RESEARCH ON AUTOMATIC GRAMMAR INFERENCE IN SPOKEN DIALOGE SYSTEMS

A Dissertation Submitted

to the Graduate School of Henan Normal University

in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

By

Zhang He

Supervisor: Prof. Wang Xiaodong

April, 2009

摘　要

基于规则的口语对话系统中的文法规则通常由领域专家和计算机语言学家手工设计，需要依赖专家的专业知识和经验，这对于普通开发人员来说是无法完成的。另外，手工设计的文法移植性差，根据某一领域设计的文法规则很难移植到其他领域继续使用，同样功能的口语对话系统对于不同领域都需要领域专家重新设计文法规则，存在大量的重复劳动，造成巨大的人力和物力浪费。随着社会需求的不断增大，系统规模空前扩大，文法规则的获取成了口语对话系统研发的主要瓶颈。

本文针对口语对话系统中语言的特点，以一种上下文无关增强文法为对象，对口语对话系统中文法规则的自动获取技术进行了研究。主要工作包括：

（1）对比分析常见形式文法的特点和性能，重点研究一种符合汉语口语特点的上下文无关增强文法，根据口语对话系统中语言的特点，选定该文法为对象进行文法规则自动推导技术的研究。

（2）提出一种基于句子分割的文法规则自动推导算法。基本思想是：用初始规则集对训练例句进行分析，若不能得到完整的语法树，则先对分析得到的片断进行消歧和归一化，然后根据顶层片断递归地推导出缺少的文法规则，并更新已有的规则集。研究片断的消歧和归一化策略，为了提高算法性能，探讨并给出一种算法的改进方案。

（3）研究面向领域任务的文法测评方法，给出一套灵活的、可领域定制的文法评测方法。使用该方法在天气预报查询领域对算法的输出文法进行评测，结果显示，输出文法的句法分析准确率在初始规则集为空时达到了 64.8%，在初始规则集只包含日期相关规则时达到了 86.4%。

关键词：口语对话系统，上下文无关增强文法，文法推导，文法评测

ABSTRACT

There are such problems for access to grammar in rule-based Spoken Dialogue Systems (SDS): Firstly, the grammar is often designed by experts, and this work depends on their professional knowledge and experience, so it is impossible for ordinary developers to finish this work. Secondly, the grammar is often designed manually, and this work is very cockamamie. Finally, the grammar designed for a domain is very difficult to apply to another domain. So it is obligatory to design new grammar for every application. With the development of the increasing social needs and scale of system, it has become a major bottleneck for development a spoken dialogue system.

Specific to the characteristic of SDS, this thesis took a type of enhanced context free grammar for Chinese spoken language as object, researched the technologies for automatic grammar inference in SDS.

1. The features and performance of common grammar were compared and analyzed, and a type of enhanced context free grammar for Chinese spoken language was researched mainly. And then took it as object according to the language characteristics in SDS to research the technologies for automatic grammar inference.

2. A method for automatic grammar inference based on sentence segmentation is proposed in this thesis. The main idea is to parse the training sentences with an initial rule set. If the parsed syntactic tree is incomplete, the top-most constituents are used to recursively infer the missing rules after disambiguation and normalization, and then the rule set is updated. The methods of disambiguation and normalization for segmentations were researched in this thesis. In order to improve the output grammar, the processing order of the training sentences is adjusted and the method's process is refined.

3. The domain-specific grammar evaluation technologies were researched in this thesis. And a set of flexible evaluation technologies was proposed which can be customized for special domain. Some experiments had been done to evaluate the performance of the algorithm proposed in this thesis in the domain of weather forecast enquiry. The parsing accuracy of the output grammar achieved 64.8% with an empty initial rule set and 86.4% with an initial rule set only including rules for date description.

Key Words: Spoken Dialogue Systems (SDS), enhanced context-free grammar, grammar inference, grammar evaluation

目　录

第一章 绪论

1.1 研究背景

在语音信号处理、语音识别、语音合成及语言理解各项技术迅猛发展的今天，口语对话系统(Spoken Dialogue Systems)具有很高的研究价值，其应用也必将带来很好的社会、经济效益。目前一批研究成果或实际系统已经出现，常见的比如旅游信息查询、电话客票服务和天气预报信息查询等。构建一个完善的对话系统，需要应用语音信号处理、语音识别、语言理解、知识表示、对话管理和文语转换等多项技术。与其它语音系统相比，对话系统面临以下几个主要问题：

（1）语音的口语性与自发性(spontaneousness)。在语音命令系统中，语音可以是孤立词；在听写机系统中，语音一般是书面语，要求发音比较规范；而在对话系统中，语音是(或者十分接近)人们日常生活中的口语，允许比较随意的发音。自发语音中包括不流利、不合语法、修改及内容不完整等口语现象，这给声学识别和语义分析带来挑战。

（2）语义分析的必要性。语音命令系统中，词表和用户意图(user's intentions)可以是简单的一一对应关系；而在对话系统中，用户意图往往必须用语义网络等更加复杂的方法来表示。此时，语义框架和语义分析模块的设计就成为必然。

（3）用户主导(User Initiative)、系统主导(System Initiative)及混合主导(Mixed Initiative)的关系处理问题。根据应用环境的不同，以及用户之间的差异，系统在对话过程中可以呈现出三种不同的主导方式；a)用户向系统主动提问或提供信息(用户主导)；b)系统向用户提问(系统主导)；c)一般情况下采取用户主导，在需要时切换到系统主导(混合主导)。

针对以上问题，清华大学的燕鹏举提出一种基于语义类的上下文无关增强文法及相应的语义分析方法[1]，较好地处理对话系统中常见的口语表达问题。然而口语对话系统中文法规则的获取却面临着以下问题：

（1）需要依赖领域专家和计算机语言学家专业的知识和经验，这对于普通开发人员来说是无法完成的。

（2）文法规则的获取目前主要采用手工方式进行，是一个相当繁琐的过程。

（3）文法移植性差。根据某一领域获取的文法规则很难移植到其他领域继续使用，

同样功能的对话系统对于不同领域都需要领域专家重新设计文法规则。

随着社会需求的不断增大，系统规模空前扩大，文法规则的获取成了对话系统研发的主要瓶颈。

针对该瓶颈，在深入分析对比国内外文法规则自动推导方法的基础上，研究汉语口语对话系统中文法规则的自动推导技术，探索一种符合汉语口语特点的文法规则自动推导算法，及一套面向领域任务的文法性能评测指标，使开发人员从专业且繁琐的手工劳动中解脱出来，提高对话系统研发效率，降低研发成本，具有很好的理论及应用价值。

1.2 研究现状

1.2.1 对话系统研究现状

对话系统，可以简单地定义为：以语音为输入输出接口，通过与用户进行交谈，实现自动信息（或其它）服务的系统。对话系统结构（图 1-1）包含四个主要功能部件，即语音识别器、语言理解器、对话管理器和语音合成器。目前，语音合成的研究已经比较成熟，其主要挑战在于如何使生成的语音更加自然与生动。一般而言，对话系统目标的实现对于语音合成自然度的依赖不是必须的，而语音识别、语言理解和对话管理是对话系统研究人员所所关注的焦点。

语音识别的目的是把人的语音转换成文字，这是许多语音系统的核心与主轴，比如听写机、语音命令系统和对话系统。与其它系统不同的是，对话系统中的语音识别的输出要付诸于语言理解，因此识别错误对语言理解的干扰是系统必须考虑的。语言理解得到语义表示后，对话管理要根据上下文语境、历史信息等，进行综合分析，以确定用户的意图，根据需要查询后台数据库，并组织应答语句等。可以看出，对话系统中这几个核心部件的关系比较紧密。

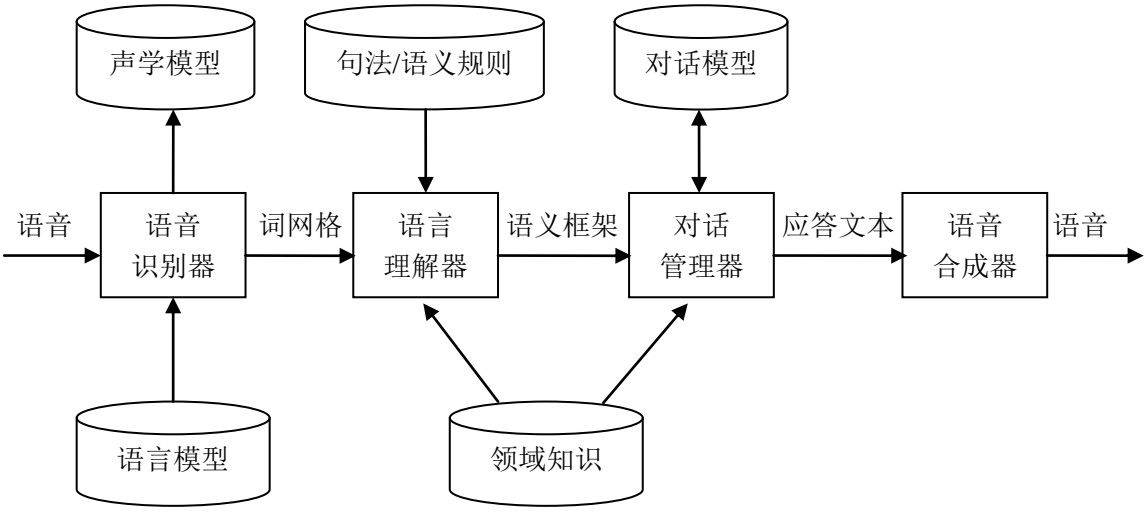

图 1-1 对话系统模型略图

对话系统的运行往往还依赖于一些模型或数据库，比如是声学模型、语言模型、句法/语义规则、领域(domain specific)知识、对话模型和领域数据库等。本节将对对话系统当前的研究现状做简要综述。

根据不同的应用，对话系统可以构建于不同的平台，有着不同的表现形式：

嵌入式平台：Hugunin 等人[2]设计了一个基于 Microsoft Excel 的嵌入式电子表格系统。该系统采用人机对话的方式进行电子表格的自动设计与填充。与使用鼠标键盘的手工输入方式和简单屏蔽鼠标键盘的语音命令方式相比，语音对话的方式提高了效率，而且使用户更加轻松。

WWW 平台：Issar[3]设计了一个用于在 WWW 网页上填充表格的语音软件。该软件使用 Java applet 作为用户接口，采用名为 Sphinx-II 的语音识别器和基于框架的语义分析器，以 plug-in 程序的方式处理语音输入输出。Issar 认为，这种基于表格的语音接口是探索分布式自然语言系统的重要一步。

机器人平台：Jijo-2[4]是一个可移动的办公室机器人，能通过语音对话的方式，完成人员查询、引路、接通特定人电话、给特定人发电子邮件等任务。这类机器人平台的系统面临的主要问题是实际使用环境中的噪音，以及系统响应的实时性。

电话平台：随着大量公有信息的出现（订票、信息查询等）以及电话的普及，基于电话的对话系统越来越多。欧洲的 ARISE 计划[5]下有法语、荷兰语、意大利语等若干系统，Els 等研究人员对各系统进行了横向比较，有助于找到不同方法的优缺点和提高研究水平。基于电话平台的系统应用前景广阔，有很好的社会效益和经济效益，其技术挑

战主要在于电话信道的窄带特性、信道之间的差异，以及现实生活中的噪音问题。

以下是国内外一些对话系统的简介。

（1）麻省理工学院的 GALAXY 系统[6]。这是一个通过口语对话获取旅游信息的系统，有大约 1500 个词的词汇量，能够提供大约 750 个城市的天气预报和大约 250 个城市的航班情况。它的语音识别器 SUMMIT，采用基于分段（Segment-Based）的识别方法，建立了 Anti-Phones 模型，词识别率为 83.9%；它直接采用了另一个对话系统 TINA[7]的自然语言理解模块，用语义框架的结构来描述语义；自然语言生成为 GENESIS；语音合成采用 Off-the-shelf 的硬件和软件。该系统的第二代 GALAXY-II[8-10]采用了 client-server 体系结构，成为美国 DARPA Communicator Program 的第一个参考体系结构。GALAXY-II 系统作为发展人类语言技术的试验平台，在其基础上，已经开发了许多不同领域、不同语言的系统，如电话天气预报查询系统 JUPITER[11]，航班订票系统 MERCURY[12]。

（2）德国的 VERBMOBIL 系统。这个对话系统用于会议的安排，可以识别并翻译大量的不同口语表达。它通过一个动态建立的上下文模型和一个建立在语料库之上的随机模型，可以预测对话某一点的下一句将会是什么。

（3）由英德法意等国共同开发的 SUNDIAL 系统[13]。这是一个提供航班和火车时刻信息的电话口语对话系统。它的词汇量为 1000 词左右，是非特定人的系统，而且具有很好的对话管理功能，通过电话进行的对话成功率达到 96%。

（4）中国科学院自动化所模式识别国家实验室的 LODESTAR 系统[14]。该系统向用户提供旅游信息，并且可以根据用户的要求计划旅游路线。它采用了大词表连续语音识别的技术，识别结果经过语义项的匹配得到有关的语义概念。它实现了对话的人机混合主导，基于模板生成系统应答，整个系统的应答准确率达到了 90.9%。

（5）清华大学智能技术与系统国家重点实验室语音技术中心的 Easy Nav 系统[15]。该系统向用户提供友好的清华大学校园导游服务，包括校园内的建筑物信息和交通信息。它考虑了口语中的省略指代现象，能处理上下文相关的对话。当信息查询结果为空时，该系统还会主动放宽某些约束条件，提供用户可能关心的信息。

总之，基于规则的口语对话系统得到了学术界的认同，并获得了很大的发展。虽然近年来也出现了基于统计的方法，但鉴于自然语言深层结构的规律性，规则方法有着统计方法不可替代的优势。众所周知，基于规则的口语对话系统中文法规则的获取是系统

研发的主要瓶颈。如何突破该瓶颈，提高系统研发效率，降低研发成本是口语对话系统研发中一个亟待解决的问题。这正是本文工作的意义所在。

1.2.2 文法规则自动推导算法研究现状

（1）有指导的学习方法

指从给定的树库（具有句法结构的语料库）中推导出句法结构知识(或文法)的方法。Brill 的基于变换的错误驱动方法[16]、Pereira 的 Inside-Outside 方法[17]和清华大学的苑春法、陈刚等提出的基于词性和语义知识的汉语文法规则学习方法[18]都属于这类方法。

（2）无指导的学习方法

指直接基于原始或者初级加工的句子，不使用人工加工后的结构信息或结构规则推导文法规则。这种方法可分成两类：

a) 基于压缩的方法。压缩方法实际上是提取“公因子”，将多次出现的多词词串代之以“成分(或称为非终结符)”。比较典型的有 Grunwald 的最小描述长度(MDL)方法[19]和 Wolff 的最小长度编码(MLE)方法[20]。但已有的研究表明，单纯的压缩方法在文法推导中并不能达到很好的效果。一个直接的原因是，貌似“公因子”的词串，实际上并不一定能够抽象为成分。

b) 基于分布的方法。按照 Harris 等语言学家的基本思想，当两个不同的词串所在的上下文具有一致的分布特点时，它们很可能就具有了可替换的特点。此时，可以将两个不同的词串用一个非终结符表示。分布方法可以分为局部分布和全局分布两种：局部分布只考虑某个词序列前后相邻的词的特征。如 Stanford 大学 Klein 和 Manning 的工作[21-22]，他们以句子的词性标注序列作为输入，通过对词性(序列)的上下文(主要是相邻的词)信息来判断两个词是否有相似。他们研究了依存结构和成分结构树的推导，分别对英语、德语和汉语进行了测试。英国 Sussex 大学的 Clark 用到了与此类似的思想[23]，在带有词性标注的语料基础上，根据词性的上下文分布将其聚类为非终结符，推导文法规则。Clark 在处理过程中结合了 MDL 方法。他们的方法对英语测试也取得了较好的结果。局部分布的最大特点是只考虑前后相邻的信息，在语料库不是非常庞大时比较适用；但在一个较小的窗口内，所得到的信息毕竟不够充分。例如，在英文中，“IN(介词)+DT(冠词)+NN(名词)”的模式，很可能将 IN+DT 归约一个结构(互信息值可能更大)，而实际情况应该是由 DT+NN 先结合。扩大词的左右窗口范围，在一定程度上可以避免这

一问题，在极端情况下，可以将范围扩展到整个句子。荷兰 Amsterdam 大学的 Adriaans 设计的 EMILE 系统[24]和英国 Leeds 大学 Zaanen 的基于对齐的学习都是以整个句子作为考察对象的[25-26]。EMILE 的基本思想是将一个句子看成 3 部分：cl+e+cr，cl 在 e 的左部，cr 在 e 的右部，称为 e 的上下文。对于一个句子，e 可以取其中的任何词串，剩下的部分就形成其上下文。在文法推导时，从句子库中抽取所有可能的模式，然后再进行聚类。而 Zaanen 的思想与 Bilkent 大学的 Cicekli 等人在翻译模板提取中的思想有很大的相似性[27]，都通过多个相同片段和不同片断交错对齐的基本方法，只是 Zaanen 进一步推导出了句子的层次结构。Zaanen 研究了英语句子结构的推导，在结构推导中，不对英语句子作任何其他预处理（如词性标注）。这种思想虽然易于实现，但如果词的词性兼类现象比较严重，而训练语料又不足够大，即使是找到了对齐，也不一定能保证是正确的对齐。如果事先对句子作适当的预加工(如词性标注和简单的语义归类)，并加入一定的对齐约束(如词性约束)，则是可以减少明显不合理推导现象发生的。

c) 一些新的思路。Tokyo Denki 大学的 Nakamura 采用一种新的思路，先构造正例集和反例集，在已有的小规模初始规则集上，用分析算法分析正例，添加新的规则，分析反例，抑制不合理的规则[28-29]。香港中文大学的 Helen Meng 等借鉴了语音识别中语言建模的思想，用统计的方法对训练语料中的词和句法结构进行聚类，若干次迭代后得到初步的上下文无关文法，再人工用语义标记代替文法中随机的类别标记[30]。清华大学的刘智博提出了一种基于主题的方法[31]：首先把领域知识划分为若干个主题，表达相同语义的不同句子属于同一个主题。算法根据预先定义好的关键词表，把某一个主题下可能的用户查询例句转化为由语义关键词类表示的模板，应用于相应的主题。这种方法得到是单句模板，而且算法需要先由人工将例句划分为不同的主题，然后才能对不同的主题分别进行处理。

d) 有关汉语的方法。汉语与西文有着不同的语言特点，处理方法也存在着较大的差异，随着汉语热的兴起，针对汉语的文法规则自动学习研究开始逐渐受到学术界的重视，主要的研究有：北京大学的王厚峰和王波设计了基于句子对齐的汉语句法结构推导的计算模型[32]。清华大学的周强、黄昌宁两位教授提出了基于元规则的汉语文法规则的自动构造方法[33]。

