



天津师范大学
硕士学位论文

天津师范大学研究生处

分类编号: _____

单位代码: 10065

学 号: 03209010

天津师范大学

研究生学位论文

论文题目: 汉语语法语料库系统的基础设计

学 生 姓 名 : 郭 鹏 申请学位级别: 工学硕士

申请专业名称: 计 算 机 应 用

研 究 方 向 : 人 工 智 能


指导教师姓名: 齐丙辰 专业技术职称: 教 授

提交论文日期: 2006 年 4 月

天津师范大学

学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的科研成果。除文中已经加以标注引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的研究成果，也不含为获得天津师范大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

作者签名： 

日期：2006年 6 月 6 日


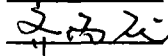
学位论文授权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权天津师范大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

- 1、保密口，在_____年解密后适用本授权书。
- 2、不保密口。

(请在以上相应方框内打“ ”)

作者签名： 
导师签名： 

日期：2006年 6 月 6 日
日期：2006年 6 月 6 日

摘 要

汉语作为自然语言处理内容的研究工作在国内已经开展数十年。根据汉语研究中所体现出的特点,有关语义分析研究也越来越多地应用到自然语言处理的研究过程中。但是由于语义分析的结果不易转化成形式化的表达方式,从而不易被计算机所识别并予以处理。因此,在研究过程中需要利用一种行之有效的方法来辅助表述语义分析过程。

语法分析是在语义分析层次之上的,它用有限的知识描述语言学中的各种语言现象,而且语法分析相对容易归纳,也容易转化成形式化的语言被计算机处理。同时,鉴于汉语研究需要在一定语言环境中进行这一特点,引入了语料库这一研究方法,并且结合汉语语法分析,实现对汉语语言研究的计算机处理。

本文总结了自然语言处理的发展情况及国内研究的重要成果。根据汉语研究的特点,结合前人的研究成果,总结出自然语言处理研究的基本过程。

本文结合自然语言处理研究的特点和汉语语法理论知识,从有利于计算机处理的角度,对汉语语法研究的元素作了重新分类,阐述了各元素间的联系,提出了特征提取的概念并且介绍了特征提取的内容和作用。

本文根据汉语语法内容以及自然语言处理过程的需要,引入了语料库的有关研究方法,详细研究了语法语料库的结构设计,并根据汉语语法中各成分的构成结构建立了语法规则库,并且依据规则库的内容设计了语料库中多个数据表之间的访问方法。另外,对语料库构建过程中的自动分词技术进行改进。

本文还为语料库添加了应用程序,实现了语料库系统主要的应用功能,并且还设置了语料库管理系统,主要用于语料库中数据表的管理,也为今后有关工作继续进行提供统一的开发环境。

关键词: 自然语言处理, 汉语语法, 语义, 语料库, 特征提取, 自动分词

Abstract

Chinese language as the study object of NLP is carried out for a lot of years. Analysis and study of the semantic is more and more put to use in NLP study course because of Chinese study's characters. In that the result about semantic analysis changes uneasily to the formational style of express, it can not be accepted and handled by computer. Then, we need find a better way to describe it in NLP.

The syntactic analysis is on the level of the semantic analysis. It describes the kinds of language phenomena, and the result of the syntactic analysis is easier to be concluded and to change to formational style. All of this is in its formational knowledge. Because the Chinese language study needs a language environment, the corpus is put into this paper study.

The paper introduces the development of the NLP and the internal important achievements. Because of the characters of the Chinese language, combined with these achievements, it concludes the base process of NLP study.

The paper reclassifies some knowledge of the Chinese syntax and describes the relationship among them, based on he characters of NLP and computer's process. And it advances a conception named Character Extract and describes its content and functions.

The paper draws the research means and techniques of the corpus into this system, based on the need of the study about the Chinese syntax and NLP. It detailed studies the structure of the Chinese syntax corpus, and builds up the Chinese syntax rules DB, and designs the visit ways among the data tables. Besides, it improves the way of the automatic words segmentation.

The paper designs the application programs about the corpus application and the DB management to perfect the whole system.

Keywords: NLP, Chinese Syntax, Semantic, Corpus, Character Extract, Automatic Words Segmentation

目 录

摘 要	I
ABSTRACT	II
目 录	IV
第一章 引 言	1
1.1 自然语言处理概述	1
1.2 国内主要成果	1
1.3 自然语言处理的基本过程	2
1.4 自然语言处理的层次	3
1.5 语义分析和语法分析	4
1.6 自然语言处理中存在的困难	5
1.7 课题背景	6
1.8 本章小结	7
第二章 汉语言语法理论知识	8
2.1 汉语的特点	8
2.2 语法概述	10
2.2.1 语法概念	10
2.2.2 语法体系	10
2.2.3 结构层次	10
2.2.4 语法性质	11
2.2.5 语法与逻辑的关系	11
2.2.6 汉语语法结构	11
2.3 词法	12
2.3.1 词类划分的原则	12
2.3.2 词类划分的依据	13
2.3.3 词类划分及语法功能	14
2.4 句法	16
2.4.1 短语	17
2.4.2 句子成分	18
2.4.3 单句类型	18
2.4.4 复句类型	19
2.5 词、短语、句子之间的关系	20
2.5.1 词与短语的关系	20
2.5.2 词、短语与句子成分的关系	20
2.5.3 词、短语与句型的关系	20
2.5.4 词、短语与复句的关系	20
2.6 语法特征提取	21
2.7 语法知识在语义分析中的应用	23
2.8 本章小结	24
第三章 汉语语法语料库的结构设计及相关技术	25

3.1 语料库历史	25
3.2 语料的采集	26
3.3 语料库的结构	26
3.4 规则库的建立	27
3.5 语料库中的数据表	28
3.5.1 数据表的分类及关系	28
3.5.2 部分数据表的内容形式	29
3.5.3 各个数据表之间的访问方法	33
3.5.4 数据表访问方法举例	35
3.6 自动分词技术	38
3.6.1 自动分词技术概述	38
3.6.2 自动分词技术种类	39
3.6.3 堆栈-最大匹配自动分词模型	39
3.6.4 堆栈-最大匹配自动分词算法	41
3.6.5 自动分词举例	44
3.7 本章小结	45
第四章 汉语语法语料库系统的实现	46
4.1 用户应用程序	46
4.2 语料库管理系统	51
4.3 本章小结	53
第五章 总结全文及今后工作的展望	54
5.1 总结全文	54
5.2 对今后工作的展望	54
参考文献	55
附录 1: 实词表及其标记	57
附录 2: 虚词表及其标记	59
附录 3: 短语表 I (按短语结构分类)	60
附录 4: 短语表 II (按短语功能分类)	63
附录 5: 句子成分列表	64
附录 6: 句型列表	66
附录 7: 复句类型列表	67
致 谢	70

第一章 引言

自然语言处理是当今人工智能中最活跃的领域之一。自然语言处理广泛应用下列多个领域：网络超容量文本数据的获取和分析；网络信息的纯洁化和安全处理；大型数据库自然语言查询；机器人语音对话；专家系统自然语言接口；CAD, CAI 和 OA 的人机交互系统；计算机自动书写、摘要提取；文档自动分类和文书管理系统；大型工业操作过程的自动化语言；机器翻译；模式识别(文字和语音)语言学再处理；话语自动翻译；文学与社会科学的文档和语料计算机自动处理；网际互联网信息过滤、主题识别、文本分类和文本挖掘；网上交叉语言和自然语言信息检索等等。

1.1 自然语言处理概述

自然语言处理 (NLP, Natural Language Processing), 也称自然语言理解或计算机语言学, 它是通过建立形式化的数学模型来分析、处理自然语言, 并在计算机上应用程序来实现分析和处理的过程, 从而达到通过计算机来模拟人的部分乃至全部语言能力的。自然语言处理扎根于计算机科学、语言学和数学等多学科。因此, 自然语言处理成为一门多学科之间的交叉学科。

从计算的角度来看语言的性质, 将语言作为研究的对象, 要求将人们对语言的结构规律认识以形式化的、可计算的方式呈现出来, 并且对一个语言片断中的语言单位进行识别, 对其结构和意义进行分析。

1.2 国内主要成果

我国的有关科研单位和专家, 从来没有停止过攻克中文信息处理难关的努力, 在国家的几个科学攻关计划中都列有信息处理项目。这些项目都是以解决计算机对自然语言进行理解问题, 也就是以开发智能型的汉语分析系统为奋斗目标。

通过 20 多年的不懈努力, 我国的自然语言处理的研究水平有了很大的进步, 并取得了丰硕的成果, 大体可以总结如下:

(1) 机器翻译: 以冯志伟教授为代表的计算语言学学者早期在机器翻译研究方面做了大量的工作, 并总结出了不少珍贵的经验和方法, 为后来的计算语言学研究奠定了基础。

(2) 语料库研究: 清华大学的黄昌宁教授领导的计算语言学实验室, 主要从事基于语料库的汉语理解。近年来, 在自动分词、自动建立知识库、自动生成句法规则、自动统计字词的使用和关联频率方面做了大量的工作并发表了不少很有价值的论文。

(3) 语篇理解研究: 东北工学院的姚天顺教授和哈尔滨工业大学的王开铸教授等在计算语言学的语篇理解方面的研究也取得了一定的成就。

(4) 受限汉语: 北京信息工程学院的周锡令教授主持的受限汉语的研究为自然语言理解提出的一种新的思路。他认为短期内计算机还很难做到真正的理解自然语言, 在继续对自然语言理解方面进行研究的同时, 应该研究受限的规范的汉语, 这样可以让研究成果较快的实用化。

(5) 知网: 由董振东先生提出的一种汉语知识表示方法。知网把客观世界看作是有很多的概念构成。概念与概念之间有各种各样的关系, 这些关系相互交织就构成了一个网。要表示一个客观世界, 就是要确定这些概念、概念的属性以及概念之间的关系。

(6) 概念层次网络(HNC): 由中科院声学所黄曾阳先生提出的一种自然语言理解的理论框架。这个理论框架是以语义表达为基础的, 它对语义的表达是概念化、层次化、网络化的, 所以称它为概念层次网络理论。该理论把认知结构分为局部和全局两类联想脉络, 认为对联想脉络的表述是语言深层(即语言的语义层面)的根本问题。这一理论的提出为语义处理开辟了一条新路。

1.3 自然语言处理的基本过程

现在的自然语言理解一般可以分为以下步骤: 原文输入、句子词语切分及词语属性特征标注、语法及句法分析、语义及语境分析、生成目标形式表示、语群及篇章理解等。

自然语言处理的核心技术是语言分析技术, 即将句子(数量无限)变换成由词语(数量可控)及其抽象形式(数量有限)构成的用某种数据结构(句法树、复杂

特征集或语义网络)表示的内部形式(数量有限)。

语言分析可以划分为词法分析、句法分析、语义分析、篇章分析等步骤。对象单元由小到大,从句子向篇章发展。实际上只有在篇章的范围内分析,省略、指代和句子的固有歧义等问题可能解决。

自然语言处理研究方式可按照下面五个基本过程进行:

- (1) 以特定的方式对自然语言的规律进行抽象,并且用形式化的语言描述有关自然语言的规律,通过计算机加工得到语言知识。
- (2) 针对特定的语言知识,根据其体现出来的所有语言现象进行归类,而后针对不同的类别进行分析并且研究相关的处理算法。
- (3) 按照自然语言的使用情况,分析不同类别的语言知识之间的关系,并且设计能描述其间关系的算法。
- (4) 根据算法编制计算机可执行的自然语言处理程序,这样的程序加上语言知识和计算机硬件系统,共同构成一个自然语言处理系统。
- (5) 用这样一个自然语言处理系统对自然语言进行分析处理,根据反馈的结果调整原来的设计。

1.4 自然语言处理的层次

语言学对语言的层次划分是这样的:

第一层	语音分析,即基本语言信号的构成
第二层	词法分析,即汉语中最小的可以独立运用的语言单位
第三层	句法分析,即词语的构成和组合的形式规律
第四层	语义分析,即语言表达的概念结构
第五层	语用分析,即语言与语言使用环境的相互作用。

表 1.1. 语言的层次结构

虽然这些层次之间并非是完全隔离的,但这种层次化的划分的确有助于更好地体现语言本身的构成,并且在一定程度上使得自然语言处理系统的模块化成为可能。

(1) 语音分析

在有声语言中,最小可独立的声音单元是音素,音素是一个或一组音,它

可与其他音素相区别。语音分析就是根据音位规则,从语音流中区分出一个个独立的音素。再根据音位形态规则找出一个个音节及其对应的词素或词。

(2) 词法分析

词法分析的主要目的是找出词汇的各个词素,从中获得语言学信息。

(3) 句法分析

句法分析是对句子和短语的结构进行分析。句法分析的最大单位是一个句子,分析的目的就是找出词、短语等的相互关系以及各自在句中的作用等,并以一种层次结构来加以表达。这种层次结构可以反映从属关系、直接成分关系,也可以是语法功能关系。自动句法分析的方法很多,有短语结构文法、格语法、扩充转移网络、功能语法等。

(4) 语义分析

理解语言的核心是理解语义。随着自然语言处理的发展,越来越多的研究者开始侧重于语义层的研究。句子是由词组成的,句子的意义与词义直接相关,但不等于词义的简单相加。因此,还应考虑句子的结构意义。语义分析就是要找出词义、结构意义及其结合意义,从而确定语言所表达的真正含义或概念。在自然语言处理中,语义愈来愈成为一个重要的研究内容。

(6) 语用分析

语用分析的任务是研究语言所存在的外界环境对语言使用所产生的影响。它描述语言的环境知识与语言使用者在某个给定语言环境中的关系,关注语用信息的自然语言处理系统更侧重于说话者/听话者模型的设定,而不是处理嵌入到给定话语中的结构信息。研究者们提出了很多语言环境的计算模型,描述说话者和他的通信目的,听话者及他对说话者信息的两组方式。构建这些模型的难点在于如何把自然语言处理的不同方面以及各种不确定的生理、心理、社会、文化等背景因素集中到一个完整的连贯的模型中。

1.5 语义分析和语法分析

语义分析是语言分析的一个分支,目的是根据上下文辨识一个多义词在指定句子中确切意义,然后根据该句子的句法结构和各词的词义推导出这个句子的句义,并用形式化的方式表达出来,从而使计算机能够根据这一表示进行推理。

语义分析是自然语言处理过程中的一个层次。从自然语言处理的应用来看,不管是信息获取、信息检索、机器翻译、自动文摘,还是人机交互,都要先对语言进行理解,确定语言所要表达的正确含义后,才能进行后续操作,并得到结果。从自然语言处理的发展来看,正是由于在实际应用中句法分析达不到令人满意的效果,研究者们才纷纷转向语义分析。

语义分析不易进行归约,但是语义分析的部分研究可以表现在语法分析层次方面,如词类的划分、词语的合理搭配、短语的构成、句子的成分划分等等。而这些方面在研究过程是比较形象具体的,比起语义分析更易于归纳。

汉语研究不同于英语。汉语只依靠语法分析并不能达到类似英语中应用语法分析的效果,可是语义分析又不能像语法分析那样很容易的进行归纳。因此,在目前自然语言处理研究的过程中依然要依靠语法分析在研究中的重要作用进行研究。如同 HNC 认为,自然语言无限的语句可以用有限的句类表示式来表达。“语句的宏观特性可以用语句的句类表示式来表达,语句的微观特性可以用语义块的构成表示式来表达。”

因此,在汉语的自然语言处理的研究的过程中,需要利用易被形式化的语法分析去归纳和表述不易被形式化的部分语义特征,同时需要对汉语言中出现的多种语法、语义现象进行统计。

1.6 自然语言处理中存在的困难

自然语言是人类在社会生活中发展出来的用来互相交际的声音符号系统,是人类历史长期发展而约定俗成的产物。

现在的计算机的智能还远远没有达到能够像人类那样理解自然语言的水平。因此,关于计算机对自然语言的理解一般是从实用的角度进行评判的。如果计算机实现了人机会话,或机器翻译,或自动文摘等语言信息处理功能,则认为计算机具备了自然语言理解的能力。

自然语言中充满歧义,在各个层次都含有巨大的不确定性。在语音和文字层次上,有一字多音、一音多字的问题;在词法和句法层次上,有词类词性、词边界、句法结构的不确定性问题;在语义和语用层次上,也有大量的由种种原因造成的内涵、外延、指代、言外之意的不确定性。

自然语言是极其复杂的符号系统,其结构复杂多样,语义表达千变万化。自然语言的语法结构和语义之间有着千丝万缕的、错综复杂的联系。一种结构可以有多种语义解释,而一种语义解释又可以由多种结构来表示。

自然语言的这些独特性和计算机使用的形式语言有很大的差异,因而应用计算机处理自然语言时无疑会遇到很多的困难。自然语言处理之所以存在困难是因为以下的原因:

(1) 目标表示的复杂性。如语义的概念依存网络表示,要从语句中提取这种表示的关键字就相当的复杂,同时还需要更多相关的客观世界的知识。

(2) 映射的类型。对于源语言到目标语言表示的映射,一对一类型是最理想的,但现实中,自然语言到目标语言表示的映射极难达到一对一的要求。

(3) 成分的交互程度。在语言中,每个语句都是由多个成分组成,若每个成分的映射与其它成分无关,那么映射过程就比较简单了。遗憾的是,自然语言中的成分交互程度相当高,句子中改变一个成分,常常会大大改变句子的整体结构,这使得映射的复杂程度大大增加。

