1a> 国际/n 形势/n 发生/v 深刻/a 变化/vn 的/ud 今天/t </PP>，/wd 作为/p 世界/n 上{shang5}/f 两/m 个/qe 最/d 大/a 的/ud 发展中国家/l ，/wd 我们/rr 都/d 肩负/v 着/uz 发展/v 经济/n 、/wu 提高/v 人民/n 生活/n 水平/n 的/ud 重任/n 。/wj

3.1.2 基于统计的介词短语边界识别

介词短语边界识别问题可以看成是一个序列标注的任务，把给定的介词所在的句子作为待观察序列，把含有介词边界的序列看作输出序列，所以可以使用CRF模型来对介词短语边界进行识别研究。

3.1.2.1 算法描述

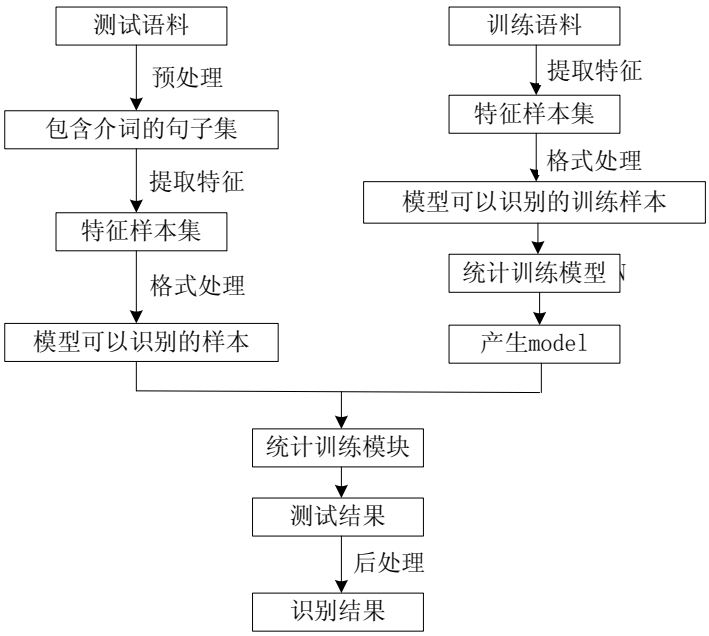


图 3.1 基于统计的介词短语边界识别处理过程

介词短语边界识别的目的是从已进行了分词与词性标注并且对介词词性的

词语进行了用法标注后的句子中识别出介词对应的介词短语结构，由于介词的词性在分词与词性标注阶段已经确定，所以介词的前边界也就确定了下来，也就是介词本身，用<PP>标记，边界识别的主要任务就是确定介词对应的后边界，用</PP>标记，此处将后边界之后的下一个词称为后词。

边界识别时，首先从句子中将介词之后出现的所有词语作为一个样本，选择合适的特征来表示该样本，利用统计模型对样本进行分类，最后在利用样本分类的结果确定介词短语的后边界。具体算法的流程图如图3.1所示。

3.1.2.2 特征选取

介词短语边界识别中最主要的是后边界的上下文语境，因此，在基于CRF模型的介词短语边界识别中，在判断当前词是否能成为介词的后边界时，选取了当前词及其词性、当前词之后的词语及其词性、介词本身用法属性来作为特征集。

在此实验中，选用了CRF2工具包来进行模型的训练与测试。同基于CRF模型的介词用法自动识别一样，在使用此工具包前，也需要将训练和测试文件转换为模型可以识别的文本格式。表3.1是将例句3.1转换为CRF工具包能够识别的训练数据。

表 3.1 例句 3.1 对应的训练数据样例

0	1	2	3	4	5	6	7
在	p	国际	n	形势	n	<p_zai4_1a>	N
在	p	形势	n	发生	v	<p_zai4_1a>	N
在	p	发生	v	深刻	a	<p_zai4_1a>	N
在	p	深刻	a	变化	vn	<p_zai4_1a>	N
在	p	变化	vn	的	ud	<p_zai4_1a>	N
在	p	的	ud	今天	t	<p_zai4_1a>	N
在	p	今天	t	，	wd	<p_zai4_1a>	Y

表3.1中的第0、1列为待识别的介词及其词性，第2、3列表示可能的后边界及其词性，第4、5列表示可能的后词及其词性，第6列表示介词本身的用法，第7列为标记状态，如果第2列和第4列可以成为待标注介词的后边界和后词，标记

为Y；否则标记为Y。

在引文[8]的实验中给出了基于规则的介词短语边界识别的准确率为57.39%，基于统计的准确率为81.59%，而未加入介词用法的基于统计的边界识别准确率为79.16%。可以看出，基于统计的方法明显优于基于规则的方法，且在基于统计方法的两组实验中，加入介词用法的边界识别效果高于不加入介词用法的边界识别效果，这也证明了介词用法在介词短语边界识别中的可用性。所以，为了突出基于用法的边界识别在句法分析中的作用，在基于介词用法的句法分析后处理研究中仅选择了基于规则的介词短语边界识别与加入介词用法的基于统计的介词短语边界识别两种方法来进行实验。

3.2 介词用法在短语结构句法分析中的应用

本文使用 Stanford 句法分析器的分析结果作为初步分析结果，通过对分析结果的分析及后处理来修改初步分析结果，并进一步提高分析准确率。

3.2.1 方法描述

在本文的对句法分析的研究中，使用了宾州中文树库(Chinese Tree Bank, CTB)[25]，该树库由美国宾尼法尼亚大学附中开发，Stanford parser 的词性标注规则和 CTB 的标注方式相同，汉语词性被划分为 33 类，主要包括 4 类动词和谓语性形容词、3 类名词、3 类限定词和数词等；因此，在此研究中句子的句法分析结果中的词性标记与前几章中有所不同。但为了介词用法识别与介词短语边界识别的方便，本文的研究是分两部分进行的，一部分是从 Stanford parser 中获得句子的句法分析结果，此结果仍然使用 CTB 的词性标注方式；另一部分是介词用法识别与介词短语边界识别，在这一部分的研究中，本文将 CTB 使用的词性标注方式转换为同介词用法识别中相同的标注方式，如例句 3.2 所示。

例句 3.2: 在_PP 民国_NN 时期_NN 的_DEG 上海_NR 生活_VV

在/p 民国/n 时期/n 的/ud 上海/n 生活/v

例句 3.2 中，第一行为 CTB 中文树库中的语料例句，第二行为转换过词性标注方式的语料例句，在进行介词用法识别及介词短语边界识别时使用了第二行的标注形式。

如图 3.2 为 Stanford parser 对例句 3.2 的句法分析结果。

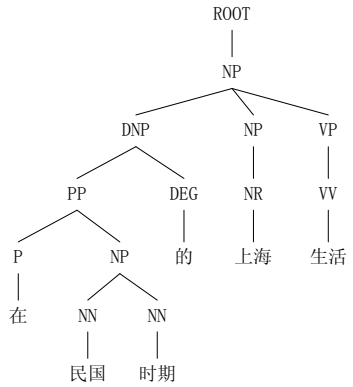


图 3.2 例句 3.2 在 Stanford 句法分析器中的结果

如图 3.2 的分析结果中，此句的介词短语为“在民国时期”，即宾语指动作发生的时间，介词“在”的用法标记为 p_zai4_1a，介词在用法；但经分析，此句中的介词短语应为“在民国时期的上海”，宾语指动作发生或事物存在的处所，用法标记为 p_zai4_2a。即对介词短语边界识别的错误导致了全句的分析也错误。相反，如果首先使用介词的用法特征及此句本身的特点进行边界识别，就可识别出例句中介词“在”宾语的边界，边界识别结果如下所示：