综合国内外文法规则自动推导的研究，我们可以发现，大部分工作以理论研究与探讨为目的，针对口语对话系统的、符合口语对话系统中语言特点的研究并不多见。因此

本文工作不仅具有很好的应用价值，而且具有一定的理论意义。

1.2.3 文法性能的评测

通过分析对比国内外相关研究中文法评测方法，可以看到，学术界主要考查文法的复杂程度（生成的规则数目及新添加的非终结符数目）和算法的时间消耗[34-38]。香港中文大学的 Helen Meng 在常用文法的基础上，结合自身算法特点评测了参数的不同取值对最终生成的文法的影响[30]。北大计算语言学研究所的王厚峰在评测中将自动推导的文法与手工标注的文法相比较，使用文法的准确率、召回率、F 值，对评测算法输出的文法规则[32]。综观这些研究，还没有一套针对领域任务需求的文法评测方法。

1.3 研究的主要内容及创新点

1.3.1 研究的主要内容

口语对话系统中文法规则的获取面临着几个主要问题：一是需要依赖领域专家和计算机语言学家专业的知识和经验，对于普通开发人员来说是无法完成的；二是采用手工方式进行，是一个相当繁琐的过程；三是文法移植性差，根据某一领域获取的文法规则很难移植到其他领域继续使用，同样功能的对话系统对于不同领域都需要领域专家重新设计文法规则。随着社会需求的不断增大，系统规模空前扩大，文法规则的获取成了对话系统研发的主要瓶颈。

针对口语对话系统中文法规则的获取，研究符合汉语口语特点的文法规则自动推导技术，研究内容如下：

（1）文法的基本概念，包括文法及语言的定义、常见的文法分析算法，分析几种常见文法的特点及优劣；重点研究一种针汉语口语特点的上下文无关增强文法，主要包括增强属性的归纳、增强规则类型形式化定义及增强文法分析算法。

（2）根据汉语口语的特点，以一种符合汉语口语特点的上下文无关增强文法为对象，研究口语对话系统中文法规则的自动推导算法。提出一种基于句子分割的文法规则自动推导算法，给出算法的形式化描述、具体步骤、片断的消歧和归一化方法，探讨并改进算法流程，以提高输出文法的性能。

（3）面向领域的文法性能评测方法。提出适用于领域需求的、可定制的文法性能评测指标。并使用这种评测方法，在天气预报查询领域对算法的输出文法进行了评测。

1.3.2 研究的创新点

（1）提出一种文法规则自动推导算法，无需人工参与的情况下，快速地从语料中学习文法规则，有助于突破口语对话系统开发的瓶颈。

（2）自动推导算法符合汉语口语特点，输出文法基本能够覆盖常见的汉语口语现象，适应了口语对话系统的需要。

（3）提出一套面向领域任务的文法性能评测方法，能够根据领域特点定制评测参数，客观地评价文法的性能。

1.4 论文的组织

本文后面的篇幅首先在第二章介绍有关文法的基本概念，包括文法及语言的定义，对比几种不同分析方法的优劣；重点介绍上下文无关增强文法——一种针对口语特点的文法，作者将以该文法为对象研究口语对话系统中的文法规则自动推导技术。第三章针对汉语口语特点，给出符合汉语口语特点的文法规则自动推导算法，包括算法基本原理、处理框架、几种不同的推导策略和片断的消歧及归一化方法，探讨并改进算法的处理流程，提高输出文法法的性能。第四章研究文法规则评测方法，给出面向领域需求的文法规则评测方法，并运用此方法对自动推导算法的输出文法进行性能评测，分析算法特点及优劣。第五章总结全文工作，指出今后的研究方向。

第二章　文法的基本概念及常见类型分析

基于规则的语言理解的核心思想是用文法来描述语言、分析语言。自 Chomsky 于 1957 年创立转换-生成语法体系[39]以来，基于规则的语言理解方法得到了语言学界的认同，并获得了很大的发展，特别是语言学与计算机结合形成计算语言学(Computational Linguistics)之后，它在自动自然语言系统中得到了广泛的应用。虽然近年来也出现了基于统计的理解方法，但鉴于自然语言深层结构的规律性，规则方法有统计方法不可替代的优势。

2.1 文法的基本概念

2.1.1 文法及语言的定义

一个文法 G 是一个四元式[40]

$$G=(V_T,V_N,S,P) \tag{2-1}$$

其中V_T是一个非空有限集，它的每个元素称为终结符。所谓终结符是组成语言的基本符号，从语法分析的角度来说，可以说终结符是一个语言不可再分的原子符号。

V_N是一个非空有限集，它的每个元素称为非终结符。非终结符是语法范畴，它代表一定的语法概念，每个非终结符也表示一定符号（包括终结符和非终结符）串的集合。

S 是一个特殊的非终结符$S\in V_N$，也称为起始符号。它代表所定义语言中的“句子”，也就是我们最终感兴趣的语法范畴。此外可定义空符号ε。

P是一个有限产生式集合，每个产生式的形式是$\alpha\to\beta$，其中$\alpha\in(V_T\cup V_N)^*$且至少含有一个非终结符，$\beta\in(V_T\cup V_N)^*$,S必须至少在某个产生式的左部出现一次。产生式是定义语法范畴的一种书写规则。

有了文法的定义之后，下面定义由文法如何生成语言。

如果$A\to\gamma$是一个产生式，且$\alpha,\beta\in(V_T\cup V_N)^*$，称$\alpha A\beta$直接推出$\alpha\gamma\beta$，记作$\alpha A\beta\Rightarrow\alpha\gamma\beta$。如果有$\alpha_1\Rightarrow\alpha_2,\alpha_2\Rightarrow\alpha_3,\cdots,\alpha_{n-1}\Rightarrow\alpha_n$，则称$\alpha_1$可推导出$\alpha_n$。用$\alpha_1\overset{+}{\Rightarrow}\alpha_n$表示从$\alpha_1$出发经过一步或若干步可推导出$\alpha_n$；而用$\alpha_1\overset{*}{\Rightarrow}\alpha_n$表示从$\alpha_1$出发经过零步或若干步可推导出$\alpha_n$。换言之$\alpha\overset{*}{\Rightarrow}\beta$表示：或者$\alpha\Rightarrow\beta$，或者$\alpha\overset{+}{\Rightarrow}\beta$。

假定G是一个文法，S是它的起始符号，如果$S\overset{*}{\Rightarrow}\alpha$，则称$\alpha$是一个句型，如果$\alpha$只含终结符，则称$\alpha$是一个句子。文法$G$所产生的句子的全体是一个语言，记为$L(G)$：

$$L(G)=\left\{\alpha \mid S\overset{+}{\Rightarrow}\alpha,\alpha\in V_T^*\right\} \tag{2-2}$$

2.1.2 文法的作用

现实世界的知识，人们往往试图通过逻辑框架去表示它们。知识表示有多种方法，比如逻辑表示法、产生式表示法、语义网络表示法、框架表示法及面向对象的表示法[41]等。广义地说，自然语言中的句子结构，以及相应的人的语言认识模型，也属于知识表示所要解决的问题。本节讨论的文法体系，本质上也是一套描述句子结构的知识表示方法。上下文无关文法描述句子结构本质上是一种树型的知识表示，这与各种知识表示方法及人类对其它事物的认识方法具有共同之处。

这种表示方法基于这样的考虑：句子可以按照一定的准则分成几个相对独立的子部分，这些子部分又可以根据类似的别的准则进行细分，直至达到所有子成分均不能再分的时候为止。

文法的作用，特别是上下文无关文法的作用，就是通过描述待研究语言的产生式规则，一方面在语言理解系统中，对于给定的输入句子，判定其相对于该文法的合法性，给出合法句子的句法结构；另一方面在语言生成系统中，根据要表达的概念，生成符合规范的自然语言句子。

下面的例子可以说明用上下文无关文法所表示的句子结构；同时该例也能说明文法的经验性概括思路。给定如图2-1所示的文法，对于“John ate the cat.”这句话来说，可以很容易地得到如图2-2所示的句子结构（句法树）；相反，如果根据传统文法得到如图2-2所示的句子结构图，也很自然地能够概括出如图2-1所示的上下文无关文法。图中，每个非叶子节点均被称为一个成分－Constituent，即分析过程中的非终结符实例。

1. S → NP VP	5. NAME → John
2. VP → V NP	6. V → eat
3. NP → NAME	7. ART → the
4. NP → ART N	8. N → cat

图 2-1 一个简单的文法

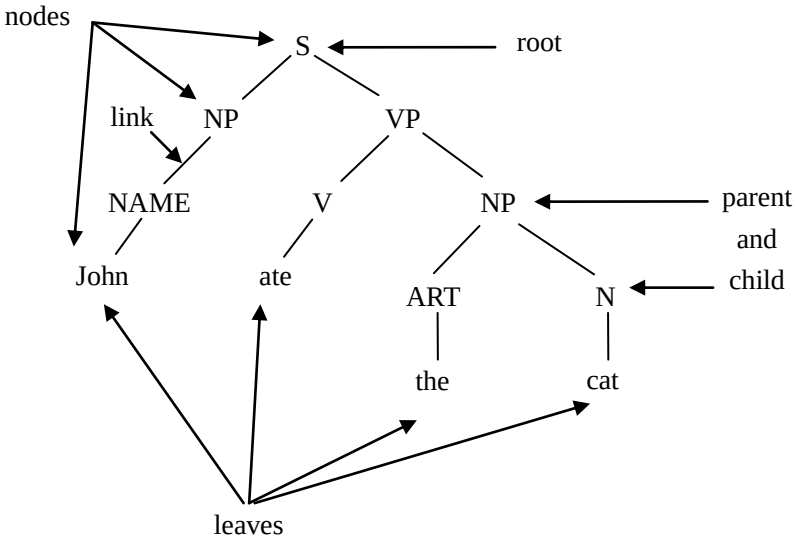

图 2-2 句子的树型表示[40]

2.1.3 文法的评价原则

文法的评价通常考虑以下几个方面[40]：

（1）一般性(generality)，指文法所能正确分析的句子范围。本文后面也用准确率来表示这个概念。对于特定系统，文法的一般性大到能够满足任务需要。

（2）选择性(selectivity)，指文法认为非句子的符号串范围。它跟Generality是互补的。

（3）可读性(understandability)，指文法本身的简单程度。良好的可读性能方便文法的移植和继承。

（4）过度生成(over-generation)。一般性提到文法能够正确分析的句子范围，但这不表明这些句子就是自然语言中合法的句子，那种自然语言中错误句子通过文法分析器的现象就称为过度生成。在追求良好的可读性的同时，往往会带来过度生成的问题。实际应用中，需要在两者间作适当折中。

2.2 Chomsky 文法体系

2.2.1 文法分类

Chomsky把文法分为四种类型，分别是0型、1型、2型和3型，它们的关系是序号小的文法的限制比序号大的文法的限制弱，从而前者的描述能力比后者的强。这四种文法构成形式语言理论中的Chomsky体系[41]。

（1）0型文法

2.1.1中关于文法的定义，如果其中产生式的重写规则不附加任何限制，则称它是一个0型文法。它是Chomsky体系中生成能力最强的文法。由这种无约束文法所定义的语言，相应地称为0型语言。它是一种可递归枚举的语言[41]。

（2）1型文法

对于文法G中的任意一个产生式$\alpha \to \beta$，如果仅要求$|\alpha| \leqslant |\beta|$，其中$|\alpha|$表示符号串$\alpha$的长度，则称该文法是一个1型文法，其生成的语言称为一个1型语言。另一种对1型文法的描述是，每一个产生式用$\alpha A\beta \Rightarrow \alpha\gamma\beta$来表示，其含义可以这样表达：只有$A$ 在上下文$\alpha _ \beta$的条件下，才能改写或被替换成γ。因此1型文法也被称为上下文有关文法。

（3）2型文法

对于文法G，如果它的任意产生式满足的$A \to \beta$形式，其中$A \in V_N$，$\beta \in (V_N \cup V_T)^*$，那么称$G$是一个2型文法，其生成的语言称为一个2型语言。直观地说，2型文法要求每一个产生式的左部是一个单独的非终结符。相对于1型文法，2型文法的推导不要求依赖于特定的上下文，因此这种文法也称为上下文无关文法。

（4）3型文法

对于文法G，如果要求它的任意产生式满足$A \to \alpha$或$A \to \alpha B$的形式，其中$A, B \in V_N$，$\alpha \in V_T^*$，则称G是一个左线性文法。而如果要求它的任意产生式满足$A \to \alpha$或$A \to B\alpha$的形式，其中$A, B \in V_N$，$\alpha \in V_T^*$，则称G是一个右线性文法。左线性文法和右线性文法分别是3型文法（或正则文法）的两种定义方式，可以证明它们是等价的。3型文法生成的语言称为3型语言。

2.2.2 各型文法的特点

正则文法在Chomsky体系中生成能力最弱，以右线性文法为例，可以这样设想，生成器每生成一个终结符后，根据产生式规则右部的第二个符号，紧接着扩展下一个非终结符，这样递归直至生成一个合法的句子。由此可见，正则文法的描述能力和确定状态有限自动机是等效的，因此这样的文法也可以称为有限状态文法(Finite State Grammar-FSN)。

正则文法的这种特性，使得生成或分析时的计算速度极快，因为它在当前状态已知的前提下可精确预测下一个状态。虽然正则文法有着计算优势，但是正则文法不能描述

自然语言中的一些常见句子，比如$S=(a\cdots abc\cdots c)$，其中a和c的个数不定，但个数相等。

上下文无关文法生成能力强于正则文法，比如上段中的例句即可用上下文无关文法加以描述。一般说来，针对现实世界中的任何一种自然语言，为其设计的上下文无关文法，可以做到覆盖该语言中的绝大部分句子构成的子集（不可数），因此目前大多数自然语言系统仍然选择上下文无关文法作为其理解工具。当然也有人认为，从理论上讲，上下文无关文法并不能完全描述自然语言，这个问题有待语言学家进一步研究，不在本论文讨论范围之内。

至于上下文有关文法及0型文法，其描述能力强于上下文无关文法，但实际系统中很少见，人们更多地只是在理论比较时谈到它们。

因此，在使用基于规则的语言理解研究中，选择什么样的文法作为工具，有两个参考点，一是其描述及生成能力，是否能够胜任特定任务对文法的要求；二是其复杂度，该文法是否有有效的分析器，过于复杂的文法将不能满足实际任务对实时性的要求。鉴于这两点，人们大多选择上下文无关文法作为描述及分析的工具。

2.2.3 文法分析器

文法分析器用来根据文法判定句子的合法性并给出句法结构。分析算法可以这样描述：给定输入句子，在文法规则各种各样的组合方式之中，找出一种可能是该句子文法树结构的组合方式的搜索过程。包括两个目标，一是给出句子是否被文法所接受，二是如果被接受，则给出句法结构。针对上下文无关文法，主要有两种类型的文法分析器，一类是自顶向下的分析算法，另一类是自底向上的分析算法。

（1）自顶向下的分析算法

简单地说，自顶向下算法的思路是：从文法的起始符 S 出发枚举文法中的规则，对当前状态中的非终结符进行推导，直至所有非终结符均已被重写成终结符，且终结符串与输入句子的词类全部匹配成功为止。算法过程中的分析状态，是指当前时刻前所有扩展操作形成的符号串结果，也就是文法定义中提到的句型[40]。

该算法用到三个术语：**可能状态列表**，它的第一个元素是**当前状态**，其余元素为**备份状态**。算法从开始状态((S)1)出发，并且不含备份状态：

1）如果可能状态列表为空，则算法失败退出；否则选取其中第一个状态C作为当前状态，并将其从可能状态列表中删去。

2）如果C包含空符号串，并且分析位置是句末位置，则算法成功退出。

3）否则根据下以下三种情况分别处理：

a）如果C中的第一个符号是终结符（词法符号），并且下一个词属于这个符号代表的词类，则把C中的第1个符号删去，更新分析位置，将新状态加入到可能状态列表中去；

b）如果C中的第一个符号是终结符，但下一个词不属于这个词类，则不做任何操作；

c）如果C中的第一个符号是非终结符，则枚举文法中所有可用的规则对该终结符进行重写，并将这些新状态加入可能状态列表中去。

4）跳至第1步。

可以看出，第1步总是选择第1个状态作为当前状态，但在第3步把新状态加入到可能状态列表中时，有两种选择，一是加到可能状态列表的后端，二是加到可能状态列表的前端，这就形成深度优先搜索和广度优先搜索两种策略。

自顶向下分析算法的特点：

1）在针对当前状态匹配下一个符号时，具有比较高的预测性，不会对输入词的多种词类作无用扩展。

2）当扩展和匹配失败时，需要回溯，此时曾经分析过的成分会被多次重复分析，效率不高。针对回溯问题，有一些改进算法，比如采用有向前看几个符号的算法，可在大多数情况下避免回溯，但理论上不能保证完全避免。

3）对待分析句子做出接受或拒绝，对于失败的句子给出的信息量太少。

（2）自底向上的分析算法

自底向上算法的思路是，从输入句子的词类出发，对相邻符号串进行归结，生成对应规则的左部符号，直至最终生成文法起始符号S。也可以这么理解，自顶向下的分析算法，是从句法树的根节点开始向叶节点，即输入句子的词类串，进行推导；而自底向上的分析算法，则是从输入句子的词类串开始，向树的根节点进行归结。

具体地说有两点：

1）将输入词重写成词类，即终结符；

2）如果一个符号串匹配上了某一条规则的右部符号串，则将该符号串用这条规则的左部符号代替。

直接按照上述方法去做，是相当耗时的，因此必须提出高效的分析算法。图表分析

器，即Chart Parser，就是这样的自底向上算法的典型代表[40]。

Chart Parser涉及到三个主要的数据结构：

1）图表－chart，它是存放当前所有已经分析得到的部分结果的数据结构，通过这个机制，可以避免已有的成分被多次地归结，实现共享。

2）活动弧－active arc，指当前已经扩展了一部分但仍没有得到最后归结的规则实例。它的表示方法与规则类似，但需在右部符号间插入一个圆点，指示下一步的匹配位置。比如$NP \rightarrow ART \circ ADJ\ N$这条活动弧，它指示下一个待扩展的符号是*ADJ*这个终结符。

3）议程表－agenda，新归结得到的成分存放在agenda中，直到它们均已被处理（被扩展）为止。正像自顶向下的分析算法一样，Chart Parser也有两种搜索策略，即深度优先和广度优先，当agenda为先进先出栈（FIFO）时，为深度优先搜索，当agenda 为先进后出队列（FILO）时，则为广度优先搜索。