1.7 课题背景

本文是属于教育部科学技术研究重点项目——机器人辅助打台球的机器人辅助教育过程中的一部分。在机器人辅助教育中,需要让机器人能够跟学生进行交流,就必须解决好人与计算机之间用语言交流各种信息的问题。对一个机器人来说,若想理解说话者语言的意思,这就涉及到自然语言处理的问题。

由于汉语本身的特点,在汉语言的研究过程中,常常需要研究深入到语义分析的研究中,而语义分析却又不易被形式化的语言所描述,这样就很难达到利用计算机处理的目的。因此,就要求在进行自然语言处理过程中,找出一种容易形式化的研究方法,用于描述汉语的种种语言现象。

无论是在自然语言处理过程中,还是在汉语言的研究过程中,无论是在某一专业知识领域,还是对所有学科的研究范围来说,汉语语法所描述的内容是整个汉语知识,所体现出来的语法内容也是。那么在自然语言处理中,面对不同专业领域的词汇,涉及到的词法分析和句法分析都是一致的。虽然词汇有专业类别的区分,但是词的语法规则却不存在专业领域差别。只要语言是汉语的形式,它就

符合汉语的句法构成规则、短语的构成规则、词语的构成规则以及词语搭配规则，而且这规则都是汉语语法研究的内容，存在于这个汉语言范围。

由于词语所体现的在语法分析层次上的研究工作是相同的，即都是使用汉语，而且相关的句式句型、短语构成、词语搭配等等，这些知识都是以当前汉语研究的成果为基础的。

综上所述，在汉语的自然语言处理过程中，需要建立起一个汉语语法研究的系统，同时配以大量的语言素材，用来研究汉语语法现象以及语法和语义分析的关系，形成汉语语法知识库。因此，围绕语料库的系统基础设计成为本文研究的主要内容。

1.8 本章小结

本章主要介绍了自然语言处理研究的内容以及国内的主要成果，概括性地归纳了自然语言处理的基本过程，总结了自然语言处理存在的不利因素。研究分析了汉语研究中语法分析和语义分析的关系，阐明了发挥两者各自的长处，结合两种层次的分析方法进行自然语言处理。最后，介绍了课题背景并引出本文研究的主要内容。

第二章 汉语言语法理论知识

汉语作为中华民族的语言有着上千年的悠久历史,经历几千年的变迁,汉语言依旧保持着它特有的风格。随着时代的进步,汉语也出现的许多的变化,一些具有当代特色的词、句子逐渐被人们所接受并且应用到实际生活中。面对中国经济的崛起,越来越多的国外友人开始认识到汉语的重要性,纷纷学习汉语,汉语热席卷全球。

汉语作为我们的母语,在我们的生活中屡屡出现,最寻常不过,但是真正的汉语言学研究并不是我们想象的那样容易,而且它本身就比较其它的语言研究要复杂的多。

2.1 汉语的特点

汉语作为四大文明古国的遗产之一,尤其经过上千年的不断变化,历朝历代对汉语言的研究和改进,同时不断与其它语言相互交织,最终形成了我们现在所见的汉语言,也称作现代汉语。

汉语属于汉藏语系,与印欧语系有着很大的差别。通过对汉语研究与汉语教学的比较,联系自然语言处理研究,发现汉语同其它语言的确有着很大的差异,且这些差异也成为汉语自然语言处理的不利因素,也是汉语的自然语言处理落后其它语言的原因。汉语言有以下特点:

(1) 汉语采用的汉字属象形文字,不像印欧语系那样是由较少的字母构成。国标码中有汉字 6763 个,假设汉语的词汇中全部是两个汉字组成一个词,那么理论上会有 P_{6763}^2 中词语构成情况,如果再加上成语这种常见的四字词情况,尽管去掉那些不恰当的词语,这个数字也是相当可观的。

(2) 汉语的词语间无间隔,不容易提取。汉语中仅仅使用标点符号将各个语义块区分开来,每个语义块中的多个词语却不易分开;而在英语中单词之间是以标点和空格将彼此分开。因此在汉语的自然语言处理中,分词成了首要问题。而且由于汉语本身有很多歧义现象存在,在分词过程中产生的歧义现象也成为汉语自然语言处理过程中的问题之一。

(3) 汉语的词缺少形态上的变化,这种变化是指表示语法意义的词类变化。一个词语其形态只有一种,如果将词用在不同的位置,所体现出来的词类意义也不同。例如“攻击是最有效的手段”和“敌人攻击那个村庄”,这里两个“攻击”形态完全一样,可是前者是名词性词语,后者是动词性词语。

而在英语中,却存在着在表示基本含义的词根上加可以体现词性的词缀,构成不同词性的单词,并且其外形也不一样。这就体现出英语词语形态的变化和语法的变化有一定的联系。这样首先从词语的形态上就可以很容易的判断词语所属的词类,进而确定其语义内容和词的用法。

由于英语单词在词语形态和语法上有一定的联系,所以在英语中还存在这样一种现象,不同词类间的搭配也比较确定。例如英语词法中,副词可以用于修饰形容词和动词,形容词修饰名词等等。在句子中也有类似情况,可以称为句子成分的构成情况。如名词可以做主语、宾语,动词作谓语等等。

如果将上述英语中的语法现象应用于汉语中,就显得有些不大合适。由于汉语缺少了形态上的变化,词类的划分就需要到一定语言环境中去考察,因此,那些涉及到词类的相关自然语言处理就变得有些困难。缺少形态的变化也使得词类之间的相互搭配关系、以及句子成分的构成情况就变得不确定了。

上面例子中的“攻击”一词,如果作名词讲,那么可以被形容词性的词或短语修饰,不能用副词来修饰;如果作动词讲,那么可以接受副词的修饰。联系作名词的情况考虑,那么就出现了形态相同的词语却由于词类而不能确定这个词是否能被形容词或副词修饰的情况。类似的,“攻击”一词可以做名词在句子中充当主语成分,由于“攻击”有动词词性,那么是否能说明动词也可以做主语呢?依此类推,那么对于其它的词语又有什么情况发生呢?因此,缺少形态变化这一特性的困扰着汉语自然语言处理研究。

结合第一章所述内容以及上述汉语的特点,对汉语的研究无论是人为工作还是计算机处理,都需要从语义的层面加以分析。这不意味着其它语言的天然语言处理就不需要涉及到语义分析这一层次,而是说,汉语的研究很大程度上是需要从更深层次的语义分析入手,同时需要可形式化的语法分析为语义分析服务。

2.2 语法概述

2.2.1 语法概念

语法是词的变化规律和组词成语的规律。语法学可以分为词法和句法两个部分。词法包括词的变化、词的构成（词语内部的构造规律）、词语的特点（词在语法方面表现出来的特点）、词的分类（词在语法方面划分出来的类别、词的用途（词在句子中担当的职务）；句法是指句子的组成规律、包括词组的构成（词语和词语互相结合的方式）、句子的分析（分析句子内部的关系）、句子的衔接（几个句子之间的关系）、句子的分类（从不同角度对句子进行分类）。

2.2.2 语法体系

语法体系就是由词的构成、词的变化、以及组词成句等现象和规律构成的语法现象和语法规律的有系统的整体¹。语法体系有两个含义：一个是指客观存在的语法事实、语法规律的系统性，每种语言的语法都是作为一个系统而客观存在的；另一个含义是语法学体系，主要是指语法学说的系统性。

2.2.3 结构层次

从语法分析的角度来说，必须有两个或两个以上的词素组合在一起，才能构成一个结构。语法结构是指两个或两个以上的语法单位的搭配和排列，即指一个词修饰另一个词，如：[[[多]数]派]。多种成分结构内部结合的先后次序叫做结构层次。有的时候一个结构的构成成分完全相同，但结构层次却可以不一样，而且由此表示出来的意义也不一样。例如“我们相信他是正确的”，这句话有两种结构层次，表示两种不同的意义，如图 2.1 所示。

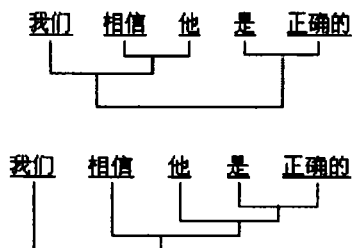


图 2.1 结构层次

¹ 吕冀平，《汉语语法基础》，商务印书馆，第6页。

因此, 辨清结构层次是语法分析的重要的一个环节。

2.2.4 语法性质

语法的性质与语音、词汇相比具有抽象性、稳固性、民族性。

(1) 语法的抽象性。语法不研究个别的、具体的词、短语、句子等内容, 而是从众多的语法单位的组合里抽象出其中的共同的组合方式或类型以及如何表达语义。语法指的是抽象出来的公式, 舍弃了个别的、具体的内容。“语法从词和句的个别和具体的东西中抽象出来, 把作为词的变化和用词造句的基础的一般的东西拿来, 并且以此构成语法规则、语法规律。”²

(2) 语法的稳固性。语法是一个由各种抽象规则交织成的有紧密联系的体系, 因此语法的变化比起语义、词汇来要缓慢得多。

(3) 语法的民族性。每种语言有明显的民族特点, 不仅表现在语音和词汇上, 同时也表现在语法上, 研究语法应当注意不同语言的共性和个性, 不能因为有共性而忽略了语法的民族特点。

2.2.5 语法与逻辑的关系

逻辑主要是指思维规律, 违背了逻辑就会做出有误的判断或推理, 必然有悖事理。

语法不是逻辑。逻辑是指思维规律, 而这种思维规律是全人类必须一致遵守的。而语法却是语言规律, 语言规律因民族不同而相异。

逻辑不是语法。但是逻辑问题在用语言方式表达出来的时候就反映在语法问题上。当一句话在逻辑上出现问题的时候, 通过对该句的语法分析可以很明确地说明逻辑问题发生的位置。

2.2.6 汉语语法结构

汉语语法结构理论知识包括很多的内容, 有词类、短语、句子成分、句类句型、复句、句群、标点符号。结合自然语言处理的特点, 本文选择词、短语、句

²斯大林, 《马克思主义和语言学问题》, 第 17 页, 人民出版社 1971.

子（包括句子成分、句型、复句等）这三个主要内容进行分析，之所以这样划分是因为：

（1）从汉语言研究的角度而言，这三个内容是构成应用汉语的基础，是应用汉语中常见的三种形式，尤其是句子应用最广，而词和短语又是句子的组成元素。

（2）短语的构成与句子的构成还是有一定的差别，而且短语构成的情况也比较复杂，需另当别论。结合词、短语、句子的类型和结构，可以看出词、短语、句子在各个方面以及各个内容上存在着多种的联系。

（3）从自然语言处理的角度而言，对汉语言来说，自然语言处理首先是对汉字的识别，进而涉及到词、短语的构成，最后形成句子，而且句子是一个完整意思的表达，三者在自然语言处理过程中是词在句子构成过程中的三个阶段的表现形式。在一定语言环境中，三者也可以同时存在。

因此，在本文的研究过程中，把短语单独作为一个语法元素考虑，与词、句子并存。

2.3 词法

词法主要研究词的变化、词的构成、词的特点、词的分类和词的用途。词语是最小的能够独立运用的语言单位，是构成短语和句子的备用单位，一部分词加上句调可以单独成句。

2.3.1 词类划分的原则

划分词类是汉语语法研究过程中的首要工作，词法中的很多内容和结构都与词类有关，因此它在自然语言处理研究过程中的作用也是非常重要的。

汉语言中，词类是只有出于说明语法规律的需要，按照词的意义和语法特点划分出来的类别才是词的语法分类，也称词类。划分词类的目的在于说明语句的结构和各类词的用法。

词类的划分有三大原则，可概括为以下几点：

（1）避免单纯按照词的词汇意义来划分词类。理论上单凭意义分类是逻辑领域的规则，如果单凭意义来划分词类，那么就会出现不同语言的词类应该相差无几或者是完全相同，因为，人们对概念上的认识是相同的，但是实际上不同语

言的分类并不相同。

(2) 坚持“词有定类，类有定词”说。仅仅凭借词汇意义划分词类还会导致另一个后果“词无定类，类无定词”，因为词与词相互结合的时候作为一个概念，其意义就有可能改变。

(3) 判断一个词具有什么样的语法特点，不应该在入句之后，而是应该在入句之前。

2.3.2 词类划分的依据

英语中，单词中就有明确的划分规则，每个单词作不同的词类使用时绝大多数有着不同的形式，从外形式上就很容易区分；另外，依照英语单词在句子中充当的成分也可以判断单词的词类归属。

汉语词语分类的依据有词的语法功能、形态和意义，其中最主要是词的语法功能。

(1) 词的语法功能是指词与词的组合能力。有三种表现：词在语句中充当句子成分的能力；一类实词与另一类实词的组合能力；虚词与实词的组合能力。

(2) 词的形态可分两种：一是指构形形态，如重叠；二是指构词形态，如词缀。

(3) 词的意义是指语法上同类词的概括意义或意义类别，主要指人或事物、动作、行为、性质、状态的。

语法功能、形态、意义三者是一个统一整体的不同表现。在运用分类标准时要注意分清主要和次要。汉语划分词类，语法功能是主要的，但是使用功能标准时必须分清主要、次要或者经常、非经常。因为汉语的实词大多是多功能的，即每类词大都能充当多种句子成分。例如，“批判”就是多功能词，即能作谓语，也能作主语、宾语、定语、状语。只因它作谓语、带宾语的用法是主要的、经常的，其它用法是次要的、有条件的，再加上它能作动词式的重叠和表示动作的意义，才把它认作动词。在同一大类里，各词的语法特征也有差异，例如说名词可以受数量短语修饰，但是这对某些名词来说不起作用，像“面目、年龄、现在、今年、东方”等就是。但是这些词跟名词的主要功能（作主语、宾语）相同，跟别的此类不同，其意义又表示事物，因其主要或多数特征相同，所以仍旧属于名

词。

2.3.3 词类划分及语法功能

汉语言中的词语大体可分类是词和虚词两大类,实虚词的划分以功能为主要依据。认为能够充当句子成分(有词汇意义、语法意义)的是实词;不能充当句子成分、只有语法意义的就是虚词。实词可分为名词、动词、形容词、区别词、数词、量词、副词、代词、拟声词、叹词(详见附录1)。虚词可分为介词、连词、助词、语气词(详见附录2)。

在词类的划分上,根据建立汉语语法语料库的实际需要,作者增加了二级词类的划分,主要是为了区别同一词类中体现不同含义类别的词语。

以下简要介绍各个词类及其语法功能。

(1) 名词

名词表示人或事物,包括表示时间、处所、方位的词在内。

名词前一般可以加上表示物量的数量短语,一般不能加副词;名词不能重叠;有的名词加“们”表示群体。

(2) 动词

动词表示动作、行为、心理活动或存在、变化、消失等。

动词能作谓语或谓语中心,多数能带宾语;动词能够前加副词“不”,多数不能加程度副词,只有表心理活动的动词和一些能断动词能加程度副词,如“很喜欢,很愿意”;一部分动词可以重叠,表示短暂。

(3) 形容词

形容词表示性质、状态。

形容词能作谓语或谓语中心语和定语,多数能够直接修饰名词,少数性质形容词能够直接修饰动词,作状语,如“快走”。

一部分形容词也能作补语,如“看清楚”。

性质形容词大都能受程度副词修饰,性质形容词的重叠式和状态形容词不受程度副词修饰。

形容词不能带宾语,有些双音节的性质形容词兼属动词,作动词时能带宾语;这些词前面加程度副词时是形容词,不能带宾语,后面带宾语时不能前加程度副

词。这些词兼属形容词和动词两类；有小部分性质形容词可以重叠；有小部分单音性质形容词可带上叠音词缀或其它词缀，如“红彤彤”。

（4）区别词

区别词表示事物的属性，有分类的作用，其属性往往有对立性质，因此区别词往往是成对或成组的。

区别词可以直接修饰名词作定语，但是不能作谓语、主语、宾语；不能前加“不”，区别词的否定式前加“非”。区别词与形容词的区别是：形容词可以充当定语、谓语、补语、和状语，能加“不”修饰；区别词只能充当“定语”，不能充当谓语、定语等，区别词也称作非谓形容词，不能加“不”。

