

博士学位论文

面向文景转换的中文浅层语义分析
方法研究

**RESEARCH ON THE METHOD OF CHINESE
SHALLOW SEMANTIC PARSING FOR TEXT-TO-
SCENE CONVERSION**

李世奇

哈尔滨工业大学

2011 年 6 月

国内图书分类号：TP391.2
国际图书分类号：681.324

学校代码：10213
密级：公开

工学博士学位论文

面向文景转换的中文浅层语义分析 方法研究

博士研究生：李世奇

导师：赵铁军 教授

申请学位：工学博士

学科：计算机应用技术

所在单位：计算机学院

答辩日期：2011年6月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.2

U.D.C: 681.324

Dissertation for the Doctoral Degree in Engineering

**RESEARCH ON THE METHOD OF CHINESE
SHALLOW SEMANTIC PARSING FOR TEXT-TO-
SCENE CONVERSION**

Candidate:	Li Shiqi
Supervisor:	Prof. Zhao Tiejun
Academic Degree Applied for:	Doctor of Engineering
Speciality:	Computer Application Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June, 2011
Degree-Conferring-Institution:	Harbin Institute of Technology

摘 要

本文针对中文浅层语义分析中的关键问题展开了全面深入的研究。浅层语义分析是自然语言处理领域里的研究要点,基于语言学特征和统计机器学习的方法是当前浅层语义分析的主流方法,该方法中最关键的因素是特征的选择和机器学习方法的优化。另外,本文中的浅层语义分析主要面向文景转换这项应用任务,文景转换是指把自然语言文本通过计算机自动转换成为相应的场景或动画,是一门具有重要理论和实际意义的新兴研究方向。本文首先对文景转换中必要的共指消解模块进行了研究;然后从特征选择角度对浅层语义分析方法进行了探索,发掘出在浅层语义分析中具有较强区分能力的句法特征;接着提出一种组合分类模型的方法对浅层语义分析进行完善;最后提出一种基于计算认知模型的方法,从更深层面对中文浅层语义分析进行了探索。具体地说,本文主要包括以下研究内容:

(1) 首先提出一种基于自适应谐振理论(ART)网络的无指导中文名词短语共指消解方法。该方法充分利用了名词短语自身特征,通过调整 ART 网络模型中的参数动态地控制聚类数量,有效解决了目前聚类共指消解中输出类别数目难以确定这一难题。另外聚类算法中还采用了一种基于信息增益率的特征选择方法,减少了区分度较弱特征给聚类所带来的干扰。该方法在保证共指消解准确率的前提下,具有较好的可移植性和鲁棒性,突破了目前文景转换中的浅层语义分析在预处理阶段的主要障碍。

(2) 本文从语言学特征层面深入地研究了中文浅层语义分析,提出一种基于多重句法特征的中文浅层语义分析方法。现有研究表明,对特征集合进行改进是目前提高浅层语义分析性能最有效的方法。本文提出将短语结构句法和依存句法两种类型的句法特征进行融合,为浅层语义分析提供了更加丰富和互补的句法信息。然后在这两个句法特征集合基础上,提出一种基于统计的组合特征选择方法,根据各个特征在语料库中的分布状况,快速有效地筛选出适于各分类阶段的组合特征。最后利用短语结构句法特征、依存句法特征以及在前两者基础上构造的组合特征进行语义分析相关的分类。实验表明,本文提出的多重句法特征集合能够有效地提高中文浅层语义分析的性能,在正确句法分析以及自动句法分析条件下均取得了较好的效果。

(3) 提出了一种基于组合分类模型的中文浅层语义分析方法,从优化

机器学习方法的层面进一步对浅层语义分析进行完善。本文在前面提出的多重句法特征集合基础上,采用五种机器学习方法:K 近邻、决策树、感知器、最大熵以及支持向量机,在训练语料上构造了五个语义角色分类模型,作为组合模型中的基本单元。接着通过一种输入相关的选通系统将五个基本分类模型有机地整合到一起,通过调整选通系统中的参数协调各个基本分类模型,控制组合模型的输出结果。最后采用 EM 算法在训练语料上对选通系统中的参数进行学习,在通用语料库上进行了相关的训练和测试,结果表明该方法能够显著地提高中文语义角色分析的效果。

(4) 最后,本文提出了探索性的基于计算认知模型的中文浅层语义分析方法,以认知理论为基本依据,通过模拟人类的语言理解过程,从本质上来研究中文浅层语义分析。首先设计了一种面向计算认知模型和文景转换的命题语义表示形式,这种命题形式能够简单高效地表达自然语言中蕴涵的语义信息。本文将该命题形式作为认知模型中的基本单元,然后在认知模型网络上模拟人脑中神经元的扩散激活机制,使符合上下文约束的命题节点不断被加强,不符合上下文约束的节点逐渐被削弱,根据当网络达到稳定状态时的最终激活命题节点,即可还原出谓词相关的语义分析结果。

关键词: 浅层语义分析; 语义角色标注; 自然语言处理; 文景转换; 计算认知模型;

Abstract

In this paper, we conduct comprehensive and deep academic research on the key issues of Chinese shallow semantic parsing (SSP). SSP is an essential research in the area of Natural Language Processing (NLP). Currently, the method based on linguistic features and statistical machine learning is the most prevalent method for SSP. The method involves two key factors: selection of linguistic features and optimization of machine learning method. Additionally, the SSP research in this paper is oriented to the Text-to-Scene conversion, which is to automatically convert the natural language text to the corresponding scene or animation by computer. It is a novel research area that has important theoretical and practical significance. Firstly, we study coreference resolution which is a necessary pre-processing module for the Text-to-Scene conversion. Secondly, we explore the issue of SSP from the perspective of linguistic feature selection and discover many discriminative syntactic features. Then, we propose a combined machine learning method to further improve the SSP. Finally, we study the issue from a deeper level and proposes a computational cognitive model-based approach to SSP. Specifically, this paper includes following contents:

(1) Firstly, we propose an Adaptive Resonance Theory (ART) network-based unsupervised noun phrase coreference resolution method for Chinese. The method makes fully use of the features of noun phrase. It can dynamically control the amount of cluster by adjusting the parameters of the ART network. Thus it provides an effective solution to the critical problem of cluster-based coreference resolution that the output cluster number, namely the number of the coreference set is difficult to determine or evaluate. Additionally, in the clustering algorithm, we use an information gain ratio-based feature selection method to reduce the interference caused by some weak clustering features. This method achieves a relative high accuracy of coreference resolution and it has good portability and robustness. It addresses the major obstacle of the pre-processing phase in the Chinese SSP in the Text-to-Scene conversion.

(2) Then, we intensively study the linguistic features for Chinese SSP and then propose a multiple syntactic features-based method. Current researches show that improving the linguistic feature set is the most effective method to enhance the performance of SSP at present. The proposed method integrates

constituent-based and dependency-based syntactic features into a basic feature set, and thus provides more extensive and complementary syntactic information to SSP. Further we propose a statistical combined feature selection method on the basis of the basic feature set. The statistical method can efficiently select discriminative combined features according to the distribution of each combined features in the corpus. Finally, we use the constituent-based syntactic features, the dependency-based syntactic features and the selected combined feature for classifications in SSP. Experiments show that the proposed method achieves better results on both gold-standard and automatic syntactic parsing.

(3) Further, we propose a combined machine learning method which is to improve SSP from the perspective of optimizing machine learning method. The proposed method is based on the above mentioned multiple syntactic features. It adopts five basic machine learning methods: K-Nearest Neighbor, Decision Tree, Perceptron, Maximum Entropy, and Support Vector Machine. We construct the five classification model using the five machine learning methods on the training corpus as the basic unit of the combined model. Then we use an input-dependent gating system to integrate the five basic classification models, and control the output of the combined model by adjusting the parameters of the gating system. Finally, we use Expectation Maximization algorithm to learn the parameters of the gating system using training data, and experimental results show that the method can significantly improve the effect of Chinese SSP.

(4) At last, this paper proposes an exploratory computational cognitive model-based Chinese SSP method. On basis of cognitive theory, the method simulates the language understanding process of human and then explores semantic analysis and calculations from fundamental properties. First we define propositional semantic representation oriented to the cognitive model and the Text-to-Scene conversion. The propositions can simply and efficiently express the semantics of natural language. We take the propositions as the neurons of the cognitive model. Then the contextually appropriate propositions will be gradually strengthened and inappropriate ones will be inhibited through iteratively spreading activations until the network stabilizes. Finally, the result of SSP can be achieved according to the activated propositions in the cognitive model.

Keywords: shallow semantic parsing, semantic role labeling, natural language processing, text-to-scene, computational cognitive model,

目 录

摘 要.....	I
Abstract	III
第 1 章 绪论	1
1.1 课题的研究背景及意义.....	1
1.2 课题的研究现状及发展趋势.....	3
1.2.1 浅层语义分析的任务描述.....	3
1.2.2 浅层语义分析的语料资源.....	4
1.2.3 浅层语义分析的基本流程和方法	8
1.2.4 浅层语义分析的评价体系.....	16
1.3 本文的研究内容及组织结构.....	20
1.3.1 本文的研究内容	20
1.3.2 本文的组织结构	22
1.3.3 论文整体与各章内容之间的关系	23
第 2 章 基于 ART 网络的聚类共指消解方法	25
2.1 引言	25
2.2 基于信息增益率的特征选择.....	27
2.3 基于 ART 网络的中文共指消解	30
2.4 实验及结果分析	34
2.4.1 实验数据	34
2.4.2 评测指标和方法	35
2.4.3 实验结果和分析	36
2.5 本章小结	40
第 3 章 浅层语义分析中的特征选择方法研究	41
3.1 引言	41
3.2 基本特征集合的构建	43
3.2.1 分类模型的选择和句法树的剪枝	43
3.2.2 短语结构句法特征集合的构建.....	45
3.2.3 依存结构句法特征集合的构建.....	48
3.3 基于统计的组合特征集合构建.....	50
3.4 实验及结果分析	53