<PP> 在/p<p_zai4_3a> 民国/n 时期/n 的/ud 上海/n </PP> 生活/v

其中<PP>表示前边界，</PP>表示后边界。

利用基于介词用法的边界识别结果来对句法分析结果进行后处理的具体操作过程如下：

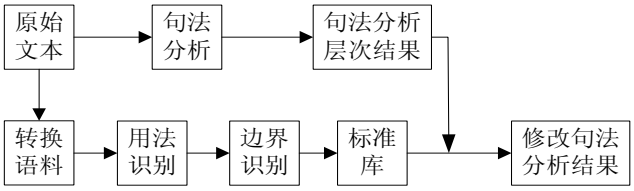


图 3.3 句法分析后处理过程

如图 3.3 所示，本文首先从树库中选取合适的语料，即选取含有介词的句子，再对这些句子利用基于统计的介词用法自动识别方法进行用法自动识别，之后在对标注过用法的句子进行边界识别。同时，使用从树库中提取出的语料作为 Stanford parser 的输入语料，进行句法分析，获得待分析语料的句法分析层次结构。最后使用已获得的边界识别结果对句法分析层次结构进行后处理，

从而获得利用介词用法修改后的新的句法分析结果。

3.2.2 获得边界识别标准库

在对句法分析结果进行后处理之前，需要对已获得的边界识别结果进行预处理，即需要将边界识别结果转换为后处理中所需要的格式。这里本文使用了一个数组形式的链表来存放边界识别中的信息，作为边界识别的标准库。如下例句：

例句 3.3: 他/rr 〈PP〉 在/p 葡萄牙/nr 〈/PP〉 里斯本/nr 念/v 大学/v 、/wu〈PP〉 在/p 德国/nr 〈/PP〉 拿/v 硕士/n ，/wd 又/d 〈PP〉 在/p 瑞典/nr 〈/PP〉 念/v 景观/n 。/wj

例句 3.4: 外商/n 投资/n 企业/n 〈PP〉 在/p 改善/v 中国/nr 出口/n 商品/n 结构/n 中/f 〈/PP〉 发挥/v 了/ul 显著/a 作用/n 。/wj

例句 3.3 中，标准库中写入的是：3222，其中 3 指句子中共 3 个介词；第一个 2 指第一个介词短语共包含两个词，即“在”和“葡萄牙”；第二个 2 指第二个介词短语共包含两个介词，即“在”和“里斯本”；同理第三个 2 指第三个介词短语包含两个介词，为“在”和“瑞典”。而例句 3.4 标准库中写入的是：17，其中 1 指句子中包含 1 个介词，7 指介词短语中共包含 7 个词，此介词短语为“在改善中国出口商品结构中”。其中提取第 n 个连词短语中所包含的成分个数 N 计算方法为：$N=P_{n+1}$，$n\in\{1,P_1\}$其中 P_i 为链表中第 i 个位置的值。本文通过这种方法将边界识别结果转换为数组链表形式的标准库，存放于内存中。

3.2.3 后处理方法

在得到边界识别标准库之后，再对句法分析初步结果进行后处理，即使用边界识别结果来修改句法分析的分析结果中的介词的边界。

如例句 3.2“在民国时期的上海生活”在 stanford 句法分析器中的句法分析层次结果为：

```
(ROOT (NP (DNP (PP (P 在)
                   (NP (NN 民国)
                       (NN 时期)))
               (DEG 的))
```

(NP (NR 上海))
(VP (VV 生活))))

而经分析得其正确的介词的宾语为“民国时期的上海”，其层次结构为：

(ROOT (NP (DNP (PP (P 在)
(NP (NN 民国)
(NN 时期))
(DEG 的)
(NP (NR 上海))))
(VP (VV 生活))))

经过对两个层次结构的分析可知，要对句法分析结果进行后处理只需要简单的修改相应括号的位置，即可修改句法分析结果。

另外，在句法修改的过程中发现，单纯的修改句法结果中的括号位置，并不能完全达到修改的效果，因为在有些句法层次结构中，介词短语部分句法标记出现的顺序也会影响到介词句法分析的正误，所以本文在修改句法层次结果的括号的过程中，对句法标记也做了一定的调整。如例句3.5：

例句3.5：在外商投资企业获得好处后

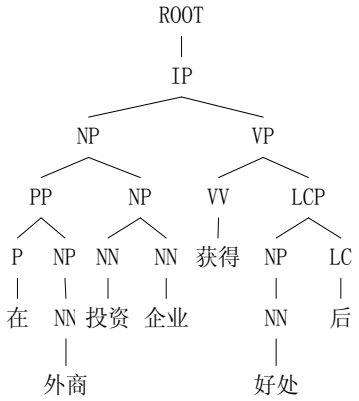


图 3.4 例句 3.5 由 Stanford parser 得出的句法树

图3.4为例句3.5经过Stanford parser的分析所得到的句法树，图3.5为标准句法树，可以看到句子“在外商投资企业获得好处后”为一个完整的介词短语，而图4a的分析结果中介词短语为“在外商”，即可以利用本文提出的方法对层次结果进行修改。例句3.5的句法分析层次结果为：

```
(ROOT (IP (NP (PP  (P  在)
                   (NP  (NN  外商)))
              (NP  (NN  投资) (NN  企业)))
         (VP  (VV  获得)
              (LCP (NP  (NN  好处))
                   (LC  后)))))
```

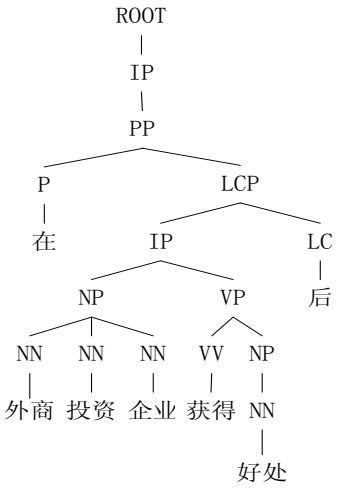


图 3.5 例句 3.5 正确的句法树

例句3.5中，分析得整个句子“在外商投资企业获得好处后”为一个完整的介词短语，而如果仅以修改括号的方法来修改句法分析层次结果的话，结果如下：

```
(ROOT (IP (NP (PP  (P  在)
                   (NP  (NN  外商))
                   (NP  (NN  投资) (NN  企业))
                   (VP  (VV  获得)
                        (LCP (NP  (NN  好处))
                             (LC  后)))))))
```

可以看到，经过对括号的修改，确实可以将介词短语的边界修改正确，但是同样带来了其他的问题，即介词短语中宾语部分句法标记混乱。对于这个问题考虑可将介词短语 PP 前的句法标记调到整个 PP 结构中的相应位置。在确定

了介词短语边界到“后”之后，可知此介词短语应以 PP 开始，在原始树中，PP 和 NP 前部的 NP 短语句法标记可调整到整个 PP 结构中,即 PP 上部的 NP 可在“外商投资企业”的 NP 调整到 PP 中时调整到 PP 的下部,“外商投资企业”的上部，并且可以根据此句中的介词用法调整其中相应的句法标记的位置，如 LCP，如此依次调整后，可得如下句法层次分析结果：

```
(ROOT (IP (PP  (P  在)
                (LCP (NP  (NP  (NN  外商))
                          (NP  (NN  投资) (NN  企业)))
                     (VP  (VV  获得)
                          (NP  (NN  好处))
                          (LC  后))))))
```

如此修改后，虽然句法层次结果并不一定是完全正确的，但相对最初的分析结果已经有了很大改进。因此，本文在修改句法分析层次结果括号的过程中，查看其句法标记，并进行相应的修改，进一步提高句法分析结果的准确率。