（5）数词

数词表示数目和次序，可以分为基数词和序数词。基数词表示数目的多少，可分为系数（一、二、……）和位数（十、百、千、……）。序数词表示次序前后。

数词通常和量词组成数量短语，作定语、状语、补语。

（6）量词

量词表示计算单位，又叫作单位词。可分为物量词和动量词。物量词表示任何事物的单位。动量词表示动作行为的单位。

量词总出现在数词后组成数量短语，其中一部分单音节量词可以重叠。

（7）副词

副词常限制、修饰动词、形容词，表示程度、范围、时间（详细分类见附录1）。

副词可以作状语；大多数副词不能单说，只有“不、没有”等少部分副词可以单说；部分副词能兼有关联作用，如“又说又笑”。

表示时间的副词和表示时间的名词有时容易混淆，它们的区别是：两者都可以作状语，但是副词不能作主语、宾语、定语，而时间名词可以做上述成分。

（8）代词

代词有代替、指示作用。

代词跟所代替、所指示的语言单位的语法功能大致相当。

（9）拟声词

拟声词单纯描摹声音,给人一种能够如闻其声的效果。它有修辞的作用,给人以身临其境的感觉。

拟声词能作状语、定语、谓语、补语、独立语等,可以单独成句;它在作状语、定语、谓语时与形容词有相似之处;它也能作独立语,而且不受程度副词和否定副词修饰。

(10) 叹词

叹词独立性很强,一般不参加句子结构;常用作感叹语,如独立语;或者单用为句子,如感叹句。

(11) 介词

介词用在名词性词语前面组成介词短语,整体修饰谓词性词语,表示跟动作、形状、有关的时间、处所、方式、原因、目的、施事、受事、对象等。

介词短语不能单独做句子成分,它总要构成介词短语作状语成分,部分介词还可以构成介词短语作补语;介词短语不能单独作谓语或谓语中心,不能加动态助词或重叠。

(12) 连词

连词用于连接词、短语、分句和句子。

(13) 助词

助词附着在实词、短语或句子上面表示语法意义。按其表示的语法意义可分为结构助词、动态助词、比况助词、其他助词。

结构组词:“的、地、得”,主要表示附加成分和中心语之间的结构关系。

动态助词:“着、了、过”,主要表示动作或形状在变化过程中的情况。

比况助词:“似的、一样、(一)般”,附着在名词性、动词性、形容词性词语后面,表示比喻。

(14) 语气词

语气词在句尾表示种种语气,也可以在句中表示停顿。

2.4 句法

汉语句法中要研究句子内部的各种关系。它的研究范围包括短语和句子的结构规律和类型。

2.4.1 短语

短语是语义上和语法上都能进行搭配的、没有句调的一组词，是造句的备用单位，大多数短语可以加上句调成为句子。

短语可分为 12 种类型，具体分类情况和结构构成见附录 3。以下简要介绍各类短语的构成情况及语法特征。

主谓短语：主语在前，谓语在后，用语序和词类表明其间的陈述关系，而不用虚词表示。

动宾短语：动语在前，宾语在后，动宾之间的支配关系、关涉关系用语序表示，不用虚词表示。

偏正短语可分两类，分别是定中短语和状中短语。

定中短语：修饰语为定语，中心语一般是名词性成分，有时用“的”。

状中短语：修饰语为状语，中心语是动词性或形容词性成分，状语之后有时用“地”。

中补短语：由中心语和补语两个成分组成，补语附加在中心语的后面，其间是补充关系，有时补语前面有“得”。

联合短语：有语法地位平等的两个或几个部分组成，其间是联合关系，可细分为并列、顺承、递进、选择等关系。一般是同词性的词语相连，整体功能与部分的功能一致。

连谓短语：有多个谓词性成分连用，共用一个主语，谓词性成分之间没有语音停顿，只有连续关系，也不需要关联词语。

兼语短语：前一动词的宾语兼作后一动词或形容词的主语，也就是说，动宾短语的宾语和主谓短语的主语重叠。

同位短语：前后各部分的词语不同但所指的内容相同，语法地位一样。

方位短语：由方位词直接附在名词性或动词性词语后面组成，主要表示处所、范围或时间，具有名词性。方位短语常与介词连用构成介词短语。

量词短语：量词短语可分两类，数量短语和指量短语。数量短语，即数词加量词；指量短语，即指示代词加量词。

介词短语：由介词附着在名词等词语前面组成。介词短语可以作状语、少数可作补语。介词短语可以用来表示动作的工具、方式、因果、施事、受事、对象

等多种语义。

比况短语可分为三类：“的”字短语、比况短语和“所”字短语。

“的”字短语属名词性短语，由“的”附在是词或短语后面组成，指称人或事物，只能作主语或宾语。

比况短语属形容词性短语，由比况助词附在名词等词语后面组成，表示比喻和推测，可以作定语、状语、谓语、补语。

“所”字短语属名词性短语，由“所”字加在及物动词前面组成，即“所+及物动词”，用来指称动作所支配或关涉的对象，作定语、状语、谓语、补语。

短语的所有类型按照功能归类又可以分成名词性短语和谓词性短语，谓词性短语是指可以充当句子成分中的谓语的短语，在谓词性短语中又可细分为动词性短语和形容词性短语，详见附录 4。

2.4.2 句子成分

句子是具有一个句调、能够表达一个相对完整意思的语言单位。按照词对句子的“贡献”而划分出来的不同“职务”称作句子成分³。根据“职务”的差别，主要可以分为七种成分，分别是：主语、谓语、宾语、定语、状语、补语、独立语。有关句子成分的内容详见附录 5。

以下是关于多层定语和多层状语的次序排列的补充说明。

多层定语的次序依次为表示领属关系的词语、表示时间处所的词语、指示代词或量词短语、动词性词语和主谓短语、形容词性词语、表示性质或类别或范围的名词动词。

多层状语的次序依次为条件、时间、处所、范围、否定、程度、情态。

2.4.3 单句类型

单句是指从整体来看只有一个主谓词组的句子。句型是根据句子结构特点分出的类型。单句句型的划分依据主要是依靠句子成分中的主谓语来进行判断的。由主语和谓语构成的单句叫做主谓句；而分不出主语和谓语的称作非主谓句，有些非主谓句需要在一定的语境中才能独立成句，详见附录 6。

³ 吕冀平，《汉语语法基础》，商务印书馆，第 60 页。

在主谓句中,根据充当谓语的词性可以再分为名词谓语句、动词谓语句、形容词谓语句和主谓谓语句。

在非主谓句中,根据构成句子的词或短语的词性再分为名词性非主谓语句、动词性非主谓句、形容词性非主谓句和叹词句。

2.4.4 复句类型

复句是由两个或两个以上相关、结构上互不包含的分句组成。复句分为联合复句和偏正复句。联合复句内各分句之间意义上平等、无主从之分;偏正复句内各分句之间意义有主有从,意义是从属关系的。分类情况及相关的关联词详见附录7。

(1) 联合复句

联合复句共分五类。

并列关系:前后分句分别叙述或描写有关联的几件事情或同一事物的几个方面,分句之间是并举或对举的关系。

顺承关系:前后的分句按时间、空间或逻辑上的顺序说出连续的动作或相关的情况,分句之间有先后承接的关系。

解说关系:分句之间有解释或说明、总分的关系。

选择关系:有的分别说出两种或几种可能的情况,让从中选择;或者是说出选定其中的一种,而舍弃另一种。

递进关系:后面的分句意思比前面分句的意思更进一层,一般由轻到重、由小到大、由浅到深、由易到难,将上述的过程反过来同样成立。

(2) 偏正复句

偏正复句也分五类。

转折关系:前后分句的意思相反或相对,通常后面分句是说话人真正想表达的意思。

条件关系:偏句提出条件,正句表示在满足条件的情况下所产生的结果。

假设关系:偏句提出假设,正句表示假设实现后所产生的结果。

因果关系:偏句说出原因,正句表示结果。有表示说明和推论的关系。

目的关系:偏句表示行为,正句表示行为的目的。

2.5 词、短语、句子之间的关系

通过对汉语言语法结构知识的了解,汉语语法可以分为词法和句法两个研究范围。结合自然语言处理的特点,我们将这两个研究范围具体到汉语语法中的三个方面,即词、短语和句子。三者 in 构成方面有着复杂的关系,具体阐述如下。

2.5.1 词与短语的关系

短语是由多个词构成,短语类型的划分是以构成短语的多个词语的词类及其排列的次序(即结构层次)为标准的。因此,对于短语的组成,可以根据在语言环境中的位置可以初步判断其是否可以构成短语,并且根据构成短语的词类属性可以判断短语的类型。

2.5.2 词、短语与句子成分的关系

首先各类词、短语都是依据各类的不同语法特征以及含义充当句子中的多种成分,但是并非所有的词类和短语都可以充当句子中的任何成分,根据句子成分的结构特点和语法特点可以对句子成分的构成材料进行了分类。同时,在词类和短语的类型上也作了部分语法解释,就是根据词类和短语类型在功能上的划分可以判断词或短语可以充当句子中的哪些成分,而不能充当哪些成分。

2.5.3 词、短语与句型的关系

句型与词、短语的类型也有密切的关系。句型的划分是按照主语、谓语的存在与否以及后者的构成情况来区分的,涉及到构成谓语材料的属性问题。在主谓句中,谓语的构成情况成为主谓句各个类型划分的唯一标准。

2.5.4 词、短语与复句的关系

在词类中,一部分连词和副词充当复句中的关联词,一些简单的短语也充当复句中的关联词。这些词和短语成为了判断复句类型的主要依据,但是需要强调一点,这些词和短语并不是复句的划分的唯一标准。在少数句子中,需要依据语气上的停顿和句义上的完整性对一些复句行划分。

2.6 语法特征提取

同英语相比,由于汉语本身缺少形态上的变化,若想通过计算机处理对判断词类、短语类型及构成、句式类型以及句子成分构成,就缺少判断或处理的依据。如果以文章中的句子为单位,分别对其中的内容提取语法特征,并且将这些特征与相关的词、短语或句子联系起来,就可承担上述过程中理论依据的角色,成为对汉语语法知识自动处理的规则。

特征提取是汉语言研究过程非常有效的手段,也是语法语料库系统中的重要过程之一,通过特征提取可以在多个数据表之间建立起了相互访问的方法,特征提取成为数据表间的纽带。

特征提取的原则:特征提取的是语法中的要素;主要通过语法中元素相互构成的规则进行,个别时候需要人为的参与。

特征提取表现在以下几个方面:

(1) 词类。汉语语法研究主要依据词类作出分析和研究的,根据词类的不同来区分语法研究中其它语法现象的构成及分类。词类的特征因陈述的语言环境而变化,但这不是无规律的变化,有些词类的变化就很少或基本无变化例。如“北京”,这是名词中的专有地名,从词类的角度,它只能是名词,不会是其它词类。如有口语化的句子“这个人很北京”,这里使用副词“很”来修饰“北京”,汉语语法中并没有明确提出副词不能修饰名词,只是说名词前一般不能加副词⁴,像这样的词与词的彼此修饰关系就是少数了,而且有一定的时代性。

在汉语言的研究中,尤其是语法结构的研究中,词类成为较为重要的参考标志。但是由于汉语本身的特点,句子成分与词类之间没有非常确定的映射关系,因而在一定程度上只能起到参考的作用。如果想建立起较为确定的映着关系的话,那就需要对这两种语法现象进行统计,通过语义分析对词语进行多级分类。

(2) 词的搭配。词的搭配在汉语语义中体现得尤为重要。一句话能不能说得通、有没有错误、在体现的意义上有没有表述不当的地方,就主要体现在词语搭配这方面。通过考察词语间的正确搭配关系能反映出两个词之间的语法结构关系,也体现着彼此之间的结合意义。另外,词语的排列次序不同所表达的含义也不同。

⁴ 黄伯荣,廖旭东,《现代汉语》,高等教育出版社,第11页。

例如“张三打李四”，如果颠倒了顺序，变成了“李四打张三”，词语的顺序改变了，动作的施事方和受事方也同时发生变化，所表达的意思也就与先前的不一样了。这是因为动词“打”在用作“打击，捶打”的意思时，有施事方和受事方的区分。

因此，通过对词语搭配情况的分析，可以区分同一词语的不同含义或者所属的不同词类。

(3) 短语类型。短语是由词语组成的，短语的组成过程就包含着词的搭配的问题。而且短语又可以充当句子成分，句子成分之间也有彼此修饰的关系。通过对短语分类及其构成情况（详见附录3）的分析，我们发现不同类型的短语在构成材料上有所差别，也正是因为构成材料的不同以及排列的次序有所差异，才分成了不同类型的短语。

实际上，各个类型短语的构成材料彼此也有部分相同之处，例如，主谓短语“今天星期三”，其构成材料是“名·名”；同位短语“首都北京”，其构成材料同样是“名·名”。从构成的词类来看完全相同，但是类型却不同。在语义方面，这四个名词的含义不一样，依据彼此间含义的涉及方面分别可以组成两个短语，根据表达的意思、陈述的结构分为两类。

这里需要提出一个问题：四个词的词类相同，但是二级词类有所不同。根据实词表，“今天”和“星期三”是表示时间的名词；“首都”可作为普通处所名词，“北京”是专有地名名词。这样看来，似乎可以这样认为，因为从意思类别上四个词分别两两二级词类相同和相近，这样就构成了搭配，分别组成了两个短语，分属两种类型。但是由于汉语的很多语言现象都是根据其语义产生出来的，所以不能贸然下结论，需要依据本系统的后续的工作，对大量的汉语文章进行统计，才能作出较为合理的判断。同时指出，由于二级词类的相同或相近而产生搭配的这种假设，也是将来自然语言处理的目标，到那时可以通过多级词类划分以及彼此之间的关系，能够较为准确地形成词语的合理搭配。

(4) 单复句的判断。单句和复句的有很大的一部分可以根据复句的关联词来判断，但是，也不是绝对的，因为在许多单句中会出现一些关联词，这些关联词与复句的关联词相同。例如有这样一个单句“这些钱只够买一瓶水或一袋瓜子”，有“或”这个词；可是，在选择关系的复句关联词当中也有这个词（见附

录 7),但是它却不是复句。

(5) 句子成分。句子成分的特征提取在句子的分析中是比较重要的,因为句式类型、各句子成分的构成情况、还有根据句子成分的彼此修饰关系所产生的词的搭配、短语的搭配等等,都需要从句子成分的语法特征中提取出来。实际上,从词的搭配、短语的搭配中提取语法特征反过来也能部分或全部影响句子的成分的判断,双方是相辅相成的。

2.7 语法知识在语义分析中的应用

由于汉语自身特点,汉语句子的语义分析对理解句子所起的作用比其它的语言要大,因此在语法分析中常常需要考虑语义分析。每个词、每个短语的具体含义都是在具体的句子中才能有所体现,不仅如此,在句中还体现词或短语间的结合意义和结构意义,这些与句子成份的判断相互影响

一个词能与另一个词形成搭配,能够形成彼此修饰关系,是因为两个词语在语义上可以形成合理和完整的意思,能表达出一种合理的概念,能被人理解。所以,在语义分析中,考察词语的搭配是语义分析的方法之一。

词语的搭配包括两种类型:一种是几个词语间的搭配构成短语;另一个是根据句子成分的修饰关系转化而来的搭配关系。

在词和词构成短语的过程中,可以通过短语类型或短语的构成情况,来确定构成短的两个词的彼此修饰关系。例如,由主谓短语“粮食丰收”,因为类型是主谓短语,又因为这个短语是由两个词语构成,根据主谓短语的构成规则就可以很容易的判断这两个词的修饰关系,那么就是“粮食”作主语,“丰收”作谓语。同时,这也体现出这两个词语在这种搭配中的词语含义。如果语序颠倒,那么短语的意思、类型等也随之变化。

在句子中,词、短语按照一定的次序组成句子,在六个主要句子成分中(除独立语外),根据语法结构的理论,我们得知定语通常是用来修饰主语或宾语的,那么什么样的材料构成的定语可以修饰什么样内容的主语或宾语呢?例如,“中华民族有着悠久的历史”这句话中,句子成分可以这样划分:

[中华][民族][有]着[悠久]的[历史]。

[定语][主语][谓] [定语] [宾语]

在这个句子中“中华”用来修饰“民族”，“悠久”这个形容词作定语来修饰“历史”。如果说“中华历史”，从语义上说仍然成立，也存在这样的修饰关系。可是如果用“悠久”直接作定语修饰“民族”，即形成“悠久民族”或“悠久的民族”。前者“中华历史”的搭配从含义上看没什么错误；而后者的搭配从语义上说似乎有些牵强，应当放到具体的语言环境中去考察正误。比如，我们通常说“历史悠久的民族”，在这里，“悠久”实际上修饰的是“历史”，两个词语结合在一起构成主谓短语来修饰“民族”一词。

再有，这句话的主语“民族”和宾语“历史”不能对调，不能说成是“悠久的历史有着中华民族”，也就是说“历史”不能拥有“民族”，这从含义上是说不通的。那么就说明，主语“民族”、谓语“有”和宾语“历史”在含义上存在从属关系，相互之间有一定的顺序不能颠倒。

实际上，主语和谓语也可以认为是彼此修饰的关系，谓语用于描述主语的动作、行为、性质、状态等等。主语是谓语行为的施事方、受事方或是当事方，能作主语的词或短语和能作谓语的词或短语构成的关系也是一种搭配关系。

综上所述，汉语的词语的搭配关系，影响着短语、句子的构成；对于一个表述正确的句子，通过对句子成分的划分可以自动反映出词语或短语间的搭配情况，而词语或短语搭配情况的研究，也是语义分析中的一个方面。

2.8 本章小结

本章详细的论述了建立汉语语法结构语料库所需要的语法知识以及语法基本要素之间的构成规律，结合自然语言处理的特点将汉语语法知识分为三个方面，即词、短语、句子，并描述了词、短语、句子三者之间联系。提出了语法特征提取这一概念并阐述了特征提取的内容和作用。本章还通过词语搭配的问题，描述了语法知识在语义分析中的体现，从中强调了语法分析是语义分析的重要手段。

第三章 汉语语法语料库的结构设计及相关技术

语料库(Corpus 或 Corpora)是按照一定的语言学原则,收集自然出现的连续的语言运用文本或话语片段,而建成的具有一定容量的大型文库。从本质上讲,语料库实际上是通过自然语言采集,获取大量的语言样本,用以进行统计性的研究的语言运用总体。

3.1 语料库历史

语料库作为一种语言学的研究方法之一,早在十八世纪就在欧洲得到了应用。由于生产力的关系,当时的语料库大多以手工方法收集,其索引以及分析过程也都是通过手工进行的,极为耗时费力。语料库的研究方法在进入十九世纪之后,在语言学研究中继续得到运用。基于语料库的研究主要集中在词典编纂和语法研究方面,许多学者将他们的研究建立在手工收集的语言材料上面,这些语料都以引用卡片的形式手工收集、整理、存放和利用。

1957年,美国语言学家 Chomsky 的《句法结构》一书出版,掀起了一场对传统的描写语言学的革命,理性主义的研究方法逐渐在语言学研究中占统治地位。作为一种经验主义的研究方法,传统的基于语料库的研究开始进入低谷。