3.4.1 实验数据及评测指标.....	53
3.4.2 组合特征选择的实验结果及分析	53
3.4.3 正确句法分析基础上的实验结果及分析	54
3.4.4 自动句法分析基础上的实验结果及分析	56
3.4.5 组合句法特征的性能分析	58
3.4.6 整体性能对比	59
3.5 本章小结	60
第 4 章 基于组合分类模型的浅层语义分析方法	62
4.1 引言	62
4.2 基于组合分类模型的浅层语义分析方法	63
4.3 基本浅层语义分析模型的构造	65
4.3.1 K 近邻(K-Nearest Neighbor, KNN)模型	66
4.3.2 决策树(Decision Tree, DT)模型	67
4.3.3 感知器(Perceptron)模型.....	68
4.3.4 最大熵(Maximum Entropy, ME)模型	70
4.3.5 支持向量机(Support Vector Machines, SVM)模型.....	72
4.4 基于 EM 算法的组合模型参数训练方法	74
4.5 实验结果及分析	76
4.5.1 实验数据及评测指标.....	76
4.5.2 正确句法分析基础上的实验结果及分析	77
4.5.3 自动句法分析基础上的实验结果及分析	78
4.6 本章小结	81
第 5 章 基于计算认知模型的浅层语义分析方法	82
5.1 引言	82
5.2 主要计算认知模型概述.....	83
5.3 命题语义表示形式的定义.....	85
5.4 基于认知模型的浅层语义分析基本方法	87
5.5 认知模型的构造和整合.....	89
5.5.1 构造候选命题	89
5.5.2 构造 LTM 网络	90
5.5.3 认知模型的构造阶段.....	91
5.5.4 认知模型的整合阶段.....	92
5.6 实验结果及分析	95

5.7 本章小结	98
结 论	99
参考文献	101
附 录	111
攻读博士学位期间发表的论文及其它成果	113
哈尔滨工业大学学位论文原创性声明及使用授权说明	115
致 谢	116
个人简历	117

Contents

Abstract (in Chinese).....	I
Abstract (in English)	III
Chapter 1 Introduction	1
1.1 Background and Significance	1
1.2 Present Research Situation and Recent Trends	3
1.2.1 Task Description of Shallow Semantic Parsing	3
1.2.2 Corpus of Shallow Semantic Parsing.....	4
1.2.3 Basic Frame and Method of Shallow Semantic Parsing.....	8
1.2.4 Evaluation system of Shallow Semantic Parsing	16
1.3 An Overview of this Dissertation	20
1.3.1 Contents of this Dissertation.....	20
1.3.2 Organizational Structure of this Dissertation.....	22
1.3.3 Relationship Between the Main Content and Each Chapter	23
Chapter 2 An Unsupervised Coreference Resolution Approach based on ART Network	25
2.1 Introduction	25
2.2 Feature Selection based on Information Gain Ratio	27
2.3 The Coreference Resolution Algorithm based on ART Network	30
2.4 Experiments and Discussion.....	34
2.4.1 Data Resources	34
2.4.2 EvaluationMetrics and Methods	35
2.4.3 Experiment Results and Discussion.....	36
2.5 Summary	40
Chapter 3 Research on Feature Seltion for Shallow Semantic Parsing	41
3.1 Introduction	41
3.2 Construction of the Basic Feature Set.....	43
3.2.1 Selection of Classification Model and the Pruning Method.....	43
3.2.2 Construction of the Phrase Structure Syntactic Feature Set.....	45
3.2.3 Construction of the Dependency Structure Syntactic Feature Set.....	48

3.3 Construction of the Combined Feature Set	50
3.4 Experiments and Discussion	53
3.4.1 Evaluation Data and Metrics.....	53
3.4.2 Experiment Results of Combined Feature Selection and Discussion	53
3.4.3 Experiment Results on Gold Parses and Discussion	54
3.4.4 Experiment Results on Automatic Parses and Discussion	56
3.4.5 Analysis of the Combined Syntactic Features.....	58
3.4.6 Comparison to Other Work	59
3.5 Summary	60
Chapter 4 Combined Classification Model-based Shallow Semantic Parsing Method.....	62
4.1 Introduction	62
4.2 The Shallow Semantic Parsing Method based on Combined Machine Learning	63
4.3 Construction of the Basic Shallow Semantic Parsing Model	65
4.3.1 K-Nearest Neighbor (KNN) Model.....	66
4.3.2 Decision Tree(DT) Model.....	67
4.3.3 Perceptron Model	68
4.3.4 Maximum Entropy (ME) Model	70
4.3.5 Support Vector Machines (SVM) Model.....	72
4.4 Parameter Estimation based on EM Algorithm.....	74
4.5 Experiments and Discussion	76
4.5.1 Evaluation Data and Metrics.....	76
4.5.2 Experiment Results on Gold Parses and Discussion	77
4.5.3 Experiment Results on Automatic Parses and Discussion	78
4.6 Summary	81
Chapter 5 Computational Cognitive Model-based Shallow Semantic Parsing Method.....	82
5.1 Introduction	82
5.2 An Overview of the Main Computational Cognitive Models.....	83
5.3 Definition of the Propositional Semantic Representation	85
5.4 The Shallow Semantic Parsing Method based on Computational Cognitive Model	87

5.5 The Construction and Integration of the Cognitive Model	89
5.5.1 Construction of Proposition Candidates	89
5.5.2 Construction of the Long-Term Memory Network	90
5.5.3 The Construction Phase of the Cognitive Model	91
5.5.4 The Integration Phase of the Cognitive Model	92
5.6 Experiments and Discussion	95
5.7 Summary	98
Conclusion	99
References	101
Appendix	111
Papers Published in the Period of PH.D. Education	113
Statement of Copyright	115
Acknowledgement	116
Resume	117

第1章 绪论

1.1 课题的研究背景及意义

在互联网和信息技术飞速发展的时代，人们每天要面对的文字信息量日益增加，使得人们对于计算机自动或辅助信息处理技术的需求越来越迫切。信息处理技术的水平和处理的信息总量已经成为衡量一个国家信息化水平的重要标志。目前计算语言学领域内普遍认为对自然语言的分析处理可层次化地分为：词法分析、句法分析、语义分析和语用分析四个层面，词法分析和句法分析技术是自然语言信息处理中的基石，经过了数十年的发展和积累，目前已经较为成熟。语义分析是自然语言处理中深层次的关键技术，实现自动语义分析是人工智能和自然语言处理领域研究者所追求的重要目标之一，而对语义分析部分研究目前尚处在初步探索阶段，还没有建立起完善的理论体系和切实有效的实现方法。因此，中文语义分析技术的研究对于提高我国信息处理技术的整体水平，改善人机交互方式，加快信息处理效率，推进信息化和现代化建设的进程都有着重要的意义。

语义分析概况地说就是研究如何将自然语言所表达的意义用某种规范的形式化方式表示出来，其中所指自然语言的单位可以是词、句子或篇章。但自然语言无论在语法层面还是语义层面都十分复杂多变，以当前计算语言学领域的理论和技术水平尚难以支撑全面和深层次的自然语言语义分析。因此目前语义分析的研究主要集中在对浅层语义的分析，它为解释复杂语言现象、探索自然语言理解过程和深层次语义分析提供了一条循序渐进式的实现途径。浅层语义分析(Shallow Semantic Parsing, 简称 SSP)是一种简化的语义分析形式，其研究内容主要是自然语言句子中的谓词及与其直接关联的成分。浅层语义分析所涉及的分析程度较浅，还比较依赖于句子的字面表达，不涉及指代、常识、推理和隐喻等深层次的语义内容。总体上看，尽管浅层语义分析属于语义层面，但在形式和方法上仍具有较多句法层面的痕迹。

浅层语义分析目前采用的形式主要是语义角色标注(Semantic Role Labeling, 简称 SRL)，其研究目标是从句子中识别出与目标谓词相关的语义角色或称为论元，并判断所识别语义角色的功能类型。根据目标谓词和语义角色之间的语义约束关系，语义角色的功能类型包括：施事、受事、与事、时间、地点、方式、目标、程度等。这些带有功能类型的语义角色信息构成了自然语言中最为基本的浅层语义信息，这些浅层语义信息能够为上层应用提供结构化的语义知识，为对信息的深入理解和分析奠定了基础。目前浅层语义分析技术已广泛应

用于自然语言处理领域的多项任务中，如信息抽取^[1]、问答系统^[2]、信息检索^[3]、故事理解^[4]和机器翻译^[5]等，并在各项任务中扮演着重要角色。

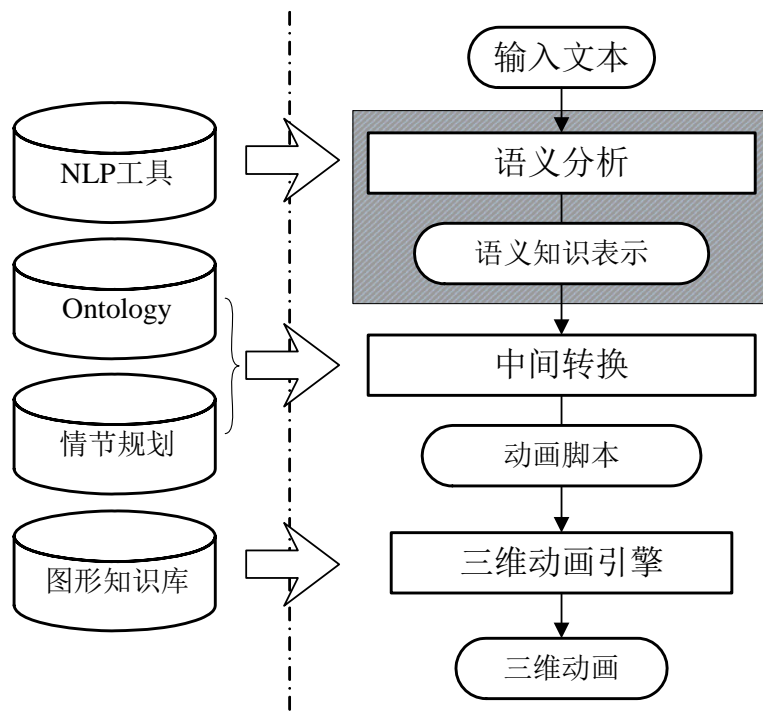


图 1-1 文景转换系统的基本框架

Fig.1-1 The basic framework of the Text-to-Scene system

本文要研究的中文浅层语义分析主要是面向文景转换任务。文景转换是指从自然语言描述到三维图形或动画的自动转换，该技术在教育、军事、影视动画、人机接口等领域有广泛的应用，是一项具有巨大实际应用价值、有待深入研究探索的新技术。对文景转换的研究可以追溯到上世纪七十年代，当时研究集中在自然语言指令驱动的图形生成方面^[6]。文景转换(Text-to-Scene)概念的正式提出是在本世纪初，由美国 AT&T 实验室的 B. Coyne 和 R. Sproat 在 2001 年 SIGGRAPH 会议上首次提出^[7]。文景转换研究涉及人工智能、自然语言处理和计算机图形学等领域理论和方法，主要研究自然语言中的信息自动获取、知识表示、故事理解、情节描述、时空推理、动画制作和自动生成等内容，其基本结构如图 1-1 所示。