Chart Parser 算法的思路：

1）从agenda中取出一个成分，称为当前成分；

2）在文法中查找以当前成分为第1个右部符号的规则，生成相应的匹配位置为1的一个活动弧；

3）枚举所有以当前成分为下一个匹配符号的活动弧，生成新的活动弧，并递进匹配位置；

4）对于以当前成分为最后一个匹配符号的活动弧，归结生成以该活动弧左项符号为符号的新成分，放入agenda中；

5）重复上述过程，直至agenda 为空为止。

自底向上分析算法的特点：

1）预测性不够，当一个新的成分生成时，均需要在文法中查找相应的规则以生成活动弧，而不管该活动弧今后能否被扩展。

2）无需回溯，对输入串仅作一遍扫描。中间生成的任何成分，均不会在以后的分析中被再次生成，实现了成分的共享。

3）不是对输入词串仅做出接受或拒绝，而是保留所有局部分析结果，因而即使对于失败的句子也能给出一定量的信息。

（3）两种方法的结合

通过分析，自顶向下的算法和自底向上的算法各有优劣，所以两者的结合将会带来一定好处。自顶向下的Chart Parser[40]就是其中有代表性的例子。

自顶向下的Chart Parser算法的主要思路是，在生成任意一个活动弧时，该活动弧匹配位置如果是一个非终结符，则连带生成所有以该非终结符为左部符号的匹配位置为1的新活动弧。这样做的好处是，不会在归结出一个符号时，生成一些以后不会被扩展到的无用活动弧；同时分析结果也能得到共享，避免了多次生成的问题。可以说结合了预测性和成分共享的优点。但是它不具备自底向上分析算法的保留所有局部分析结果的特点，而这些局部分析结果可能是有用的。

2.3 上下文无关增强文法

口语语言理解是一个口语对话系统中最重要的组成部分，其性能的好坏对对话系统的性能有关键性的影响。尽管有将统计知识用于语言理解的方法出现，但目前最为常见的口语语言理解的方法仍然是基于规则的分析方法。

众所周知，与书面语不同，口语对话系统中用户的语句是很随意的，其中充满了垃圾、碎片、犹豫、纠正、重复、省略、词序混乱和病句等现象。汉语的表意性及口语现象问题，是汉语口语对话系统中语言理解所面对的重要课题，清华大学的燕鹏举提出了上下文无关增强文法的概念，使用语义符号来编写文法，对解决这一问题具有较好的效果。本节将重点介绍上下文无关增强文法的定义及其分析算法。

2.3.1 文法的形式化定义

[定义 2-1]上下文无关增强文法：一个上下文无关增强文法 G 是一个四元式[1]：

$$G=(V_T,V_N,S,P) \tag{2-3}$$

其中 V_T、V_N 是两个非空有限集，V_T 的每个元素称为终结符，V_N 的每个元素称为非终结符。$S\in V_N$ 是一个特殊的非终结符，也称为起始符号。P 是一个有限的产生式集合，每个产生式的形式是 $A[rule\_type]\rightarrow\beta$，其中 A 是一个非终结符，称为规则左部符号；S 必须至少在某个产生式的左部出现一次；$\beta\in(V_T,V_N)^*$，称为规则右部，其中各个符号称为规则右部符号；*rule_type* 为增强属性，作为可选项置于规则产生符 $\rightarrow$ 之前。图 2-3 从文法的书写角度给出了其描述，该描述本身是一个传统的上下文无关文法实例。

rule_text → *rule_list*
rule_list → *rule* | *rule rule_list*
rule → *symbol [rule_type] ’* → *’symbol_list*
symbol_list → *symbol* | *symbol symbol_list*
symbol → *symbol_prefix* | *symbol_prefix symbol_suffix*
symbol_prefix → *alphabetic*
symbol_suffix → *alphanumeric* | *alphanumeric symbol_suffix*
alphanumeric → *alphabetic* | *numeric*
alphabetic → ’_’ | ’a’ | ’A’ | ’b’ | ’B’ | … | ’z’ | ’Z’
numeric → ’0’ | ’1’ | … | ’9’
rule_type → ’*’ | ’@’ | ’#’ | ’~’

图 2-3 上下文无关增强文法编写的形式化描述[1]

从定义2-1及图2-3中可以发现，增强文法与传统文法的主要不同，在于对规则附加了增强属性，即用*rule_type*的标识的单元。但文法本身并不描述增强属性的功能，需要从实际应用的角度进行归纳，这使得文法具有一般性及功能可扩展性。

2.3.2 增强属性的归纳及规则类型的定义

根据口语对话系统中用户语言的特点，文献[1]对增强属性做如下归纳：

（1）对于出现于成分中间的口头习语、礼貌用语、声学垃圾、语言垃圾或识别错误等，如果在归结时能够以跳过一部分的形式越过它们，则这部分问题可以得到解决。

（2）对于口语中短语和其它成分以任意顺序的出现的问题，如果对不同顺序的组合用一条规则来描述，在归结时不考虑规则右部符号在时间上的先后顺序，那么这个问题也可以得到解决。

（3）一些概念（这里概念可以定义成与任务相关的最小语言单元）或“短语”在空间上有着长程关系，比如“有……吗”和“是……吗”，与第1种情形不同，它们中间所跨跃的部分在句意上是至关重要的，而且其跨跃范围也相对较长，那么如果有一种规则属性使其可在这种情况下进行长程归结，那么这种概念或短语的检出也能解决。

（4）涉及长程型概念的组合，由于概念在空间上的跨跃性，归结时不能象传统分析方法那样要求子成分在空间上是不交叉的，这时需要在规则属性上加以体现。

（5）相对于上述功能来说，要有一种规则属性能够描述传统上下文文法所能描述的现象，即要求子成分间在空间位置上紧密相连。

因此，规则的增强属性在本文中具体化为规则类型，我们可以得到五种类型的规则，分别是苛刻型（*up-tying*）、跳跃型（*by-passing*）、长程型（*long-spanning*）、无序型

（*up-messing*）以及交叉型（*over-crossing*）。

在说明这5种类型的规则定义之前，先介绍一些相关的术语[1]。

[定义2-2]句子-Sentence：一个句子是一个由语言中的基本单元组成的串，这里的基本单元指某种意义上的词类（包括补白类）。可以写成 $sent=(K_0,K_1,\cdots,K_{n-1})$，其中$n$表示句子长度，$K_i$ 是第i个词（包括补白）的词类，$0\le i\le n$。同时 K_i 也是文法的终结符。

[定义2-3]位置-Position：句子 $sent=(K_0,K_1,\cdots,K_{n-1})$ 中终结符 K_i 的位置定义为 $P_{K_i}=[i,i+1)$，其中$0\le i\le n$。而非终结符C的位置定义为$P_C=[p_1,p_2)$，其中 $p_1=\min\{b_i\}$，$p_2=\max\{e_i\}$，这里$[b_i,e_i)$ *is the position of a leaf node of* C，$0\le i\le n$。

[定义2-4]占位-Occupation，占位的冲突或交叠：句子 $sent=(K_0,K_1,\cdots,K_{n-1})$ 中终结符 K_i 的占位定义为集合 $O_i=\{i\}$。而非终结符C的占位定义为集合 $O_c=\{o\mid o \text{ is the occuptation of a leaf node of } C\}$。称两个占位$O_1$与$O_2$相互冲突，当且仅当 $(\exists l\in O_1,\exists m\in O_2,l=m)$；称 O_1 与 O_2 相互交叠，当且仅当 $(\exists i,1\le i\le 2,(\exists l,m\in O_i,\exists n\in O_{2-i},(l<n<m)))$。

如果一个文法规则形如$Y[rule_type]\to Y_0Y_1\cdots Y_{m-1}$，$C_j$是$Y_j$的一个分析时实例成分，$C_j$的位置为$P_j=[p_{j,1},p_{j,2})$，占位为$O_j$，其中$0\le j<m$，那么上述5种规则类型可定义如下[1]。

[定义2-5]苛刻型(*up-tying*)规则：称$Y[rule_type]\to Y_0Y_1\cdots Y_{m-1}$为一个苛刻型的规则，如果 $rule\_type=$'e'，对于所有 $\forall 0\le j<m$，C_j 均不是交叉型成分，并且 $(\forall 1\le j<m-1,p_{j,2}=p_{j+1,1})$。此时称$C$为一个苛刻型成分。

[定义2-6]跳跃型(*by-passing*)规则：称$Y[rule_type]\to Y_0Y_1\cdots Y_{m-1}$为一个跳跃型规则，如果 $rule\_type=\varnothing$（空），对于所有 $\forall 0\le j<m$，C_j 均不是交叉型成分，并且 $(\forall 1\le j<m-1,0\le p_{j+1,1}-p_{j,2}\le thres)$，其中$thres$是一个预先给定的阈值。此时称$C$为一个跳跃型成分。

[定义2-7]长程型(*long-spanning*)规则：称$Y[rule_type]\to Y_0Y_1\cdots Y_{m-1}$为一个长程型规则，如果 $rule\_type=$'~'，对于所有 $\forall 0\le j<m$，C_j 均不是交叉型成分，并且 $(\forall 1\le j<m-1,p_{j,2}\le p_{j+1,1})$。此时称$C$为一个长程型成分。

[定义2-8]无序型(*long-spanning*)规则：称$Y[rule_type]\rightarrow Y_0Y_1\cdots Y_{m-1}$为一个无序型规则，如果$rule_type=='@'$，对于所有$\forall 0\le j<m$，$C_j$均不是交叉型成分，并且$\forall 1\le j,k<m, j\ne k$，$O_j$和$O_k$均不互相交叠。此时称$C$为一个无序型成分。

[定义2-9]交叉型(*long-spanning*)规则：称$Y[rule_type]\rightarrow Y_0Y_1\cdots Y_{m-1}$为一个交叉型规则，如果$rule_type=='\#'$，对于所有$\forall 1\le j,k<m, j\ne k$，$P_j$和$P_k$均不互相冲突。此时称$C$为一个交叉型成分。

通过考察上述定义，各种类型规则的功能可以直观地描述为：（1）苛刻型规则就是传统方法的规则，表示连续句法成分组成更高级成分的关系。（2）跳跃型规则允许组成更高级成分的各句法成分之间有少量的其它成分，用以解决口语中出现的停顿和无意义语气词现象。（3）长程型规则允许组成更高级成分的各句法成分之间有任意数目的其它成分，用以解决口语中出现的重复和修正现象。（4）无序型规则允许组成更高级成分的各句法成分先后顺序任意，用以解决汉语口语语序随意的问题。（5）交叉型规则允许组成更高级成分的各句法成分在句子中占位交叉但不冲突，且各句法成分先后顺序任意，用以解决汉语中某些特定句型，如“有……吗？”和“是……吗？”。

2.3.3 语义文法

Chomsky的形式语言理论基本思路是：句子由短语和词组成，短语由短语和词组成，而词可以按功能分成不同的词类，因此可以由词类的组合方式来描述句子的结构。相对于印欧语，语言学界认为汉语语法的主要特点有[42]：(1)缺乏严格意义的（狭义的）形态变化。这是最重要的一个特点，汉语语法的其它特点都跟这点有关。形态指词的单复数、人称、性等变化形式。(2)词类和句法成分间的关系错综复杂。比如动词、形容词可作主语和宾语，形容词可作谓语和状语，名词可作定语，一定条件下还可作谓语等。(3)语序显得比较重要。汉语的词缺乏形态变化，句意主要靠语序的不同来体现。如果说英语（或其它拼音语言文字）是一种良好的结构化的语言[43]，那么汉语在很大程度上，则是一种结构化不明显的表意语言。

因此，文献[1]舍弃于词类的句法文法，用语义类来编写文法，在特定领域任务的汉语对话系统中能表现出好的性能。

语义文法所涉及的终结符和非终结符在不同的对话系统中会随着领域的不同而不同，但有如下共同规律：

(1)语义原子作为终结符。语义原子指领域中不可再分的知识或信息的最小载体。终结符实际上是领域内的关键词类，一个关键词类内含有一个或多个关键词，所有关键词组成系统词表。

(2)较小的语义单元按领域内固有关系组成较大的语义单元，作为非终结符。

(3)具有独立完整交际功能的语义单元成为句子，也即文法所能描述的最高级别符号。

2.3.4 增强的文法分析器

2.2 节介绍的分析算法均是针对传统上下文无关文法提出的，不能用以分析这里的增强型上下文无关文法，需要选取合适的传统分析算法作为基础，针对增强型上下文无关文法的特点进行增强与改进。如 2.2 节介绍，上下文无关文法的分析算法主要有三种策略，一是自顶向下，二是自底向上，三是两者一定程度上的结合，这 3 种策略各有优劣。

对话语音的自发性及口语现象，不仅要求文法描述能力的提高，也要求在分析算法上加以考虑，以最大限度地利用可能得到的信息。尽管上下文无关增强文法提高了文法的覆盖度，但如果期望实际应用背景中的语音符合整句规则，仍然是不现实的。文献[1]对自底向上的 Chart Parser 算法进行了改进，设计了 Marionette 分析器，能够满足上下文无法增强文法的跨成分特性，并尽可能保留部分分析结果以备后用。算法描述如下图 2-4 所示。

进行以下步骤直至 agenda 为空：

1) 检查句子中的下一个词 *W*，根据其关键词类生成相应的成分 C_W，存入 agenda 中。将 C_W 的位置赋成 *W* 在句子中的位置（*p1,p2*）。对于兼类的情况，生成与所有词类相对应的成分。
2) 只要 agenda 不为空，循环取出 agenda 中第一个成分 *C*，进行下面两步：
   2a) 调用起始孤生成过程 Algorithm2.2；
   2b) 调用孤扩展过程 Algorithm2.3。

图 2-4 Marionette 算法总流程-Algorithm2-1

在给出子过程之前，先给出活动孤位置和占位的定义。

[定义2-10]活动孤的位置-Position：活动孤 $Y[rule\_type] \rightarrow Y_0 Y_1 \cdots Y_{l-1} \circ Y_{m-1}$ 的位置定义

为区间 $P_Y=[p_1=\min\{b_j\}, p_2=\{e_j\})$，$0\le j\le l$。

[定义 2-11]活动孤的占位-Occupation：活动孤 $Y[rule\_type]\rightarrow Y_0Y_1\cdots Y_{l-1}\circ Y_{m-1}$ 的占位定义为集合 $O_Y=\bigcup_{0\le j<l} O_{Y_j}$，其中集合 O_{Y_j} 是 Y_j 的占位，$0\le j\le l$。

起始弧生成过程Algorithm2-2如图2-5所示，其中起始弧指的是匹配位置为0的活动弧。弧扩展过程Algorithm2-3如图2-6所示。

对于位置为(p_1, p_2)的成分 C，以及文法中的任意规则 $X[rule\_type]\rightarrow X_1X_2\cdots X_n$：

1) 如果规则形如 $X*\rightarrow CX_2\cdots X_n$，则在($p_1$, p_2)位置添加一个苛刻型活动孤 $X*\rightarrow\circ CX_2\cdots X_n$；
2) 如果规则形如 $X\rightarrow CX_2\cdots X_n$，则在($p_1$, p_2)处添加一个跳跃型活动孤 $X\rightarrow\circ CX_2\cdots X_n$；
3) 如果规则形如 $X\sim\rightarrow CX_2\cdots X_n$，则在($p_1$, p_2)处添加一个长程型活动孤 $X\sim\rightarrow\circ CX_2\cdots X_n$；
4) 如果规则形如 $X@\rightarrow X_1X_2\cdots C\cdots X_n$，则在($p_1$, p_2)处添加一个无序型活动孤 $X@\rightarrow\circ X_1X_2\cdots C\cdots X_n$；
5) 如果规则形如 $X\#\rightarrow X_1X_2\cdots C\cdots X_n$，则在($p_1$, p_2)处添加一个交叉型活动孤 $X\#\rightarrow\circ X_1X_2\cdots C\cdots X_n$。

图 2-5 Marionette 中起始孤生成过程-Algorithm2-2

对于位置为(p_1, p_2)的成分 C，以及活动孤 $Y[rule\_type] \rightarrow Y_1 Y_2 \cdots Y_m$，如果 k<m-1：（其中 th1 和 th2 是两个预先认定的正阈值）

1) 对于所有位置为(p_0, p_1')的活动孤 $Y^* \rightarrow Y_1 Y_2 \cdots Y_k \circ C \cdots Y_m$，如果 $0 \le p_1 - p_1' < th1$，则在(p_0, p_2)位置处添加新活动孤 $Y^* \rightarrow Y_1 Y_2 \cdots Y_k C \circ \cdots Y_m$；
2) 对于所有位置为(p_0, p_1')的活动孤 $Y \rightarrow Y_1 Y_2 \cdots Y_k \circ C \cdots Y_m$，如果 $0 \le p_1 - p_1' < th2$，则在(p_0, p_2)位置处添加新活动孤 $Y \rightarrow Y_1 Y_2 \cdots Y_k C \circ \cdots Y_m$；
3) 对于所有位置为(p_0, p_1')的活动孤 $Y \sim\!\!\rightarrow Y_1 Y_2 \cdots Y_k \circ C \cdots Y_m$，如果 $p_1' < p_1$，则在(p_0, p_2)位置处添加新活动孤 $Y \sim\!\!\rightarrow Y_1 Y_2 \cdots Y_k C \circ \cdots Y_m$；
4) 对于所有位置为(p_0, p_1')的活动孤 $Y@ \rightarrow Y_1 Y_2 \cdots Y_k \circ Y_{k+1} \cdots Y_{l-1} C Y_{l+1} \cdots Y_m$，如果 $p_1' < p_1$，并且 C 和 Y 的位置互不交叠，则在(p_0, p_2)位置处添加新活动孤 $Y@ \rightarrow Y_1 Y_2 \cdots Y_k C \circ Y_{k+1} \cdots Y_{l-1} Y_{l+1} \cdots Y_m$；
5) 对于所有位置为(p_0, p_1')的活动孤 $Y\# \rightarrow Y_1 Y_2 \cdots Y_k \circ Y_{k+1} \cdots Y_{l-1} C Y_{l+1} \cdots Y_m$，如果 C 和 Y 的位置互不冲突，则在(p_0, p_2)位置处添加新活动孤 $Y\# \rightarrow Y_1 Y_2 \cdots Y_k C \circ Y_{k+1} \cdots Y_{l-1} Y_{l+1} \cdots Y_m$；
6) 上述步骤中如果 $k = 0$，即活动孤是起始孤，则在操作完成后将该活动孤删除；
7) 上述步骤中如果 $k = m-1$，不是添加新活动孤，而是在(p_0, p_2)位置处添加新归结成分 Y，并存入 agenda 和 chart 中。