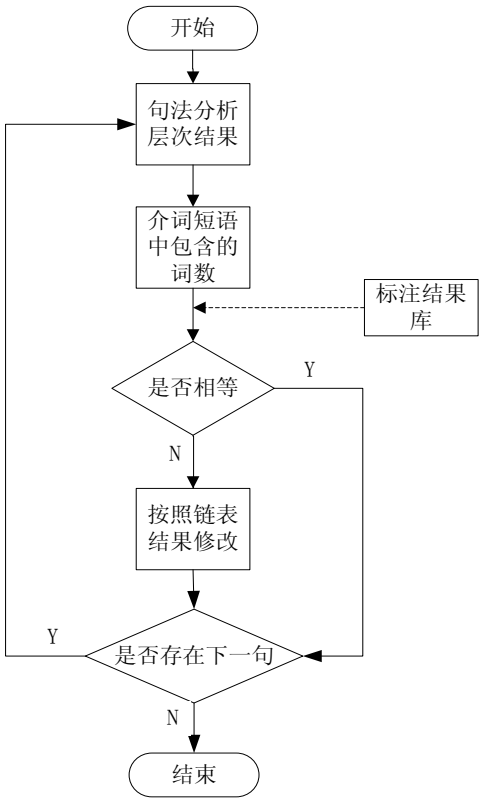


图 3.6 修改句法分析结果流程图

句法分析具体的后处理方法如图 3.6 所示，步骤为：

步骤 1：获得 stanford 句法分析器的句法层次结果，计算此结果中介词短语所包含的词的个数，如例句 1 中介词短语为“在民国时期”，则个数为 3。

步骤 2：获得原始句子的边界识别结果，将此结果设定为标准结果，计算此标准结果中介词短语所包含的词的个数，将其写人标准结果库中。如例句 1 中正确的介词短语为“在民国时期的上海”，则标准结果库中为 15。

步骤 3：将从两个结果中获得的结果进行对比，如 3 和 5，若相等，则不对实验结果进行操作；若不相等，则按照标准结果修改实验结果中括号的位置。对比结果相等或修改完成后，判断是否还存在下一个句子，若存在，则进入步骤 1；若不存在，则退出程序。

其中，句子中第 n 个介词短语的后处理的具体算法描述如图 3.7 所示。

```
输入：边界识别标准库S
      句法分析层次结果T
输出：处理后的句法分析层次结果T1；
Begin
     for i∈[1,T.length] do
       if T中介词第n次出现 then
         for j∈[i,T.length] do
           C←连词前结构中包含的词数
           if C!=Pn+1 then
             k1←介词短语结束的位置
             k2←位置i之后第Pn+1+1个词的位置
             T1←修改T中k1位置的‘）’到k2位置
             T1←修改T中PP前最后一个‘)’到PP
之间的最后一个结果标记到PP之后，此标记对应的
‘)’在k1之后
     return T1
End
```

图 3.7 第 n 个介词短语后处理的算法描述

3.3 实验结果及分析

3.3.1 实验语料

本文选用Chinese Tree Bank5.1来作为实验语料，该树库共包含了507216个词，824975个汉字，18782个中文句子，890个文件。该语料库主要有四个组成部分：原始文本、分词文本、词性标注文本、句法文本[25]。如下为语料库中四个部分中同一句子的格式：

原始：在新增的储蓄存款中

分词：在 新增 的 储蓄 存款 中

词性标注 ：在_P 新增_VV 的_DEC 储蓄_NN 存款_NN 中_LC

句法：((IP (PP (P 在)

(LCP (NP (CP (CP (IP (VP (VV 新增)))

(DEC 的)))

(NP (NN 储蓄)

(NN 存款)))

(LC 中)))))

在本文的实验中，以中文树库中的原始文本作为Stanford parser的输入文本来获得最初句法分析结果，同时以此原始文本作为介词用法识别和介词短语边界识别的输入文本，以句法文本来作为句法分析标准结果。

3.3.2 实验评价指标

为了评价基于介词用法的句法分析后处理的性能，本文在实验中使用了准确率和召回率来作为评价指标。其中，准确率为句法分析结果中正确的短语个数所占的比例，召回率为句法分析结果中正确的短语个数占标准分析树中全部短语个数的比例，其具体定义如下：

$$准确率=\frac{分析得到的正确短语个数}{分析得到的短语总数}\times 100\% \quad (3.1)$$

$$召回率=\frac{分析得到的正确短语个数}{标准树库中的短语总数}\times 100 \quad (3.2)$$

3.3.3 实验结果

文本的实验分为两组，第一组考察句子中介词短语句法分析的分析效果，第二组考察整个句子句法分析的分析效果。其中，介词短语句法分析结果由整句的句法分析结果中得来，即每得到一个句子的句法分析层次结果，就从中提取出此整句句法分析中所分析出的介词短语部分。当对 Stanford parser 初步分析结果进行后处理之后，我们同样使用上述方法提取句法分析结果中所包含的介词短语部分进行第一组的实验。

第一组中计算介词短语的句法分析准确率和召回率时，如例句3.2的最初分析结果中，从中提取的介词短语句法分析部分如下所示：

(PP (P 在)

(NP (NN 民国) (NN 时期)))

而标准结果中提取的介词短语句法分析部分为：

(PP (P 在)

(NP (NN 民国) (NN 时期))

(DEG 的)

(NP (NR 上海)))

为了研究方便，本文先对介词“在”进行了实验，共从语料库中提取出含有介词“在”的长度小于 40 个字的句子 1257 个，表 3.2 为本文实验所得到的介词“在”句法分析的结果，其中原始结果为 stanford 句法分析器分析出的初步分析结果，规则模型、CRF 模型分别为进行了相应的后处理之后的分析结果。从总体结果可看出，边界识别结果越好，后处理结果也就越好，即以边界识别来后处理句法分析层次结果，在句法分析领域可以取得一定的成果。因此，对介词边界识别的研究不论是在边界识别方面，还是在句法分析方面都显得更加重要。

表 3.2 后处理之后各结果对比

模型	准确率(一)	召回率(一)	准确率(二)	召回率(二)
原始结果	78.75	78.16	70.33	71.15
规则模型	81.81	82.94	70.80	71.56
CRF 模型	85.74	87.27	71.63	72.40

从表 3.2 中的实验结果可以看出，本文所用的方法在两组实验中都有一定的提高，而在第一组，即介词短语的句法分析方面效果相对较好，而在整个句子的句法分析方面效果较弱。因为本文的实验中只考虑了边界识别在句法分析中的影响，而介词边界在整个句子的句法分析中所占比例很小，且不足以完全体现介词“在”用法的特点。在今后的研究中，会进一步考虑介词“在”对句法分析有影响作用的其他方面，进一步体现出介词“在”用法特点在句法分析中的影响，提高句法分析的准确率。

表 3.3 为实验语料与标准语料中短语相同的个数的对比，其中原始结果为 Stanford 句法分析器分析出的初步分析结果与标准语料短语相同的个数，规则

模型、CRF 模型分别为经过相应的后处理之后的实验结果与标准语料短语相同的个数，标准结果为标准语料中所有句子所包含的短语数总和。

由表 3.3 可看出，无论是第一组实验还是第二组实验，短语相同的个数都有一定的提高。从整体上来看，规则模型的后处理方法中结构相同的个数与原始结果相比，在两组实验中分别提高了 180 和 290 个相同结构。而 CRF 模型的后处理方法，在两组实验中分别提高了 438 和 537 个相同结构。由此可见，无论是哪一种后处理方法，在对介词短语所进行的边界修改的过程中都没有影响到整个句子的句法分析结果，而是在提高介词短语句法分析准确率的基础上提高整个句子的句法分析准确率。

表 3.3 相同结构数对比

模型	第一组	第二组
原始结果	4407	18701
规则模型	4587	18991
CRF 模型	4845	19238
标准结果	6171	27260

表 3.4 词性标注正确时各分析结果对比

模型	准确率(一)	召回率(一)	准确率(二)	召回率(二)
原始结果	90.03	89.16	87.99	88.91
规则模型	90.88	91.11	88.07	89.02
CRF 模型	93.50	93.75	88.59	89.54

无论表3.2还是表3.3中的实验结果，都是直接在Stanford parser上得出的，并没有考虑到词性标注的正确与否，所以在表3.4中又给出了介词“在”在词性标注正确情况下的实验结果，即从Stanford parser分析出的所有句子的词性标注都与标注语料库中相同。

从表3.4中可以看出，词性标注正确率越高，句法分析正确率也会越高。其实，词性标注正确率越高，介词用法自动识别正确率也会越高，相对应的介词短语边界识别也会越高，所以，可以看到，在词性标注完全正确，即句法分析正确率越高时，对句法分析后处理后的正确率仍会提高。这也说明了本文所提