本文研究的浅层语义分析是作为文景转换过程中的自然语言信息处理模块，在整个系统中扮演着提供数据源的重要角色，其效果会直接影响整个文景转换系统的性能，图 1-1 中阴影部分即为本文要研究的内容。我们从实践角度出发，通过对文景转换关键环节的分析，探索适于文景转换任务的中文浅层语义分析方法。从整体上看，本文研究的面向文景转换的中文浅层语义分析与通用的浅层语义分析在研究目标、内容和方法上基本一致，都是为了将自然语言


中谓词相关的语义信息用一种形式化方式表示出来。不同之处在于文景转换所需的语义信息表示形式更加具体，主要采取命题序列的形式，以动词作为命题的谓词，动词相关的施事、受事、时间、地点等作为命题的论元。因此本文的浅层语义分析中更加关注对于在文景转换中作用最为重要的动作相关实体、时间和地点类型论元的分析。另外由于文景转换任务的特殊性，通常浅层语义分析中不涉及对共指消解的处理，但作为系统中唯一的语义分析模块，本文将共指消解这一文景转换中所必须的语义分析技术也作为本文研究内容的一部分。

1.2 课题的研究现状及发展趋势

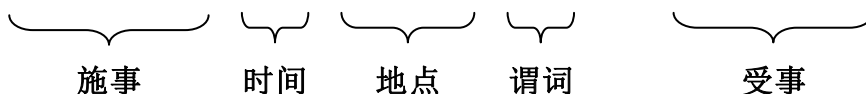
本文面向文景转换的浅层语义分析所采用的主要形式是语义角色标注 (Semantic Role Labeling, 简称 SRL)，中文语义角色标注是指从中文句子中识别出与目标谓词相关的语义角色并判断其类型。根据目标谓词和语义角色之间的约束关系，可以把语义角色分为若干个类型，如施事、受事、与事、时间、地点等等。中文计算语言学的发展以及中文自然语言处理中底层技术的逐渐成熟，如分词、词性标注、句法分析，都为中文语义角色标注奠定了基础。本文主要研究面向文景转换的中文浅层语义分析方法，正是研究如何自动获取自然语言描述中所蕴涵的语义角色信息，并将其应用在文景转换具体任务上。为了从整体上把握正确的研究方向，下面首先对国内外浅层语义分析方面的研究现状和发展趋势进行全面的介绍和分析。

1.2.1 浅层语义分析的任务描述

语义角色标注作为目前浅层语义分析的主要实现形式，其目标是从自然语言句子中识别出与目标谓词相关的全部语义角色并标出它们各自的语义角色类型。语义角色，又称论元或题元，这一概念来源于 Gruber 在词汇关系研究中提出的题元关系^[8]和 Fillmore 格语法理论中的语义格^[9]，其后 Chomsky 在转换生成语法中又对题元角色进行了研究^[10]，分析了谓词语义和句法结构之间的关系。根据目标谓词和语义角色之间的约束关系，可把语义角色分为若干个类型，包括施事(Agent)、受事(Theme)、时间(Time)、地点(Location)等，分别以英文：“John put the book on the shelf yesterday.” 和中文句子：“中美两国企业 28 日在芝加哥签订了 28 个项目的合同。” 为例，语义角色标注后的形式为：

1. [John] [put] [the book] [on the shelf] [yesterday].

Agent Predicate Theme Location Time

2. [中美两国企业] [28 日] [在芝加哥] [签订] 了 [28 个项目合同]。



在语义角色标注中，每个语义角色都与谓词(Predicate)直接相关，如在上例中文句子中，施事“中美两国企业”是谓词“签订”的发出者，受事“28 个项目合同”是谓词“签订”的承受者，时间论元“28 日”和地点论元“在芝加哥”则分别代表谓词“签订”发生的时间和地点。可见，浅层语义分析是一种以谓词为核心的语义分析形式，这里谓词是指动词和一些带有动作性的名词。可以看到语义角色标注具有目标明确、形式简洁、便于标注和评测等优点，因此成为浅层语义分析目前的主要形式。

在目前计算语言学界公认的自然语言分析的四个层面，即词法分析、句法分析、语义分析和语用分析中，语义分析位于句法分析的上层，浅层语义分析领域内的相关实验也证明了句法信息对浅层语义分析的必要性^[11,12]。所以目前浅层语义分析通常要在句法分析基础上进行，而句法分析的形式取决于其采用何种语法体系，目前常被浅层语义分析所采用的句法分析中的语法体系主要有：概率上下文无关语法、短语结构语法、依存语法和范畴语法。浅层语义分析的分析单元主要是句子，而语义角色的基本标注单元则有多种，主要有词、基本短语和句法成分三种，这取决于语义分析中所采用的句法信息形式，其中词主要用于基于依存句法分析的语义角色标注系统，基本短语主要用于基于组块分析的语义角色标注系统，句法成分主要用于基于短语结构句法分析的语义角色标注系统。本文与目前大多数的浅层语义分析研究一样，采用句法成分作为语义角色基本标注单元，主要原因在于一个语义角色应该是在句法和语义上较为完整的实体或描述，这与句法成分在粒度上最为接近。这种将句法成分作为标注单元的方法目前的不足之处在于对句法分析技术的依赖较强，对句法分析中的错误也较为敏感，而在句法分析技术较为成熟的语言上能够表现出良好的效果。中文计算语言学的发展以及中文自然语言处理中词法分析及句法分析技术的逐渐成熟，如分词、命名实体识别、词性标注、句法分析，都为中文浅层语义分析奠定了坚实基础。

1.2.2 浅层语义分析的语料资源

基于语料库的统计学习方法是目前自然语言处理领域中最常用和有效的方法，其核心思想是首先建立一个能够全面地描述某种语言现象的语料库，然后通过设计统计学习模型，在语料库上对该语言现象进行充分地抽象学习，最后学习获得的模型或规则便有能力对新文本中的该语言现象进行分析。在该体

系中，语料库资源是支撑统计学习方法的基石，语料库的形式、规模和质量都会直接影响到统计处理和分析的结果。浅层语义分析也不例外，需要以形式简单、规模大、质量高的语料库作为基础。目前，浅层语义分析领域中的语料库可以分成两大体系，一是以 FrameNet^[13]为代表的辞典型语料库，还包括 VerbNet^[14]等，其语义角色标注单元为词；二是以 PropBank^[15]为代表的应用型语料库，该语料库以句法树库为依托，其语义角色单元为句法成分，还包括 NomBank^[16]等。我们首先介绍 FrameNet 和 PropBank 这两个最具代表性的语料库，并借此来了解语义角色标注的整体情况。

表 1-1 FrameNet 的结构示例

Table 1-1 Demonstration of the architecture of FrameNet

Placing Frame:	
框架定义	Generally without overall (translational) motion, an <i>Agent</i> places a <i>Theme</i> at a location, the <i>Goal</i> , which is profiled. In this frame, the <i>Theme</i> is under the control of the <i>Agent/Cause</i> at the time of its arrival at the <i>Goal</i>
框架元素	Core: <i>Agent, Theme, Cause, Goal</i> . Non-core: <i>Area, Beneficiary, Cotheme, Degree, Distance, Duration, Manner, Path, Place, Results, Speed, Time, ...</i>
词汇单元	<i>archive.v, arrange.v, bag.v, bestow.v, billet.v, bin.v, bottle.v, box.v, place.v, placement.n, plant.v, plunge.v, pocket.v, position.v, pot.v, put.v, rest.v, rub.v, set.v, ...</i>
标注实例	1. [The waiter] _{AGENT} [carefully] _{MANNER} placed [the food] <i>Theme</i> [on the table] _{PLACE} . 2. [John] _{AGENT} put [the book] _{THEME} [on the shelf] _{PLACE} [yesterday] _{TIME} .