二十世纪八十年代中期是语料库研究的复兴时期。现代语料库是指以电子文档为主要构成的应用数据库访问及存储技术的大型计算机语料库。最近三十年中,随着计算机技术的飞速发展,自然语言处理的研究深入,基于现代语料库的研究再次兴起。机器的存储量越来越大,运算速度越来越快,这样的客观条件使得大容量的机器可读语料库的建设成为可能。同时,一些新的、更好的统计语言模型也开始出现。而且,随着自然语言处理系统的不断实用化,知识获取问题已经成为一个瓶颈,基于规则的自然语言处理系统在处理大规模的真实文本中遇到种种困难,尤其是在汉语的自然语言处理研究中。因此督促广大研究人员去探索和采用一种新的研究思想。所有这些因素,推动了基于语料库的研究方法。语料库越来越多地应用到机器翻译、语音识别以及信息检索等应用研究中去。在自然语言处理领域,语料库的建设和利用具有越来越重要的意义。

3.2 语料的采集

在语料库构建过程中,我们首先应该考虑的就是语料的采集问题。语料库的建设是一项工作量极大的工作。这是因为,构建一个有实际应用价值的语料库要求其中的内容有一定的代表性,并且需要体现这种类型语料库的特点。

语料的采集应该遵循以下几个原则:

- (1) 样本选取和代表性
- (2) 有限的容量
- (3) 信息的时效性

在建设语料库的同时,我们还需要考虑两个因素:平衡和代表性。理论上,语料的采集应该完全涵盖给定语言的所有特性,但是,由于自然语言是无穷的,因此需要斟酌语料的平衡,旨在采集的语料信息具有最大限度的代表性。语料库的平衡与适当的取样范围有着密切联系。对一个面向综合性语言研究的语料库,它的语料数据常常来得范围都非常广泛。

由于不可能对自然语言中进行全部的取材,那么就需要语料数据有较强的代表性,无论是对常见的还是罕见的语言现象都需要一定数量信息来反映他们在整个自然语言中的分布情况。

在实际设计一个语料库时,还应当考虑语料库的实际应用。根据实际应用情况以及对后续研究的展望,设计出符合实际应用的相关功能,使得整个语料库系统更具有实际应用和开发的价值。

就本文研究的内容来讲,是以汉语语法为建立语料库的核心内容的。这是由于相关汉语语法的研究可以很好服务于汉语语义的研究过程中,可以形象地、具体的反映汉语的语言现象,能够用形式化的表达方式被计算机接纳处理。同时,汉语语法的研究是汉语言研究中的有效手段,也是自然语言处理过程中的重要工具之一。因此,我们需要在汉语语法的框架之下,将语法内容进行分类,并且按照各个语法要素间的相互关系建立连接。

3.3 语料库的结构

根据汉语语法的特点,结合自然语言处理过程,本系统语料库的结构设计如图 3.1 所示。

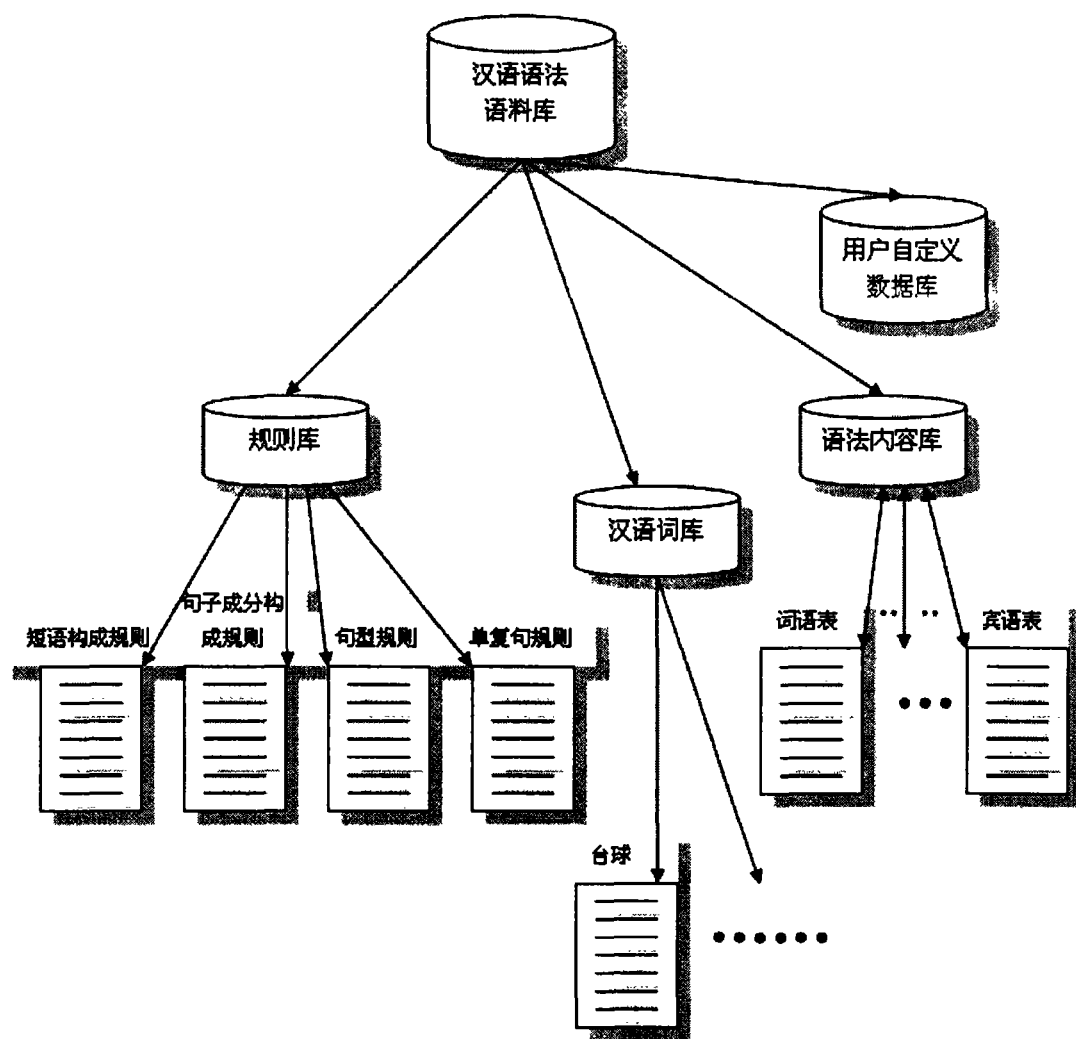


图 3.1 语料库的结构设计

整个语料库按数据表的类别又可以分为规则库、汉语词库、语法内容库和用户自定义库。其中规则库主要用于描述语法现象中语言单位的结构和构成规则；汉语词库主要存放分词后的词语内容以及词义、搭配等等，单个词语表通过语法内容库中的“词语表”作为索引；另外，作者还根据今后工作和程序开发的需要，设置了用户自定义数据库。

3.4 规则库的建立

在语料库的构建过程中，语料的采集尤为重要，而语料的输入也是相当繁重的，为了简化输入的过程，提高整个系统的运行效率，本系统特地设计规则库，根据汉语语法的规则辅助语料输入时的语法特征的提取。同时，在语料库构建之后，也可以作为语法的规则集，允许将未纳入规则的语法现象加入到规则库中。

另外,规则库在应用语法语料库时,为用户提供所查内容的详细的语法结构知识信息。

汉语语法结构规则库主要有两个过程构成。一个过程是词构成短语的过程;另一个过程是词或短语构成句子成分的过程。其中第一个过程的规则相对容易一些,词构成短语的形式(参见附录3),而第二个过程(参见附录5、6、7),从整体上看由于句子成分的构成材料与词类、短语类型有一定的联系,但却不是较为确定的映射关系,因此带有许多不确定性,不过在复句类型划分和句型的划分过程中,句子与词的内容、词类、短语类型的联系还是比较确定的。表3.1描述的是短语构成规则的数据表形式的内容形式。

项目	成员
主谓1	名+名
主谓2	名+动
...
联合1	既A又B
联合2	A和B
...	...

表 3.1 规则表

3.5 语料库中的数据表

3.5.1 数据表的分类及关系

汉语语法结构语料库是根据汉语语法知识构建成的,而相应的语法知识以数据表的形式构成。根据本语料库系统的需要,将汉语语法知识大致分了三个方面,即词、短语、句子三个汉语语法的基本元素。围绕这三个方面的内容,构成了多个汉语语法基本元素的数据表,具体的数据表名称及其之间的关系如图3.2。

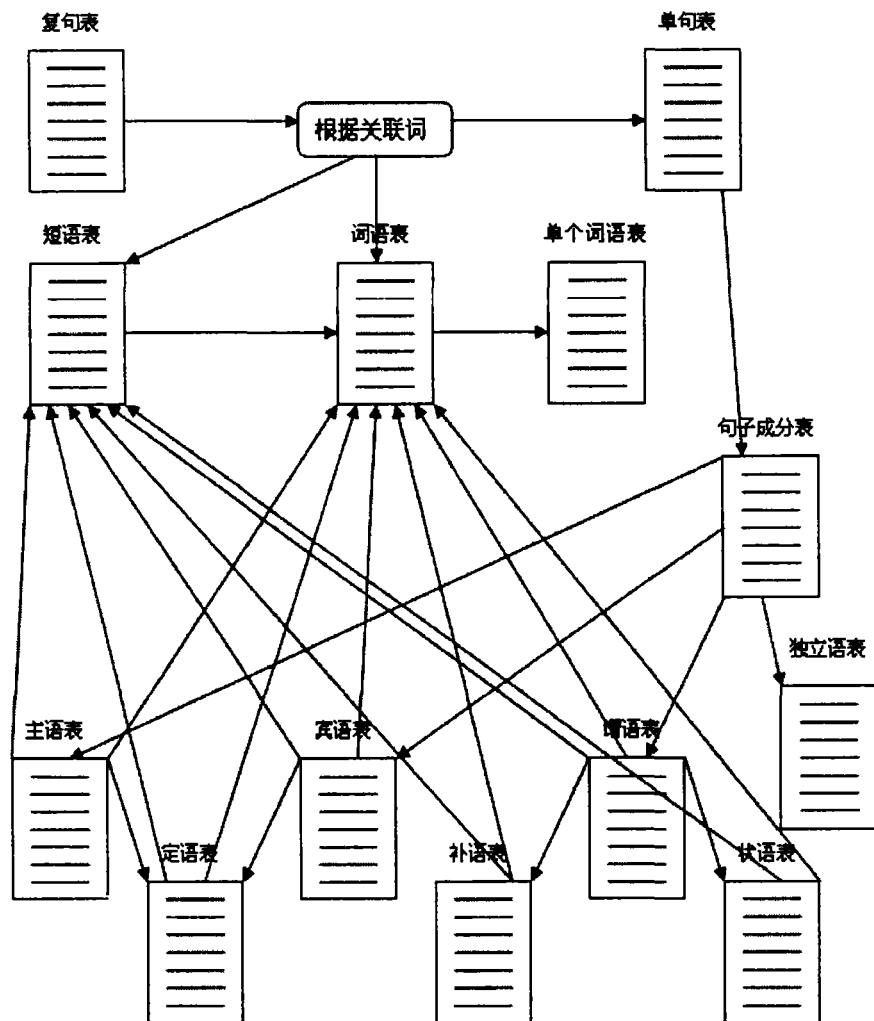


图 3.2 数据表关系图

应当指出的是，在句子成分表中，只包含了四个句子成分，即主语、谓语、宾语、独立语。之所以这样安排，是因为其他的三个句子成分，即定语、状语、补语同前三种成分是修饰的关系，或者说是依存的关系，没有主语、谓语、宾语也就没有定语、状语、补语。当然有些时候未必有了主语、谓语、宾语就一定有定语、状语、补语。而这些数据表纷纷与词语表、短语表构成关系，这是因为，具体分析到某一句子成分的构成情况时，那么只有短语和词这两种情况。而涉及到具体的短语和词的类型、以及相关的词的含义，就需要到短语表和词语表中考察，甚至到具体的单个词语表中去研究和分析。

3.5.2 部分数据表的内容形式

由于涉及汉语语法知识比较多，因此在整个系统中需要设置多个数据表，用

于存放不同内容的汉语语法知识。以下介绍部分数据表的内容形式。

(1) 词语表

词语表是本系统最重要的数据表之一,这是因为词语是汉语语法中最小的独立运用的语言单位,且短语的构成元素和句子的构成元素是从词语表中而来。词语表主要担当的责任是负责短语构成、句子成分中的词语的索引,同时也是单个词语表的索引表。它的内容如表 3.2 所示。

项目内容	成员
词语编号	Id
词语内容	Content
词性	Type
出现频率	Freq

表 3.2 词语表

其中词语编号的构成方式是采用“汉字编号+汉字编号+…”的形式。汉字编号的计算过程是这样的⁵:

计算机中的中文字符是用 ASCII 代码表示的,每一个汉字字符(包括中文标点)占用两个字节的 ASCII 值。国标码汉字的编码空间是 B0A1~F7FE,共有 $72 \times 94 = 6768$ 个汉字,实际有 6763 个汉字。两个字节的序号取值范围分别是,第一个字节 176~247,第二个字节 161~254。

因此有下列的汉字编号计算公式:

$$id = (c_1 - 176) \times 94 + (c_2 - 161)$$

其中, id 是汉字在国标码中对应的下标,也是本文中提到的汉字编码; c_1 和 c_2 分别代表两个字节的 ASCII 码。例如,汉字“安”的两个字节分别是[176, 178],通过上述公式的计算就得出其对应的数组下标为 17,也是“安”的汉字编号。由于国标码中的汉字上千,为方便起见,将编号“17”转化成“0017”。同理,汉字“排”的两个字节分别是[197, 197],其编号为“2031”。那么,词语“安排”的编号就为“0017+2031”,其它的单字词或多字词的编号与此类似。

(2) 单个词语表

单个词语表在语料库中是以单个词语的内容作为数据表名称的,依靠词语表

⁵ 引自陈小荷《现代汉语自动分析——Visual C++》,第 51 页。

作索引的,它详细记载了这个词语在所有出现的句子中的使用情况。表中包括它的不同含义,同时还记录了与该词语的前后的搭配或修饰关系。它的内容如表 3.3 所示。

项目内容	成员
含义种类	Id
词性	Type
含义内容	Meancontent
出处	Location
充当角色	Part
前搭配	Lastword
后搭配	Nextword

表 3.3 单个词语表

“充当角色”用来表明该词语时作句子中的某一成分,或者是某短语中的一部分。“前/后搭配词”主要是依靠“充当角色”的具体内容,根据句子成分的彼此修饰规则或者是短语的构成规则,找到另一个修饰或被修饰的内容,并以此作为“搭配”一项的内容。

(3) 短语表

短语表中存放着已录入文章中出现的所有短语,以及短语所属的类型,另外还存有该短语在句中所充当的角色,据此,可对不同类型的短语所作的成分进行统计。它的内容如表 3.4 所示。

项目	类中成员
出处	Location
内容	Content
短语类型	Type
构成材料	Compose
充当角色	Part
前搭配	Lastword
后搭配	Nextword

表 3.4 短语表

其中的“充当角色”、“前/后搭配”等内容与单个词语表的相同。

(4) 整句表

整句表, 主要用于存放录入文章中的所有句子, 也就是说, 以句义终结符(如句号、叹号等)为界定的句子。在数据表中可以通过“句式类型”的具体内容看出是属于单句还是复句。复句的类型可以通过判断关联词是否出现作初步判断。它的内容如表 3.5 所示:

项目	类中成员
出处	Location
内容	Content
句式类型	Snttype

表 3.5 整句表

(5) 主语表

主语表中存放着所有充当句子主语成分的内容, 主语表中各数据彼此之间的区别就是主语内容的出处, 如果是同一复句中的几个主语也可以根据分句号不同来加以区分。另外, 宾语表、状语表等与主语表的内容大致相同, 在此不再赘述。它的内容如表 3.6 所示。

项目	类中成员
出处	Location
主语内容	Subject
构成材料	Type
定语	Attribute
谓语	Predicate

表 3.6 主语表

(6) 谓语表

谓语表示所有表中内容较为繁多的一个表。因为它牵扯的句子成分最多, 以它为研究的方面也是最多的。作为谓语, 向前接有主语, 向后接有宾语, 状语和补语也是修饰它的。再有, 句式类型中主谓句中的类型划分也是根据谓语而来的, 所以在所有数据表中, 尤其是体现句法的研究过程中, 谓语表是非常重要。它的内容如 3.7 所示。

项目	类中成员
出处	Location
谓语	Predicate
中心词	Keycontent
构成材料	Type
主语	Subject
宾语	Object
主谓关系	Relation
状语	Adverbial
补语	Complement

表 3.7 谓语表

3.5.3 各个数据表之间的访问方法

在拥有众多数据表的语料库中,如何将多个数据表按照汉语语法的要求形成一个的体系尤为重要。除了用户可以自定义的一些数据表外,绝大多数的数据表在语料库构建的过程中就需要确定各个数据表之间的关系以及访问的方法。

汉语中,词是最小的可以独立运用的语言单位,因此,从词的语法特征提取入手,根据提取的内容,对应相关的规则库,在多个数据表之间建立联系。从词的语法特征提取开始的过程如图 3.3 所示。