出的后处理方法，对句法分析的发展是很有帮助的。

表3.5为介词用法词表中一些常用介词在本文实验中的准确率。

表 3.5 常用介词实验结果

词语	词数	原始结果(一)	规则模型(一)	CRF 模型(一)	原始结果(二)	规则模型(二)	CRF 模型(二)
按	53	65.83	79.40	84.96	64.00	65.60	66.27
按照	42	61.61	67.52	75.10	63.88	64.10	75.10
比	126	86.05	86.66	91.08	68.87	68.75	69.69
除	60	71.91	74.76	75.61	60.96	61.05	61.12
除了	145	60.18	67.11	70.54	51.35	52.07	52.77
从	276	64.30	70.75	81.22	62.82	65.31	66.36
当	335	49.97	50.64	60.27	50.69	51.43	52.01
到	170	74.77	79.33	85.75	60.66	61.56	61.97
对	689	67.65	74.92	84.25	68.81	69.23	75.55
给	146	92.07	93.77	95.02	77.08	77.73	78.62
据	127	79.60	83.17	85.48	69.25	70.59	71.62
通过	139	69.24	75.91	79.37	76.16	78.62	79.93
为了	72	49.35	52.05	61.24	55.25	56.32	57.19
向	255	80.93	86.27	87.34	71.05	72.08	72.40
以	256	67.61	77.13	81.06	66.79	72.63	76.69
由	182	68.71	74.60	85.17	51.63	54.14	55.32
于	156	81.88	83.23	89.51	79.07	80.14	81.43
在	1257	78.94	83.54	87.31	70.45	71.60	72.43

3.4 本章小结

这一章主要介绍了介词用法属性在短语结构句法分析中的应用。首先介绍了基于规则和基于CRF统计方法的介词短语边界识别，又详细介绍了介词用法在

短语结构句法分析中的应用研究方法。本章的实验中分别用两种介词短语边界识别方法对短语结构句法分析进行了后处理，并给出了实验结果，且实验中两种后处理方法，特别是基于CRF模型的统计方法，都使句法分析的准确率与召回率有所提高，且无论在准确率与召回率方面，还是正确的短语个数方面，还是对词性标注正确句子的实验，都达到了本文预定的效果，证明了本文提出的介词用法在短语结构句法分析中的研究策略是有效的。

4 连词用法在短语结构句法分析中的应用

为了验证第三章提出的句法分析后处理方法的适用性，本文将介词用法在短语结构句法分析中的研究方法运用到了连词中，提出了连词用法在短语结构句法分析中的应用研究方法。因此，本章主要研究连词在句法分析中的边界，即使用连词的边界识别结果来修改句法分析中连词的结构，以此来考察连词用法特征在句法分析中的影响。而根据连词与介词用法的不同，对句法分析后处理方法也进行了一些必要的修改。

4.1 连词短语边界识别

连词短语边界识别[48]是在连词用法识别的基础上得到的。连词是具有连接作用的虚词，可以连接词、短语、小句、句子和句群，能表示选择、并列、递进、转折、因果、目的等多种连接关系。连词短语边界识别指的是使机器自动识别出连词短语边界，本文分别使用了基于规则和基于统计两种方法来实现连词短语边界识别。

4.1.1 基于规则的连词短语边界识别

在边界识别方面，连词与介词有很大差别，对介词来说，通常在确定了介词的时候也确定了介词短语的前边界，介词的短语边界识别只需要确定其后边界即可。而对于连词来说，其短语边界识别则涉及到了前后两个边界的确定。本文通过对连词用法词典与连词用法规则库的研究，及对各个连词用法的考察，查找与每个用法对应的连词短语结构的边界或形式化表示，抽取其中具有可操作性判断条件的特征，得到了连词短语边界识别规则。本文用“CP”和“/CP”来表示连词短语的前后边界，用“bl”、“xz”、“bc”等表示结构之间的连接关系，即各种关系的汉语拼音缩写。如下例句表示并列关系：

<CP_bl> 改革、发展和稳定 </CP_bl> 的任务非常繁重。

表4.1为连词用法词典中连词“和”的部分属性说明，包括连词“和”的用法编码、释义、用法及例句。如用法“c_he2_1”表示连接三个以上的成分，可知，其第一个成分是其短语边界的前边界，最后一个成分是其后边界。用法

c_he2_2”用在“无论”、“不管”、“不论”之后，即可确定“无论”、“不管”、“不论”之后的第一个词即为此连词短语边界的前边界，而后边界也可根据连词“和”之后所跟成分来确定。

表 4.1 连词“和”用法词表中的部分属性说明

用法编码	释义	用法	例句
c_he2_1	表示平等的联合关系。<b>	连接类别或者结构相近的并列成分。<b>	老师和同学都赞成这么做<b>
c_he2_1a	表示平等的联合关系。<b>	连接三项以上，“和”放在最后两项之间，前面的成分用顿号连接。<b>	一切事物都有发生、发展和消亡的过程<b>
c_he2_1b	表示平等的联合关系。<b>	有几个层次的多项并列成分。<b>	爸爸、妈妈和哥哥、姐姐都不在家<b>
c_he2_1c	表示平等的联合关系。<b>	连接做谓语的动词短语、形容词短语时。<b>	事情还要进一步调查和了解<b>
c_he2_2	表示选择，相当于“或”。<x>	常用于“无论、不论、不管”后。<b>	在沪注册的企业不论所有制和归属，都可以享受这一政策。<r>

如此，根据对连词的各个用法的分析，得到了连词短语边界识别规则，如下为连词“和”的边界识别规则：

$和

@<c_he2_2>→MN ^M→(无论|不论|不管)<CP> ^N→</CP>(, |。)

@<c_he2_1>→B¯B ^B→n|d

@<c_he2_1a>→B、{B、}¯B ^B→a|v|n

@<c_he2_1a>→MN ^M→X、 ^N→</CP>(等|的)

@<c_he2_1c>→MN ^M→v ^N→n

@<c_he2_1c>→B¯B ^B→a|v

用法<c_he2_2> 规则中“无论”、“不管”、“不论”之后加上<CP>表示前边界在这些词的后面，“，”、“。”前面加上</CP>表示后边界在这些词语之前；B表示同词性不同词；X表示任意的词性。其中规则“@<c_he2_2>→MN ^M

→(无论|不论|不管)<CP> ^N→</CP>(，|。)”即表示，在连词“和”前面出现的“无论”或“不论”或“不管”后面标注<CP>，而在连词“和”后面的“，”或“。”之前标注</CP>。

在基于规则的连词短语边界识别系统中，以行为单位处理，其具体流程如下：

(1)以行读取文本，以逗号、冒号、分号、句号、叹号、问号将文本分割成小句。

(2)判断小句中是否含有连词。若含有则记录该连词出现的位置及用法编码；否则，处理下一句。

(3)根据用法编码对规则文件中的规则进行解析，得到对应的连词短语边界的规则表示。若无法找到对应的规则，将句子中用法编码左边的字符串写入结果文件，右边的字符串设为新的句子，执行步骤(2)。

(4)从对应的规则中获取连词短语左右边界的定义及特征，并按照是否含有“<CP>”、“</CP>”来确定边界的位置。

(5)根据连词短语左右边界的特征在连词前后进行匹配。若匹配成功，则根据连词用法词典，获得对应的关系标记x，左边界插入“<CP_x>”，右边界插入“</CP_x>”，最后将连词左边的字符串写入结果文件，右边的设为新的句子，执行步骤(2)；若匹配不成功，对连词对应的下一个边界规则进行解析，执行步骤(3)。

4.1.2 基于统计的连词短语边界识别

连词短语的边界识别可以被看成为文本中词语和词性序列选择标记和确定边界的过程。因此，可以选择CRF统计模型来确定连词短语的边界，识别出连词结构短语。

4.1.2.1 算法描述

连词短语边界识别的目的是从已进行了分词与词性标注并且对连词词性的词语进行了用法标注后的句子中识别出连词对应的连词短语。

与介词短语边界识别不同，连词短语边界识别时，要将句子中的所有词语作为一个样本，选择合适的特征来表示该样本，利用统计模型对样本进行分类，最后在利用样本分类的结果确定连词短语的前、后边界。具体算法过程如图4.1