FrameNet 语料库由美国加州大学伯克利分校开发，是以 Fillmore 格框架语法为基础、高度详尽的英文谓词语义辞典，采用语义框架(Semantic Frame)作为谓词语义的描述形式，并通过大量计算机辅助人工标注的真实文本对语义框架中的各个元素进行实例化描述。作为 FrameNet 中的基本单元，语义框架相互之间并非独立，而是通过继承、使用、自框架等多种关系相互关联。每个语义框架由四个部分组成：框架定义(Frame Definition)、框架元素(Frame Elements)、词汇单元(Lexical Units)和标注实例(Annotation)。其中框架定义是指框架元素的含义描述；框架元素是指构成该框架所必须的核心元素以及可选的非核心元素；词汇单元描述了能够触发该语义框架的谓词及其对应的词

义；标注实例则是指带有该语义框架标注信息的自然语言句子。表 1-1 以动名词“Placing”框架为例给出了 FrameNet 语料库中的基本标注形式。目前英文 FrameNet 资源中定义了 960 个语义框架，其中包含 6,800 个完全标注词汇单元和 150,000 个标注的句子，除英语版本外，FrameNet 还有汉语、德语、法语、日语和西班牙语等多种语言的版本。

表 1-2 PropBank 的结构示例
Table 1-2 Demonstration of the architecture of PropBank

Predicate Put:			
框架描述	Frameset: put.01 location.	Frameset: put.02 say.	Frameset: put.03 result, attributive.
	Arg0: putter Arg1: thing put Arg2: where put	Arg0: speaker Arg1: thing said	Arg0: putter Arg1: thing put Arg2: attribute of arg1
标记实例	1. [John] _{Arg0} [put] _{Predicate} [the book] _{Arg1} [on the shelf] _{Arg2} [yesterday] _{ArgM-Loc} . Arg0: John Arg1: the book Arg2: on the shelf ArgM-TMP: yesterday Rel: put		

PropBank 语料库由美国宾夕法尼亚大学创建，是一个更偏向应用型的语料库，它在句法分析领域应用广泛的 Penn TreeBank 这一带有句法成分标注信息的树型语料库基础上，加入了谓词相关的语义角色标注信息，采用一种由谓词及其论元组成的命题形式来表达句子的浅层语义信息。PropBank 语料库在结构上分成两部分：框架描述和标注实例。框架描述部分描述了一个谓词的可能语义框架；标注实例部分是在句法树上对充当谓词语义角色的句法成分所进行的标记。表 1-2 以谓词 ‘put’ 为例说明了 PropBank 中的基本标注形式。PropBank 把语义角色分成两类：一类是核心语义角色，标记为 Arg0-Arg5，Arg0 通常表示谓词动作的施事，Arg1 通常表示谓词动作的受事，Arg2-Arg5 根据在不同的语义框架中具有不同的含义；另一类是修饰性语义角色，标记为 ArgM，其中包括时间、地点、副词、方式、目的、条件等 13 个子类型(详见附录 1)，因此还要另外附加上其子类型标记，如表 1-2 的标记实例中时间类修饰型语义角色 ‘yesterday’ 在 PropBank 中标记为 ‘ArgM-TMP’。

在目前浅层语义分析应用中，PropBank 较 FrameNet 更为常用，主要是由于 PropBank 自身的两个特点使其更适合于统计学习方法，一是其标记形式更

加泛化，大大地减少了需要标记的语义角色类别，能够有效地减轻数据稀疏现象；二是建立在带有词法和句法标注的 TreeBank 基础上，蕴含了十分丰富的语法信息，能够为语义分析所用。如图 1-2 所示，图中句法树中括号标记的内容是 PropBank 中的标记，其余部分均为 TreeBank 中提供的词法和句法标注信息，这些信息能够为浅层语义分析提供丰富有效的特征。PropBank 的不足之处在于仅标注动词相关的语义角色，而 FrameNet 标注了动词、名词、形容词、副词等相关的语义角色。因此，纽约大学按照 PropBank 中动词的标注方法，对该语料库中所有动作性名词进行了语义角色标注，建立了 Nombank 语料库，弥补了 PropBank 仅提供动词标注信息的局限。目前 PropBank 中包含了 3,600 个动词，5,050 个框架集合和 110,000 个语义角色。除英语版本外，PropBank 目前还有汉语、俄语、西班牙语等版本。

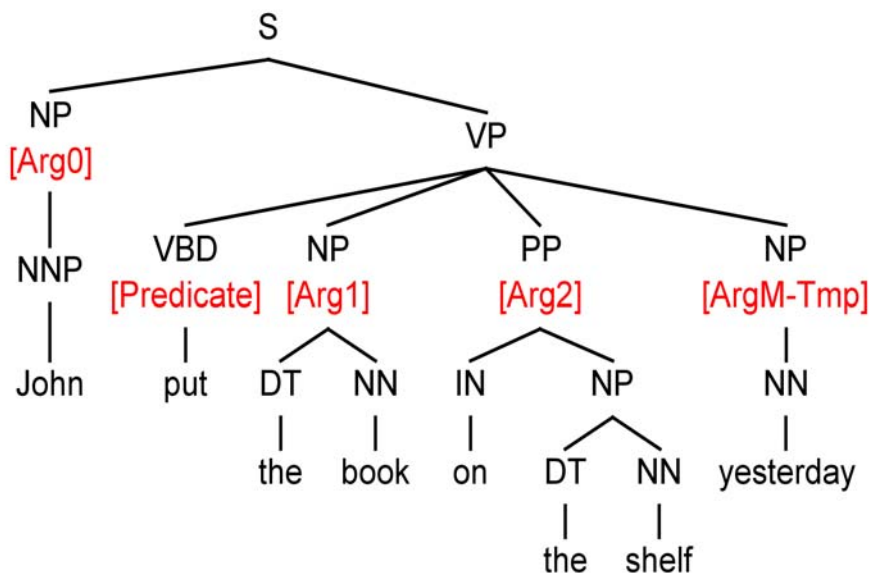


图 1-2 基于短语结构句法分析的浅层语义分析

Fig.1-2 Phrase structure parsing based shallow semantic parsing

受到上述英文语料库的启发，中文研究方面也涌现出一些浅层语义分析语料库，主要包括山西大学构建的汉语框架语义知识库(Chinese FrameNet)^[17]、清华大学构建的句法语义链接知识库^[18]、美国宾夕法尼亚大学在中文树库基础上构建的 Chinese PropBank^[19]和 Chinese Nombank^[20]。但从整体规模上来看，目前中文语料库的规模仍然远小于英文，建设大规模真实可靠的语料库资源还需要较长时间的积累以及中文语言学研究者的努力协作。

根据前面的描述，本文采用最适于统计学习方法的 Chinese PropBank 语料库作为浅层语义分析的实验数据，该语料库不但具有句法信息丰富、语义类别较少、适于机器学习方法等特点，另外其标注质量较高且规模较大，是目前中

文浅层语义分析研究领域应用最为广泛的语料库，因此在该语料库上可以进行充分的对比实验，来验证本文提出方法整体上的有效性。

1.2.3 浅层语义分析的基本流程和方法

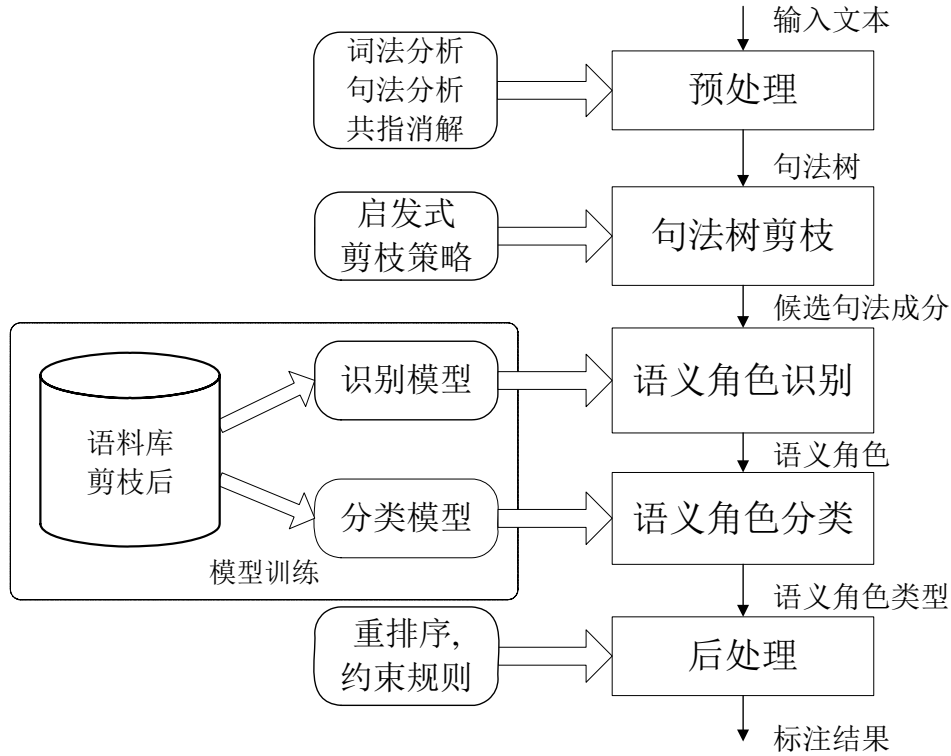


图 1-3 浅层语义分析基本流程

Fig.1-3 The basic procedure of shallow semantic parsing

统计学习方法是目前国内外浅层语义分析研究领域的主流方法，基于统计机器学习方法、以句法成分为标注单位的浅层语义分析整体流程可以分为五个步骤，如图 1-3 所示。首先是输入文本的预处理，主要采用分词、词性标记、名实体识别、句法分析等自然语言处理技术对输入文本进行处理，目的是得到自然语言文本的完整句法分析树，这些非语义相关的工作不在本文的研究范围之内。另外，共指消解并不是一般浅层语义分析中必要的预处理步骤，但由于本文面向文景转换任务的特殊性，必须获得实体之间的指向关系。因此，作为系统中唯一的语义分析模块，必须在语义分析预处理中将共指消解作研究内容的一部分；其次是句法树的剪枝，该步骤利用启发式方法过滤掉句法分析树中很多明显不是语义角色的句法成分结点，以缩小候选范围，降低后续分类步骤的复杂性，提高语义分析准确率^[21]；第三步是语义角色识别，该步骤逐个判断候选句法成分是否为目标谓词的语义角色，是一个典型的二值分类问题；第四步语义角色分类，针对语义角色识别步骤中识别出的语义角色，标记出其对

应的语义角色类型，语义角色类型十分有限(见附录 1)，因此该步骤通常被视为一个多值分类问题；最后一步是后处理，由于在第三步语义角色识别和第四步语义角色分类中，目标谓词相关的语义角色及其类型均是通过机器学习方法逐个识别的，彼此之间相互独立，因而忽略了同一谓词所辖的多个语义角色之间的约束关系。这一阶段主要是针对这种约束关系对语义角色标注结果进行修正，通过制定约束规则来对一些明显的错误现象进行更正，比如某一类型核心语义角色重复出现以及多个语义角色在位置上相互重叠。另外一些全局模型还要把一个目标谓词相关的所有可能语义角色进行组合，通过计算整体概率的方法对语义角色标注结果进行重排序^[22]。