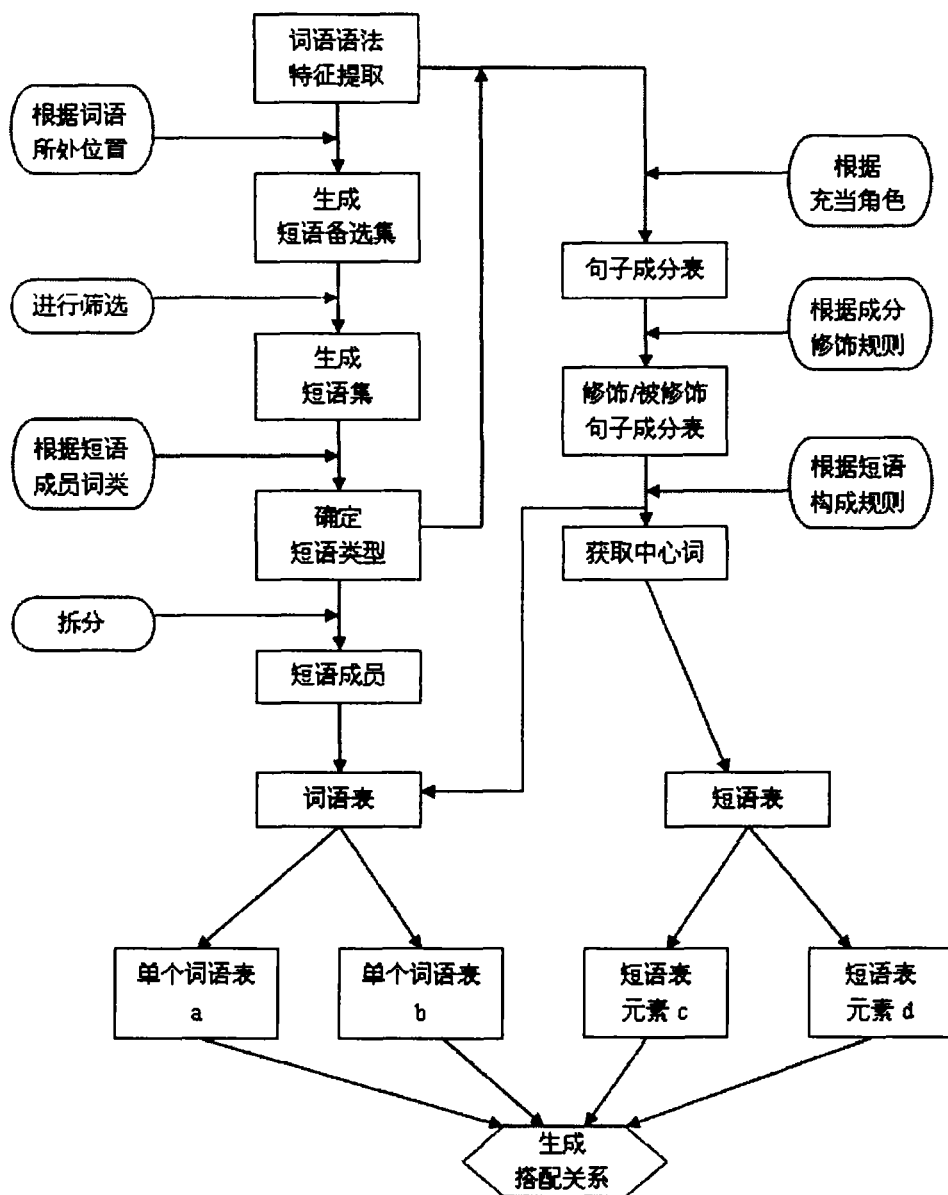


图 3.3 主要数据表的访问过程

在上图中，“具体的某一句子成分表”指的是主语、谓语、宾语表中的某一个。由于句子成分和短语的构成有词和短语两种情况，因此，最后都需要到某单个词语表或短语表中实现词语搭配、短语搭配的确定。

根据上图于是就有如下的数据表之间的访问方法，涉及到词语表、短语表、句子表、七个句子成分表和规则库，共八个步骤：

- (1) 词语语法特征提取，特征内容有词类、充当角色；
- (2) 设有词语 word，提取特征内容词类 word.type 和充当角色 word.part，根据词语的充当角色进行分析；

(3) 如果充当角色的内容是“短语”，则在 block 语义块中找与该词语邻近的词，即前一个词 $\text{lastword.location} = \text{word.location} + \text{lastword.Length}$ ；后一个词 $\text{nextword.location} = \text{word.location} + \text{word.Length}$ ；

(4) 检查这个词的“充当角色”这一内容，如果也是“短语”，则将这些词语按照顺序构成短语，“lastword + word”或“word + nextword”，将这些短语存入短语备选表中；

(5) 在短语备选项中筛选出正确的短语，根据构成短语各个成员的词类 type，结合短语构成规则库中的构成规则确定该短语类型的范围；

(6) 根据构成短语的成员，在短语表中查找短语，或者通过词语表找到具体的单个词语，根据搭配关系，填入到单个词语表中的“搭配”一项中；

(7) 如果“充当角色”的内容是某一具体的句子成分，根据句子成分的内容确定其修饰或被修饰的关系，随即确定另一个句子成分，根据词语的“出处”，在句子成分表中找到相应内容，形成修饰的关系；

(8) 如果修饰双方都是词，那么就分别在这两个词的表中的“搭配”一项分别填入对方的词语；如果彼此修饰双方只有一方是词，或者都是短语，则先找出短语中的中心词，如果有中心词就以中心词为主，形成搭配的关系，将中心词填入到“搭配”一项中，如果是没有中心词的短语，则将整个短语填入到“搭配”中。

这种访问方法是在构建语法结构语料库的过程中所采用的，可以在语料采集的过程中将各个语法知识的内容根据多个数据表之间的联系，填入到相应的数据表中。

语料库中众多的数据表根据语法构成规则形成彼此交错的复杂联系，如图 3.2。上述的方法只是构建数据表之间联系的最主要的一种，其它的数据表之间访问过程及联系方法在此不作赘述。

3.5.4 数据表访问方法举例

根据上述数据表的访问过程，下面以一句话为例介绍语法特征在多个数据表中的传递过程。有这样一句话“目标球很难落袋”，通过自动分词后，对该句中的词语进行特征提取，在单个词语表中内容如图 3.4 所示（数据表中只包括与数

据表访问相关的内容，下同)。

“目标”	内容	“球”	内容	“很”	内容
词性	N2	词性	N2	词性	D1
出处	0	出处	2	出处	3
充当角色	短语	充当角色	短语	充当角色	短语
前搭配	空	前搭配	空	前搭配	空
后搭配	空	后搭配	空	后搭配	空

“难”	内容	“落”	内容	“袋”	内容
词性	A1	词性	V1	词性	N2
出处	4	出处	5	出处	6
充当角色	短语	充当角色	谓语	充当角色	宾语
前搭配	空	前搭配	空	前搭配	空
后搭配	空	后搭配	空	后搭配	空

图 3.4 各单个词语表中的具体内容

根据词语的“出处”内容的连续性以及“充当成分”，生成以下的短语备选集，如图 3.5 所示。

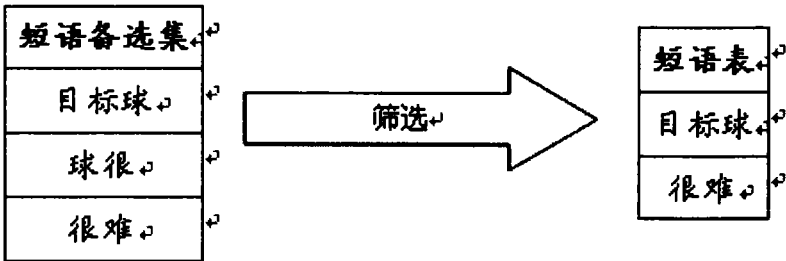


图 3.5 筛选生成短语表

那么去掉不合适的短语“球很”，就剩下两个短语，根据短语构成规则，分析短语备选集中的短语构成情况，并且由此确定短语类型范围，如表 3.8 所示。

短语备选集	构成情况	短语类型
目标球	N2+N2	主谓、偏正、同位
很难	D1+A1	状中

表 3.8 短语备选集中的内容

随即确定两个短语的类型，并确定这两个短语的充当角色。此时，结合单个词语表，整个句子的句子成分就都被确定了。表 3.9 是短语表的内容截取。

出处	内容	短语类型	充当角色	前搭配	后搭配
0	目标球	偏正	主语	空	空
3	很难	状中	状语	空	空

表 3.9 短语表中的部分内容

当短语表的内容完成后，根据短语表修改单个词语表的搭配一项；同时根据单句表的中主谓宾的内容，并且依据规则库中句子修饰成分的规则，找到各个成分，填入搭配的内容，最终完成有关该句的短语表和多个单个词语表的内容的获取，如图 3.6、表 3.10 所示。

“目标”	内容
词性	N2
出处	0
充当角色	短语
前搭配	空
后搭配	+球

“球”	内容
词性	N2
出处	2
充当角色	短语
前搭配	目标+
后搭配	+落

“很”	内容
词性	D1
出处	3
充当角色	短语
前搭配	空
后搭配	+难

“难”	内容
词性	A1
出处	4
充当角色	短语
前搭配	很+
后搭配	+落

“落”	内容
词性	V1
出处	5
充当角色	谓语
前搭配	目标球+
后搭配	+袋

“袋”	内容
词性	N2
出处	6
充当角色	宾语
前搭配	+落
后搭配	空

图 3.6 产生搭配后的各单个词语表

短语表	出处	内容	短语类型	充当角色	前搭配	后搭配
短语 i	0	目标球	偏正	主语	空	+落
短语 i+1	3	很难	状中	状语	空	+落

表 3.10 生成搭配后的短语表部分内容

主语表和谓语表的内容分别如表 3.11、表 3.12 所示。

主语表项目	成员
出处	0
主语内容	目标球
构成材料	偏正短语
定语	空
谓语	落

表 3.11 主语表的内容

谓语表项目	成员
出处	5
谓语	落
构成材料	V1
主语	目标球
宾语	袋
主谓关系	施事
状语	很难
补语	空

表 3.12 谓语表的内容

实际上,如果将“目标”认为是句子的定语也是可行的,词语间的搭配关系完全一样。

3.6 自动分词技术

3.6.1 自动分词技术概述

在书面汉语中,字与字、词与词是连写的,词在句中并没有明显的区分标记。理解汉语的首要任务就是把连续的汉字字符串分割成词的序列,将一个没有间隔标志的汉字字符串序列转化为词串序列,这就是自动分词。

词是汉语中最小的能够独立运用的语言单位,是构成短语和句子的备用单位,同时,在自然语言处理中也是汉语语法所考察的最小单位⁶,也是构成短语、句子的基本元素。因此在自然语言处理过程中,分词技术必不可少。目前中文分词技术主要是借用一个词库,按词库收录的词语进行分词。

在汉语语法语料库的构建过程中自动分词也是一个重要的环节。在将采集的语料输入到语料库的过程中,如果采用自动分词技术,首先将语料信息进行自动分词,为随后单个词语的语法知识获取作了必要的准备工作,大大节省了构建语料库的时间,提高了工作效率。自动分词也是语料库系统的应用功能之一。

⁶ 汉语语法中最小的考察单位是语素。上述说法是在本文研究的系统中考虑的。

3.6.2 自动分词技术种类

本节以机械分词方法为主简要介绍一下自动分词技术的种类。

机械分词方法的思路是先查词库进行匹配,然后再适当利用部分词法规则进行歧义校正。机械分词法之所以称之为“机械”,是因为它的切分过程是依赖于词库进行。词库中词条的数目、词条的选择直接影响到最后的分词效果。机械分词法加歧义校正属于机械分词法的一种改进,它主要利用词法规则对歧义进行校正,以提高切分精度,事实证明这种改进是有效的,而且这种改进最终导致了知识分词方法的出现。

机械分词法主要有最大匹配法、逆向最大匹配法、逐词遍历法、设立切分标志法、最佳匹配法、有穷多层次列举法、部件词典法、二次扫描法、高频优先分词法、双向扫描法等。

3.6.3 堆栈-最大匹配自动分词模型

堆栈-最大匹配自动分词技术主要是结合最大匹配法(MM法)和堆栈技术对文章中的词进行自动切分,是对最大匹配法的改进。

最大匹配法的主要思想是在词库中检索与原句中能够匹配成功的字符长度最长的那个词作为分词的结果。其流程图如图3.7所示。

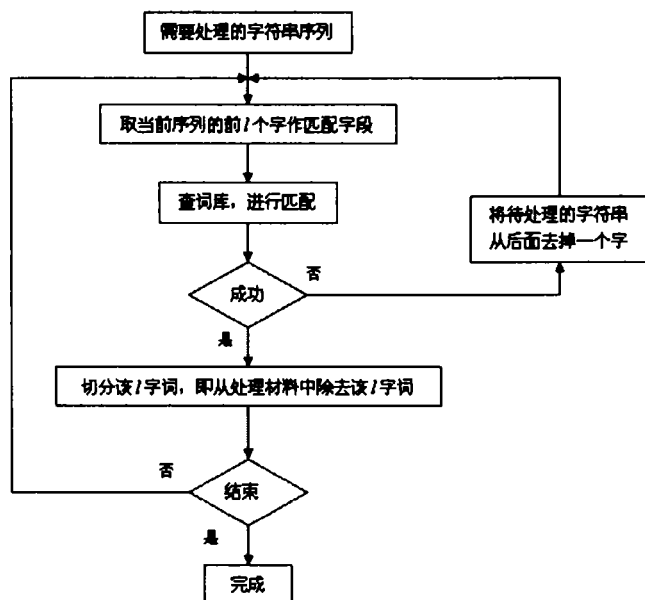


图 3.7 最大匹配法流程图

最大匹配法重视的是字符长度,如果遇到在分词过程中后面字符串出现不可

分的情况,能自动弹栈回退,并且重新检索出另一个成功匹配的词作为分词结果,就有可能解决后面字符串不可分的窘境。其基本设计思想是:

首先按照文章中的标点符号将文章内容切分成语义块,每个语义块就是一个字符串,针对每一个字符串作循环。每次只处理一个汉字,将该汉字假设为词首,并且在词库中检索以该汉字为词首,检索该汉字后的字符匹配情况。如果在语义块中,该汉字与其后面的若干字符形成的字符串在词库中存在,就是说这个字符串可以构成词语的形式,则将符合该条件的词语全部检索出来。根据检索出来的词作为分词结果的备选项,按长度排列,首先取出长度最长的那个词,即最大匹配,假设这个词就是以该汉字为首的分词结果,加入到这个语义块的分词结果栈中,然后继续该词语位置之后的下一个汉字的处理。

按照此类方式继续在语义块中进行分词,如果遇到以某一个汉字为首的词在词库中并不存在,或者词库中存在的词语与语义块中该汉字为首的字符串并不匹配,则认为是前半部分的分词出现了歧义,产生了问题。于是从该语义块的分词结果栈中弹出上一次压栈的结果。如果上一次的分词结果还有其它备选项,则取分次备选项中长度第二长的词为分词结果,并且压栈。如果上一次的分词结果备选项中只有一个,那么则继续弹栈。如果出现将全部分词结果都被弹出的情况,则说明先前在词库中找不到匹配的那个汉字未录入词库,应该先略过不分,等其它字符全部分词结束后,由用户判断正确的分词结果,然后再录入词库。

在该方法实现的过程中,作者将语义块中已经分词成功的那部分字符串在压栈的同时,从语义块中去掉,也就是说,语义块每次都是从字符串开头进行字符匹配的处理。如果分词结果栈中出现分词歧义需要弹栈时,将弹出的结果加在原来语义块字符串的首部。这样就不需要在每得到一个分词结果后计算下一个即将处理的汉字的位置了。

以下是堆栈-最大匹配自动分词模型。

假设存在有文章模型 $T = \sum_{i=0}^n b_i p_i$, b_i 表示语义块, p_i 表示间隔的标点符号。

$b_i = \sum_{j=0}^m a_j$, 也就是说, 每个语义块 $b_i = a_0 a_1 \cdots a_m$ 。 a_j 表示 b_i 中的单个汉字, a_0 表示这个语义块的首字符。

有这样一个集合表示词语集, 即词库,

$w_i \in Q\{w_i | i = 0, 1, \dots\}$, 其中 $w_i = \sum_{j=0}^n c_j$ 。在字符匹配的过程中, 将 $\sum_{j=0}^k a_j$ 与 $\sum_{j=0}^n c_j$ 进行匹配检验, 其中 $0 \leq k \leq m$, 如果形成匹配, 即 $\sum_{j=0}^k a_j = \sum_{j=0}^n c_j$, 且 $n = k$, 则将匹配的结果, 也就是词语集中的 w_i 加入到分词结果备选项集 E 中, $E = \{w_i | \sum_{j=0}^k a_j = \sum_{j=0}^n c_j\}$, 同时将 b_i 中 $\sum_{j=0}^k a_j$ 去掉, 重新生成 b_i ; 集合 R 是分词结果的集合, $R = \{w_i | \sum_{j=0}^k a_j = \sum_{j=0}^n c_j\}$ 。

如果分词结果备选集 $E = \Phi$, 首先判断 a_0 是否存在于集合 $U = \{c_i | u_i = c_i \sum_{i=0}^n b_i, u_i \notin Q\}$, 如果 $a_0 \notin U$, 则将 a_0 加入到 U 中, 然后将上次加入到分词结果集 R 中的字符串 w_{k-1} 从 R 集中取出, 加入到 b_i 的首部; 如果 $a_0 \in U$, 则 $b_i = \sum_{j=1}^k a_j$, 略过 a_0 。

如果分词结果备选集 E 中只有一个元素, 则说明该元素就是该汉字的相关分词的结果, 并且加入到结果集 R 中。

如果分词结果备选集 E 中不止一个结果, 则选择集合 E 中长度最长的那个词作为分词结果加入到集合 R 中, 其中 $w_i = \sum_{j=0}^n c_j$, 其中 n 为集合 E 中最大值。