所示。

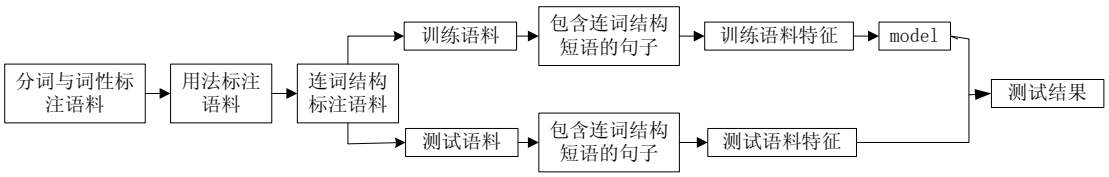


图4.1 基于统计的连词短语边界识别过程

4.1.2.2 特征抽取

在连词短语的边界识别中，可以利用有连接功能的连词和顿号来对边界识别进行研究。因此，选择了词语、词性和连接功能标记来作为本文的特征。在本节的特征集中，连词本身使用连词用法编码表示，顿号使用Y来标记，其他情况使用N来标记。

连词短语边界识别的标记本文参考了王东波[49][50]的研究方法，利用公式4.1来计算语料中连词短语的平均长度，将标注集确定为七词位标注集。

$$L=\frac{1}{N}\sum_{i=3}^{k} i N_i \tag{4.1}$$

其中，N_i为长度为i的连词短语的个数，K为连词短语的最大长度，N为连词短语的总个数。具体的七词位标注集为$T=\{B, S, T, F, M, E, O\}$，其中$B$表示连词短语的开始词，$S$表示第二个词，$T$表示第三个词，$F$表示第四个词，$M$表示包括第五个在内的、第五个以上的词，$E$表示介词短语结尾的词，$O$表示连读短语外部的词。例句4.1的特征如表4.2所示。

例句4.1：各地开展的节水灌溉、打井和灌区节水等工作。

在引文[47]的实验中给出了基于规则的连词短语边界识别的准确率为48.67%，基于统计的准确率为72.73%，而未加入介词用法的基于统计的边界识别准确率为71.27%。可以看出，基于统计的方法明显优于基于规则的方法，且在基于统计方法的两组实验中，加入连词用法的边界识别效果高于不加入连词用法的边界识别效果，这也证明了连词用法在连词短语边界识别中的应用价值。所以，与第三章相同，本章中为了突出基于用法的边界识别在句法分析中的作用，在基于连词用法的句法分析后处理中选择了基于规则的连词短语边界识别与加入连词用法的基于统计的连词短语边界识别两种方法来进行试验。

表 4.2 例句 4.1 的特征表示

词语	词性	连接结构标记	识别标记
各地	rz	N	O
开展	v	N	O
的	ud	N	O
节水	vi	N	B
灌溉	v	N	S
、	wu	Y	T
打井	vi	N	F
和	c	<c_he2_1a>	M
灌区	n	N	M
节水	vn	N	E
等	u	N	O
工作	vn	N	O

4.2 连词用法在短语结构句法分析中的应用研究

本章的研究同样使用 Stanford 句法分析器的分析结果作为初步分析结果，通过对句法结果的分析及后处理来修改初步分析结果。

4.2.1 方法描述

因为本章基于连词用法的句法分析研究是从第三章的基于介词用法的句法分析研究上延伸出来得，所有基本的研究方法与第三章相同。如例句 4.2：

不管是地方文化、民间文化**或者**精致文化。

如图 4.2 为 Stanford parser 对例句 4.2 的句法分析结果。

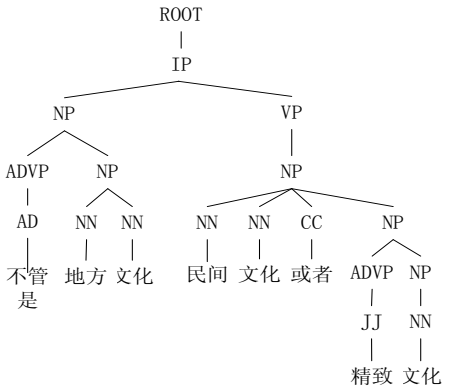


图 4.2 例句 4.2 在 Stanford parser 中的结果

如图 4.2 的分析结果中，此句的连词短语为“民间文化或者精致文化”，表示两个成分之间的选择关系，此时的连词“或者”的用法为<c_huo4zhe3_1>。而分析整个例句 4.2 可得，其正确的连词短语应为“地方文化、民间文化或者精致文化”，为三个成分之间的选择关系，且与前面的“不管是”相对应，此时连词“或者”的用法为<c_huo4zhe3_1c>。即对连词短语边界识别的错误导致了全句的分析也错误。相反，如果首先使用连词的用法特征及此句本身的特点进行边界识别，就可识别出例句中连词“或者”的边界，边界识别结果如下所示：

不管是/d <CP_ xz> 地方/n 文化/n 、/wu 民间/n 文化/n **或者**/c 精致/a 文化/n </CP_ xz >

其中<CP>表示前边界，</CP>表示后边界。

利用基于连词用法的边界识别结果来对句法分析结果进行后处理的具体操作过程如图 4.3 所示。

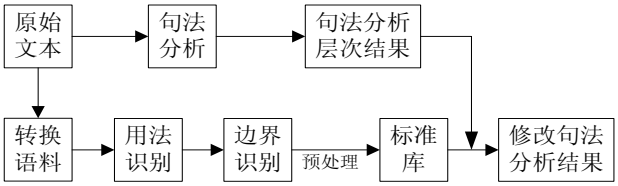


图 4.3 基于连词的句法分析后处理过程

本文首先从树库中选取合适的语料，即选取含有连词的句子，再对这些句子利用基于统计的连词用法自动识别方法进行用法自动识别，之后对标注过用法的句子进行边界识别，再对边界识别结果对连词进行相应的处理。同时，使用从树库中提取出的语料作为 Stanford parser 的输入语料，进行句法分析，获得待分析语料的句法分析层次结构。最后使用处理后的边界识别结果对句法分析层次结构进行后处理，从而获得利用连词用法修改后的新的句法分析结果。

4.2.2 获得边界识别标准库

根据连词的特性，为了方便对句法分析进行后处理，对连词的边界识别结果进行了一些修改，如对例句 4.2 的边界识别结果的修改结果如下：

不管是/d <CP_ xz> <c_1> 地方/n 文化/n </c_1> 、/wu <c_2> 民间/n 文化/n </c_2> **或者**/c <c_3> 精致/a 文化/n </c_3> </CP_ xz >

此处分别用<c_n>、</c_n>标注出了连词短语中的不同成分，其中 n 指句子

中的第 n 个连接成分。有了这个标注，就可以更加详细的描述连词短语，从而能更详细的修改句法分析结果。

与介词的句法分析修改过程相同，在对连词的句法分析结果进行后处理之前，同样需要对已获得的边界识别结果进行预处理，即需要将边界识别结果转换为后处理中所需要的格式。这里本文使用了一个数组形式的链表来存放边界识别中的信息，作为边界识别的标准库。如下例句：

名列/v <CP_b1> <c_1> 第二/m 位/q </c_1> 与/c<c_yu3> <c_2> 第三/m 位/q </c_2> </CP_b1> 的/ud 企业/n 分别/d 为/v <CP_b1> <c_3> 摩托罗拉/nr （/wkz 中国/nr ）/wky 电子/n 有限/a 公司/n </c_3> 和/c<c_he2_1> <c_4> 广东/nr 核电/n 合营/n 有限/a 公司/n </c_4> </CP_b1> 。/wj