在基于统计机器学习的浅层语义分析流程中，语义角色识别和语义角色分类是浅层语义分析中的核心模块，也是统计机器学习方法应用的集中体现。统计机器学习方法简要地说就是能够自动地从数据或者经验中获得知识并以此提高自身性能的一种算法^[23]，是继专家系统之后人工智能领域应用中所广泛采用的方法。机器学习方法避免了专家系统中依赖大规模知识库、缺乏学习和发现的能力、扩展和移植代价较大等缺点。该方法仅需要一种语言现象的一些已知数据作为训练语料，通过某种机器学习算法对训练语料进行学习，就可以实现对该语言现象的分析和预测。根据所需训练语料的标注情况，机器学习方法可以分成三种：有指导学习(Supervised Learning)，即训练语料全部需要人工标注，该方法指导性最强，学习效果比较好，但缺点在于人工标注大规模语料库所需开销很大，此类机器学习算法主要有：决策树、朴素贝叶斯、K 近邻、最大熵、支持向量机等；无指导学习(Unsupervised Learning)，即训练语料不需要人工标注，节省了人工标注的过程，学习效果略差，此类机器学习算法主要有：K 均值聚类、层次聚类、DBSCAN 算法、CLIQUE 算法、STREAM 算法等；半指导学习(Semi-Supervised Learning)，是有指导和无指导方法的折中，即小部分训练语料需要人工标注，大部分训练语料不需要人工标注，以有标注的样本为核心逐渐向无标记的样本扩散，此类机器学习算法主要有：EM 算法、Co-training 算法、Transductive SVM、Graph-based 半指导算法等等。

基于机器学习的浅层语义分析方法中，根据样本之间距离度量方式的不同又可以分为两类：基于特征向量(Feature-based)的方法和基于核函数(Kernel-based)的方法^[24]。基于特征向量的方法基本思想是采用人工定义的语言学层面的特征抽象地表示样本，然后将样本转换成为特征向量形式，间接通过向量之间的距离来度量样本之间的距离；基于核函数的方法基本思想是通过设计好的核函数来直接衡量样本之间的距离，通过比较两个样本结构之间的相似程度来衡量样本之间的相似度。基于特征向量的方法直观、有效、快速，是领域内目

前广泛采用的方法。特征选择是该方法的核心，如何选择区分度高的特征，以及如何构造特征模板是该方法所面对的主要问题。基于核函数的方法较为新颖，它引入了更加丰富结构化的信息，减轻了句法特征的数据稀疏问题，由于核函数自身的性质还能够融合基于特征向量的方法，但该方法计算量更大、模型更加复杂、实用性稍差。下面本文通过对这两种方法中代表性成果的回顾，来对国内外同类研究的现状进行分析和总结。

1.2.3.1 基于特征向量方法的研究

基于特征向量的浅层语义分析方法中，主要有以下这些具有代表性的研究工作。2002 年，Gildea 和 Jurafsky^[25]开创性地采用统计概率模型进行浅层语义分析，在标注过句法成分的 FrameNet 语料库基础上，利用多种统计学习技术进行了语义角色的识别，在已标注句法成分的测试样本上分析取得了 82% 的较高准确率。此外，该研究中提出的浅层语义分析系统的七个基本特征：谓词(predicate)、句法类型(phrase type)、次范畴框架(subcategorization)、路径(path)、位置(position)、语态(voice)和中心词(head word)，为后续研究所广泛采用，这七个特征也被称为“标准特征”；随后，Gildea 和 Palmer^[11]采用同样的模型和特征又在 PropBank 语料库上重新进行了实验，在已标注句法成分的测试样本上取得了 80.5% 的准确率，略低于前者。该研究的另外一个贡献在于详细分析了路径和中心词两个句法特征在浅层语义分析中的作用，证实了句法成分信息对于浅层语义分析的必要性。

随后的研究工作集中在对文献[25]中提出的标准特征集合的扩展上，人们不断尝试设计更具区分度的特征和利用更加丰富的句法信息。2003 年，Chen 和 Rambow^[26]尝试用树邻接语法从句法信息中挖掘出深层的语言学特征，利用 C4.5 决策树算法在自动依存句法分析的 PropBank 语料库上取得了 56% 的 F 值；Surdeanu 等^[27]在标准特征基础上，引入了短语内容核心词(constituent content word)、短语内容核心词词性、核心词的名实体类型、中心词词性等语义级的深层次特征，并设计了一套提取内容核心词的启发式规则，此处的内容核心词不同于标准特征里的中心词，中心词是指句法的中心，而内容核心词是指语义内容的中心。然后利用决策树 C5 模型把在正确句法分析上的浅层语义分析准确率提高到了 83.74%；Pradhan 等^[28]类似地在标准特征基础上，引入了名实体、中心词词性、谓词类别、部分路径、时间指示词等 12 个新特征，采用支持向量机模型在 PropBank 上取得了 84% 的精确率和 75% 的召回率，并详细评估了这些特征各自在语义角色识别和分类中所起的作用。

2004 年，Xue 和 Palmer^[21]深入讨论了 Gildea 和 Palmer 所提出的标准特征对于浅层语义分析各个阶段的贡献，并提出了一系列新的词法和句法特征：句

法框架、词汇成分类型、词汇中心词、语态位置组合、介词短语中心词等，并使用最大熵模型在 PropBank 语料库上进行了实验，实验结果表明采用新的特征后系统性能有了显著提高，基于人工标注句法树的 F 值达到了 88.51%，并详细分析了各个特征对语义分析的贡献。除此之外，这篇文章还有两个十分重要的贡献：一是提出了一个新型高效的句法成分剪枝算法，该算法基于启发式规则，简洁有效，能够将剪枝效率和召回率提高到 90.5% 和 97.9%；二是率先尝试使用在标准特征集合基础上构成的组合特征，他们在语义角色标注中采用了“谓词+短语类型”和“谓词+短语中心词”和“语态+位置”这三个组合特征，实验结果发现这些组合特征能明显提升语义角色标注效果，这两项改进都被后来的研究者所广泛采用。

随后，Xue^[29]又发现了一系列新的组合特征，包括“谓词类别+论元短语中心词”和“谓词类别+论元短语类型”两个有效组合特征；Ding 和 Chang^[30]提出了一种层次化特征选择策略，以短语结构句法分析为基础，定义了由基本特征和组合特征组成的模板，提出一种贪心选择算法从中提取出最有效的特征集合；Zhao 等^[31]类似地采用贪心特征选择算法从大规模依存句法分析特征中提取出有效的特征模版，在 CoNLL-2009 语义依存分析评测中取得了较好的成绩；Boxwell 等^[32]提出一种基于丰富特征的 SRL 方法，其中采用了组合范畴、短语结构和依存三种句法分析的特征，但多种句法分析带来了丰富信息的同时，也带来了较大的噪声。

2008 年，Toutanova 等^[22]提出一种基于概率的全局模型进行语义角色标注，区别于在分类过程中仅考虑谓词及单个语义角色的局部模型，全局模型还同时兼顾了与语义框架内其它语义角色之间的相互关系。该方法采用一种基于概率的局部语义角色标注模型和一种基于 log-linear 的重排序模型。该方法简要描述如下，SRL 整体标注概率 $P_{SRL}(L|t, v)$ 可表示成：

$$P_{SRL}(L|t, v) = \left[P_{SRL}^l(L|t, v) \right]^\alpha \cdot P_{SRL}^r(L|t, v) \quad (1-1)$$

式中 t ——短语结构句法树；

v ——目标谓词；

L ——对目标谓词的语义角色标注。

公式(1-1)中 $P_{SRL}^l(L|t, v)$ 表示局部模型， $P_{SRL}^r(L|t, v)$ 表示重排序模型。局部模型由语义角色识别(ID)和语义角色分类(CLS)两个步骤，因此 $P_{SRL}^l(L|t, v)$ 可由扩展成公式(1-2)，其中 n_i 表示句法树 t 上的第 i 个节点，

$$P_{SRL}^l(L|t, v) = \prod_{n_i \in t} P_{ID}(Id(l_i)|t, v) \cdot \prod_{n_i \in t} P_{CLS}(l_i|t, v, Id(l_i)) \quad (1-2)$$

式中 n_i ——句法树 t 上的第 i 个节点。

基于 log-linear 的重排序模型 $P_{SRL}^r(L|t, v)$ 可表示为公式(1-3)。其中 $\Phi(t, v, L)$ 表示特征映射函数, W 表示 log-linear 模型的参数向量, L_1, \dots, L_n 表示前 n 种可能的全局标注。

$$P_{SRL}^r(L|t, v) = \frac{e^{\langle \Phi(t, v, L), W \rangle}}{\sum_{j=1}^N e^{\langle \Phi(t, v, L_j), W \rangle}} \quad (1-3)$$

式中 n_i ——句法树 t 上的第 i 个节点;

$\Phi(t, v, L)$ ——特征映射函数;

W ——模型的参数向量;

L_j ——第 j 种可能的全局语义角色标注。

最后, 使整体标注概率 $P_{SRL}(L|t, v)$ 最大的全局标注 L 可由公式(1-4)求得。

$$L = \arg \max_{L \in L_1, \dots, L_N} \alpha \log P_{SRL}^l(L|t, v) + \log P_{SRL}^r(L|t, v) \quad (1-4)$$

作者采用上述方法在 Propbank 和 CoNLL 语料库上进行了语义角色标注实验, 在带有正确句法标注的 Propbank 语料库基础上 F 值达 91.2%, 较传统的局部模型提高了 2.8%; 在自动句法的 CoNLL 语料库基础上 F 值达 68.1%, 较传统的局部模型提高了 3.3%, 证明了该模型的有效性。