图 2-6 Marionette 中孤扩展过程-Algorithm2-3

与传统的Chart Parser相比，*Marionette*有明显的特点：(1)部分分析的特点：对句子不作接受或拒绝的简单判断，而是保留分析过程中得到的所有部分结果，提供最大信息以便后续处理。(2)跨成分的归结特性：不拘泥于传统算法中成分间位置关系的紧密相连性与严格偏序性，而是根据规则类型的不同采取更为灵活的归结策略，容许更为自由的口语通过分析器。

2.4 本章小结

本章介绍了文法的基本概念及常见的几种文法，分析了3种分析算法的步骤及优劣；重点介绍了一种针对汉语口语特点提出的上下文无关增强文法，描述了其形式化定义和5种增强的规则类型，最后介绍了针对这种上下文无关增强文法改进的Chart Parser算法——Marionette分析器，分析了它的主要特点。上下文无关文法覆盖了大部分口语现

象，下一章将根据汉语口语对话系统中语言的特点，以该文法为对象，研究文法规则的自动推导算法。

第三章　文法规则自动推导算法

研究口语对话系统中文法规则自动推导技术，首先要根据语言特点选择适当文法作为的算法的推导对象。本章首先分析汉语口语对话系统中语言的特点，根据这些特点选定合适的文法作为算法的推导对象，开展自动推导技术的研究。

3.1 汉语口语对话系统中语言的特点

3.1.1 汉语的特点

（1）结构化不明显

相对于印欧语，语言学界认为汉语语法有如下主要特点[42]：（1）缺乏严格意义的（狭义的）形态变化。这是最重要的一个特点，汉语语法的其它特点都跟这点有关。形态指词的单复数、人称等变化形式。（2）词类和句法成分间的关系错综复杂。比如动词、形容词可作主语和宾语，形容词可作谓语和状语，名词可作定语，一定条件下还可作谓语等。（3）语序显得比较重要。汉语的词缺乏形态变化，句意主要靠语序的不同来体现。如果说英语或其它拼音语言文字是一种良好的结构化语言[43]，那么汉语在很大程度上是一种结构化不明显的表意语言。

（2）表意性

汉字是表意语言文字。汉字的造字法有六种，分别是象形、指示、会意、形声、转注、假借，其中象形是其它造字法的基础。基于这个特性，每个汉字均有其独立的含义，即使不在句子或上下文中也是如此。基于汉字的汉语自然也具有表意性。汉语的表意性，使得汉语句子的组成方式跟其它拼音文字有所不同。绝大多数拼音语言文字，词有比较确定的词类，兼类比较少，词组成句子需要遵循比较简单的、严格的语法。然而汉语不同：一是没有确定的词概念；二是词的词类很难确定，大多数要依其所在句子的实际情况而定；三是句子结构不易描述，词类和短语类的在句子中的位置非常灵活。

可以简单地用一个例子来说明问题。如果要对“这时候从车上跳下来一个美国人”这个句子进行分析，学习过中文语法知识的中国人对此会有不同的见解。有人认为这句话没有主语，有人认为主语是“这时候”，而另外一些人则认为主语是“美国人”。然而如果是一句表达同样含义的英语“Then an American jump off the truck.”，其语法成分则很明显，“an American”是主语，“jump off”是谓语，“truck”是宾语，“then”

是状语，没有分歧。

根据这些特点，用基于词类的文法来描述汉语，应该说仍没有形成在数学上比较简单的体系；或者没有英语文法体系那么成功。从另一个角度来说，基于词类的分析方法，描述的是语言的表面形态（句法层面的表面形式），而汉语的表意特性在一定程度上使汉语句子的结构更接近于其深层的交际功能的模型结构。

3.1.2 口语的特点

文献[1]依托一个电话航班信息系统收集真实场景中的人-人对话语料，研究了汉语口语的独特现象。将汉语口语特点概括如下：

（1）包含与要表达的语义没有关系的成分，如礼貌用语等。例如：喂，你好，请问是中关村航空客运代理处么？

（2）对话中包含思考时的重复或为强调所作的重复。例如：我问一下那个四月三十，呃，四月三十号北京到……

（3）存在基于上下文的省略，对话中表现为语义相对单一的短句。例如：

C：我问一下那个四月三十呃四月三十号北京到福州的机票最后一班还有么?

O：只有一班有。

C：那个那五月一号的下午三点有么？（省略起飞与到达时间，C表示客户语句，O表示接线员语句）

（4）信息充足的前提下，成分以任何顺序出现都不影响表达。例如：……五点二十五国航飞深圳的……（时间、航空公司、地点或其它信息可以以任何顺序出现）。

（5）经常包含习语或不必要的成分。例如：那，那个八点二十那个是去什么机场的呀？

（6）存在包含所有信息长句。例如：哎，您好，这样那个我订一张那个明天下午五点四十五去北京到上海的那个机票的。

3.1.3 语音识别器导致的问题

对话系统的语言理解面对的输入是语音识别器的输出。目前世界上比较先进的语音识别器的性能，即使在实验室环境下，词的正确率也只有90%左右，整句的正确率自然不高，识别输出存在这样一些问题：

（1）插入错误，指识别器输出中存在原有语音中不存在的词或单元。

（2）删除错误，指识别器输出中漏掉了原有语音中存在的词或单元。

（3）替换错误，指识别器输出中的词或单元与原有语音中的词或单元在时间位置上重叠，但互不相同。

（4）此外，自发语音中存在一些不可用文字表达的语音，比如咳嗽声、笑声、咂嘴声、拖音、不流利、噪声等（本质上说，“嗯”“啊”等标志说话者思考或犹豫的语音也属此类），识别器对它们的输出或者是一些有意义的词或单元，或者在关键词检出的方法中作为补白输出。

出现这些识别错误时，原本被认为是合乎文法的句子或短语将被单元的插入、丢失或替换所干扰，而被认为是不正确的。因此，语音识别器输出的句子，有相当大的可能会被文法分析器所拒绝。

3.1.4 本节小结

上述的语言特点给口语对话系统中的文法带来较大的挑战。但是，幸运的是，这些挑战中最主要的是无意义成分（包括语气词、重复、识别误插入等）的出现和汉语口语顺序的任意性特点。解决了这两类主要问题也就解决了汉语口语对话系统中语言特点带来的大部分挑战。

3.2 算法的推导对象

传统文法体系的发现和使用，始于对西文拼音语言文字的处理，在上世纪初随着其它西方科学被介绍到中国，并用于对中文的描述。形式语言理论自Chomsky创立并用于处理西文语言之后，也被借用于对中文的计算。它的基本思路是，认为句子由短语和词组成，短语由短语和词组成，而词可以按功能分成不同的词类，因此可以由词类的组合方式来描述句子的结构。根据3.1对汉语口语特点的分析，这种基于词类的描述及分析体系对于汉语口语来说似乎并不太适合。

鉴于汉语口语的特点和2.3节对上下无关增强文法特点的分析，作者选定上下文无关增强文法为算法的推导对象，开展汉语口语对话系统中文法规则自动推导算法的研究。本章后面部分将给出一种基于句子分割的文法规则自动推导算法，包括基本原理、形式化描述、几种不同的推导策略、消歧及归一化方法和对算法处理流程的改进。

3.3 基于句子分割的文法规则自动推导算法

3.3.1 算法基本原理

对于一个例句，如果已有的文法规则对它来说是完备的，则用文法规则分析这个句

子后能够得到一棵完整的语法树。反之，如果已有的文法规则是不完备的，则分析结果一定是由若干棵完整语法树的子树组成的森林，其中每一棵子树都相当于例句的某一个子部分，后文把这些子部分称为例句的片断。例如句子“John ate the cake.”，如果已有规则集对它来说是完备的，则分析之后得到的语法树是完整的（如图 3-1），反之得到的语法树则是不完整的（如图 3-2）。显然，如果能够设法以这些不完整的分析结果为基础推导出缺不的文法规则，把规则集补充完整，再来分析句子就可以得到完整的语法树。由此，我们的文法自动推导思路是：用已有规则集对训练集中的例句逐个进行分析，如果能够成功分析出句法结构，则直接进行下一个例句的处理；否则根据不完整的分析结果推导出缺少的规则，并更新已有的规则集。

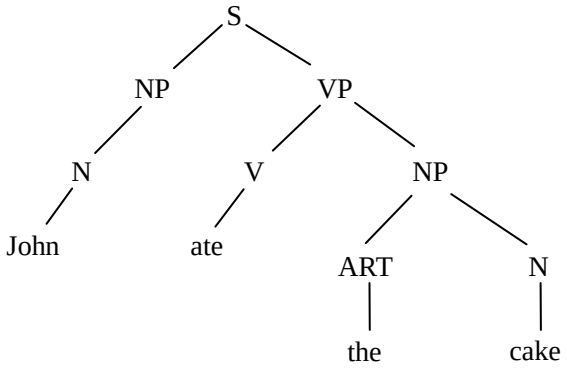

图 3-1 成功分析得到的语法树

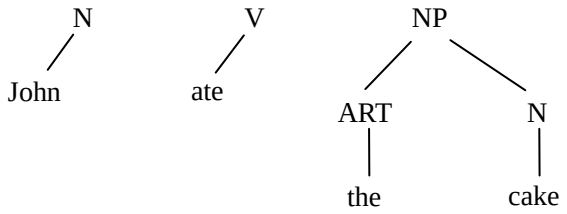

图 3-2 未成功分析得到的语法树

通过 3.1 的分析，我们知道，汉语口语对话系统中大部分的语言现象都可以用跳跃型规则和无序型规则来表示，而无序型的规则又可以看作是跳跃型规则的“压缩”版本，即多条跳跃型的规则可用一条无序型的规则表示。所以，目前本文工作只考虑苛刻型和跳跃型两种类型的规则推导，今后将逐步扩展算法以支持其他几种类型的规则。算法的基本原理如图 3-3 所示，输入是特定领域内的例句集合和一个初始规则集（可为空），输出是最终推导出的文法规则。图 3-3 中，句法分析（Parsing）过程利用初始规则集对

当前例句进行句法分析，文法推导（Reinforcing）过程根据当前句子的不完整分析结果推导出缺少的规则，完善当前规则集。调用句法分析过程之前需要根据预定义的领域关键词表对例句进行分词，例句经预处理后结果如图 3-4 所示。

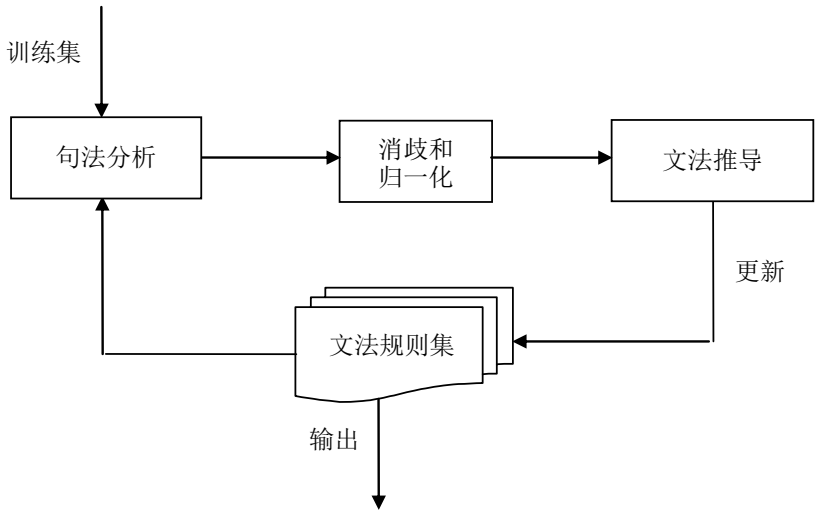

图 3-3 文法规则推导算法流程

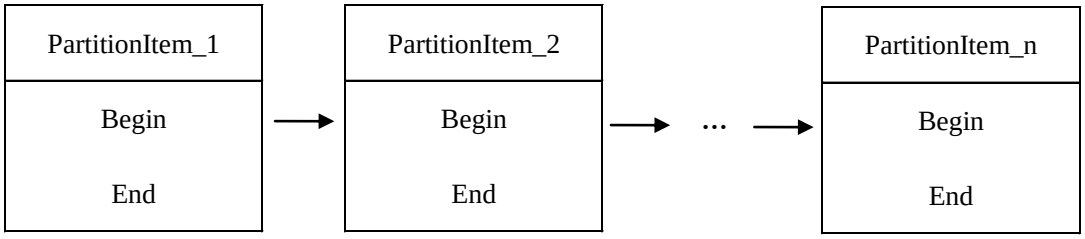

图 3-4 例句分词结果

3.3.2 相关术语定义

在给出文法推导过程之前先给出相关的术语定义。

[定义 3-1] 句子：一个例句经分词之后可看作由词类（包括表示领域背景下无意义词的垃圾词类）组成的一个串，写成句子 $sent=(K_0,K_1,\ldots,K_{n-1})$，其中 n 表示句子中的词类总数，K_i（$0\le i<n$）是第 i 个词（包括无意义词）的词类。垃圾词类以外的词类构成了文法中的终结符。

[定义 3-2] 位置，位置交叠，位置重叠：句子 $sent=(K_0,K_1,\ldots,K_{n-1})$ 中词类 K_i 的位置定义为区间：

$$P_{K_i}=\begin{cases}[0,L_0), & i=0\\ [\sum_{j=0}^{i-1}L_j,\sum_{j=0}^{i}L_j), & 0<i<n\end{cases} \tag{3-1}$$

其中 L_j 表示 K_j 的长度。终结符成分的位置就是所对应词类的位置。非终结符成分 C 的位置定义为 $P_C=[p_1,p_2)$，其中 $p_1=\min\{b_i\}$， $p_2=\max\{e_i\}$，$[b_i,e_i)$ 是构成 C 的终结符成分 K_i（$0\le i<n$）的位置。如果两个成分的位置区间交集非空，则称这两个成分**位置交叠**。如果两个成分的位置区间交集等于这两个成分的位置，则称这两个成分**位置重叠**。

[定义 3-3] 片断：一个句子的片断是句子中连续的若干词类，可以写成 $seg=(K_i,K_{i+1},...,K_{i+j})$，其中位于首尾的 K_i 和 K_{i+j} 是文法的终结符，其余位于中间的 K_{i+x}（$0\le x\le j$）是终结符或垃圾词类。可见，终结符成分是句子的一个片断，非终结符成分也对应句子的一个片断。

[定义 3-4] 片断的位置，片断位置交叠、片断位置重叠：片断 $seg=(K_i,K_{i+1},...,K_{i+j})$ 的位置定义为区间 $P_{seg}=[p_1,p_2)$，其中 $p_1=\min\{b_i\}$， $p_2=\max\{e_i\}$，$[b_i,e_i)$ 是构成 seg 的 K_x（$0\le x\le j$）的位置。若两个片断的位置区间交集非空，则称这两个**片断的位置交叠**。如果两个片断的位置区间交集等于这两个片断的位置区间，则称这两个**片断的位置重叠**。

[定义 3-5] 片断的间隔：对于两个位置不重叠的片断 S_1 和 S_2，位置分别为 $P_{s_1}=[p_m,p_n)$ 和 $P_{s_2}=[p_i,p_j)$，且 P_{s_1} 在 P_{s_2} 之前，则定义 $uGap=P_i-P_n$ 为片断 S_1 和 S_2 的间隔。片断之间的间隔决定着新生成规则是否为跳跃型规则以及可能的跳跃距离。

3.3.3 文法推导算法

从图 3-3 可以看到，我们的文法自动推导算法包括两个核心过程：句法分析和文法推导。其中句法分析过程先采用 2.3 节介绍的 Marionette 分析器对例句进行分析，然后使用一定的策略对分析结果进行消歧和成分归一化处理。文法推导过程是算法的核心过程，用于从例句中推导文法规则。

我们将一个句子经句法分析后结果中那些没有父结点的片断称为顶层片断。文法推导过程的输入是例句分析结果中按先后顺序排序的顶层片断，对应着已有规则集中的若干终结符或者非终结符。由于规则推导只是对经过句法分析不能得到完整语法树的例句进行处理，所以文法推导过程的输入至少包含两个顶层成分。

在特定领域应用中，人们在口语对话时所加入的无意义成分（如语气词等）的数量一般不会太多，反映在跳跃型规则中就是可能的跳跃距离存在一个最大值。在文法推导过程中，我们针对领域估计并预设一个允许的最大跳跃距离，作为算法的一个参数。算法的具体步骤如图 3-5 所示。该算法输出的规则右边只包含一个或两个符号。

步骤1： 添加一个新的非终结符代表整个句子，将整个句子作为当前处理的片断；

步骤2： 若非终结符 *A* 代表当前处理的片断，对于当前处理片断中位置最前面的两个顶层片断 C_1 和 C_2，计算它们的间隔 *uGap*；

2.1 若当前处理的片断只有两个顶层片断，且 *uGap* 等于 0，则添加形如 $A \rightarrow C_1\ C_2$ 的苛刻型规则；

2.2 若当前处理的片断只有两个顶层片断，且 *uGap* 大于 0 且小于允许的最大跳跃距离，则添加形如 $A \rightarrow C_1\ [uGap]\ C_2$ 的跳跃型规则；

2.3 若当前处理片断有两个以上顶层成分，且 *uGap* 大于等于 0 且小于允许的最大跳跃距离，则把当前处理的片断分割成两个子片断：第一个子片断是 C_1，第二个子片断从 C_2 开始到当前处理的片断的最后一个子片断；添加一个新的非终结符 β 代表第二个子片断；

a) 当 *uGap* 等于 0 时添加形如 $A \rightarrow C_1\ \beta$ 的苛刻型规则，对第二个子片断递归调用步骤 2；

b) 当 *uGap* 大于 0 时添加形如 $A \rightarrow C_1\ [uGap]\ \beta$ 的跳跃型规则，对第二个子片断递归调用步骤 2；

图 3-5 文法推导过程的具体步骤-Algorithm3-1

3.3.4 不同的推导策略

图 3-6 是对分析后包含 5 个顶层片断的例句调用文法推导过程的示意图。可以看到，文法推导过程本质上是对句子自顶向下逐步分割并推导出规则的过程，因此我们称这种算法为基于句子分割的文法规则自动推导算法。图 3-6 中对句子分割并推导规则是按从左到右的顺序依次进行的，每一次分割总是把左边第一个顶层成分作为一个片断，剩余

部分作为另一个片断。我们称这种推导策略为“左部优先”策略。Algorithm3-1 也可以调整为每一次分割总是把右边第一个顶层成分作为一个片断，剩余部分作为另一个片断，称之为“右部优先”策略，如图 3-7 所示。这两种策略对片断进行组合的先后次序不同，会导致最终得到不同的文法。如果不考虑应用领域的语言特点，这两种策略并无优劣之分。考虑到在汉语中更多地出现“修饰语在前，中心语在后”的现象，所以 Alogrithm3.1 中我们采用了图 3-6 的“左部优先”策略。