在算法描述时, 将 E 应用为数组的形式; 将 R 应用为堆栈的形式。

3.6.4 堆栈-最大匹配自动分词算法

根据堆栈-最大匹配自动分词方法的基本思想和模型, 形成了相应的堆栈-最大匹配自动分词算法。堆栈-最大匹配自动分词算法的流程如图 3.8 所示。

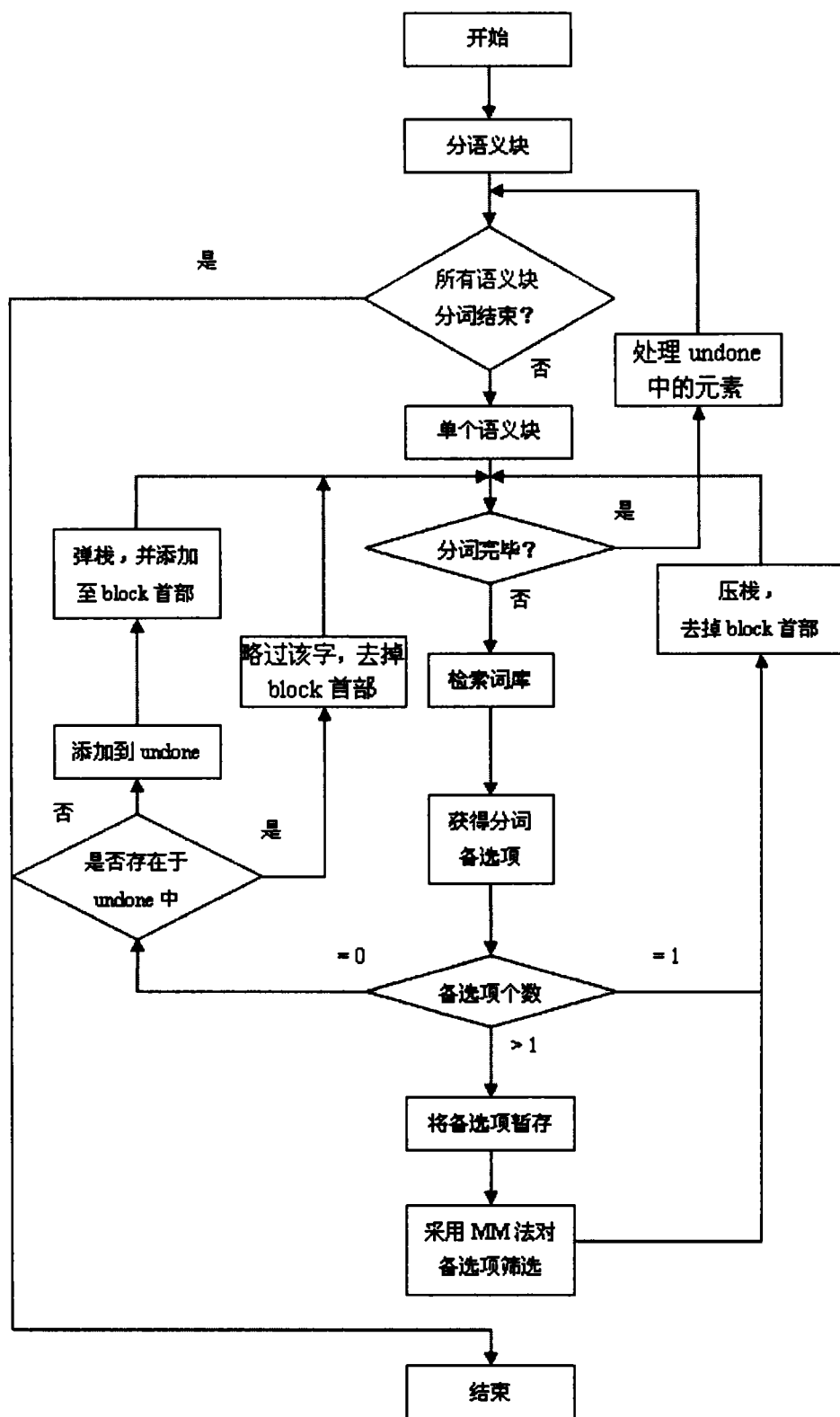


图 3.8 堆栈—最大匹配自动分词法流程图

堆栈-最大匹配自动分词的核心算法如下:

- (1) 在现有的句子中以标点符号为标界, 且分成多个语义块 **block**, 存为

字符串数组；设置另一个字符串数组 `result`，存放单个 `block` 的分词结果；设整型数组 `undone`，用来记录不可分的汉字的出现位置。

(2) 循环字符串数组，对数组中每个语义块 `block` 进行步骤 3，直到整个字符串数组被处理完毕；

(3) 对单个的语义块每次都是从 `block` 的首个汉字开始进行分析，执行下一步；

(4) 如果 `result` 的总长度与原语义块的长度相等，或者是 `block` 的长度为零，说明该语义块分词完毕，执行步骤 10；如果分词结果栈 `result` 为空，`block` 长度与原语义块相等，且不可分数组 `undone` 有新近填入的元素，则说明该句中有不可分的字符，那么开始重新分词，根据不可分数组 `undone` 中的最新记录，当分词过程遇到该汉字时，将该汉字暂时略过；执行步骤 3；

(5) 取 `singleword = block.SubString(0,1)`，继续；

(6) 在词语表中查找以 `singleword` 为首词语，存为一个字符串数组 `temp`，作为分词的备选项，继续以下判断；

(7) 如果 `temp` 的长度为零，即 `if(temp.Length == 0)`，则说明不存在以该字为首的词语；比较该汉字的位置是否在不可分数组 `undone` 中有记录，如果有，则略过该汉字，执行步骤 3；如果没有记录，则将该汉字的位置记录到不可分数组 `undone` 中，弹出栈顶元素加至 `block` 字符串的首部，执行步骤 3；

(8) 如果 `temp` 的长度为 1，即 `if(temp.Length == 1)`，则说明在词语表中只有一个分词结果备选项，那么该结果就是所要的分词结果，从 `block` 首部取出该词语压入分词结果栈中 `result` 数组中，执行步骤 3；

(9) 如果 `temp` 的长度大于 1，即 `if(temp.Length > 1)`，则说明分词结果备选项中存在多个结果，按照 `temp` 数组中的字符串长度的次序由小到大排列，取数组最后一个元素的字符串，也就是长度最长的字符串作为分词的结果，假设为分好的一个词，并且在 `block` 首部去掉该词，压入分词结果栈 `result` 中，执行步骤 3。

(10) 如果不可分数组 `undone` 不为空，则对数组中的元素和分词结果中的元素进行人为干预，将新词录入词库，执行下一步；

(11) 开始下一个语义块的分词，将上一个语义块的分词结果输出，并且

将分词结果栈 result 清空, 执行步骤 2。

3.6.5 自动分词举例

假设在文章的句子中, 已经有了切分好的语义块。例如, 有一句话“这些学生会都来了”。词库中已经有以下的词语了:

这些 学生 学生会 会员 都 来 了

那么, 应用上述的自动分词算法, 依次对该句的汉字进行分析, 其详细过程如下:

(1) 检索“这”, 发现“这些”在词库中并且与原文匹配。

(2) 检索“学”, 发现有两个匹配, 分别是“学生”和“学生会”, 取字符长度最长的那个匹配项“学生会”;

(3) 检索“员”, 发现词库中没有以“员都”或“员”这样的词语, 因此不存在匹配, 于是将先前的栈顶元素弹出, 压入第二长的分词备选项“学生”;

(4) 检索“都”, 这是一个副词, 在词库中;

(5) 同理, “来”和“了”依次被分出来。

上述过程的分词结果栈的存储过程如图 3.9 所示。

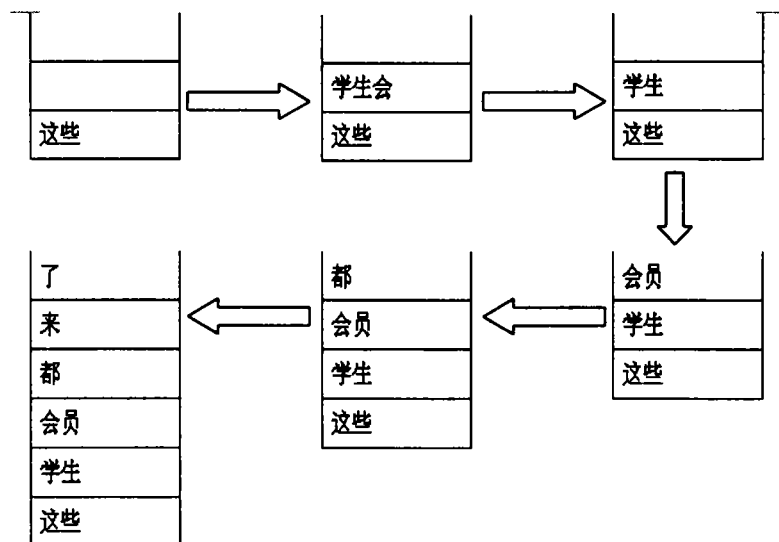


图 3.9 栈存储过程图

3.7 本章小结

本章主要介绍了构建汉语语法语料库过程中的几个主要的组成部分。首先阐明了应用语料库研究方法在语言学研究过程中的重要性,强调了预料采集的注意事项;引入规则库作为建立众多数据表联系的桥梁;描述了数据表之间的关系,介绍了部分数据表的内容,建立了数据表间的访问方法;根据建立汉语语法结构语料库的实际需要,还引入的自动分词技术,对原有的分词技术作了改进。

第四章 汉语语法语料库系统的实现

在汉语语法理论知识和语料库构建知识的指导下,我们采用 Microsoft 公司的 SQL Sever 和 Visual C# .NET 开发工具分别制作开发了语料库和应用程序,汉语语法结构语料库系统结构如图 4.1 所示。

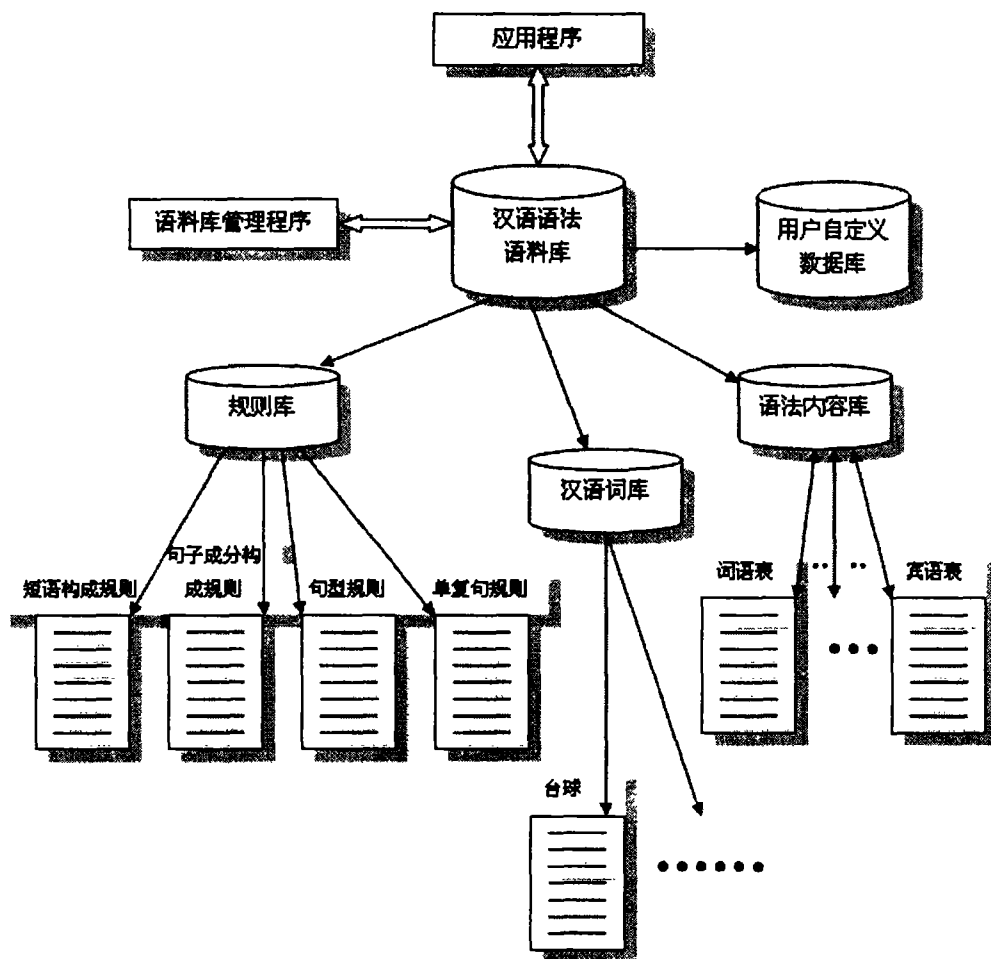


图 4.1 汉语语法语料库系统结构图

整个系统共分三个部分,分别是用户应用程序、语料库管理程序和汉语语法语料库。语法语料库在第三章已经作了详细的说明,在本章就不再重复了。本章主要介绍一下用户应用程序和语料库管理程序。

4.1 用户应用程序

用户应用程序作为前台界面,负责与后台语料库的连接,它有两个主要功能。

(1) 构建语料库时的录入功能。设计此项功能的目的是,就是在构建汉语语法语料库的时候,将语料信息输入到语料库的过程中,可以根据应用程序中给出的信息提示,填入内容;此过程中有部分内容可以按规则库自动生成。工作过程如图 4.2 所示。