前面所述研究工作主要集中在有指导的机器学习方法, 在半指导和无指导浅层语义分析方法研究方面主要有以下代表性的工作。2004 年, Swier 和 Stevenson^[33]率先提出一种无指导的语义角色标注方法, 采用 bootstrapping 算法先利用 VerbNet 动词辞典标注语料库中一些明显无歧义的语义角色, 然后利用这些已标注样本和 backoff 概率模型对未标记的样本中有充分证据的部分进行标注, 这样不断迭代地扩展已标注数据的规模直到样本全部被标注, 最终该方法在通用语料库上取得了 87% 的准确率。2009 年, Furstenu 和 Lapata^[34]提出一种半指导的浅层语义分析方法, 其核心思想是先采用一个小规模的标注语料作为种子集合, 然后针对种子集合里的每个句子, 通过计算句子相似度的方法找出未标注集合中与之最相似的数个句子, 然后把种子句子的语义角色标注信息投影到未标注的句子, 完成对其的标注, 不断地扩展已标注的数据直到完成对整个语料库的标注。2010 年, Abend 和 Rappoport^[35]提出一种无指导的方法用来区分核心语义角色与修饰性语义角色, 先用一种基于语法和词性的推理算法对语义角色进行预测, 然后采用一种基于集成和自训练的方法将这些预测组合起来, 在 PropBank 语料库上的分类准确率达 80%。半指导和无指导方面的研究工作还包括[36-40]等。

1.2.3.2 基于核函数方法的研究

基于核函数的浅层语义分析较为新颖，相对于基于特征向量的方法，该方法相关研究的数量还比较少。在统计学习理论中，基于核函数方法研究较为广泛，主要是针对实际应用中大量存在的线性不可分问题，隐式地将低维空间线性不可分问题映射到高维空间，使之线性可分，然后通过核函数在原空间中就可以计算在高维空间的内积。基于核函数的浅层语义分析方法基本思想是通过定义一些核函数来直接衡量结构化样本之间的距离，目前浅层句法分析领域常用的核函数主要是基于树核函数的，通过比较两个字符串或者树结构中相同子结构的数量来衡量两个样本之间的距离。基于核函数的浅层语义分析研究方面，主要有以下代表性工作：

Moschitti^[41,42]最早提出了将核函数方法用于浅层语义分析。该方法在 Collins 和 Duffy 的卷积树核^[43]基础上提出了 PAK 核(Predicate-Argument feature Kernel)来提取谓词和论元之间部分句法树的结构化信息，通过比较两个句法结构之间相同子树的个数来两个结构之间的相似性。同时还利用核函数的特点，把基于特征向量的多项式核和基于句法结构的 PAK 核结合起来，使得句法结构信息与语言学特征相融合，在 CoNLL 和 PropBank 数据集上的均取得了优于单纯基于特征或核函数方法的结果；Che 等^[44]在 Moschitti 的 PAK 核基础上，进一步将其细分为两部分：路径核(Path Kernel)和成分结构核(Constituent Structure Kernel)，然后采用线性组合方法把这两个核组合成为一个混合核，在 CoNLL-2005 语料的 WSJ 部分上取得了 66.01% 的 F 值，比 PAK 核的结果高出 1.63%；Zhang 等^[45]在文献[44]工作的基础上进行了改进，提出一种基于语法驱动的混合卷积树模型，解决了传统卷积树核函数进行子结构比较时大多基于精确匹配，因而无法分辨相似或近义的句法成分的问题，例如“buy a car”和“buy a red car”，而基于语法驱动的卷积树核方法能够通过语言学知识进行节点和树结构的近似匹配，解决了上述问题，该方法在 CoNLL-2005 数据集上准确率达 87.96%，结合了基于特征的多项式核之后准确率高达 91.97%。

基于核函数的浅层语义分析方法的优点在于能够较好地反映全局信息和结构化的句法信息，同时还能兼顾语言学特征，而且核函数能够很好地融入支持向量机、感知器等学习算法，具有很强的实际应用能力；其缺点主要是针对实际问题设计合理有效的 Kernel 函数难度较大，另外结构化数据在计算时的复杂性较高，因此在学习和预测过程中的效率较低。

1.2.3.3 中文浅层语义分析的研究

中文语义角色标注的研究从方法上来讲与英文相类似，但目前的研究队伍和研究资源相对较少，中文方面代表性的研究工作主要有如下内容，美国科罗

拉多大学波尔得分校的 Sun 和 Jurafsky^[46]率先将英文中短语结构句法特征集合移植到中文语义角色标注上, 利用在宾州中文树库上训练的 Collins 句法分析器进行句法分析, 并利用 SVM 分类器在手工标注的小规模语料上进行了实验, 在正确和自动句法分析基础上分别取得了 83.1%和 76.7%的 F 值。美国宾夕法尼亚大学的 Xue 和 Palmer^[47]按照英文 PropBank 的标注方法, 在中文宾州树库基础上人工标注了中文命题语料库(Chinese PropBank, 简称 CPB), 其中共包含 760 篇文章、4854 个谓词、10364 个句子以及 92959 个语义角色。接着以 CPB 语料库为依托, 采用若干语言学特征和组合特征, 利用有指导最大熵分类模型, 在正确和自动句法分析基础上分别取得了 92.7%和 61.2%的 F 值。此外, CPB 语料库的创建对中文浅层语义分析研究产生了很重要的影响, 为后续的研究工作奠定了一个良好基础。

Xue^[29]在 CPB 语料库上对基于特征向量的中文动词和动作性名词的浅层语义分析进行了深入分析, 探索了词性标记和句法标记在中文浅层语义分析领域的重要作用。另外还发掘了一系列新特征, 包括把被信息、动词类别、短语话题等单一特征, 以及“谓词类别+论元短语中心词”和“谓词类别+论元短语类型”两个组合特征。研究结果表明在句法分析完全正确的情况下, 中文浅层语义分析的 F 值可以达到 92%这样令人满意的结果, 与当前英文浅层语义分析的水平相当; 然而在自动句法分析的情况下, 中文浅层语义分析的 F 值下降到 67%, 与英文的差距明显。主要原因是浅层语义分析对于句法分析中的错误较为敏感, 而中英文短语结构句法分析性能上尚有较大差距。

德国萨尔兰德大学 DFKI 研究中心的 Sun 在语言学特征方面进行了深入研究^[48], 在基于传统特征方法的基础上, 又定义了多个新的词汇以及句法特征, 采用 SVM 分类器在 CPB-1.0 语料库上取得了 93.49%的 F 值, 较[29]中 92%的结果又提高了 1.49%。

西北大学的安强强和张蕾提出一种基于依存树的中文语义角色标注^[49], 先将由中文短语结构分析树转化成的依存树作为数据集, 然后在选取特征集时, 引入了知网^[50]概念作为语义特征, 最后采用最大熵分类器进行实验, 在正确句法分析条件下 F 值达 90.68%, 较基于短语结构句法分析的系统提高了 0.2%。在基于依存分析的研究方面, 还有王步康等人^[51]对比了在 CPB-1.0 和 CoNLL-2009 两个语料库上训练的系统性能的差别, 在标准谓词方面 CPB-1.0 语料库取得了较好的结果, 在自动识别谓词方面 CoNLL-2009 语料库略优。

苏州大学的李军辉等^[52]探索了中文动词性谓词 SRL 对中文名词性谓词 SRL 的影响, 在传统基于特征向量的动词性谓词浅层语义分析基础上, 进一步提出了名词性谓词浅层语义分析相关的特征集合, 并且联合谓词自动识别实现

了中文名词性谓词的浅层语义分析。在中文 NomBank 的实验结果显示, 基于正确句法树和正确谓词识别的中文名词性谓词 SRL 性能 F 值为 72.67%, 而基于自动句法树和自动谓词识别的中文名词性谓词 SRL 性能 F 值仅为 55.14%。

北京大学的丁伟伟和常宝宝^[53]提出一种基于语义组块分析的中文语义角色标注方法。该方法将汉语语义角色标注从一个句法树节点的分类问题转化为序列标注问题, 使用了条件随机域模型取得了较好的结果, 同时省去句法分析的步骤, 减少语义角色标注对句法分析的依赖, 降低了浅层语义分析的复杂性。该方法在北京大学和中文树库上取得了 63% 的 F 值。

邵艳秋等人^[54]提出一种基于词汇语义特征的中文语义角色标注研究方法, 通过引入一些词汇语义特征对一些仅依靠句法分析难以解决的浅层语义分析问题进行处理。该文基于北京大学的语义词典 CSD, 引入了配价数和主客体语义类等词汇语义特征。在 CPB 语料库上进行十折交叉验证的结果显示, 引用词汇语义特征使浅层语义分析的整体 F 值比仅使用句法特征提高了 1%。

哈尔滨工业大学的车万翔^[55]提出了一种语法驱动的卷积树核方法进行中文语义角色标注, 该方法结合了语言学信息和结构化信息, 是语义角色标注研究领域中的重要创新; 山西大学的李济洪^[56]提出了基于汉语框架语义网的中文语义角色标注, 通过建立了类似 FrameNet 的汉语框架语义网资源, 采用词和基本块特征结合 CRF、SVM、ME 等统计学习模型进行了相关研究; 苏州大学王红玲^[57]采用基于特征向量的方法在短语结构句法分析和依存分析上分别进行中英文的语义角色标注; 刘怀军等^[58]和丁金涛等^[59]探索了中文语义角色标注的特征工程, 主要涉及特征的组合以及优化问题; 北京邮电大学的刘娜^[60]研究了中文四种特殊句式中的语义角色标注问题, 包括‘把’字句、‘被’字句、‘是’字句和‘使’字句, 详细分析了这些句式中的语义情景, 并通过机器学习的方法来对其进行实现。

1.2.3.4 面向文景转换的浅层语义分析研究

文景转换是一项较新的研究课题, 目前相关研究并不充分。浅层语义分析是文景转换过程中的自然语言信息处理模块, 在动画系统中扮演数据源的重要角色, 并能直接影响文景转换系统的输出结果。下面对国内外具有代表性的文景转换系统中的语义分析模块进行简要介绍。

美国 AT&T 实验室研发的 WordsEye 系统是目前最具代表性的文景转换系统^[7]。它由一个规模较大的三维模型库支持, 能够根据文本的简单描述生成对应的静态的三维场景。该系统的语义分析流程是, 首先输入文本经过词性标注、基于头驱动统计模型的句法分析处理, 转换成一种依存结构; 然后, 依存结构被解释为语义表示。接着根据语言学家的空间介词语义研究, 手工编写语