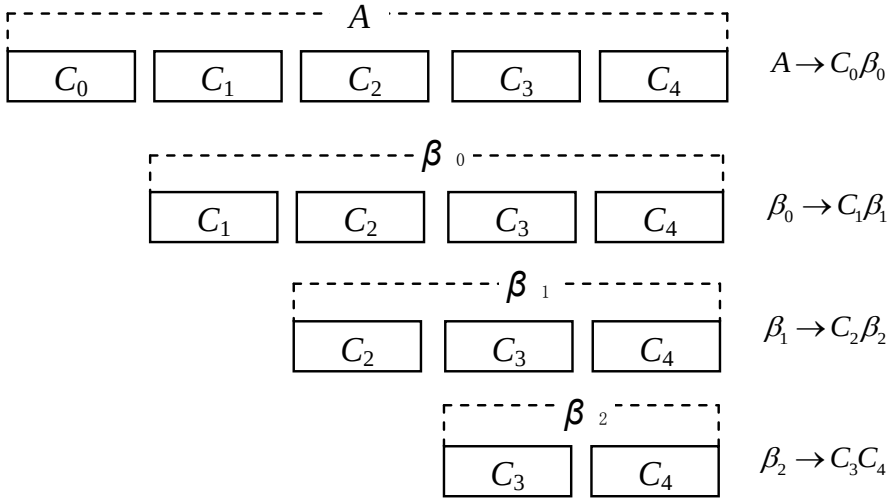

图 3-6 “左部优先”策略示意图

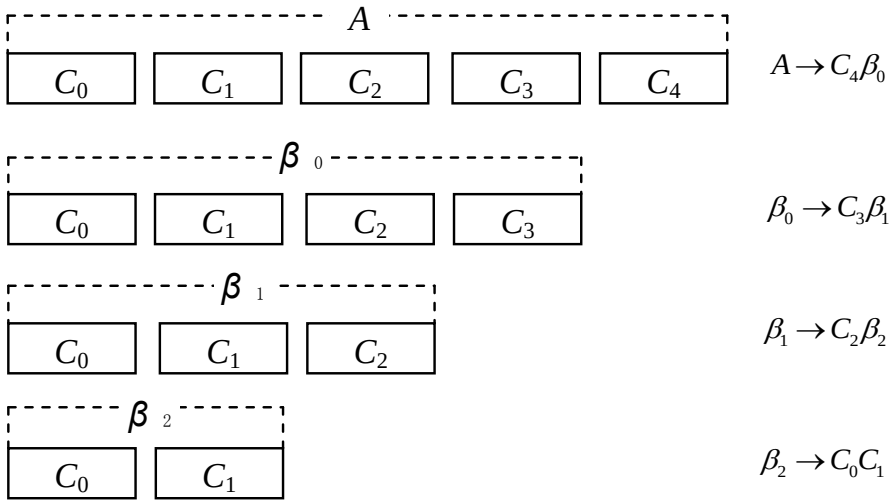

图 3-7 “右部优先”策略示意图

图 3-6 与图 3-7 所示的文法推导过程对例句的处理是一个“自顶向下、逐步分割”的递归过程，直到分析当前例句的所有文法规则推导完毕递归过程结束。文法推导过程

也可以采用非递归的处理：对于一个例句直接取图 3-6 或图 3-7 中最下面的两个顶级片断得到一条文法规则，然后用这条规则分析所有未成功分析的例句，这实际上是把例句中的公共子串用新的非终结符替换的过程。上述过程直到所有的句子都能够被覆盖为止。这是一个“自底向上、逐步分割”的处理过程。同理，也可以有“左部优先”和“右部优先”两种推导策略。从文法推导的角度看这两种处理过程本质上是等价的，最终输出不同的文法规则，但是性能基本没有差异。

3.3.5 歧义片断的消除与归一化

在句法分析过程中相同的片断可能会有不同的归结方法，产生不同的分析结果，这些分析结果的位置可能交叠，也可能重叠。这是由文法本身的歧义或知识的匮乏引起的。文献[1]根据对话系统中常见的歧义类型以及日常生活中的经验总结，提出了歧义消解的假设和准则，设计了对分析结果进行歧义消解的策略（如表 3-1 所示）。

表 3-1 分析结果的消歧策略[1]

高	低
顶级结点	非顶级结点
非顶级节点相等	
比值(声学得分/覆盖)大者（因为得分为负，因而可认为表示声学得分高或覆盖大者）	比值小者
顶级规则节点	非顶级规则节点
对顶级规则节点，节点数小者	节点数大者
对顶级规则节点，深度小者	深度大者
句序大者	句序小者
归结顺序大者	归结顺序小者

在调用文法推导过程之前需要对相互冲突的片断进行消歧处理：（1）用表 1 中策略判定冲突片断的优劣；（2）算法直接剔除次优的分析结果，保留所有相互不冲突的顶级分析结果。例如分析结果 A -> X B、C -> B D、D -> E F，且片断 A 比片断 C 优，算法会剔除片断 C，而与 A 并不冲突的 D 被保留，成了一个新的顶级片断。因为在文法推导时，丢失了任何一个相互不冲突顶级片断就会导致最终文法的不完整性，因此片断 A 和片断 D 都是文法推导不可或缺的基础。

对于位置相互冲突的片断在大多数情况下都只需保留一个较优的即可，除非两个顶

层片断相互重叠并且都出现或都不出现在其它规则的右侧。文法推导算法是以所有的顶级片断为输入的，当顶级片断位置相互重叠时，上述的消歧策略会剔除位置相互重叠的片断，只保留其中一个，这会使得推导算法输入的不完整性，导致最终文法规则的缺失。

解决这一问题可以有两种方法：一是循环地让位置重叠的片断每次只有一个与其它顶层片断一起进入文法推导过程，直到位置重叠的片断都参与完成了文法推导过程。假设某例句的不完整分析结果共有 n 个顶层片断，其中有 m 个片断相互重叠，则依次让 m 个片断中的某一个和剩下的 n-m 个片断一起进入文法推导过程，每次都可得到 n-m 条规则，最终可得到 m*(n-m)条规则。另一种方法是：先添加 m 条形如 $A \to a$ 的规则，从而以顶层片断 A 来代表所有冲突的 m 个片断，然后让 A 和剩下的 n-m 个成分一起进入文法推导过程，再得到 n-m 条规则，最终可得到 n 条规则。当不完全分析的结果中有若干组片断位置相互重叠，则第二种方法得到的最终结果规则数量要少得多。本文在 Algorithm3-1 中采取了第二种方法，并将其称为歧义片断的归一化，如图 3-7 所示。

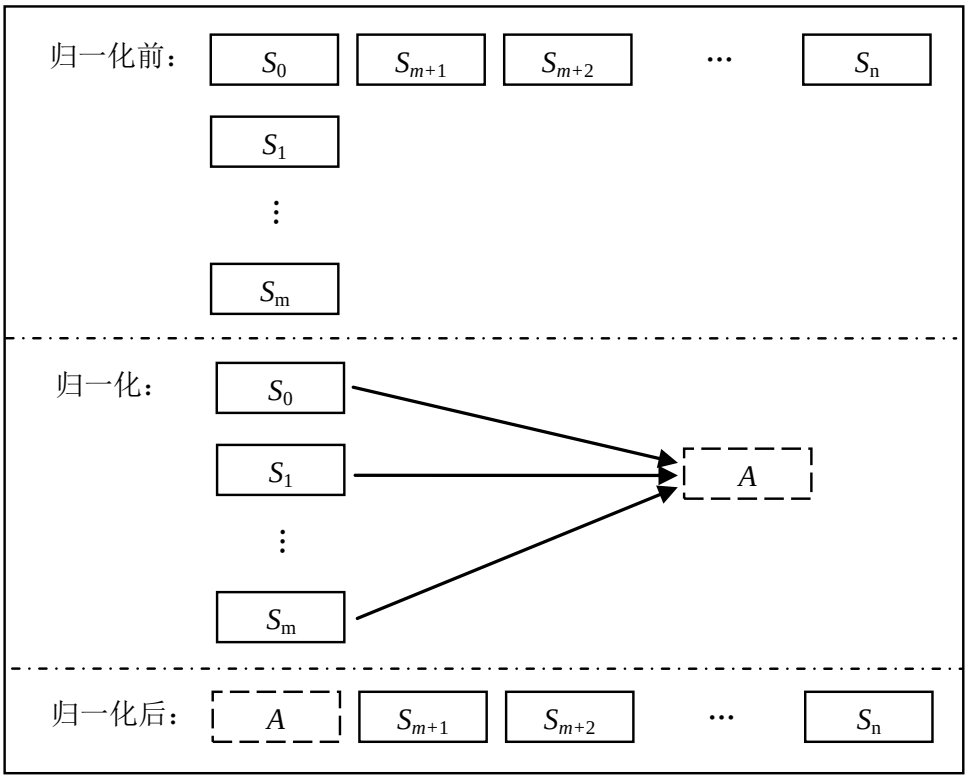

图 3-7 歧义片断归一化过程

3.4 算法流程的改进

从 3.3.2 节对算法的介绍和分析可知，以何种方式将句子分割为不同的片断，决定了推导出什么样的规则，换句话说只有合理正确地将句子分割为各个片断，才能保证最

终文法的正确性。但是在执行文法推导过程时，并没有额外的语义信息或领域知识可以参考，很难保证句子分割的正确性。考虑到在实际口语对话过程中，特别是在一些信息的确认问答中，基于上下文语境，人们经常会使用较短的句子表达单一的语义，因此语义单一且完整的短句和语义复杂的长句都可以收集到。如果文法推导过程先对语义单一的短句进行处理，则容易得到可正确分析单一语义的规则，而后在处理复杂的长句时用更新后的规则集进行分析，则所包含的各语义将被归结为顶层成分，从而尽可能保证分割的正确性。因此，算法应尽可能按照“由简到繁”的顺序进行。

在图 3-3 所示步骤中，新规则是苛刻型还是跳跃型，跳跃距离多大，是由例句的实际情况决定的，所得到的规则类型和跳跃距离在领域范围内不一定准确。也许两个成分间可跳过最多 5 个汉字并归结为另一成分，但是在某一例句中这两个成分的位置恰巧相接，结果推导出一条苛刻型规则，而在另一例句中这两个成分的位置间隔为 2，结果推导出一条跳跃距离为 2 的跳跃型规则。如：例句“北京啊明天”和“北京那个明天”应该生成一条跳跃型的规则。在算法实际处理中，对于第一个例句会生成一条跳跃距离为 1 的规则，当处理第二个例句时由于垃圾成分长度是 2，刚刚生成的规则不能用于该句子的分析，所以又生成一条跳跃距离为 2 的规则。这两条规则除了跳跃距离之外其他部分完全相同，应该合并成一条跳跃型的规则。同理，还存在着前面例句中生成的苛刻型规则不能用于后面含垃圾词的例句的现象。因此，算法应在推导出规则后仍允许对规则类型和跳跃距离进行修正。

为了解决以上两个问题，我们对图 3-3 的算法流程进行了改进：1）用初始规则集对所有例句进行分析之后，先按照消歧和归一化后的顶层片断的数量由少到多排序，再依次调用文法推导过程；2）对一个例句调用文法推导过程推导出新规则之后，要应用新规则来尝试分析其它未处理例句的顶层成分，在分析时以只增不减的方式修正规则的类型以及跳跃距离（不可超过预先设定的最大跳跃距离）。改进后的算法流程如图 3-5 所示。图中的句法分析(1)过程与图 3-3 中的句法分析过程相同，句法分析(2)过程只对未处理例句的顶层成分进行分析，并且加入了规则类型和跳跃距离的自动修正功能。改进后算法的性能有了明显的提升。

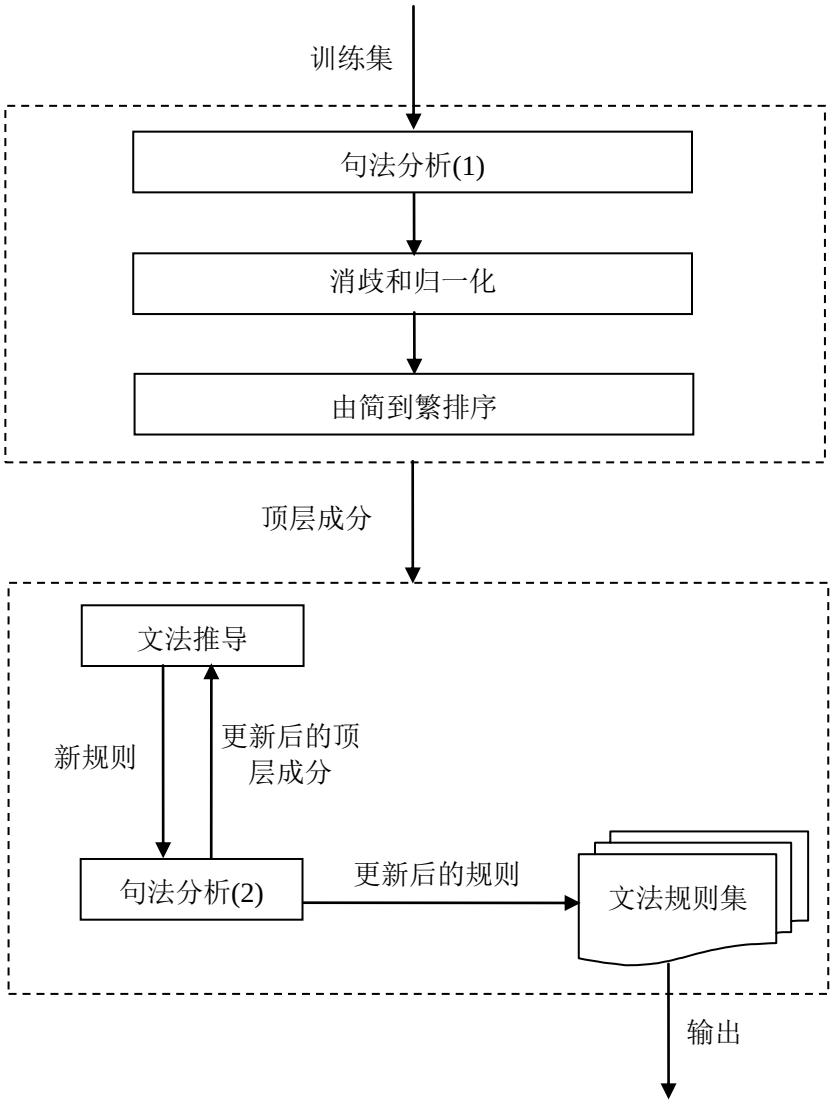

图 3-5 改进后的算法流程

3.5 本章小结

本章从 3 个方面分析了汉语口语对话系统中语言的特点。鉴于这些特点，选定一种符合汉语口语特点的上下文无关增强文法作为对象，深入开展汉语口语对话系统中文法规则自动推导算法的研究，主要成果有：提出了一种基于句子分割的文法自动推导算法，详细讨论了算法的基本原理、形式化描述及算法步骤，分析了算法处理过程中的歧义片断的消解及归一化方法：一是按照由简到繁的顺序处理例句；二是允许在文法规则自动推导过程中修正前面得到的不合理的文法规则。最后讨论了例句的处理顺序对输出文法的影响，根据这些影响对算法处理流程进行了改进。下一章将研究文法性能评测方法，给出面向领域任务的评测指标，并对本文算法推导得到的文法性能进行评测，以考查自动推导算法的效果。

第四章　算法评测与分析

4.1 评测指标的定义

正如 1.2.3 节所述，学术界对文法的评测主要考查文法的复杂程度（规则数目、非终结符数目）和算法的时间消耗。还没一套能够较好适应领域任务需求的文法评测方法。因此有必要根据领域任务的特点探索一套合理的文法性能评测方法。

在口语对话系统应用中，句法分析的目的是为了进一步的语义分析和理解，因此句法分析得出的句子的语法树，应能够体现出语义的结构，不能将完整的语义割裂开。例如“十月五日的天气”这句话，合理的句法分析应将“十月五日”这一日期语义归结为一个片断（如图 4-1），不合理的句法分析可能将它割裂开（如图 4-2）。

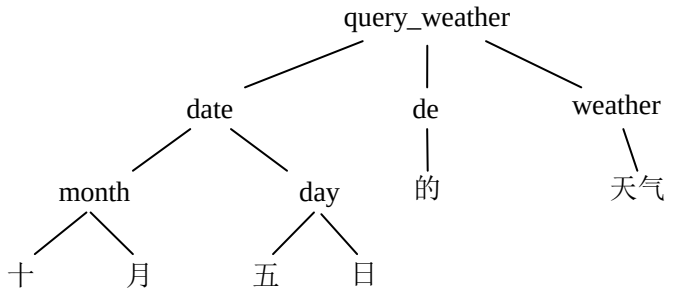

图 4-1 合理的分析结果

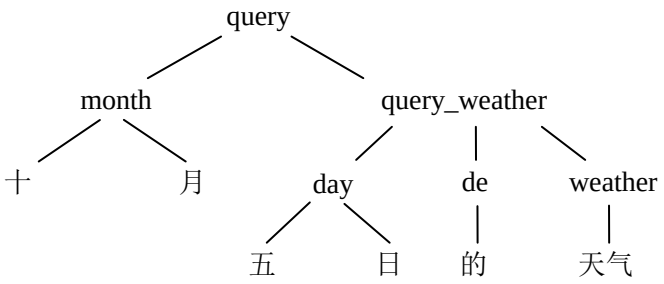

图 4-2 不合理的分析结果

[定义 4-1]语义单元或语义原子：表达领域知识的不可再分的最小单位称为领域语义单元或语义原子。语义单元按照表达的内容类型可分为知识型语义单元和结构型语义单元。知识型语义单元表示领域的某一类知识，如地点、日期、时间等；结构型语义单元表示句子的结构信息，如疑问词、语气词等。知识型语义单元按照表达内容的重要程度可分为核心语义单元和普通语义单元。核心语义单元是指对句子语义表达起关键作用

的语义单元；普通语义单元指对句子语义表达不起关键作用的语义单元。一个句子可能包含若干个语义单元，其中必有核心语义单元，否则将影响句子语义表达。

[定义 4.2]分析的准确率：用文法规则分析一个句子，如果可以得到一棵完整的语法树，且其中正确包含了句子所有的核心语义单元，则在实际领域对话系统应用中可认为正确分析了句子的句法结构。由此，我们可以定义分析准确率来衡量领域文法性能的优劣：

$$\text{准确率}=\frac{\text{正确分析出句法结构的句子数}}{\text{测试集句子总数}} \tag{4-1}$$

对于不同的领域任务，我们可以定制不同的核心语义单元，对不同领域文法的性能进行评测。因此，这里给出的文法性能评测指标具有领域可定制性，能够满足不同领域任务的需求。

与其他研究一样，我们仍采用规则数目与非终结符数目衡量文法复杂度。

通常，口语对话系统对文法推导的即时性要求并不很高，随着计算机软硬件技术的飞速发展，一般的系统应用对算法效率并没有过高的要求。因此，我们给出的评测方法重点关注算法输出的文法对语义解理的贡献，对算法本身的性能没有做过多考查。