例句中共有 2 个连词，分别为“与”、“和”，标准库中写入的是：2222275，其中第一个 2 指句子中共含有 2 个连词；第二位的 2 指第一个连词短语共有两个连接的成分；第三位的 2 指第一个连词短语的第一个成分中有 2 个词，同理第四位的 2 指第一个连词短语的第二个成分中有 2 个词；其中提取第 n 个连词短语第 m 个部分中的成分个数在链表中所在的位置 $Q_{n,m}$ 的计算方法为：

$$\begin{cases} Q_{n,m} = \sum_{i=1}^{n-1}(P_i + 1) + m + 2 \\ P_i = S_{W_i} \\ W_i = W_{i-1} + S_{W_{i-1}} + 1 \end{cases} \tag{4.2}$$

其中，P_i 为第 i 个连词包含的结构数，W_i 为第 i 个连词在链表中存储的起始位置，W_1 值为 2，S_j 为链表中第 j 个位置所对应的值，S_1 为句子中包含的连词总个数。知道了每个连词短语的成分数，就知道了相对于的每个成分中所含的词数，在此标准库中，第 W_n 个位置后面所跟的 S_{Wn} 个值即为第 n 个连词中每个成分中的词数。本文通过这种方法将边界识别结果转换为数组链表形式的标准库，存放于内存中。

4.2.3 后处理方法

在得到边界识别标准库之后，再对句法分析初步结果进行后处理，即使用边界识别结果来修改句法分析的句法结果中的连词的边界。

如例句 4.2“不管是地方文化、民间文化或者精致文化”在 stanford 句法分析器中的句法分析层次结果为：

```
(ROOT (IP  (NP (ADVP  (AD 不管是))
               (NP (NN 地方) (NN 文化)))
           (VP (NP (NN 民间)
                   (NN 文化)
                   (CC 或者)
                   (NP (ADJP (JJ 精致)) (NP (NN 文化)))))))
```

而标注树库中连词短语为“地方文化、民间文化或者精致文化”，其层次结构为：

```
(ROOT (IP (ADVP (AD 不管是))
      (NP (NP (NN 地方) (NN 文化))
          (NP (NN 民间) (NN 文化))
          (CC 或者)
          (NP (ADJP (JJ 精致)) (NP (NN 文化))))))
```

同样，经过对两个层次结构的分析可知，要对连词的句法分析结果进行后处理只需要简单的修改相应括号的位置，即可修改句法分析结果。

因为连词短语中，有部分成分在连词前部，有部分成分在连词后部，所以对连词句法分析的后处理分为三部分：前部、后部、整体。此处，设置了一个变量i来表示连词前部成分中已经被修改的成分，初始值为0，此值主要用在连词总成分数大于二的情况；设置N=总成分数-1。具体修改步骤如下：

(1)获得Stanford parser句法分析层次结果，作为最初结果，从边界识别标准库中提取该连词短语的成分数。

(2)判断i是否小于N，若小于则表示连词前部成分中还未修改完，进入步骤(3)；若i等于N，则表示连词前部所有成分已修改完成，进入步骤(4)。

(3)连词前部成分修改过程。从最初结果中获得当前成分中包含的词数，同时从边界识别标准库中提取该成分中正确的词个数，判断两者是否相等，若不相等，则修改最初结果；若相等或者修改完成后，i值加1，即表示连词前部成分修改了一次，之后进入步骤(2)。

(4)连词后步成分修改过程。从最初结果中获得当前成分中包含的词数，同时从边界识别标准库中提取该成分中正确的词个数，判断两者是否相等，若不

相等，则修改最初结果；若相等或者修改完成后，进入步骤(5)。

(5)连词短语修改过程。从最初结果中获得整个连词短语中包含的词数，同时从边界识别标准库中提取该连词短语中正确的词个数，判断两者是否相等，若不相等，则修改最初结果；若相等或者修改完成后，则判断是否还存在有下一句，若存在，进入步骤（1）；若不存在，则退出程序。

句法分析后处理中对一个连词短语修改的过程如图4.4所示。

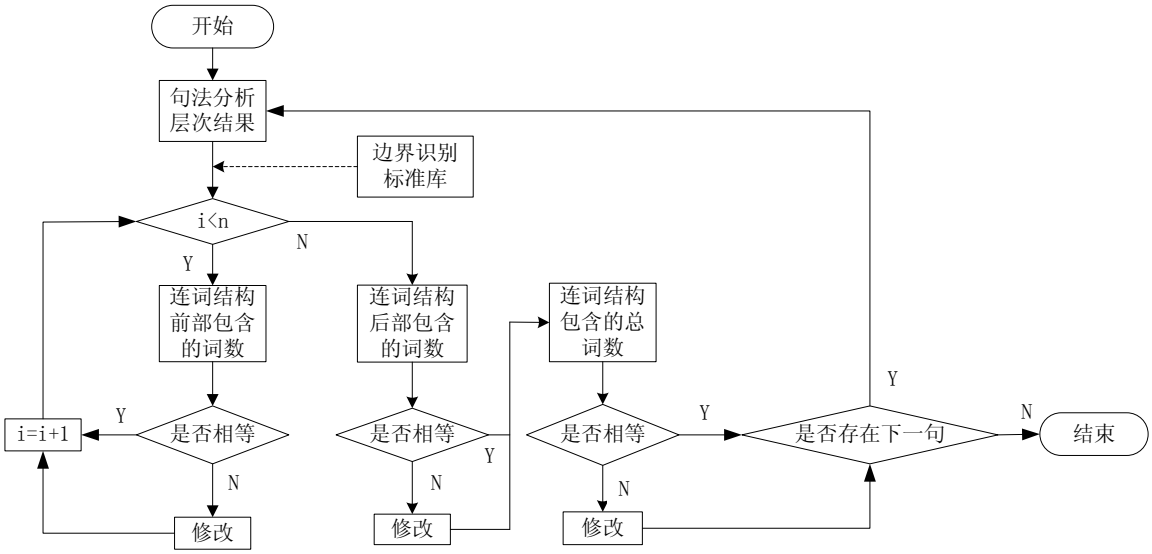


图4.4 基于连词用法的句法分析后处理流程图

其中，对句子中第n个连词后处理的具体算法描述如图4.5所示。

输入：边界识别标准库S
　　句法分析层次结果T
输出：处理后的句法分析层次结果T_1;
Begin
 for i∈[1,T.length] do
  if T中连词第n次出现 then
   if W_n=2 then
    C←连词前结构中包含的词数
    if C!=S_{Wn+1} then
     k_1←连词前结构的起始位置
     k_2←连词前S_{Wn+1}个词的起始位置
     T_1←修改T中k_1位置的‘(’到k_2位置
   if W_n>2 then
    for j∈[1,S_{Wn}-1] do
     C←连词前第j个结构中包含的词数
      if C!=S_{Wn+j} then
       k_1←连词前第j个结构的起始位置
       k_2←连词前第j个结构中S_{Wn+j}个词的起始位置
       T_1←修改T中k_1位置的‘(’到k_2位置
 C←连词后结构中包含的词数
  if C!= S_{Wn+SWn} then
    k_1←连词后结构的结束位置
    k_2←连词后第S_{Wn+SWn}个词的结束位置
    T_1←修改T中k_1位置的‘)’到k_2位置
 C←连词结构中包含的词数
  if C!= S_{Wn+1}+S_{Wn+2}+…+S_{Wn+SWn} then
    k_1←连词结构的结束位置
    k_2←连词结构中S_{Wn+1}+S_{Wn+2}+…+S_{Wn+SWn}个词的结束位置
    T_1←修改T中k_1位置的‘)’到k_2位置
  return T_1
 End

图 4.5 对第 n 个连词结果后处理的算法描述

4.3 实验结果及分析

因为本章由介词用法在短语结构句法分析中应用研究中延伸所得，所以本章实验中所用的实验语料与评价指标皆与第三章中相同。

本章实验分为两组，第一组考察句子中连词短语句法分析的分析效果，第二组考察整个句子句法分析的分析效果。第一组中计算连词短语的句法分析准确率和召回率时，如例句 4.2 的最初分析结果中，从中提取的连词短语句法分析部分如下所示：

(NN 民间)

(NN 文化)

(CC 或者)

(NP (ADJP (JJ 精致)) (NP (NN 文化)))