义函数将依存结构转换为空间关系，并将此关系表示成为一个语义表示片段；随后把语义表示转换为三维描述器，这些描述器可以标识实体的以下信息：框架、形状、部分、颜色、透明、默认大小、活动属性和空间方位标签，也能够描述人物的动作或姿势，包括单独人物姿势，特殊用途姿势，一般用途姿势，抓取姿势等。最后根据这些描述器添加背景生成输出场景。

瑞典 Lund 大学研发的 CarSim 系统可将一篇关于交通事故的报告，转化为对于当时情景的三维动画^[61]。其过程可划分为语言的分析 and 场景的建立两部分，这两步之间通过一种形式化语义描述联系起来。为了更好的描述交通事故建立的模型，它的定义包括静态对象、动态对象和发生碰撞的对象。静态对象包括两个属性参数：一个描述对象的自然属性，另一个描述对象的位置；动态对象也包括两个参数：初始方向和事件链，即该对象的一系列顺序的运动。碰撞是发生在肇事者和受害者之间的，肇事者必须是动态物体，而受害者可以是动态的也可以是静态的，还要定义冲撞的坐标和事故中汽车发生冲撞的部分。然后根据上述内容定义目标模板，并采用信息抽取的方法提取出上述生成碰撞动画所需的语义信息。

我国中科院陆汝钐院士带领研开的天鹅系统是国内最具代表性的动画自动生成系统^[62]。其技术路线是，将人工智能技术和基于知识的方法引进动画生成的全过程，它的目标是只要有了一个适当的故事，以受限自然语言的形式把它输入计算机里，从此时开始，直到最终生成动画，每一步都是在计算机辅助下完成的。该系统主要包含五个步骤：首先是对故事文本作自然语言理解。故事文本是受限自然语言，句子的结构主要为“主语+谓语+宾语”的形式；使用自然语言单句理解系统对其分析，包含基于最大匹配法的分词、基于 CATN 的语法分析、格框架树表示语义分析；基于常识的指代分析和常识检查；故事情节理解。提出角色图的浅方法、基于高维和上下文有关文法或者角色动态特征的深方法；把故事改编为分场景剧本；根据分场景剧本作动画设计，包括角色、背景、动作等设计，其中体现了时间、空间规划；根据上述设计和规划，利用事先构造好的动画素材库和声音库，生成完整的动画。

综上所述，面向文景转换的浅层语义分析与通用的浅层语义分析在研究目标内容和方法上基本一致，都是为了将自然语言中谓词相关的语义信息形式化地表示出来。但由于文景转换任务的特性，其中的浅层语义分析更加关注文本中提及的动作、与动作相关实体、时间和地点等类型语义论元的分析。

1.2.4 浅层语义分析的评价体系

1.2.4.1 浅层语义分析的评价指标

为了能够客观地反映自动浅层语义分析系统的性能，对浅层语义分析方法做出正确评价，建立合理、有效、规范、量化的评价指标和评价方法是十分必要的。目前国内外浅层语义分析领域中，常用的评价指标主要有四个：准确率(Accuracy, 简称为 A)、精确率(Precision, 简称为 P)、召回率(Recall, 简称为 R)以及 F 值(F_1 -score, 简称为 F)，它们的具体计算方法如下：

$$Accuracy = \frac{\text{正确标注的实例个数}}{\text{测试数据中实例总数}} \quad (1-5)$$

$$Precision = \frac{\text{正确标注的语义角色个数}}{\text{分类器预测为语义角色的个数}} \quad (1-6)$$

$$Recall = \frac{\text{正确标注的语义角色个数}}{\text{测试数据中语义角色总数}} \quad (1-7)$$

评价指标 F_β -score 是精确率和召回率的调和平均值， β 通常取值为 1，即为 F_1 -score，其计算方法如公式(1-8)所示。

$$F_\beta\text{-score} = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (1-8)$$

另外值得注意的一点是单纯地比较上述几项指标的大小还不能完全准确地衡量系统之间性能是否具有统计性的差异，还跟数据规模的大小有关。例如，一个在 10 个数据上测试准确率为 80% 的系统，并不一定优于在 100 个数据上测试准确率为 75% 的系统。因此，在浅层语义分析研究中通常还需要一种衡量统计显著性的指标。卡方检验是领域内应用较为广泛的假设检验方法，属于非参数检验范畴，其主要功能是比较两个或两个以上样本率或构成比之间的差别是否具有统计意义，根本思想在于比较理论频数和实际频数的吻合程度或拟合优度问题。卡方检验的统计量是卡方值，它是样本类别中实际频数 A 与理论频数 T 差值平方与理论频数之比的累计和，可表示为公式(1-9)。

$$\chi^2 = \sum \left[(A - T)^2 / T \right] \quad (1-9)$$

表 1-3 四格表形式卡方检验

Table 1-3 Chi-square test in the form of four-fold table

		参考系统	
		True	False
标注结果	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

浅层语义分析中的卡方检验采用通常采用四格表形式，如表 1-3 所示。其中四个格子分别代表正确标注的正例数 TP 、错误标注的正例数 FP 、错误标注的反例数 FN 、正确标注的反例数 TN ，四格表卡方检验的卡方值可表示为公式(1-10)。

$$\chi^2 = \frac{(TP \cdot TN - FP \cdot FN)^2 (TP + FP + FN + TN)}{(TP + FP)(TP + FN)(FP + TN)(FN + TN)} \quad (1-10)$$

采用上述四格表形式同样能够方便计算准确率(Accuracy)、精确率(Precision)和召回率(Recall)三个评价指标，在实际应用中较为常用，具体计算方法如公式(1-11)、公式(1-12)和公式(1-13)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1-11)$$

$$Precision = \frac{TP}{TP + FP} \quad (1-12)$$

$$Recall = \frac{TP}{TP + FN} \quad (1-13)$$

可见，卡方统计显著性检验能够兼顾系统之间的性能和数据规模的大小，避免了单纯比较系统准确率所带来的片面性。在浅层语义分析问题中，卡方的自由度通常为 3，在一定的显著性水平 p 下，卡方值越大表明系统之间的差异越明显。卡方值可通过查表来获得，当卡方值大于一定数值时，两个系统显著不同的概率是 $1-p$ 。例如，当显著性水平 $p=0.05$ 时，查表可知如果卡方值大于 7.815，两个系统的差别有 95% 的概率具有统计显著性^[55]。

1.2.4.2 浅层语义分析相关的国际评测

标准化评测是促进一项研究技术快速发展和提高效率的重要手段。从 2004 年起，国际上对浅层语义分析研究的重视程度逐渐提高，涌现出了许多浅层语义分析相关的国际会议及评测，这些评测为浅层语义分析研究提供了完备的数据集和评价标准，为该项研究提供一个良好的平台，提高了研究的整体水平。这些国际评测中最具影响力的当属每年一届的 CoNLL 国际会议 (Conference on Computational Natural Language Learning)，以及每三年一届的国际评测 Senseval (Workshop on Sense Evaluation)，后续扩展为 Semeval (Workshop on Semantic Evaluation)。与本文研究直接相关的有 CoNLL-2004^[63] 和 CoNLL-2005^[64] 设立的语义角色标注任务，CoNLL-2008^[65] 设立的联合依存句法-语义分析任务和 CoNLL-2009^[66] 的多语言句法和语义分析任务；2004 年 Senseval-3^[67] 中的语义角色自动标注任务，Semeval-2007^[68] 设立的 FrameNet 式语义结构抽取任务，以及 Semeval-2010 设立的中文新闻句子中事件检测也都

是与浅层语义分析研究直接相关的评测项目^[68]。

在 CoNLL-2004 评测上来自世界各国的十家单位使用了感知器、Winnow 算法、最大熵和支持向量机等多种统计机器学习方法参加语义角色标注的评测。其中美国科罗拉多大学的 Hacıoglu 等人^[70]以短语为语义角色标注单元，用 SVM 算法在不使用全局特征的条件下，在测试集合上取得了 61% 的 F 值，获得了该项评测中的最好成绩；在 CoNLL-2005 上有 19 家机构参加了评测，美国伊利诺大学香槟分校的 Koomen 等人^[71]使用 SNoW 工具包中的 Winnow 算法^[72]，综合多种深层句法分析结果，加上整数线性规划后处理方法取得了最好的成绩，测试集合上的 F 值接近 80%；

随后 CoNLL-2006 和 CoNLL-2007 两届评测的内容主要是依存句法分析，人们发现句法分析和语义分析之间有着十分密切的联系，在接下来的 CoNLL-2008 和 CoNLL-2009 两届评测中的主要任务是联合依存句法和语义分析，结果表明联合学习确实能够在一定程度上提高依存句法分析和语义分析的结果。在 CoNLL-2008 上共有 25 个系统提交的结果，封闭测试中瑞典隆德大学的 Johansson 等人^[73]采用基于在线 passive-aggressive 算法^[74]的句法模型以及基于特征和逻辑回归分类器^[75]的语义模型，取得了最好成绩，结果 F 值达 85.49%；开放测试中德国 DFKI 实验室的 Zhang 等人^[76]组合了目前领域内经典的两个依存句法分析器 MST Parser^[77]和 MaltParser^[78]作为句法模型，然后采用类似基于特征的分类方法作为语义模型，取得了 79.61% 的最好结果；CoNLL-2009 评测进一步将联合依存句法和语义分析扩展到了多国语言上，在提交的 13 个系统中，Zhao 等^[79]采用一种计算复杂度较高的大规模特征模版选择方法，在领域内语义分析封闭测试中取得了 80.47% 的最高 F 值；哈尔滨工业大学信息检索研究室的车万翔等人^[80]采用基于支持向量机模型的谓词词义分类、基于最大熵模型的语义角色分类和基于整数线性规划的后处理方法，在领域外语义分析开放测试中获得了 80.06% 的最高 F 值。