4.2 实验领域及步骤

4.2.1 实验领域

作者所在的课题组已经开发完成了天气预报查询的口语对话系统，系统日志收集了大量用户查询例句。因此，作者以天气预报查询为具体领域，选取真实的用户查询例句作为实验数据开展实验，考查算法输出文法的性能。

4.2.2 实验数据

在天气预报查询领域，作者选取了 1000 条真实的用户查询例句作为实验数据，例句示例如图 4-3，图中的“啊”、“嗯”为用户语音停顿时语音识别器自动填充的词。选取例句时主要遵循两个准则：（1）尽可能全面地选取领域对话句型，包括语义单元及其数目变化、语序变化等。（2）兼顾长短句，数据集中应包含语义单一的短句和语义相对复杂的长句。定义该领域的核心语义单元为城市、日期和天气类型。数据集中，包含一个核心语义单元的句子有 100 句，包含两个核心语义单元的句子有 300 句，包含三个核心语义单元的句子有 600 句。采用开放测试的方法，随机选取其中 500 句用作训练集，其余 500 句用作测试集。

郑州啊十二月二十啊气温嗯如何
包头啊啊周四天气咋样
周日啊啊大连啊啊天气嗯嗯怎么样
星期天郑州天气怎么样
长春啊啊啊昨天啊啊啊温度嗯嗯嗯几度
周二包头气温嗯嗯嗯嗯嗯怎么样
天气怎样啊郑州星期二
天气嗯热吗二十号啊长沙
北京二十二号啊气温高吗
天气嗯嗯嗯嗯热吗七月三十啊啊啊啊长春
气温嗯嗯咋样长春啊啊今天
这几天啊啊长春啊啊温度怎么样
天气暖和吗啊啊啊啊三十一号啊啊啊啊长春
前天大连温度多少度
温度嗯嗯嗯咋样啊啊啊成都啊啊啊前天
大连三日天气嗯嗯热吗
天气暖和吗啊元月十七啊郑州
五月三十一
五月三
温度嗯嗯嗯嗯嗯几度近段啊啊啊啊啊成都
温度低吗近几天啊长春
大连昨天天气热吗
十二月二十二日
气温怎么样后天啊啊啊大连
礼拜日
长春气温高吗
郑州礼拜天啊啊天气如何
近几天啊包头啊温度几度
二十四号
长沙二十四号气温嗯多少度
元月三十号啊啊成都天气嗯嗯热吗
温度几度啊啊长春这几天
昨天啊啊啊郑州温度几度
近段时间啊啊啊郑州啊啊啊天气嗯嗯嗯如何
明天啊啊啊啊包头啊啊啊啊天气怎样
温度怎么样长春啊啊近两天
成都四月三十一天气嗯嗯嗯咋样
天气嗯咋样大连十二月三十一
天气嗯嗯嗯暖和吗十一月三十一日成都
温度嗯怎么样啊这些天长沙
礼拜日啊啊成都天气嗯嗯热吗
大连啊啊七月三十号气温嗯嗯高吗
北京啊啊这两天天气怎样
温度多少度北京近几天
天气暖和吗啊元月十七啊郑州
成都啊啊啊三日天气嗯嗯嗯怎样
天气嗯嗯怎么样大后天啊啊郑州
大连啊啊啊啊啊今天啊啊啊啊啊气温多少度
气温高吗长春二十日
天气嗯怎么样啊前天啊成都
天气嗯嗯咋样近两天北京
礼拜天大连啊啊啊天气嗯嗯嗯怎么样
前天大连啊啊啊啊啊气温高吗
温度嗯嗯嗯嗯高吗啊啊啊啊成都
包头啊啊近两天气温嗯嗯怎么样
天气嗯嗯怎么样郑州
……

图 4-3 训练例句示例

预定义的关键词语义类基本涵盖了该领域所有可能的关键词（见附录 A）。

4.2.3 实验步骤

为了考查算法的效果，作者设计了 6 种实验，具体步骤如下：

（1）文法性能的评测实验

随机选取训练集中 20%、40%、60%、80%、全部的例句进行文法推导，实验共分 5 组，采用改进后的算法流程，初始规则集为空，使用“左部优先”的推导策略，用测试集对每次得到的文法进行测试，用文法的分析准确率考查训练集规模与文法性能的关

系。

（2）文法复杂度评测实验

采用改进后的算法流程，初始规则集为空，使用“左部优先”的推导策略，分别随机选取训练集中 20%、40%、60%、80%、全部的例句参与文法推导，然后用测试集对每次得到的文法进行测试，用生成的规则数目与非终结符数目来考查训练集规模与文法复杂程度的关系。

（3）初始规则集对文法影响的评测

采用改进后的算法流程，使用“左部优先”的推导策略，分别随机选取训练集中 20%、40%、60%、80%、全部的例句参与文法推导，然后用测试集对每次得到的文法进行测试，每组实验都包括初始规则集为空和初始规则集只包含日期相关规则两种情况。通过文法的分析准确率和复杂程度来考查初始规则集对文法的影响。

（4）“左部优先”策略与“右部优先”策略对比实验

采用改进后的算法流程，初始规集包含日期相关规则，分别随机选取训练集中 20%、40%、60%、80%、全部的例句参与文法推导，然后用测试集对每次得到的文法进行测试，每组实验都分别使用“左部优先”和“右部优先”两种推导策略。通过文法的分析准确率和复杂程度来对比这两种推导策略的效果。

（5）“自顶向下”策略与“自底向上”策略对比实验

采用改进后的算法流程，使用“左部优先”的推导策略，初始规集包含日期相关规则，分别随机选取训练集中 20%、40%、60%、80%、全部的例句参与文法推导，然后用测试集对每次得到的文法进行测试，每组实验都分别用“自顶向下”和“自底向上”两种处理过程。通过文法的分析准确率和复杂程度来对比这两种推导策略的效果。

（6）算法改进前后效果对比实验

初始规则集为空和包含日期相关规则两种情况下，分别随机选取训练集中 20%、40%、60%、80%、全部的例句参与文法推导，采用“左部优先”的策略，然后用测试集对每次得到的文法进行测试，每组实验都分别使用改进前后两种处理流程，通过文法的分析准确率和复杂程度来对比算法改进前后的效果。

4.3 实验结果及分析

4.3.1 文法性能的评测

按照 4.2.3 所述实验步骤，实验结果如图 4-3 所示。可以看到文法的分析准确率随

着训练集句子数增加而增加，当训练数据达到数据集的半数时，文法的分析准确率达到了 64%。在实际应用中，训练集应尽可能地涵盖领域内的不同句型，训练集涵盖的领域句型越全面，算法输出的文法的分析准确率将会越高。

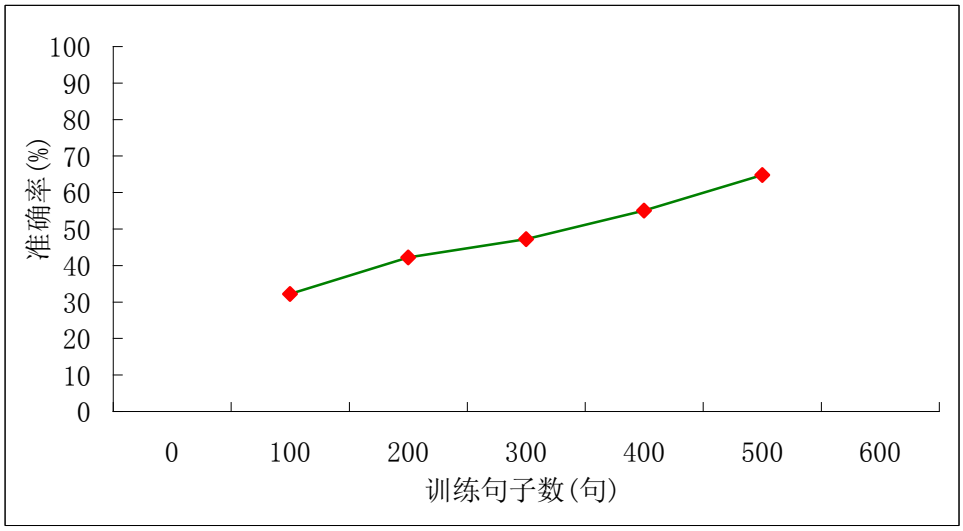

图 4-3 文法的分析准确率

4.3.2 文法复杂程度的评测

按照 4.2.3 所述实验步骤，实验结果如图 4-4 所示。可以看到，随着训练集句子数的增加，非终结符的数目与规则数据同时增加。但是，5 组实验中文法复杂程度始终小于训练集句子数，显示了算法良好的性能。理论上讲，当文法规则覆盖了领域中所有句型后复杂程度将不再增加。

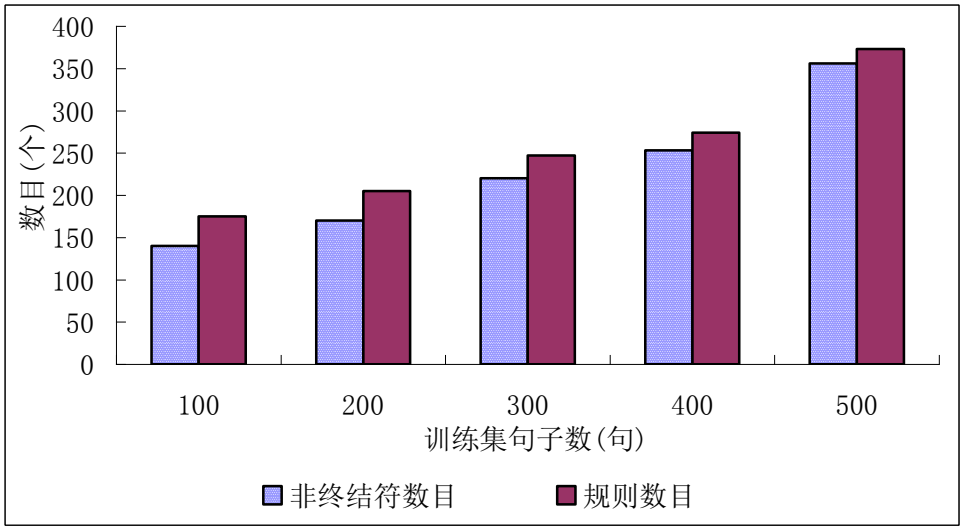

图 4-4 文法的复杂程度

4.3.3 初始规则集对文法影响的评测

按照 4.2.3 所述实验步骤，实验结果如图 4-5、4-6 所示。可以看到，引入日期分析相关的规则后，最终文法的分析准确率有 20%左右的提升，同时输出文法的复杂程度大大地降低。实验说明算法具有较强的利用预先知识的能力。现实世界中有些知识如日期、时间、电话号码、身份证号码等，对于不同的领域是通用的，所以把这类知识存储起来作为算法初始知识，只使用算法推导领域相关的文法可以较大地提升最终文法的性能。人工收集这类知识的代价是非常小的，而且是一劳永逸的，从这个意义上讲算法具有较好的工程应用价值。

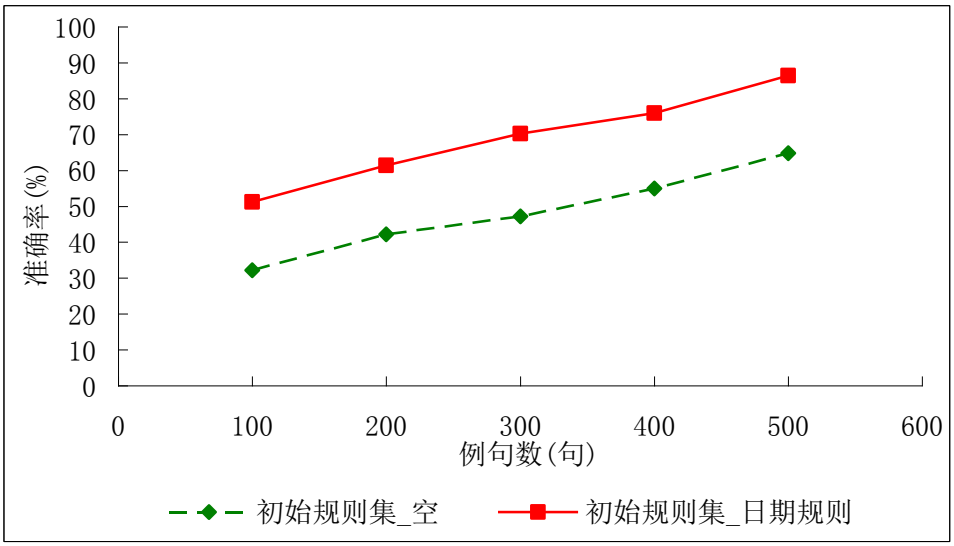

图 4-5 加入日期规则前后文法的分析准确率

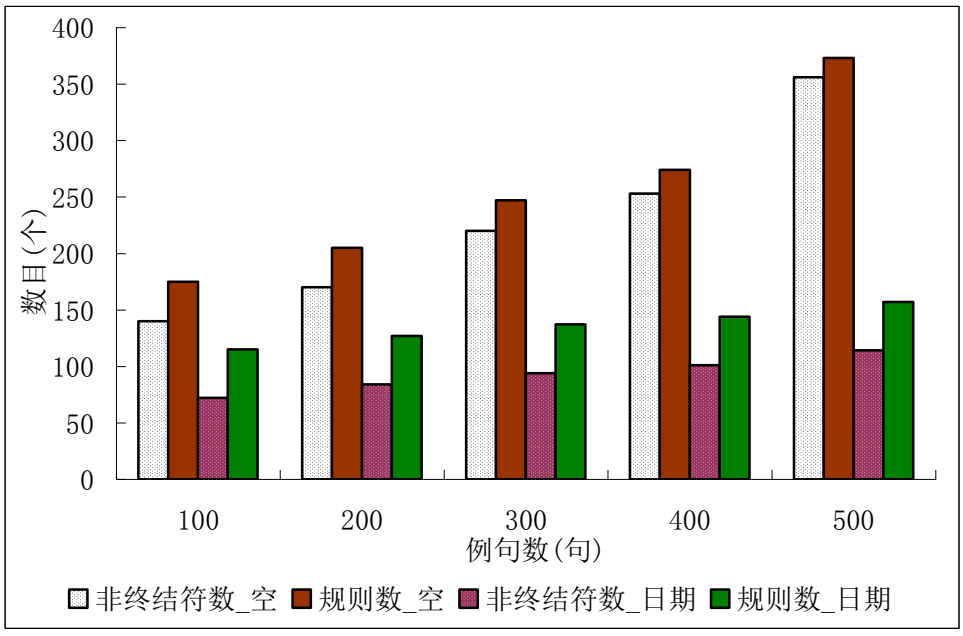

图 4-6 加入日期规则前后文法的复杂程度

4.3.4 “左部优先”策略与“右部优先”策略对比

按照 4.2.3 所述实验步骤，实验结果如图 4-7、4-8 所示。可以看到，“左部优先”与“右部优先”两种策略推导出的文法的分析准确率与复杂程度并没有太大的差异，这两种策略只是对片断的组合次序不同，如果不考虑领域语言特点，它们并没有本质上的优劣之分。实验结果再次验证了前面我们的分析。

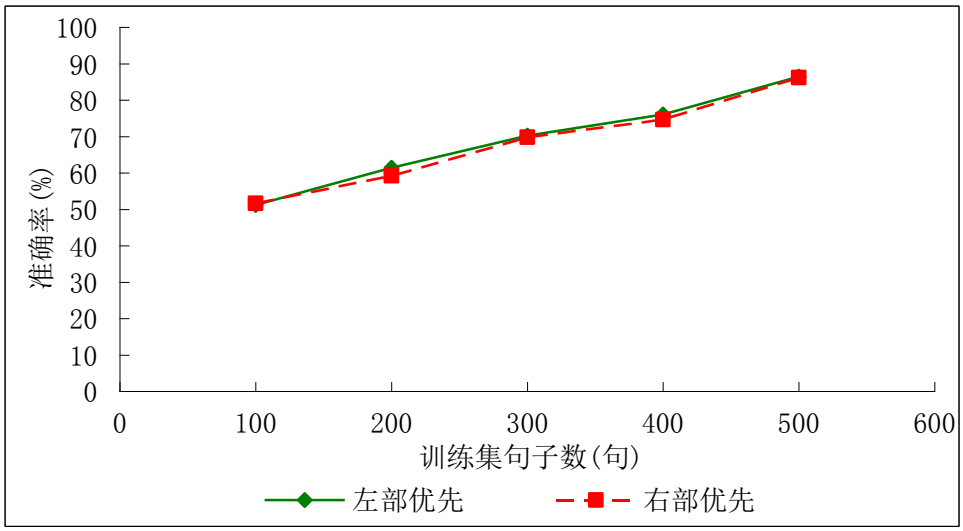

图 4-7 “左部优先”与“右部优先”文法的分析准确率对比

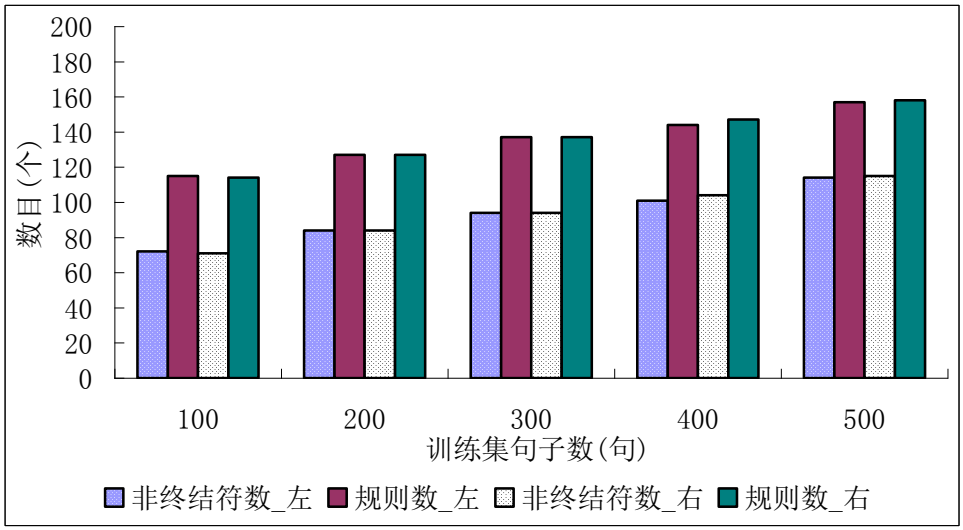

图 4-8 “左部优先”与“右部优先”文法的复杂程度对比

4.3.5 “自顶向下”策略与“自底向上”策略对比

为了考查“自顶向下”与“自底向上”两种策略的效果，开展了对比实验，结果如图 4-9、4-10 所示。可以看到，“自顶向下”与“自底向上”两种推导策略的输出文法的分析准确率与复杂程度并没有太大的差异，这两种策略只是例句中公共片断的替换次