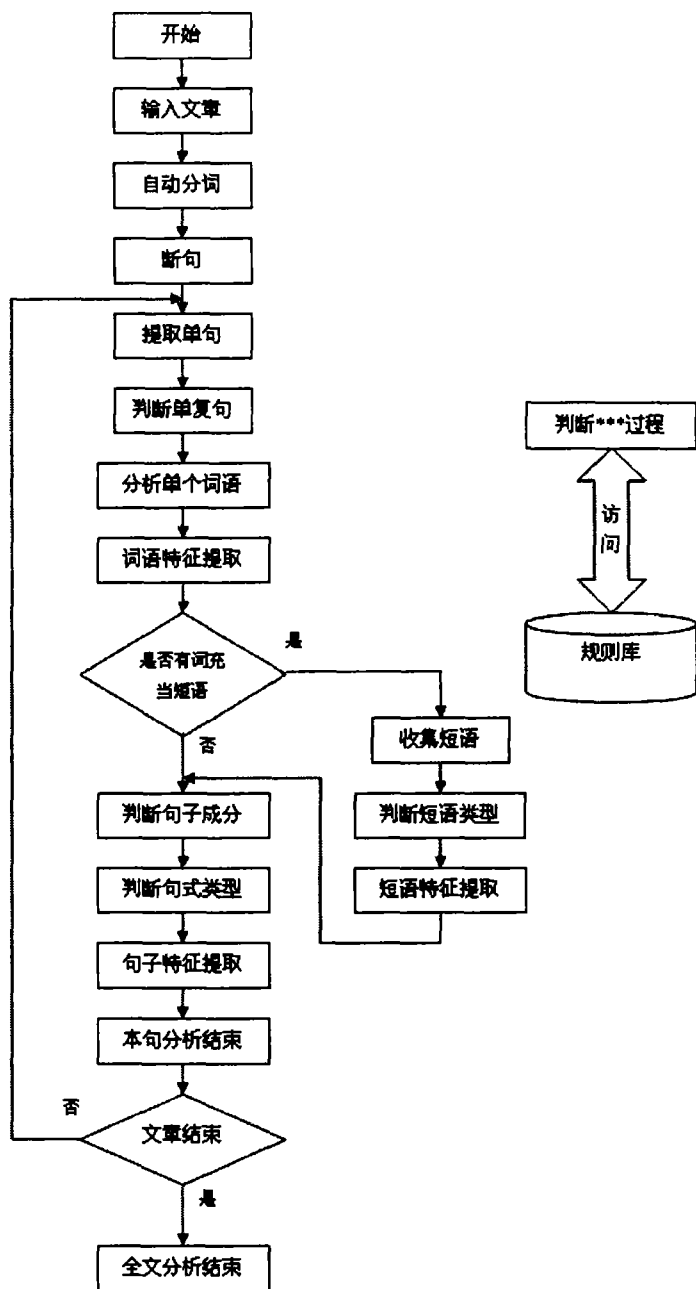


图 4.2 语料库系统工作过程图

整个工作过程中的判断过程是通过访问语料库中的相关规则库来实现的。如图 4.3~图 4.7 所示。

第一步：选定一个句子并且自动分词；

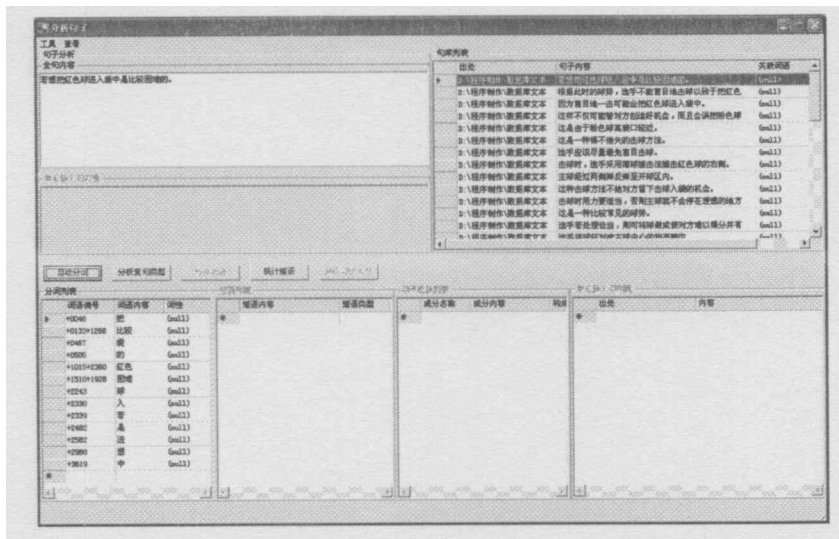


图 4.3 选句、自动分词

第二步：分析单个词语；

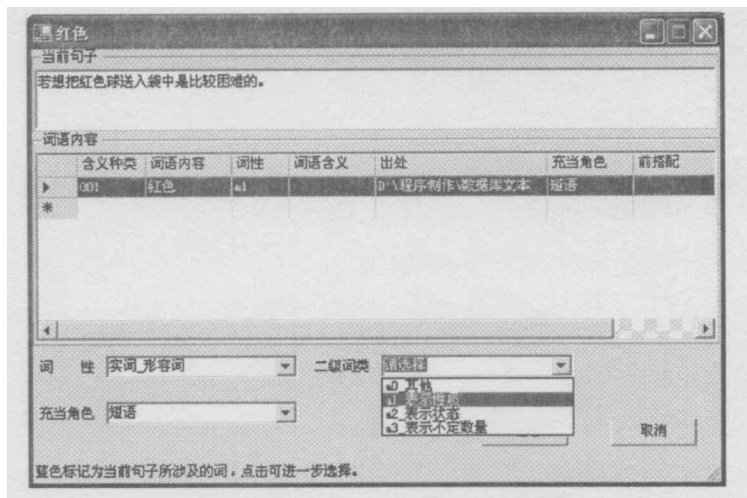


图 4.4 分析单个词语

第三步：统计短语，同时对每个短语进行分析；

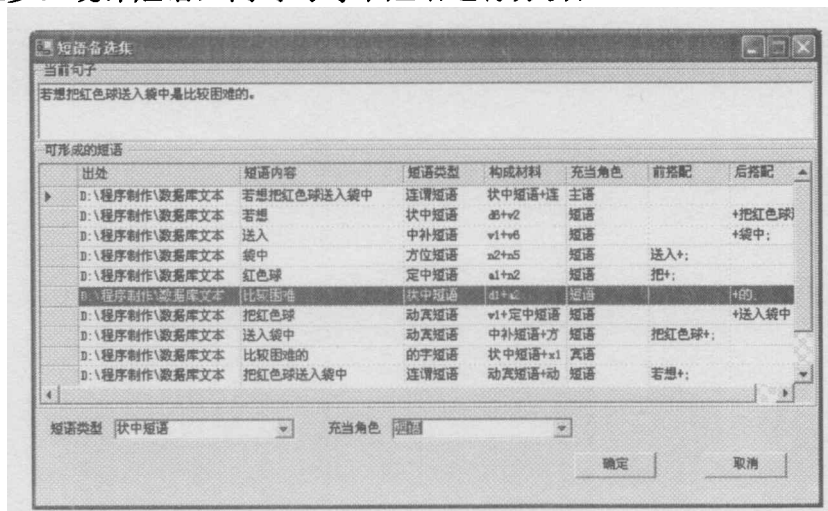


图 4.5 分析短语

第四步：分析句子成分，并作补充；

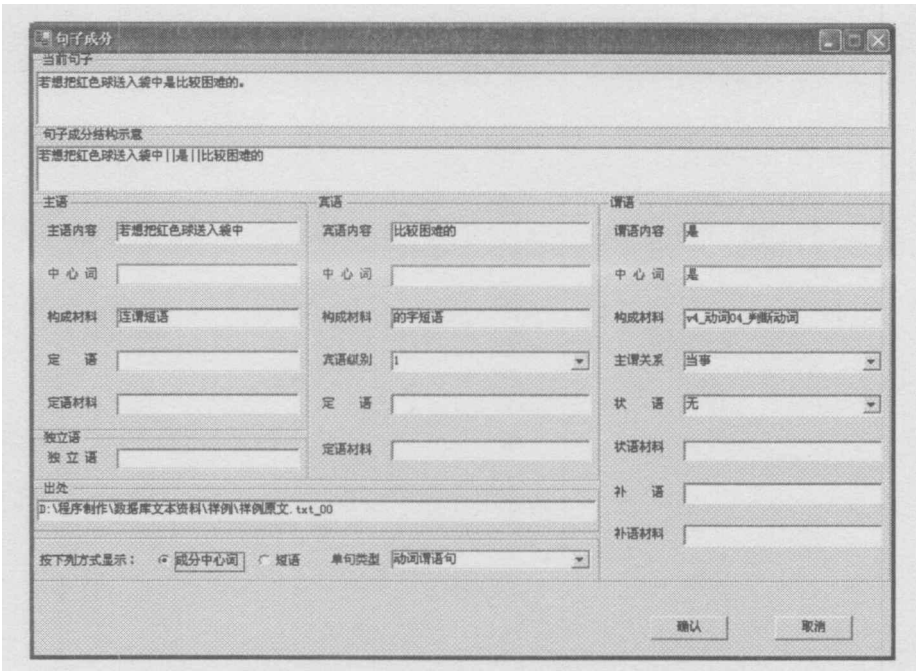


图 4.6 分析句子成分

第五步：将结果呈现在列表中，并更新相应的数据库，同时形成前后搭配。

分词列表			短语列表			句子成分列表		
词语编号	词语内容	出现频率	出处	前搭配	后搭配	成分名称	成分内容	构成
+0048	把	1	D:\程序制作			谓语	是	V_动
+0133+1268	比较	1	D:\程序制作		+把红色球送	主语	若想把红色球送入袋	连谓
+0467	袋	1	D:\程序制作	把红色球+;	+袋中;	宾语	比较困难的	的字
+0505	的	1	D:\程序制作	送入+;				
+1015+2360	红色	1	D:\程序制作	把+;	+送入袋中;			
+1510+1928	困难	1	D:\程序制作		的;			
+2243	球	1	D:\程序制作	若想+;	+送入袋中;			
+2330	入	1	D:\程序制作	把红色球+;				
+2339	若	1	D:\程序制作					
+2482	是	1	D:\程序制作	若想+;				
+2582	送	1						
+2988	想	1						
+3619	中	1						
*								

图 4.7 结果列表

通过计算机处理由用户填入的内容，自动提取相应的语法特征，可以大大减轻以往在构建语料库的过程中的录入过程。但是，为了尽可能多的统计的汉语中的各种语法现象，仍需要后期大量的工作，以丰富汉语语法结构语料库的内容。

(2) 应用语料库时的检索功能。建立汉语语法结构语料库的目的就是以统计的方式对汉语言的语法结构现象进行研究，根据大量的统计资料，来说明某个汉语语法现象，并且对该现象的产生以及应用情况进行归纳，从而为汉语语义的研究辅以有效的手段。

系统的检索过程如下图所示。

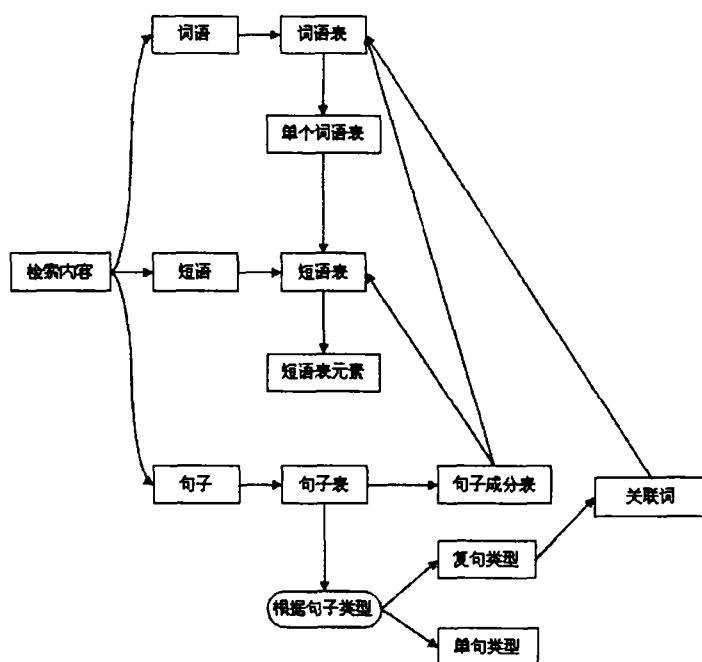


图 4.8 检索功能的数据流向图 a

除了上图之外，还有一些内容如下图所示，下图表示的是根据提取的语法特征内容为线索进行检索的数据流向图：

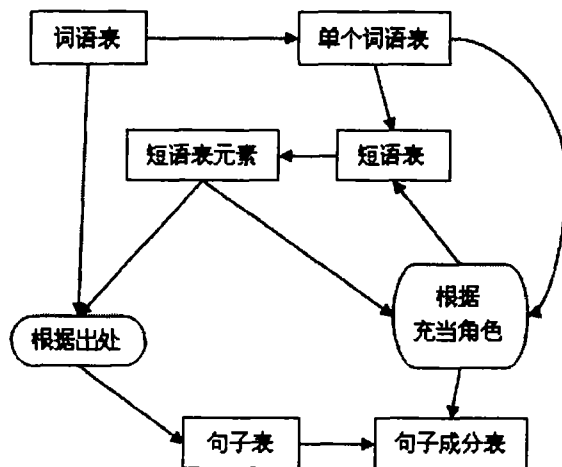


图 4.9 检索功能的数据流向图 b

以下是查询检索功能的图示：

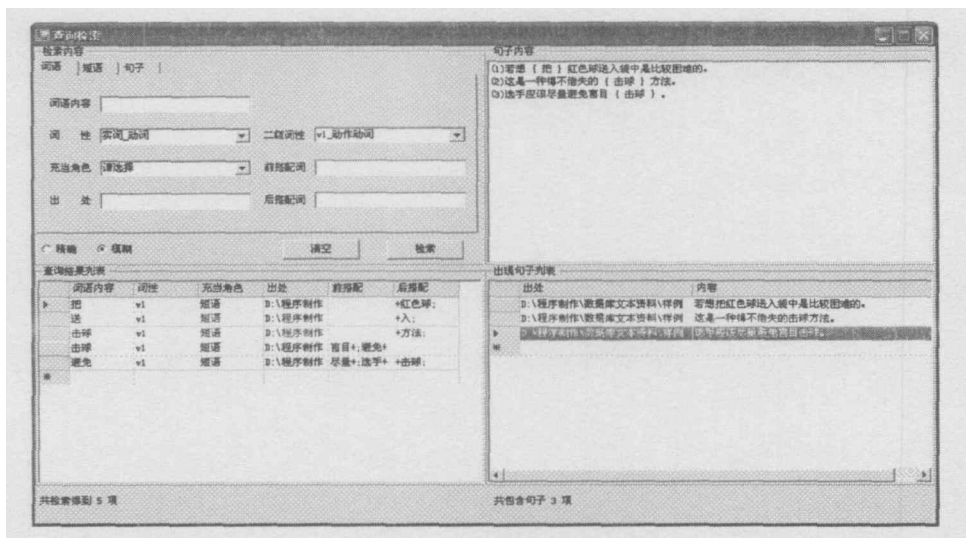


图 4.10 查询检索功能

在检索的时候，需要将上述两种检索方式结合起来。

4.2 语料库管理系统

在汉语语法结构语料库中，数据表是整个语料库系统核心，数据表繁多成为汉语语法语料库的主要特点。那么，对这些数据表建立起一套的管理系统，不仅可以有效地管理语料库中的数据表，使整个语料库的数据表按不同的分类存放，也可以使用户根据不同需要自定义一些数据类和数据表。

以下是语料库管理系统图例。图 4.11 表示的是数据类的管理，用户或程序员可以根据不同的需求，在“添加新数据类名称”或“添加新内容”中增加设置一新数据类，这一数据类中有可以容纳多个属于该类的数据表。

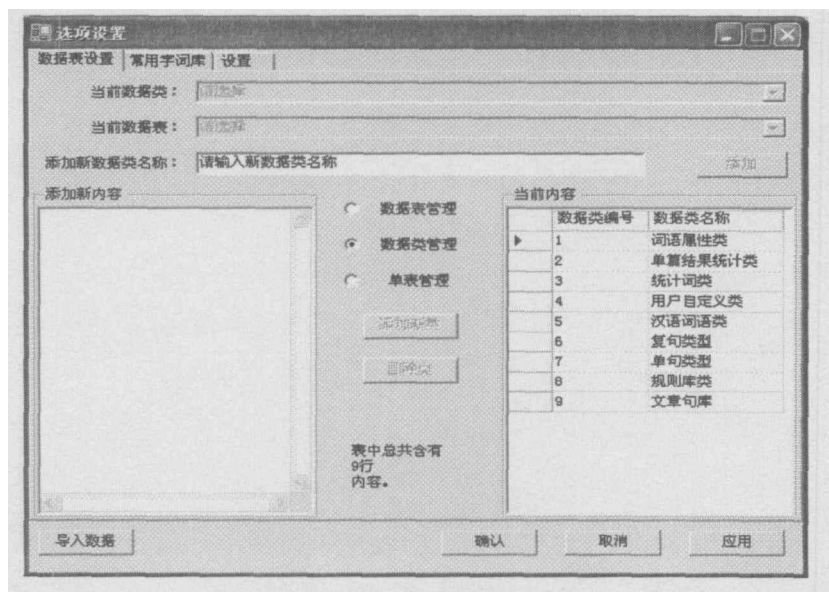


图 4.11 数据类管理

图 4.12 显示的是在某一数据类下的数据表的管理，根据用户和程序员开发的需要可以在“添加新内容中”增加新的数据表。

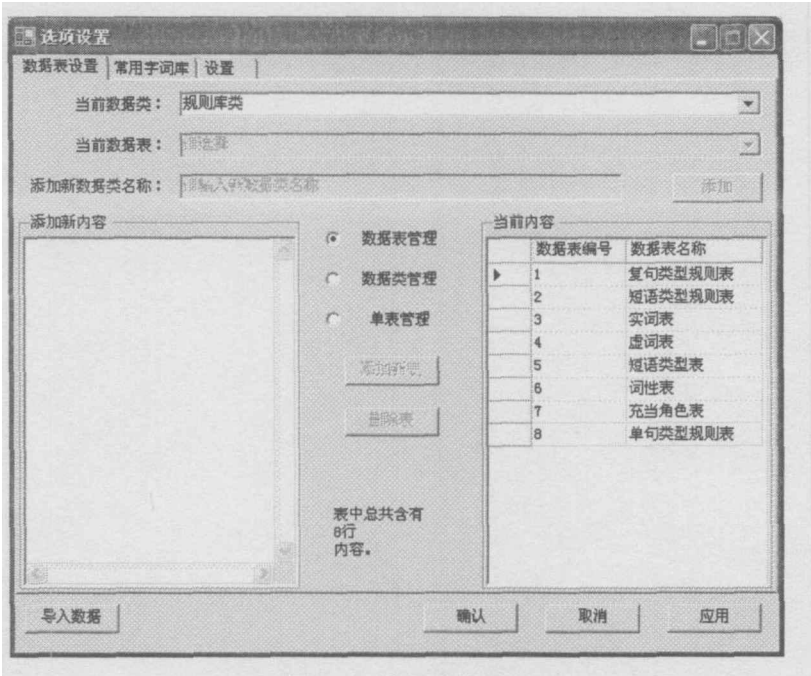


图 4.12 数据表管理

图 4.13 表示的是某一数据表的详细内容，如“规则库”中的“短语类型规则表”。

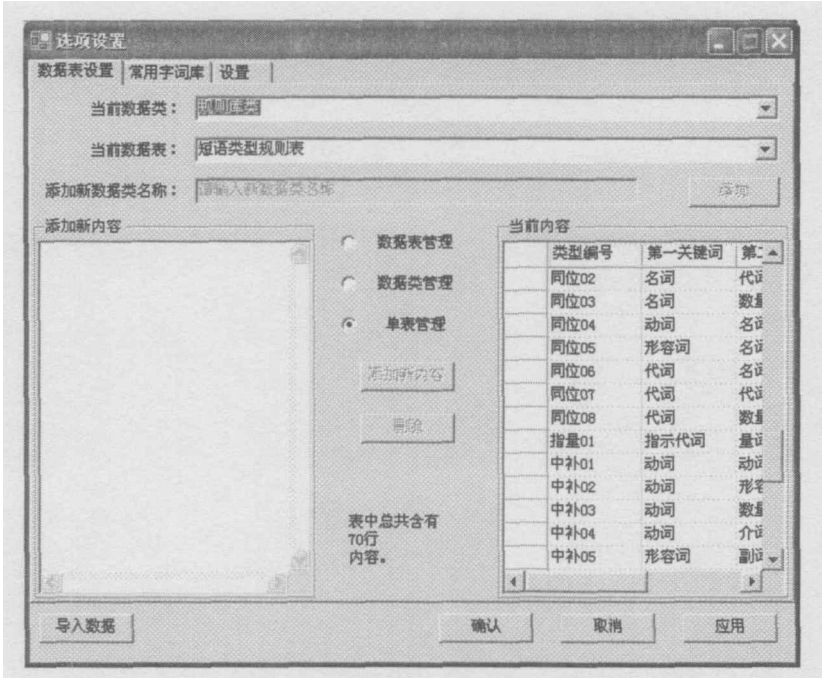


图 4.13 短语类型规则表管理

这个语料库管理系统的建立，对后续的研究工作提供了数据表管理的方便。无论是作为用户需要自定义一些数据表，还是程序员在开发过程中需要建立新的数据表内容都可以使用该数据表管理系统。