而标准结果中提取的连词短语句法分析部分为：

(NP (NN 地方) (NN 文化))

(NP (NN 民间) (NN 文化))

(CC 或者)

(NP (ADJP (JJ 精致)) (NP (NN 文化)))

本文对几个常用连词进行了研究，分别为“和”、“而”、“或”、“因为”、“那么”、“而且”。表 4.3 为本文实验所得到的常用连词句法分析的结果，其中原始结果为 Stanford 句法分析器分析出的初步分析结果，规则模型、CRF 模型分别为进行了相应的后处理之后的分析结果。从总体结果可看出，边界识别结果越好，后处理结果也就越好，即以边界识别来后处理句法分析层次结果，在句法分析领域可以取得一定的成果。因此，对介词边界识别的研究不论是在边界识别方面，还是在句法分析方面都显得更加重要。

表 4.3 后处理之后的各分析结果对比

模型	准确率(一)	召回率(一)	准确率(二)	召回率(二)
原始结果	60.48	50.38	72.43	70.70
规则模型	62.19	62.76	72.92	72.32
CRF 模型	64.77	67.95	73.32	73.41

从表 4.3 中的实验结果可以看出，本文所用的方法在两组实验中都有一定

的提高，而在第一组，即连词短语的句法分析方面结果提高较大，而在整个句子的句法分析方面提高较少，主要是因为对连词来说，在句子中的运用较灵活，而本章的研究方法仅针对句子中的连词短语。

表 4.3 的结果是直接由 Stanford parser 得出的，并没有考虑到词性标注的正确与否，所以在表 4.4 中又给出了几个常用连词在词性标注正确情况下的实验结果，即从 Stanford parser 分析出的所有句子的词性标注都与标注语料库中相同。

表 4.4 词性标注正确时各分析结果对比

模型	准确率(一)	召回率(一)	准确率(二)	召回率(二)
原始结果	68.11	59.71	83.94	79.69
规则模型	68.84	69.71	84.12	81.83
CRF 模型	71.88	79.13	85.07	84.76

在表4.4中，词性标注完全正确，即句法分析正确率越高时，对句法分析后处理后的正确率仍会提高。也就是说，本文提出的句法分析后处理方法不仅对含介词的句子有用，对含有连词的句子同样有用，即利用边界识别结果对句法分析结果进行后处理，对句法分析是有帮助的。这也说明了本文所提出的后处理方法，对句法分析的发展是有利的。

4.4 本章小结

这一章主要介绍了连词用法属性在短语结构句法分析中的应用。首先介绍了基于规则和基于CRF统计方法的连词短语边界识别，又详细介绍了连词用法在短语结构句法分析方法中的应用研究。本章的实验中分别用两种连词短语边界识别方法对短语结构句法分析进行了后处理，并给出了实验结果，且实验中两种后处理方法，特别是基于CRF的统计方法，都使句法分析的准确率与召回率有所提高，且无论在准确率与召回率方面，还是对词性标注正确句子的实验，都达到了本文预定的效果，进一步证明了本文提出的短语结构句法分析后处理方法的适用性。

5 结论与展望

5.1 结论

本文构建了“三位一体”的现代汉语介词用法知识库，包括现代汉语介词用法词典、现代汉语介词用法规则库、现代汉语介词用法语料库。在此知识库与基于规则的介词用法自动识别的基础上，研究了基于统计的现代汉语介词用法自动识别。分别选用了条件随机场模型、最大熵模型和支持向量机模型三种统计模型进行了该实验，实验结果表明，上下文窗口与模型的选择都对介词用法的识别有很大影响。在统一了上下文窗口时，三种不同的统计模型中，支持向量机模型效果最好，条件随机场模型次之，而最大熵模型效果相对最差，但在总体上，基于三种统计模型的识别准确率都要高于基于规则的介词用法识别的准确率。

在介词用法与介词短语边界识别的基础上，本文提出了一种短语结构句法分析后处理方法。即将Stanford parser分析结果与基于介词用法的介词短语边界识别结果相结合，获得一个更准确的句法分析结果。本文共选用了两种介词短语边界识别方法进行实验，包括基于规则的介词短语边界识别和基于统计的介词短语边界识别，另外在实验中，分别在介词短语和整句上进行了对比验证。实验证明，两种后处理方法所得的结果，都比Stanford parser的原始结果有所提高，证明了该后处理方法的可用性。

为了进一步证明介词用法在短语结构句法分析中的应用研究方法的适用性，本文又将基于介词的方法运用到连词中，提出了连词用法在短语结构句法分析中的应用研究，且在原来的基础上，根据连词用法的特征做出了相应的修改。实验中选用了基于规则和基于统计两种连词边界识别方法，分别在连词短语和整句上进行对比验证。该实验结果的提高证明了本文提出的用法属性在短语结构句法分析中的应用的研究策略的有效性，为短语结构句法分析的研究提出了新的研究思路。

5.2 展望

本文提出的介词用法自动识别中，虽然基于统计的介词用法自动识别方法比基于规则的识别方法效果要高，但是在实验结果中也可以看到，并不是所有的基于规则的方法都比基于统计的低，特别是有些稀有的用法。对此，我们发现，在自然语言处理研究中，不能单纯依靠规则，也不能单纯依靠统计，应当把两者更紧密地联系起来，彼此取长补短。

在介词用法在短语结构句法分析中的应用研究中，本文的研究已经取得了一定的提高，但是在此后处理方法中，本文只考虑了介词短语边界对句法分析的影响，并没有具体的考虑介词本身在句法分析中的影响，所以，今后的研究中，需要进一步将介词用法本身融入到句法分析中，更强的体现介词用法在句法分析中的作用。

最后，对介词用法在短语结构句法分析中的应用研究方法的适用性的验证，本文只在连词用法上进行了验证，并且也得到了较好的效果，下一步的研究中也会将本文的后处理的方法运用到其他虚词中。

参考文献

[1] 马金山，刘挺.汉语自动句法分析的理论与方法[J].当代语言学，2009，第二期:100-112

[2] Allen, J. Natural Language Understanding.2nd edition. Menlo Park, CA: Benjamin Cummings，1995

[3] 俞士汶，朱学锋，刘云.现代汉语广义虚词知识库的建设[J].汉语语言与计算学报，2003，2（1）:89～98

[4] 刘云.汉语虚词知识库的建设[D].北京大学博士后出站报告，2004

[5] 昝红英，朱学锋.面向自然语言处理的汉语虚词研究与广义虚词知识库构建[J].当代语言学，2009，11（2）:124～135

[6] 刘锐.基于规则的现代汉语副词用法自动识别研究[D].[硕士学位论文].郑州：郑州大学，2009

[7] 昝红英，张军珲，朱学锋，俞士汶.副词“就”的用法及其自动识别研究[J].中文信息学报，2010，24（5）:10～16

[8] 袁应成.基于用法属性的现代汉语介词短语边界识别研究[D].[硕士学位论文].郑州：郑州大学，2011

[9] 周丽娟.现代汉语连词用法的自动识别及应用研究[D].[硕士学位论文].郑州：郑州大学，2012

[10]Lingling Mu，Yiya Pang，Hongying Zan.Studies on Automatic Recognition of Preposition BA’s Usages Based on Statistics[C]. In: 2012 2nd IEEE International Conference on Cloud Computing and Intelligence Systems，2012:1875～1879

[11]T. L. Booth，R. A. Thompson. Applying Probability Measures to Abstract Languages[J]. IEEE Transactions on Computers. 1973, C-22(5):442～450

[12]K. Lari，S.J.Young. The Estimation of Stochastic Context-free Grammars Using the Inside-Outside Algorithm[J].Computer Speech and Language.1990, 4(1):35～56

[13]M. Collins.A New Statistical Parser Based on Bigram Lexical Dependencies[C]. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. Santa Cruz. 1996:184～191

[14]M. Collins.Three Generative, Lexicalised Models for Statistical Parsing[C]. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Comference of the European Chapter of the Association for Computational Linguistics, Madrd, 1997:16～23