序不同，并没有本质的区别。

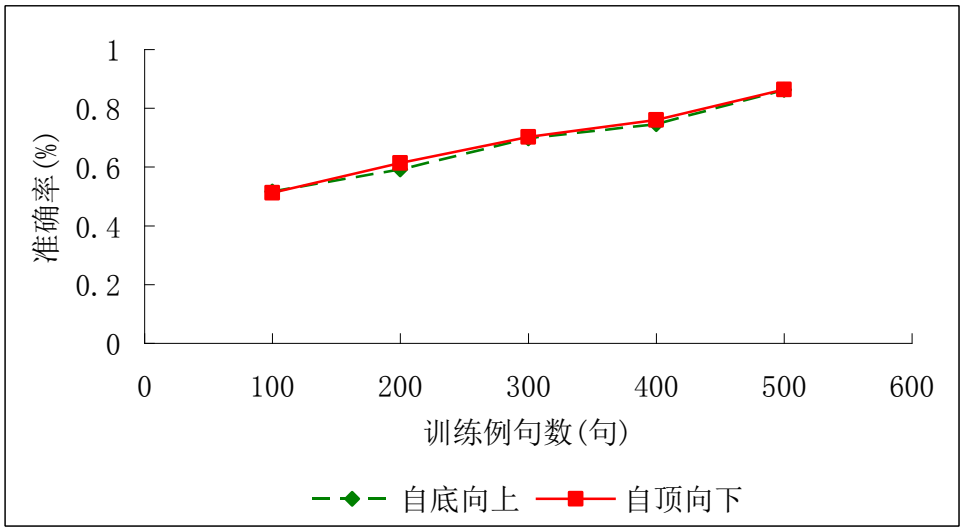

图 4-9 “自顶向下”与“自底向上”两种策略的文法准确率

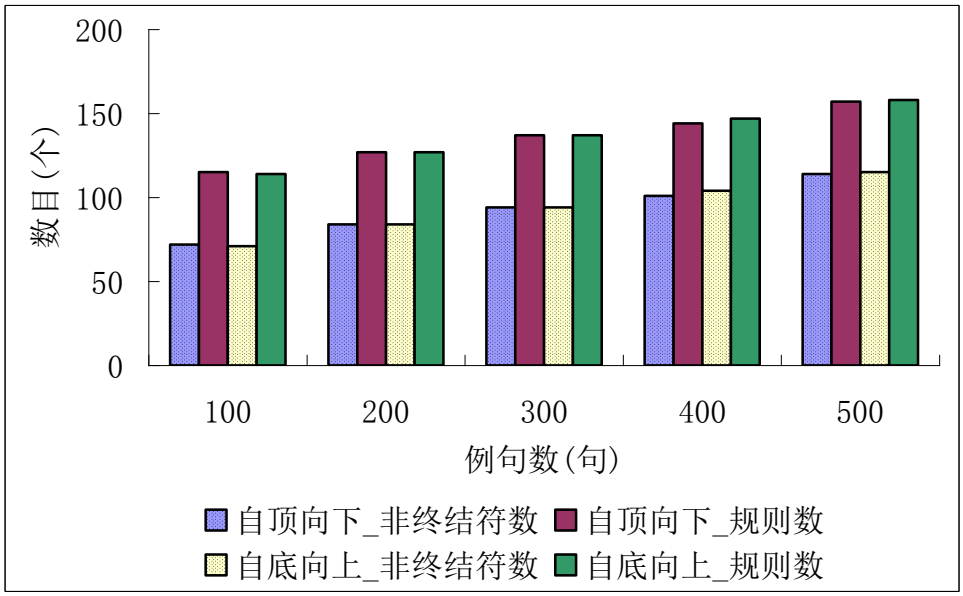

图 4-10 “自顶向下”与“自底向上”两种策略的文法复杂度

4.3.6 算法改进前后效果对比

如 4.2.3 所述，我们对算法改进前后的性能进行了比较。图 4-11 给出了初始规则集为空和初始规则集包含日期相关规则两种情况下，算法改进前后输出文法的分析准确率曲线。可以看到，算法流程改进后输出文法的分析准确率基本不变。

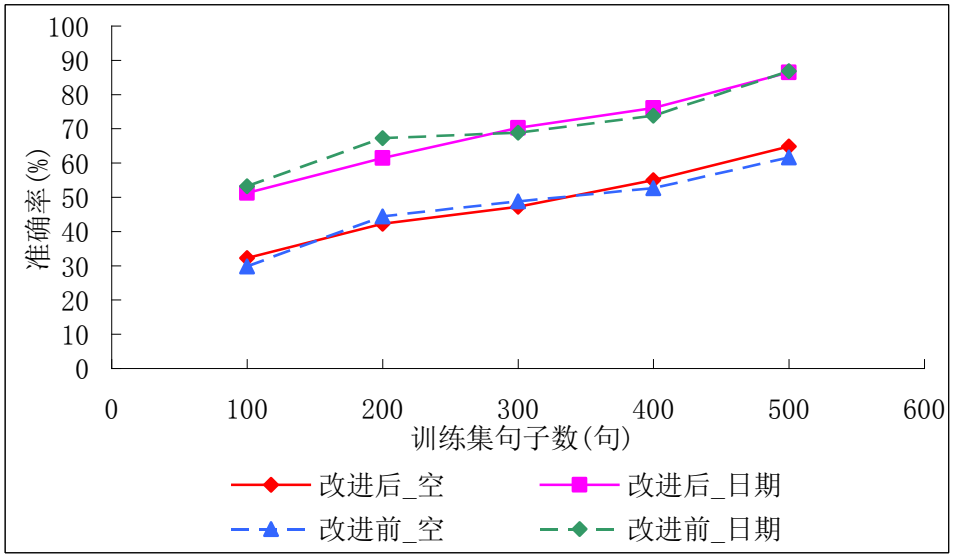

图 9 算法改进前后文法分析准确率

图 4-12 给出了用 500 句例句进行文法推导时，算法改进前后输出文法的非终结符数和规则数。可以看到，算法流程改进后，文法的复杂程度大大下降。用其它数量的例句进行实验，结果也是类似的。说明改进后的算法能够在降低文法分析准确率的前提下大大降低文法的复杂程度，较好地提升了算法的整体性能。

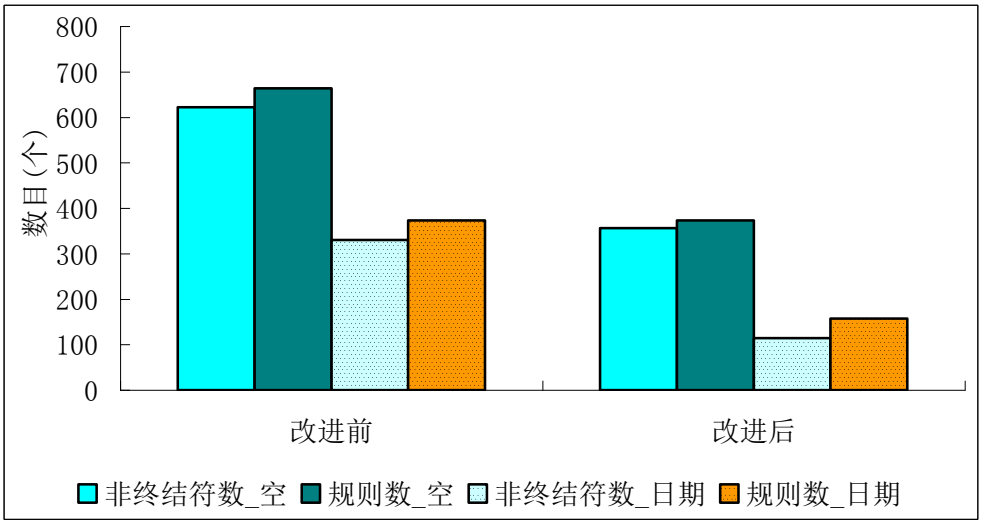

图 4-10 算法改进前后文法复杂程度

4.4 本章小结

本章结合口语对话系统领域任务需求给出了一套文法性能评测方法，该方法具有较好的灵活性和领域可定制性，不同领域的评测只需定义相应的核心语义单元即可。

使用这种评测方法，作者设计了 5 种实验评测，通过对是否包含初始规则的对比、“左部优先”和“右部优先”的对比、算法改进前后的对比，全面考查了本文提出的文

法规则自动推导算法的性能。结果表明，算法具有如下特点：

（1）算法充分考虑了口语对话系统中语言的特点，输出的文法能够较好地覆盖口语对话系统中主要的语言现象。

（2）算法具有较好的学习能力，能够在较少训练数据基础上获取分析准确率较高的文法，使开发人员人专业且繁琐的劳动中解脱出来，降低了研发成本，提高了研发效率。

（3）算法具有较强的利用已有知识的能力，加入初始规则集后，输出文法的性能获得了大大的提升，同时文法的复杂程度明显降低。

第五章 总结与展望

5.1 本文工作总结

本文在深入对比分析形式文法特点的基础上，根据口语对话系统中语言的特点，选定一种符合汉语口语特点的上下文无关增强文法作为对象，开展口语对话系统中文法规则自动推导技术的研究，提出了一种文法规则自动推导算法；研究了面向领域任务的文法评测，提出了一套灵活的、领域可定制的文法评测方法，使用该方法验证了算法的有效性。本文的主要工作包括如下几个方面。

（1）对比分析了常见的各种文法的特点和性能，根据汉语口语对话系统中语言的特点，选定上下文无关增强文法进行口语对话系统中文法规则自动推导算法研究。

（2）针对汉语口语对话系统中语言的特点，提出了一种基于句子分割的文法规则自动推导算法，给出了算法的形式化描述、具体步骤、推导过程中的歧义片断的消歧和归一化策略，探讨了算法流程的改进。

（3）研究了面向领域的文法性能评测方法，提出了适用于领域需求文法性能评测指标。用这种评测指标，在天气预报查询领域对算法的输出文法进行评测。使用 500 句训练数据得到的文法的分析准确率达到了 86.4%，显示了算法较好的性能。

5.2 相关问题讨论

（1）算法遵循由简到繁的顺序处理例句，尽可能地保证最终文法的正确性，因此训练集中应尽可能全面地包含领域内语义单一且完整的短句。

（2）对于一组句型结构相同的例句，算法只需处理其中任何一句，就可覆盖这一组例句。换言之，训练集中包含的不同句型结构的数目才是真正影响文法性能的因素。所以训练集还应尽可能全面地包含领域内的不同句型。

（3）实验表明，有效的初始规则集能够明显地提升最终文法的性能。现实世界中有些语义的表达如日期、身份证号码等对于不同的领域是通用的，所以可人工积累这些语义相关的文法规则，在自动推导实际领域的文法时将它们加入初始规则集，有利于得到较好的最终文法。

5.3 未来的研究方向

（1）现有的算法推导出的文法规则右项只可能包含一个或两个符号，与领域专家人工给出的文法规则形式差别较大，虽然不影响对领域句子的成功分析，但是不便于在实际应用中人工为文法规则附加语义处理。因此，怎样改进算法使得输出文法尽可能地接近领域专家人工给出的文法规则形式，增加后续语义处理的便利性是需要今后研究的问题。

（2）整个算法过程基于例句的分词结果，因此算法依赖一个较完善的领域词表。如何在规则推导过程中自动识别未登录的领域关键词，是我们后续研究中的一个重要课题。

（3）目前算法只考虑了上下文无关增强文法的苛刻型和跳跃型规则，其它类型规则的自动推导方法也是我们今后继续研究的方向。

参考文献

[1] 燕鹏举.对话系统中自然语言理解研究[D]. 北京：清华大学, 2002.

[2] Hugunin J., Zue V. On the Design of Effective Speech-based Interfaces for Desktop Applications. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech'97). Rhodes, Greece, 1997. 1335~1338.

[3] Issar S. A Speech Interface for Forms on WWW. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech'97). Rhodes, Greece, 1997. 22~25.

[4] Asoh H., Matsui T., Fry J., et al. A Spoken Dialog System for a Mobile Office Robot. In:Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, 1999. v3, 1139~1142.

[5] Os E.D., Boves L., Lamel L., et al. Overview of the ARISE project. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, 1999. v4, 1527~1530.

[6] Goddeau D., Brill E., Glass J., et al. GALAXY: A Human Language Interface to Online Travel Information. In: Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'94). Yokohama, Japan, 1994. 707~710.

[7] Seneff S. TINA: a Natural Language System for Spoken Language Applications. Computational Linguistics, 1992, 18(1): 61~86.

[8] Seneff S., Hurley E., Lau R., et al. Galaxy-II: A Reference Architecture for Conversational System Development. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98). Sydney, Australia, 1998. 931~934.

[9] Seneff S., Lau R., Polifroni J. Organization, Communication, and Control in the GALAXY-II Conversational System. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, 1999.1271~1274.

[10] Polifroni J., Seneff S. Galaxy-II as an Architecture for Spoken Dialogue Evaluation. In:Proceedings of the International Conference on Language Resources and Evaluation (LREC2000). Athens, Greece, 2000. 725~730.

[11] Zue V., Seneff S., Glass J., et al. JUPITER: A telephone-based conversational interface for weather information. IEEE Transaction on Speech and Audio Processing, 2000, 8(1):100~112.

[12] Seneff S., Polifroni J. Dialogue Management in the Mercury Flight Reservation System. In:Proceedings of the ANLP/NAACL 2000 Workshop on Conversational Systems. Seattle,USA, 2000. 1~6.

[13] Simpson, A., Fraser, N. Black Box and Glass Box Evaluation of the SUNDIAL System. In:Proceedings of the 3rd European Conference on Speech Communication and Technology (EuroSpeech'93). Berlin, Germany, 1993. 1423~1426.

[14] Huang C., Xu P., Zhang X., et al. LODESTAR: A Mandarin Spoken Dialogue System for Travel Information Retrieval. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99). Budapest, Hungary, 1999. v3,1159~1162.

[15] 黄寅飞，郑方，燕鹏举，等．校园导航系统 EasyNav 的设计与实现．中文信息学报，2001,15(4): 35~40.

[16] Brill E. Automatic grammar induction and parsing free text: A transformation-based approach. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. 1993. 259-265.

[17] Pereira F, Schabes Y. Inside-Outside reestimation from partially bracketed corpora. In: Pros. of the 30th Annual Meeting of the Association for Computational Linguistics. 1992. 128−135.

[18] 苑春法，陈刚，黄昌宁．基于词性和语义知识的汉语文法规则学习．中文信息学报，2000, 15(3):1-8.

[19] Grunwall P. A minimum description length approach to grammar inference. In: Wermter S, Riloff E, Scheler G, eds. Proc. of the Symbolic, Connectionist and Statistical Approaches to Learning for Natural Language Processing. LNCS 1040, Springer-Verlag, 1996. 203−216.

[20] Wolff GJ. Unsupervised grammar induction in a framework of information compression by multiple alignment, unification and search. In: de la Higuera C, Adriaans P, van Zaanen M, Oncina J, eds. Proc. of the Workshorp at ECML/PKDD2003: Learning Context-Free Grammars. 2003. 114−124.
[21] Klein D, Manning CD. A generative constituent-context model for improved grammar induction. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 128−135. http://acl.ldc.upenn.edu/P/P02/
[22] Klein D. The unsupervised learning of natural language structure [Ph.D. Thesis]. Stanford University, 2005.
[23] Clark A. Unsupervised induction of stochastic context-free grammars using distributional clustering. In: Daelemans W, Zajac R, eds. Proc. of the CoNLL 2001. Morgan Kaufmann. 2001. 105−112.
[24] Adriaans P, Trautwein M, Vervoort M. Towards high speed grammar induction on large text corpora. In: Hlavac V, Feffrey G, Wiedermann J, eds. Proc. of the SOFSEM-2000, Theory and Practice of Informatics. LNCS 1963, Springer-Verlag, 2000. 173−186.
[25] van Zaanen M. Bootstrapping syntax and recursion using alignment-based learning. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann. 2000. 1063−1070.
[26] van Zaanen M, Adriaans P. Alignment-Based learning versus EMILE: A comparison. In: Krose B, de Rijke M, Schreiber G, van Someren M, eds. Proc. of the Belgian-Dutch Conf. on Artificial Intelligence (BNAIC). 2001. 315−322.
[27] Cicekli I, Guvenir HA. Learning translation templates from bilingual translation examples. In: Applied Intelligence, vol.15. 2001. 57−76.
[28] Katsuhiko Nakamura. Incremental Learning of Context Free Grammars by Bridging Rule Generation and Search for Semi-optimum Rule Sets [A]. Proc. of the 8th International Colloquium on Grammatical Infer-ence (ICGI 2006) [C]. Heidelberg: Springer, 2006. 72-83.
[29] Katsuhiko Nakamura1 and Takashi Ishiwata. Synthesizing Context Free Grammars from Sample Strings Based on Inductive CYK Algo-rithm [A]. Proc. of the 5th International Colloquium on Grammatical Inference (ICGI 2000) [C]. Heidelberg: Springer, 2000. 186-195.
[30] Helen M. Meng and Kai-Chung Siu. Semiautomatic Acquisition of Semantic Structures for Understanding Domain-Specific Natural Lan-guage Queries [J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2002, 14(1): 172-181.
[31] 刘智博, Michael Brasser, 郑方, 徐明星. 一个基于文本输入的口语对话系统的新的实现策略[J].计算机科学, 2006, 22(11): 205-209.
[32] 王厚峰, 王波. 基于句子对齐的汉语句法结构推导的计算模型[J]. 软件学报, 2007, 18(3):538-546.
[33] 周强, 黄昌宁. 汉语文法规则自动构造方法研究. 中文信息学报, 1997, 12(3):1-5.
[34] P. Langley and S. Stromsten, Learning Context-Free Grammars with a Simplicity Bias, Machine Learning: ECML 2000, LNAI 1810, Springer-Verlag, pp. 220-228,2000.
[35] C. de la Higuera and J. Oncina, Inferring Deterministic Linear Langauges, Computational Learning Theory; 15th Annual Conference on Computational Learning Theory (COLT 2002), LNCS 2375, Springer-Verlag, pp. 185-200, 2002.
[36] Sophia Katrenko and Pieter Adriaans.Grammatical Inference in Practice:A Case Study in the Biomedical Domain[A]. Proc. of the 8th International Colloquium on Grammatical Inference (ICGI 2006) [C]. Heidelberg: Springer, 2006. 188-200.
[37] Katsuhiko Nakamura. Incremental Learning of Context Free Grammars by Bridging Rule Generation and Search for Semi-optimum Rule Sets [A]. Proc. of the 8th International Colloquium on Grammatical Infer-ence (ICGI 2006) [C]. Heidelberg: Springer, 2006. 72-83.
[38] Oates, T., Armstrong, T., Harris, J., Nejman, M.: On the relationship between lexical semantics and syntax for the inference of context-free grammars. In: Proceedings of AAAI. (2004) 431–436.
[39] 诺姆·乔姆斯基.句法结构邢公畹等据 1957 年本译[M], 北京: 中国社会科学出版社, 1979.
[40] Allen J., (Allen95) "Natural Language Understanding", 2nd edition, Benjamin/Cummings Publishing Company, Redwood City, California, 1995.
[41] 石纯一, 黄昌宁, 王家庆. 人工智能原理[M], 北京: 清华大学出版社, 1993.
[42] 范晓. 三个平面的语法观[M].北京: 北京语言文化大学出版社, 1996.