4.3 本章小结

本章主要介绍了汉语语法语料库系统的构成情况，从功能上共分两种用途：语料库构建过程中的语料信息智能输入过程和应用语料库时的检索功能。并且根据语法语料库系统的特点对上述两种工作过程进行了介绍。

第五章 总结全文及今后工作的展望

5.1 总结全文

本文围绕当前自然语言处理过程中所存在的问题以及汉语的特点,论述了汉语的自然语言处理程序需要以语义分析为主,并辅以语法分析的手段来进行。这是因为语法分析过程是比较容易被形式化的语言所描述,容易被计算机接纳并处理这些信息。与此同时,由于汉语研究工作需要在一定的汉语语言环境中进行的特点,提出了建立语料库的方法来辅助研究。本文的主要研究工作如下:

(1) 在总结前人在自然语言处理研究成果基础上,结合汉语语言特点,总结了汉语作为自然语言处理内容研究的过程;阐述了语法分析和语义分析的关系。

(2) 将汉语语法知识与自然语言处理研究结合起来,并且根据计算机处理的需要,对汉语语法知识做了新的分类,分为词、短语、句子三个方面,并且归纳了三者之间的多重关系;提出了语法特征提取这一方法,并对其内容及作用作了描述。

(3) 鉴于目前研究工作的需要,提出了引入构建汉语语法语料库设计思想,并且详细描述了建库的主要过程;根据汉语语法中的构成规则建立了规则库,为数据表之间的相互访问提供了依据;结合语法特征提取和规则库的内容,设计了数据表间的访问方法;并且根据建库的需要,在原有的最长匹配法自动分词技术基础上加入了堆栈设计,优化了原先的算法。

(4) 在系统中分别设置了语料信息智能录入和语料库信息的检索功能,初步实现了汉语语法语料库系统的基础设计工作。

5.2 对今后工作的展望

根据本文中提出的汉语的自然语言处理过程的研究思想,目前这个系统只是阶段性的工作内容之一。在今后的研究工作中首先需要采集大量的语料信息以充实语料库的内容;对本文提到的自动分词技术进行优化;并且在原有的基础上增加更多的智能功能,使整个系统运作更加简洁、省力;完善对语料库管理系统的功能设计,加强对语料库中诸多数据表的访问和管理。

参考文献

- [1] E Brill, Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics. 1993.
- [2] Microsoft Corporation. MSDN Library - Visual Studio .NET 2003 Edition.
- [3] P Resnik, Probabilistic Tree-Adjoining Grammar as a Framework for Statistical Natural Language Processing In Processing, 1992.
- [4] Benny Johansen, Matthew Reynolds, 张哲峰译, Windows 应用高级编程—C#编程篇, 清华大学出版社[M], 2003 年 1 月第 1 版.
- [5] 陈小荷, 现代汉语自动分析—Visual C++实现[M], 北京语言文化大学出版社, 2000 年 3 月第 1 版.
- [6] 代建英, 汉语自动分词系统的研究与实现, 硕士学位论文, 重庆大学, 2005.
- [7] 段恩泽, 基于统计的汉语自动分词系统, 硕士学位论文, 电子科技大学, 2004.
- [8] 冯志伟, 基于短语结构语法的自动句法分析方法[J], 中文信息学报, 2000.05 期.
- [9] 黄伯荣, 廖序东, 《现代汉语》下册[M], 高等教育出版社, 2002 年 7 月, 第三版.
- [10] 黄曾阳, HNC 概念层次网络理论[M], 清华大学出版社, 1999.
- [11] 黄昌宁, 中文信息处理中的分词问题[M], 语言文字应用, 1997.
- [12] 黄昌宁、李涓子, 语料库语言学[M], 商务印书馆, 2004 年第一版.
- [13] 亢世勇, 面向信息处理的现代汉语语法研究, 上海辞书出版社[M], 2004 年 12 月第 1 版.
- [14] 李香敏, SQL Server2000 Programmer's Guide 程序员指南[M], 北京希望电子出版社[M], 2000 年 12 月第 1 版.
- [15] 刘华, 基于关键短语的文本内容标引研究, 博士学位论文, 电子科技大学, 2005.
- [16] 刘颖, 计算语言学, 清华大学出版社[M], 2002.
- [17] 陆俭明, 词的具体意义对句子意思理解的影响[J], 汉语学习, 2004.02 期.
- [18] 陆俭明, 关于句处理中所要考虑的语义问题[J], 语言研究, 2001.01 期.
- [19] 吕冀平, 《汉语语法基础》, 商务印书馆[M], 2000 年 1 月, 第一版.
- [20] 皮晓峰, 基于概率上下文无关语法的句注分析研究与实现, 硕士学位论文, 电子科技大学, 2005.

- [21]齐丙辰等,“计算机辅助教育”与“机器人辅助教学”[J],天津师范大学学报(自然科学版),2003年3月.
- [22]齐丙辰等,现代教育技术的新领域—机器人辅助教育[J],机器人技术与应用,2000年第1期.
- [23]齐沪扬,现代汉语短语[M],华东师范大学出版社,2000.
- [24]孙茂松,汉语自动分词研究中的若干理论问题[J],语言文字应用,1995.04期.
- [25]孙茂松,语言计算与基于内容的文本处理[M],清华大学出版社,2003.
- [26]王小捷,常宝宝,自然语言处理技术基础[M],北京邮电大学出版社,2002.
- [27]吴雪敏,汉语语句的计算机分析,硕士学位论文,电子科技大学,2001.
- [28]姚天顺,自然语言理解——一种让机器懂得人类语言的研究[M],清华大学出版社.
- [29]俞士汶,语法知识在语言信息处理研究中的作用[J],语言文字应用,1997.04期.
- [30]詹思瑜,自然语言的计算机处理,硕士学位论文,电子科技大学,2003.
- [31]张普,关于控制论与动态语言知识更新的思考[J],语言文字应用,2002.04期.
- [32]赵志靖等,智能人机交互中自动分词技术的实现[J],扬州大学学报,2005.03.
- [33]郑逢斌,关于计算机立即自然查询语言的研究,博士学位论文,西南交通大学,2001.
- [34]周舫,汉语句子相似度计算方法及其应用的研究,硕士学位论文,河南大学,2002.
- [35]朱德熙,语法讲义[M],商务印书馆,1982.
- [36]朱钦隼,计算机汉语理解的初步实践,硕士学位论文,电子科技大学,2001.

附录 1: 实词表及其标记

词语大类	二级词类	标记
名词(N)	普通人	N1
	普通事物	N2
	时间	N3
	普通处所	N4
	方位	N5
	中文人名	N6
	西文人名	N7
	专有地名	N8
	其他	N0
动词(V)	动作动词	V1
	心理活动动词	V2
	表示存在、变化、消失	V3
	判断动词	V4
	能愿动词	V5
	趋向动词	V6
	其他	V0
形容词(A)	表示性质	A1
	表示状态	A2
	表示不定数量	A3
	其他	A0
区别词(F)		F
数词(M)	基数词	M1
	序数词	M2
	其他	M0
量词(Q)	专用物量词	Q1
	借用物量词	Q2

	专用动量词	Q3
	借用动量词	Q4
	复合量词	Q5
	其他	Q0
副词(D)	表示程度	D1
	表示范围	D2
	表示时间、频率	D3
	表示肯/否定	D4
	表示情态、方式	D5
	表示语气	D6
	其他	D0
代词(R)	人称代词	R1
	疑问代词	R2
	指示代词	R3
	其他	R0
拟声词(S)		S
叹词(E)		E

附录 2：虚词表及其标记

词性大类	二级词类	标记
介词(P)	表示时间、处所、方向	P1
	表示方式、方法、依据、工具、比较	P2
	表示原因、目的	P3
	表示施事、受事	P4
	表示关涉对象	P5
	其他	P0
连词(C)	连接词语、短语（和、跟、同、与、及、或）	C1
	连接词语、分句（而、而且、并、并且、或者）	C2
	连接复句中的分句（可以参考附录 7）	C3
	其他	C0
助词(X)	结构助词（的、地、得）	X1
	动态助词（着、了、过、来着）	X2
	比况助词（似的、是的、一样、般、一般）	X3
	其他（所、给、连、们）	X0
语气词(T)	表示陈述语气	T1
	表示疑问语气	T2
	表示祈使语气	T3
	表示感叹语气	T4
	其他	T0

附录 3: 短语表 I (按短语结构分类)

名称	分类	构成形式	举例
主谓短语		名·名	昨天 元旦
		名·是·名	明天 是 元旦
		名·动	夜幕 降临
		名·形	灯火 辉煌
动宾短语		动·名	拿 杯子
		动·动	拒绝 施舍
		动·形	喜欢 凉快
		动·数量	买 两双
		动·代	爱 你
偏正短语	定中短语:	名·名	昨天 的结果
		名·动	领导 的视察
		名·形	工程 的浩大
		动·名	投掷 的要领
		动·动	求学 的希望
		动·形	设计 的新颖
		形·名	好 人
		形·动	合理 的规划
		区别·名	野生 动物
		代·名	你 的座位
		代·动	他 的考虑
		代·形	它 的残暴
		数量短语·名	五个 儿童
	状中短语:	名·动	明天 出发
		动·动	绕道 走
		形·动	快 跑
		区·动	大型 规划

		代·动	那么 想
		代·形	这么 宽
		副·动	立刻 出发
		副·形	很 好
		拟声·动	咚咚 地敲
		数量短语·动	一口一口 地喝
		数量短语·形	十米 深
		介词短语·动	到学校 学习
		方位短语·动	家里 商量
中补短语		动·动	救 活
		动·形	跑 得快
		动·数量短语	想了 一晚
		动·介词短语	跑 到山顶
		形·副	漂亮 极了
联合短语	表示并列词：“和”、“而”、“既 A 又 B”、“又 A 又 B”、“A 不 A”		张三和李四
			既美丽又大方
			好不好
			又大又圆
	表示选择词：“或”、“或者”、		一个或两个
			小张或者你
连谓短语		动·动	上天 揽月
		动·形	想着 难受
兼语短语			请他进来
同位短语		名·名	省城 太原
		名·代	渤海湾 那里
		名·数量短语	甲乙 两类
		动·名	下海 这种事
		形·名	困难 这个词
		代·名	他们 学生

		代·代	我们 大家
		代·数量短语	你们 几个
方位短语		名·方	操场 上
		动·方	唱歌 前
		数量短语·方	五年 后
		主谓短语·方	天亮 后
量词短语	数量短语: 数词加量词		两 个
	指量短语: 指示代词加量词		这 件
介词短语	有介词附着在名词等词语前面组成。		略
助词短语	的字短语	只能作主语或宾语	有 _大 的 _和 小的 _的
	比况短语	作定语、状语、谓语、补语	略
	所字短语	所+及物动词	略

附录 4: 短语表 II (按短语功能分类)

名词性短语:	联合短语 (名词性成分联合)
	偏正短语 (定中短语)
	方位短语
	同位短语
	量词短语 (用名量词)
	“的”字短语
	“所”字短语
谓词性短语 (谓词性短语 还可再分为动词性短语 和形容词性短语)	联合短语 (谓词性成分联合)
	偏正短语 (状中短语)
	动宾短语
	中补短语
	连谓短语
	兼语短语
	比况短语

附录5: 句子成分列表

句子成分	分类	构成材料	意义类型	
主语	名词性主语	名词、数词、名词性代词、名词性短语	施事 受事 当事	
	谓词性主语	动词、形容词、谓词性代词、动词性短语、形容词性短语		
谓语	谓词性谓语	动词、形容词		
	名词性谓语	构成名词谓语句		
	主谓短语谓语	构成主谓谓语句		
动语	动词性词语	名宾动词：只能带名词性宾语		
		谓宾动词：只能带谓词性宾语		
		名宾兼谓宾动词：可带名词性和谓词性宾语		
	兼属动词形容词			
宾语	名词性宾语	名词、数词、名词性代词、名词性短语		
	谓词性宾语	动词、形容词、谓词性代词、动词性短语、形容词性短语		
	主谓短语宾语	主谓短语作宾语		
定语	限制性定语	名词性词语、动词性词语、区别词		
	描写性定语	形容词性词语		
状语	可分为限制性和描写性	时间名词、能愿动词、形容词、副词、介词短语、量词短语；一般名词、动词不作状语		
补语	补语通常由谓词性词语、数量短语、介词短语			
	结果补语	表示动作、行为产生的结果，与中心语有因果关系		
	程度补语	这样的词语非常有限，有如下词语“极、很、透、慌、死、坏”和“一些、一点”		
	状态补语	由动作、性状而呈现出来的状态		

	趋向补语	表示动作的方向或事物随动作而活动的方向，用趋向动词充当
	数量补语	又可分为动量补语、时量补语
	时间、处所补语	多用介词短语来表示动作发生的时点和处所
	可能补语	用“得”或“不得”充当；在结果补语或趋向补语和中心语之间插入“得/不”
独立语	可分为插入语、称呼语、感叹语、拟声语	

附录6：句型列表

类别	子类别	举例
主谓句	名词谓语句	明天国庆节
	动词谓语句	你有课么？
	形容词谓语句	这儿真好！
	主谓谓语句	他个儿真高
非主谓句	名词性非主谓句	哪里？
	动词性非主谓句	下雪了？
	形容词性非主谓句	真好！
	叹词句	哎哟！

附录7: 复句类型列表

复句大类	复句类型	子类型		关联词
联合复句	并列关系	并举	合用	既 A, 也 (又) B 又 (也) A, 又 (也) B 有时 A, 有时 B 一方面 A, (另、又) 一方面 B 一边 A, 一边 B 时而 A, 时而 B 一会 A, 一会 B
			单用	也 又 还 同时 同样 另外
		对举	合用	是 A, 不是 B 不是 A, 而是 B
			单用	而 而是
	顺承关系	合用		首先 A, 然后 B 起先 A, 后来 B 刚 A, 就 B
			单用	就 便 才 又 再 于是 然后 后来 接着 跟着 继而 终于
	解说关系			
	选择关系	数者选一	合用	或 A, 或 B 或者 A, 或者 B 或是 A, 或是 B 是 A, 还是 B
			单用	或者 或是 或 还是
		二者选一	合用	不是 A, 就是 B 要么 A, 要么 B
			单用	还不如 倒不如

	递进关系	先取后舍	合用	宁可（宁肯、宁愿）A，也不（决不、不）B
		一般递进	合用	不但（不仅、不只、不光、非但）A，而且（还、也、又、更、就连）B 不但（不但不、非但没）A，反而（反倒还、相反还、偏偏还）B
			单用	而且 并且 何况 况且 甚至 更 还 甚至于 更何况
		衬托递进	合用	尚且A，何况（更不用说、还）B 别说（慢说、不要说）A，连（就是）B
			单用	尚且 何况 反而
偏正复句	转折关系	重转	合用	虽然（虽是、虽说、虽则、虽、尽管）A，但是（可是、然而、但、却、还、也、而）B
			单用	虽然 虽 但是 但 然而
		轻转	单用	可是 可 却
		弱转	单用	只是 不过 倒
	条件关系	充足条件	合用	只要（只需、一旦）A，就（都、便、总）B
			单用	便 就
		必要条件	合用	只有（唯有、除非）A，才（否则、不）B
			单用	才 否则 要不然
		无条件	合用	无论（不论、不管、任、任凭）A，都（总、总是、也、还）B
	假设关系	一致	合用	如果（假如、假使、假若、假设、倘若、倘使、若是、若、要是）A，就（那么、那、便、则、也）B
			单用	那 那么 就 便 则

		相背	合用	即使（就是、就算、纵使、纵然、哪怕） A，也（还）B 再A，也B
			单用	也 还
	因果关系	说明	合用	因为（由于）A，所以（才、就、便、 于是、因此、因而、以致）B 之所以A，是因为（是由于、就在于）B
			单用	因为 由于 是因为 是由于 所以 因此 因而 以致 致使
		推论	合用	既然A，那么（就、有、便、则）B
			单用	既然 既 就 可见
	目的关系	达到目的	单用	以 以便 以求 用以 借以 好 好让 为 的是
		避免什么	单用	以免 免得 省得 以防

致 谢

随着论文的完成，我也即将结束三年的研究生生活，准备开始人生的另一个起点。回首走过的岁月，三年的研究生学习生活使我受益匪浅。在人生的旅途中，师长、朋友及家人的帮助才有了我今天的成绩。

首先要由衷地感谢我的导师齐丙辰教授，此篇论文从选题到撰写、修改、定稿的每一步，都得到了齐老师的鼓励与支持。齐老师严谨的治学态度、实事求是的作风的、开创性的研究理念都给我留下了非常深刻的印象，我从他身上学到许多宝贵之处。

同时，感谢冯舜玺、李学武两位资深教授在学习和实践中给与我的莫大的关心和支持，两位教授的一言一行深深地影响着我今后的学习、工作和生活。

我还要感谢马希荣、王慧芳、孙华志教授在这三年的研究生生活和学习中给与我的关心和帮助。

感谢天津师范大学国际交流学院的研究生支建刚、王业奇、郭晓玮三位同学在我作论文期间给我的热情帮助。

感谢我的同窗好友张永、魏雪峰、王刚、单国全、曹帷在研究生期间给与我的帮助与鼓励。

感谢我的女朋友李睿及其家人对我的关心、理解与支持。

感谢我的家人在我的漫漫求学生涯中对我的无私奉献与支持，使我得以顺利地完成学业。

衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授！

作者

2006. 4. 20