[15]E. Charniak.A maximum-entropy-inspired parser[C].In Proceedings of the North-American Chapter of the Association for Computational Linguistics, New Brunswick NJ,2000:132～139

[16]Michael Collins.Discriminative Reranking for Natural Language Parsing[C]. In

Proceedings of the 7th International Conference on Machine Learning, Stanford, CA,2000:175～182

[17]Eugene Charniak，Mark Johnson. Coarse-to-fine N-best Parsing and Maxent Discriminative Reranking[C].In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics, Michigan, 2005:173～180

[18]Mark Johnson，Ahmet Engin Ural. Reranking the Berkeley and Brown Parsers Human Language Technologies[C].In Proc. The 2010 Annual Conference of the North-American Chapter of the Association for Computational Linguistics, 2010:665～668

[19]Zhou Q. A statistics-based Chinese Parser[C]. In Proceedings of the 5th Workshop on Very Large Corpora, New York, 1997:4～15

[20]Zhang Y, Xu B and Zong C Q. Chinese Syntactic Paring Based on Extended CLR Parsing Algorithm with PCFG[C].In Proceedings of the 19th International Conference on Computational Linguistics, Taipei, 2002:1308～1332

[21]曹海龙.基于词汇化统计模型的汉语句法分析研究[D].[博士学位论文].哈尔滨：哈尔滨工业大学，2006

[22]Wang M. W，Sagae K，Mitamura T.A Fast, Accurate Deteministic Parser for Chinese. In Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006:425～432

[23]Xiao chen, Changning Huang, Mu Li and Chunyu Kit. Better Parser Combination[J]. 中文信息学报，2009

[24]徐文智，王小捷. 一种结合汉字信息的 LPCFG 汉语句法分析器[J].中文信息学报（主办中文信息学会，清华大学），2009

[25]宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008

[26]中国社会科学院语言研究所词典编辑室编. 现代汉语词典(第 5 版)[M]. 北京:商务印书馆，2005

[27]吕叔湘. 现代汉语八百词(增订本)[M]. 北京：商务印书馆，2007

[28]张斌. 现代汉语虚词词典[M]. 北京：商务印书馆，2001

[29]陈火旺. 程序设计语言编译原理(第三版)[M]. 国防工业出版社，2001

[30]昝红英，张坤丽，柴玉梅，等. 现代汉语副词用法的形式化描述[C]//第八届汉语词汇语义学研讨会论文集. 香港理工大学，2007

[31]刘锐.基于规则的现代汉语副词用法自动识别研究[D].[硕士学位论文].郑州：郑州大学，2009

[32]Hongying Zan，Junhui Zhang. Studies on Automatic Recognition of Chinese Adverb CAI’s usages Based on Statistics[C]. In：Proceedings of International Conference on Natural Language Processing and Knowledge Engineering(NLP-KE2009). Da Lian，2009.393～397

[33]袁应成，昝红英，张坤丽，周溢辉基于规则的虚词用法自动标注算法设计与系统实现[C]. 第 11 届汉语词汇语义学研讨会论文集. 苏州大学：2010. 163～169

[34]昝红英，朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建[J]. 当代语言学，2009

[35]J Lafferty,A McCallum,F Pereira. Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]. In：International Conference on Machine Learning， 2001:282～289

[36]T. Cohn, P. Blunsom. Semantic Role Labeling with Tree Conditional Random Fields. Proceedings of the Ninth Conference on Computational Natural Language Learning[C]. Ann Arbor, Michigan: Association for Computational Linguistics, 2005

[37]苗雪雷. 基于条件随机场的汉语词义消歧方法研究[D].[硕士学位论文]. 沈阳：沈阳航空工业学院，2007

[38]洪铭材，张阔，唐杰等.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学，2006

[39]Jaynes E T. Information theory and statistical mechanics[J]. Physies Reviews. 1957，106(4):620～630

[40]彭其伟. 基于统计方法的中文文本情感倾向分类研究[D]. [硕士学位论文]. 山西：山西大学，2007

[41]王江伟. 基于最大熵模型的中文命名实体识别[D]. [硕士学位论文]. 南京：南京理工大学，2005

[42]陈笑蓉,秦进. 基于最大熵原理的汉语词义消歧[J].计算机科学,2005,32(5):174-176

[43]李素建，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003（12）:1722～1727

[44]李国正，王猛，曾华军.支持向量机导论[M].电子工业出版社，2005

[45]卢志茂，刘挺，李生. 统计词义消歧的研究进展[J]. 电子学报，2006，34（2）：333～343

[46]John Hughes. Automatically acquiring a classification of words[D]. PhD dissertation.Paris：University of Leeds，1994

[47]陈昌来.介词与介引功能[M]. 安徽教育出版社，2002

[48]昝红英，周丽娟，张坤丽. 基于用法的现代汉语连词结构短语识别研究[J]. 中文信息学报，2012，26（6）

[49]王东波，陈小荷，年洪东. 基于条件随机场的有标记联合结构自动识别[J]. 中文信息学报，2008，22（6）:3～7

[50]Dongbo Wang, Danhao Zhu, Xinning Su, Jing Xie. Automatic Identification of Parallel Structure Based on Conditional Random Field[C]. In：Proceedings of the Third International Symposium on Computer Science and Computational Technology(ISCSCT ’10)，Jiaozuo，2010:400～408

个人简历　在学期间发表的学术论文及研究成果

个人简历

庞熠雅，女，1987 年 11 月 7 日生，河南省卫辉市人。
2010 年 6 月毕业于河南科技学院新科学院，计算机科学与技术专业，工学学士；
2013 年 6 月毕业于郑州大学，计算机应用技术专业，工学硕士。

在学期间发表的学术论文

Lingling Mu，Yiya Pang，Hongying Zan. Studies on Automatic Recognition of Preposition BA's Usages Based on Statistics[C]. In: 2012 2nd IEEE International Conference on Cloud Computing and Intelligence Systems，2012: 1875-1879

致谢

时光飞逝，弹指之间，研究生的三年生活转眼就要过去。遥忆初到这所大学时，意气风发。而如今即将毕业，感慨万千，奋斗和辛劳成为了甜美的记忆，各种欢笑也历历在目。在我的论文即将完成之际，我谨向三年来所有给予我关怀、帮助和支持的人们表达最诚挚的深深的感谢与美好的祝福！

首先衷心的感谢我的恩师穆玲玲副教授和郑州大学自然语言处理实验室的负责人咎红英教授，她们治学严谨，对工作认真负责，也为我指明了研究方向，在生活上和学习上给我无私的关怀和帮助，在我从论文选题到收集资料，从写稿并反复的修改的过程中给予了我精心的指导，并给我的课题提出了建设性的意见。能够遇到这样两位好老师，我觉得自己非常的幸运，并且非常感激穆老师和咎老师对我的指导和帮助。

感谢郑州大学自然语言处理实验室的柴玉梅老师、韩英杰老师、张坤丽老师、贾玉祥老师、赵丹老师给予我的指导和建议，你们的每次指导都使我获得很大的收获。

感谢郑州大学信息工程学院的周清雷老师、李向丽老师、钱晓捷老师、叶会英老师、徐江峰老师等老师在我学习上的指导和帮助。感谢研究生办公室的庞军老师和高明磊老师在生活上给予的帮助。

感谢郑州大学自然语言处理实验室的兄弟姐妹：王作飞、周溢辉、袁应成、李国华、周丽娟、陈明、张腾飞、张静杰、梁猛杰、夏静、李钰、范庆虎、任丽君、娄鑫坡，和你们在一起的快乐时光令我难忘。

感谢我的朋友亲人，在我多年的求学路上，他们默默地为我付出，给予我关怀、照顾和支持，他们是我人生最坚强的后盾和道路上的指明灯！在未来的时间和日子里，我会努力的工作和学习，不会辜负父母对我的期望，我会将他们一如既往的鼓舞和支持转化为我努力的精神支柱和力量源泉。

最后，衷心的感谢所有在百忙之中抽出来时间对本文进行审阅、评议和参加本文答辩而付出辛勤劳动和提出宝贵意见的各位专家和教授！谢谢！