[43] Schadle I., Antoine J.V., Memmi D.. Connectionist Language Models for Speech Understanding: The Problem of Word Order Variation[J], Proceedings of EruoSpeech'99[A].1999.

附录 A 预定义的天气预报领域关键词表

- 注：中括号中是关键词类名，同一类下每一行为一个关键词，左部为汉字形式，右部为拼音形式。

//城市名
[mat_city_name]Copy
北京->bei3 jing1
包头->bao1 tou2
长春->chang2 chun1
长沙->chang2 sha1
成都->cheng2 du1
大连->da4 lian2
郑州->zheng4 zhou1
……

//气候类型
[mat_weather_type1]
天气->tian1 qi4

[mat_weather_type2]Copy
气温->qi4 wen1
温度->wen1 du4

//气候状态
[mat_weather_status]Copy
下雨->xia4 yu3
刮风->gua1 feng1
有风->you3 feng1
有雨->you3 yu3
晴->qing2
阴->yin1

//与天相关时间
[mat_date_rel_day]Copy
大前天->da4 qian2 tian1
前天->qian2 tian1
昨天->zuo2 tian1
今天->jin1 tian1
明天->ming2 tian1
后天->hou4 tian1
大后天->da4 hou4 tian1

//关于时间的约数词
[mat_date_about]Copy
这几天->zhe4 ji3 tian1
这些天->zhe4 xie1 tian1
这两天->zhe4 liang3 tian1
近几天->jin4 ji3 tian1
近两天->jin4 liang3 tian1
近期->jin4 qi1
近段->jin4 duan4
近段时间->jin4 duan4 shi2 jian1

[tag_i_want]Copy
请告诉我->qing3 gao4 su4 wo3
我想问一下->wo3 xiang3 wen4 yi2 xia4
我想知道->wo3 xiang3 zhi1 dao4

[tag_may_i_ask]Copy
麻烦查一下->ma2 fan2 cha2 yi2 xia4
请查一下->qing3 cha2 yi2 xia4
请问一下->qing3 wen4 yi2 xia4
请问->qing3 wen4

[tag_exist]Copy
有->you3
会->hui4
能->neng2
可能->ke3 neng2
是->shi4

[tag_exist_or_not]Copy
会不会->hui4 bu2 hui4
可不可能->ke3 bu4 ke3 neng2
是不是->shi4 bu2 shi4

[tag_y_m_y]
有没有->you3 mei2 you3

[tag_question_mark]Copy
吗->ma0
不会->bu2 hui4
不可能->bu4 ke3 neng2

[tag_how_many]Copy
多少->duo1 shao3
几->ji3

//温度度量
[tag_measure_w]
度->du4

//风力度量
[tag_measure_f]
级->ji2

//描述天气状况
[tag_weather_description1]Copy
冷->leng3
热->re4
暖和->nuan4 huo0

//描述气温状况
[tag_weather_description2]Copy
高->gao1
低->di1

[tag_what_about]Copy
怎么样->zen3 me0 yang4
怎样->zen3 yang4
如何->ru2 he2
咋样->za3 yang4

[tag_de]
的->de0

[tag_probability]
可能->ke3 neng2

[ato_month]
月->yue4

[ato_day]Copy

号->hao4
日->ri4

[ato_week]
礼拜->li3 bai4
星期->xing1 qi1

[ato_week_zhou]
周->zhou1

[ato_week_tian]
天->tian1

[ato_dgt_week]Copy
一->yi1
二->er4
三->san1
四->si4
五->wu3
六->liu4
日

[ato_1_m]
元->yuan2

//用于日期或时间
[ato_1_dt]
一->yi1

[ato_2]
二->er4

[ato_3]
三->san1

[ato_1_2]Copy
一->yi1
二->er4

[ato_2_3]Copy
二->er4
三->san1

[ato_1_10]Copy
一->yi1
二->er4
三->san1
四->si4
五->wu3
六->liu4
七->qi1
八->ba1
九->jiu3
十->shi2

[ato_1_9]Copy
一->yi1
二->er2
三->san1
四->si4
五->wu3
六->liu4
七->qi1
八->ba1
九->jiu3

[ato_10]
十->shi2

附录 B 包含日期相关规则的初始规则集

```
// File Name: WeatherForecast1.grm

[Lexical Analysis]
0

[Rules]
//月份
dgt_m ->ato_1_m
dgt_m ->ato_1_10
dgt_m *->ato_10 ato_1_2

//日份
dgt_d ->ato_1_10
dgt_d *->ato_10 ato_1_9
dgt_d *->ato_2_3 ato_10
dgt_d *->ato_2 ato_10 ato_1_9
dgt_d *->ato_3 ato_10 ato_1_dt


//月
sub_month *-> dgt_m ato_month
//日
sub_day *-> dgt_d ato_day
//月日日期
sub_month_day -> sub_day
sub_month_day *-> sub_month dgt_d
sub_month_day *-> sub_month sub_day
month_day -> sub_month_day
month_day->sub_month_day tag_and sub_month_day

//周日期
sub_week_day *-> ato_week ato_dgt_week
sub_week_day *-> ato_week_zhou ato_dgt_week
sub_week_day *-> ato_week ato_week_tian
week_day->sub_week_day
```

附录 C 算法输出的文法规则

//****以下为从语料中学到的规则****
qmE5ir *-> ato_week Ambiguity3
uZ4V0M *-> ato_week Ambiguity7
xADKC6 *-> ato_1_m Ab4Alr
Ab4Alr -> ato_month [1] Ambiguity7
DGDx3L -> mat_city_name [5] Hh3mF5
Hh3mF5 -> mat_weather_type1 [3] tag_what_about
KUCcni -> mat_weather_type2 [5] Nn216D
Nn216D -> tag_what_about [4] mat_city_name
QZBZIX -> Hh3mF5 [5] mat_city_name
UB2Pqh -> mat_city_name [5] X4BE9C
X4BE9C -> mat_weather_type2 [5] tag_what_about
_H1BSX -> mat_weather_type1 [5] cisrtg
cisrtg *-> tag_weather_description1 gM0gbB
gM0gbB -> tag_question_mark [5] mat_city_name
jor5VO *-> mat_date_rel_day UB2Pqh
m_03w7 *-> Ambiguity12 qurTes
qurTes *-> Ambiguity14 t5_IYN
t5_IYN *-> Ambiguity7 ato_day
wHqxy7 -> mat_city_name [5] zbZvhr
zbZvhr -> mat_weather_type2 [5] DNpk_M
DNpk_M *-> tag_weather_description2 tag_question_mark
GpZ9B6 -> mat_city_name [4] JTp_kj
JTp_kj -> mat_weather_type2 [5] MuQX2D
MuQX2D *-> tag_how_many tag_measure_w
Q6oMEY -> mat_weather_type2 [4] TAPCmi
TAPCmi *-> tag_weather_description2 gM0gbB
Wcor5C -> KUCcni [1] mat_date_rel_day
ZOPoOX -> _H1BSX [5] mat_date_about
cineph -> mat_city_name [5] fUO38C
fUO38C *-> mat_date_rel_day ivm0RW
ivm0RW -> mat_weather_type1 [5] m_NRs8
m_NRs8 *-> tag_weather_description1 tag_question_mark
pBmGat *-> _H1BSX mat_date_rel_day
sdNvUO *-> zbZvhr vHdsu7
vHdsu7 -> mat_date_about [5] mat_city_name
ziMids *-> JTp_kj vHdsu7
CVc7XN -> mat_city_name [3] FoLXx6
FoLXx6 -> mat_date_about [3] zbZvhr
I0cVgr -> Hh3mF5 [3] MCLKZE
MCLKZE *-> qmE5ir mat_city_name
P5bzAZ -> mat_city_name [5] SIKpii
SIKpii -> mat_date_about [5] ivm0RW
Vjam1D *-> KUCcni ZOJbDY
ZOJbDY -> Ambiguity11 [2] Ambiguity3
bpa1li -> KUCcni [3] e0BZ4C
e0BZ4C *-> ato_week_zhou Ambiguity7
hv9ONX -> Q6oMEY [2] mat_date_rel_day
l6AEo9 -> mat_date_rel_day [3] GpZ9B6
oJ8t6u -> vHdsu7 [3] ivm0RW
rczqQO *-> JTp_kj vO8fr8
vO8fr8 *-> mat_date_rel_day mat_city_name
yqz59t -> mat_city_name [5] BU7VTN
BU7VTN -> mat_date_rel_day [5] JTp_kj
EwySt7 -> JTp_kj [2] I76Ics
I76Ics *-> mat_city_name mat_date_rel_day
LBxxWN -> Hh3mF5 [5] OdZmwZ
OdZmwZ *-> ato_week RPxkfj
RPxkfj *-> ato_week_tian mat_city_name
VjY9YE *-> KUCcni YVwZzZ
YVwZzZ *-> ato_week Ambiguity11
axXXhi *-> Ambiguity14 d0wM0D
d0wM0D *-> Ambiguity11 hCXBCY
hCXBCY -> ato_month [1] kevrkh
kevrkh -> Ambiguity14 [2] Ambiguity7
nIWo3u *-> Ambiguity7 rkudMP
rkudMP *-> Ab4Alr uWV3n9
uWV3n9 -> Ambiguity14 [4] Ambiguity11
xxmT5t *-> ato_week A1VQPO
A1VQPO *-> ato_week_tian UB2Pqh
EDlFq8 *-> mat_city_name HfUv8t
HfUv8t *-> Ambiguity7 KJkkSN
KJkkSN *-> Ambiguity3 ivm0RW
NkThs_ -> JTp_kj [1] RXk7bk
RXk7bk *-> ato_week UqTXUE
UqTXUE -> Ambiguity12 [3] mat_city_name
X2jUvZ *-> ato_week_zhou _ESKej
_ESKej *-> UqTXUE ivm0RW
d7izXE -> JTp_kj [2] gKJoyY
gKJoyY -> mat_city_name [2] jlimgi
jlimgi *-> ato_week Ambiguity7
nQJb_D *-> Ambiguity12 qrh0IQ
qrh0IQ *-> Ambiguity3 GpZ9B6
t2IRj9 *-> ato_week_zhou wxgO2u
wxgO2u -> Ambiguity3 [3] wHqxy7
A8HDLP -> JTp_kj [4] DLgsm9
DLgsm9 -> mat_city_name [4] uZ4V0M
GeHi4t -> QZBZIX [4] JQffOO
JQffOO *-> Ambiguity12 NsG4o8
NsG4o8 *-> Ambiguity14 ato_day
QW6V7k *-> Ambiguity7 TyFKRF
TyFKRF *-> Ambiguity14 W96Hr_
W96Hr_ *-> ZOJbDY DGDx3L
_EFxak -> X4BE9C [5] cf5mTE
cf5mTE *-> Ambiguity12 fREjuZ
fREjuZ *-> Ambiguity14 jl49cj
jl49cj *-> Ambiguity7 mXEZWE
mXEZWE -> Ambiguity3 [5] mat_city_name
pz4OxQ -> _H1BSX [1] s2DMfa
s2DMfa -> Ambiguity12 [2] wE3BZv
wE3BZv -> Ambiguity14 [5] Ambiguity3
Fmud1u *-> KUCcni JY22KP
JY22KP *-> Ambiguity12 MrtSl8

MrtSl8 *-> kevrkh Ambiguity3
P31I3t -> mat_city_name [1] SFsFNG
SFsFNG *-> Ambiguity14 W90un0
W90un0 -> ZOJbDY [1] Hh3mF5
ZLsk6k *-> Q6oMEY bm0hQF
bm0hQF *-> Ambiguity7 eRr6q_
eRr6q_ *-> ato_month isSX9k
isSX9k -> Ambiguity12 [2] Ambiguity14
l4qMSE -> zbZvhr [2] UqTXUE
oyRJtZ -> mat_city_name [3] s9qzbb
s9qzbb *-> xADKC6 vMRoVv
vMRoVv *-> uWV3n9 Hh3mF5
yfpdwQ *-> JTp_kj BSQbea
BSQbea *-> m_03w7 mat_city_name
Fto0Yv *-> ivm0RW IXPQGP
IXPQGP *-> Ambiguity7 LzoFh9
LzoFh9 *-> Ab4Alr OaPD_u
OaPD_u -> Ambiguity14 [4] mat_city_name
SFfsJH *-> mat_city_name VgOhk0
VgOhk0 *-> xADKC6 YTef2l
YTef2l -> ato_day [5] ivm0RW
amN4MG -> JTp_kj [2] eYeUm_
eYeUm_ -> mat_city_name [2] hANK5k
hANK5k *-> xADKC6 uWV3n9
uTMbRw *-> Ambiguity7 xnc8sR
xnc8sR *-> Ab4Alr BZDZab
BZDZab *-> Ambiguity14 wxgO2u
EBbOUv -> mat_city_name [1] H4CDCQ
H4CDCQ *-> Ambiguity14 KGaAda
KGaAda *-> isSX9k OiCqXu
OiCqXu *-> ZOJbDY X4BE9C
//****以下是消歧及归一化的规则****
Ambiguity7 -> ato_1_10
Ambiguity7 -> ato_1_9
Ambiguity7 -> ato_2_3
Ambiguity7 -> ato_3
Ambiguity7 -> ato_dgt_week
Ambiguity3 -> ato_day
Ambiguity3 -> ato_dgt_week
Ambiguity14 -> ato_10
Ambiguity14 -> ato_1_10
Ambiguity11 -> ato_1_10
Ambiguity11 -> ato_1_2
Ambiguity11 -> ato_1_9
Ambiguity11 -> ato_1_dt
Ambiguity11 -> ato_dgt_week
Ambiguity12 -> ato_1_10
Ambiguity12 -> ato_1_2
Ambiguity12 -> ato_1_9
Ambiguity12 -> ato_2
Ambiguity12 -> ato_2_3
Ambiguity12 -> ato_dgt_week
Ambiguity7 -> ato_1_10
Ambiguity7 -> ato_1_9
Ambiguity7 -> ato_dgt_week
Ambiguity7 -> ato_1_10
Ambiguity7 -> ato_1_9

致　谢

本文是在导师王晓东教授的悉心指导下完成的。他以精深的专业造诣给了作者很好的学术指引，以严谨求实的治学态度培养了作者的科学精神。他忘我的工作精神和扎实的工作作风将是作者永远的学习榜样。在此作者谨向导师致以诚挚的谢意！

衷心地感谢所有在学习科研道路上给予过作者指导的老师，他们渊博的学识和严谨的学术风格奠定了作者学习科研的基础，尤其是清华大学语音与语言技术中心的郑方教授和邬晓钧老师在作者的研究过程中倾注了大量的心血和汗水，衷心地感谢他们！最后，感谢课题组所有同学的支持和帮助！

同时感谢各位专家在百忙之中对此文的审阅和赐教！

张　合

2009 年 4 月 8 日

攻读学位期间发表的学术论文目录

[1] X.D. Wang, L. Zhang, H. Zhang. Researches on The Path Optimization in Logistics Transport Network. Proc. of International Conference of Chinese Logistics and Transportation Professionals, VOL.4, July 2008.(EI)

[2] H. Zhang, X.D. Wang, H.T. Wang etc. Research on Ontology-based Environmental Quality Monitoring and Knowledge Management. PROC. OF ITESS 2008, VOL.2, PP. 756-760. (ISTP)

[3] H.T. Wang, X.D. Wang, H. Zhang etc. Design and implementation of the pollutant source's monitoring data acquisition system base on GPRS. PROC. OF ITESS 2008, VOL.2, PP. 701-706. (ISTP)

[4] C.L. Li, X.D. Wang, H.T. Wang, H. Zhang The implementation of GIS-based management system of contaminating and environmental quality monitoring. PROC. OF ITESS 2008, VOL.3, PP. 95-99. (ISTP)

[5] 张合，邬晓钧，王晓东，郑方．一种基于句子分割的文法规则自动推导算法．清华大学学报（自然科学版），已录用.（EI 源，全国中文核心期刊）

[6] 王晓东，王岁花，张合等. IT 课程目标及其语义 Web 应用．计算机应用与软件, 2008, 25(2): 86-88.（全国中文核心期刊）

[7] 王晓东，张合，王红涛．基于本体的语义检索模型研究．计算机工程与设计，2008，29(11): 2939-2941.（全国中文核心期刊）

[8] 张合，王晓东，杨照岩. Ontology 驱动的面向主题的网页关系识别．河南师范大学学报(自然科学版），已录用.（全国中文核心期刊）

[9] 王晓东，张合．基于 Ontology 的知识管理在环境质量监测中的应用．计算机应用与软件，已录用.（全国中文核心期刊）

[10] 张合，王晓东，杨建宇等．一种基于层叠条件随机场的古文断句与句读标记方法．计算机应用研究，已录用.（全国中文核心期刊）

[11] 张合，李淑平，胡伟强等．自然语言理解中文法规则自动推导，计算机科学,2008,35(8A), P193- P197.

[12] 王红涛，张合，王晓东．一种分布式开发本体的系统模型．计算机研究新进展, 2007.

独 创 性 声 明

本人郑重声明：所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得河南师范大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：_____日期：_____

关于论文使用授权的说明

本人完全了解河南师范大学有关保留、使用学位论文的规定，即：有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权河南师范大学可以将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、汇编学位论文。（保密的学位论文在解密后适用本授权书）

签名：_____导师签名：_____日期：_____