2004 年第三届 Senseval 评测上，首次设立了浅层语义分析相关任务，香港理工大学的 Ngai 等人^[81]采用基于特征向量的方法并集成了 SVM、boosting 和最大熵三种机器学习方法，在无约束的语义角色标注评测上取得了较好的效果，精确率和召回率分别达到 87.4% 和 86.7%；随后在 Semeval-2007 评测设立的 FrameNet 语义结构抽取任务上，Johansson 和 Nugues 提出的一种基于 WordNet 相似度的扩展方法^[82]，有效地减轻了 FrameNet 中的数据稀疏现象，利用 SVM 分类器在测试语料取得了最好成绩；Semeval-2010 评测首次设立了中文新闻句子中的事件检测项目，该项目实质上是浅层语义分析直接相关的应用任务，清华大学在此项评测中取得了 53.76% 的最高整体准确率^[69]。

1.3 本文的研究内容及组织结构

1.3.1 本文的研究内容

通过对当前浅层语义分析方法的研究和综述发现, 基于语言学特征和统计机器学习的方法是目前国内外浅层语义分析领域的主要研究方向。基于统计机器学习的浅层语义分析方法中有三个关键要素: 一是特征的设计和选择, 二是机器学习方法的优化和改进, 三是训练语料库的规模和质量。其中还有很多问题有待深入研究, 包括高区分度特征的设计、句法信息的应用、组合特征的选择、机器学习方法的优化以及新模型的探索等等。本文针对这些关键问题, 结合本研究面向的文景转换具体任务, 对中文浅层语义分析方法进行了探索 and 实验, 主要的研究内容和关键技术包括以下方面:

(1) 篇章共指消解预处理的研究

共指消解是一种深层语义理解技术, 在面向文景转换任务的浅层语义分析中具有非常重要的意义。共指是指两个语言单位指向现实世界中同一实体, 是自然语言描述中一种极为普遍的现象, 确定指向和被指向的语言单位的过程就是共指消解。本文研究主要面向文景转换任务, 其目标是将文本转换成为三维动画, 生成动画所需的最主要信息是动作及动作相关的人物或实体, 在自然语言描述中这些人物或实体经常由代词或名词短语表达, 只有明确代词或名词短语在现实中指代的对象才能确定动作的发出者或承受者。因此, 篇章中的代词或名词短语共指消解是必须解决的问题, 在此基础上才能进行浅层语义分析。

共指消解也是一个语义层级的问题, 在研究方法方面目前采用的主要是基于标注语料库机器学习的方法, 而中文共指消解方面缺乏面向文景转换任务且规模较大的标注语料库, 人工标注这样的语料库工作量过于巨大。在这样的客观情况下, 采用无指导聚类方法是一种较为合理的选择。然而在基于聚类的共指消解问题中面临的主要问题是类别数目无法事先预知且难以估计, 因此本文提出一种基于 ART 网络的聚类共指消解方法, 该方法以自适应谐振理论为基础, 能够通过调节网络参数, 动态生成聚类数量, 能够有效地解决上述问题, 同时还具有较好的通用性和可移植性。

(2) 浅层语义分析中句法特征选择的研究

特征选择是浅层语义分析的核心问题, 如何设计有效的、区分度较高的句法特征, 如何选择特征之间的组合方式, 如何定义特征模版等问题都是特征选择中的重要问题。目前, 在基于特征的浅层语义分析研究中, 对于单一句法形式和单一特征的研究已经较为充分, 然而对多重句法形式和组合特征的研究还很不充分, 只有基于单一句法形式和少数几个被实验证实有效的组合特征被采用, 仍有许多能够提高系统性能的组合特征尚未发掘, 缺乏快速有效的组合特

征选择方法是该问题的主要障碍。

因此,本文提出一种基于多重句法特征的中文浅层语义分析方法,该方法融合了短语结构句法和依存句法两种类型的句法特征,并在其基础上构造出一些高效的组合特征用于浅层语义分析过程。首先总结了现有的系统中被证实有效的短语结构句法特征和依存句法特征作为基本特征集合,然后提出一种基于统计的组合特征选择方法能够高效地在基本特征集合构造出组合特征。我们通过一个统计量在语料库上估计每个组合特征对语义角色识别和分类过程所产生的影响,高效地筛选出有助于语义角色识别和分类的组合特征。最后将由短语结构句法特征、依存句法特征和组合句法特征共同构成的多重句法特征集合用于语义角色相关的分类过程。

(3) 浅层语义分析中机器学习方法的研究

机器学习方法是浅层语义分析问题中的关键,但目前从机器学习方法方面来提高浅层语义分析的效果比较困难,因为机器学习算法在浅层语义分析中的研究已经比较充分,数量较为有限机器学习模型几乎都已经被尝试过,这方面的研究处于一种难以突破的瓶颈状态,使得目前大多数工作都集中在特征选择方面。针对这种情况,本文提出一种基于组合分类模型的浅层语义分析方法,从机器学习方法的角度进一步对浅层语义分析进行完善。本文在前面提出的多重句法特征集合基础上,采用 K 近邻、决策树、感知器、最大熵以及支持向量机五种机器学习方法,在训练语料上构造了五个语义角色分类模型,作为组合模型中的基本单元。接着通过一种输入相关的选通系统将五个基本分类模型有机地集成到一起,通过调整选通系统中的参数协调各个基本分类模型,控制最终组合模型的输出。最后采用 EM 算法在训练语料上对选通系统中的参数进行学习,并在通用的标准语料库上进行了训练和测试。

(4) 基于认知模型的浅层语义分析方法

最后,本文提出了探索性的基于计算认知模型的中文浅层语义分析方法,该方法以语言理解的认知理论为基础,能够抽象地描述和表现人类认知的表征和过程,包括语义处理单元、信息加工的阶段、各个要素之间的相互关系以及阶段过渡过程中的行为等,从理解的本质出发来研究语义分析处理技术。首先设计了一种面向计算认知模型和文景转换任务的命题语义表示形式,这种命题形式能够简单高效地表达自然语言中蕴涵的语义信息。将该命题形式作为认知模型中的基本单元,也就是神经元,然后在认知模型上模拟人脑中神经元的扩散激活机制,使符合上下文约束的命题节点不断被加强,不符合上下文约束的节点逐渐被削弱,根据当网络达到稳定状态时的最终激活命题节点,就可以通过谓词的语义框架直接获得谓词相关的浅层语义分析结果,最后我们同样在标

准语料库上对该方法进行了相关实验，并与上述机器学习方法进行了比较。

1.3.2 本文的组织结构

本文在组织结构上共分为五章，除第一章绪论部分外，其余四章的内容分别按上一小节中所述的四项研究内容进行展开。各章节具体内容安排如下：

第一章，绪论中首先介绍了本文的研究背景和意义，接着总结了浅层语义分析的研究现状以及发展趋势，对近些年该领域的主要研究方法进行了较为全面的综述，最后概述了本文的主要研究内容和组织结构。

第二章，提出了一种基于自适应谐振网络的篇章聚类共指消解方法，该方法能有效地解决聚类的共指消解中面临的类别数目无法确定这一难题。其中采用了基于信息增益率的聚类特征选择方法，以减少从有指导分类中直接引入的区分度较弱特征给名词短语聚类所带来的干扰。实验证明，该聚类共指消解方法在同类研究中取得了较好的效果。

第三章，首先总结了现有的语义角色标注系统中有效的短语结构句法特征和依存结构句法特征，以此为基础构造基本特征集合。然后提出一种基于语料库的统计方法估计每个组合特征划分类别的能力，筛选出有效的组合特征用于语义角色识别和分类步骤，将由基本特征和组合特征构成的特征集合集合 **SVM** 分类器进行浅层语义分析。最后在 **CPB** 标准语料库上对上述方法进行了充分的实验，验证了该方法的有效性。

第四章，提出一种基于组合分类模型的浅层语义分析方法。首先构造多个基本机器学习模型，然后采用一种输入相关的选通系统将这些基本机器学习模型整合在一起，通过调整选通系统中的参数使在对每个样本的分类过程中能够充分发挥各分类算法的优势，减少单一分类模型产生的错误，提高整体准确率。最后采用 **EM** 算法根据语料库来调节组合分类模型中选通系统的参数，并在 **CPB** 语料库上对该方法进行了评价。

第五章，提出一种基于计算认知模型的中文浅层语义分析方法。首先设计了一种面向计算认知模型和文景转换任务的命题语义表示形式，将该命题形式作为认知模型中的基本单元。然后根据上下文和长时记忆库构造认知模型网络，并在网络上模拟人脑中神经元的扩散激活机制，使符合上下文的命题节点不断被加强，不符合上下文的节点逐渐被削弱。最后根据网络达到稳定状态时的激活命题节点，还原出谓词相关的语义分析结果。

最后，结论部分总结了本文研究中的主要成果、创新点和不足之处，描述了本文的主要贡献，并提出了下一步的工作计划。

1.3.3 论文整体与各章内容之间的关系

本文的主要内容是面向文景转换的中文浅层语义分析方法研究，这里文景转换中的“文”是指篇章级别的叙述文本，“景”是指三维动画，本文研究的是文景转换中的浅层语义分析部分，如图 1-1 所示，其目标是将篇章级的文本自动转换成一种以描述动作为主的语义表示形式。例如，句子：“中美两国企业 28 日在芝加哥签订了 28 个项目的合同。”，可表示成如下语义框架：

表 1-4 例句对应的语义框架

Table 1-4 Semantic frame of the example sentence

谓 词	签订
施 动	中美两国企业
受 动	28 个项目合同
动作时间	28 日
发生地点	在芝加哥
动作程度	——

这种语义表示将在之后的中间转换模块中补充动作相关的具体参数转换成动画脚本，最后通过动画引擎结合模型库将动画脚本生成相应的三维动画。这些后续模块并不在本文的研究范围之内。本文所研究的浅层语义分析正是要研究如何通过浅层语义分析来自动地从文本中获得这种谓词语义的表示形式。图 1-4 说明了本文各个章节之间的核心内容以及逻辑关系。如图所示，首先，第一章绪论部分交代了本文的研究背景和意义，以及论文的主要研究内容。第二章针对文景转换中关键的共指消解预处理模块，提出一种基于 ART 网络的聚类共指消解方法。接下来在第三章和第四章分别针对特征选择和机器学习方法这两个浅层语义分析中的核心问题展开了研究，并进行了详细的论述。第五章前面相关工作的启发下，从认知层面对中文浅层语义分析进行了探索性的研究，提出一种基于计算认知模型的中文浅层语义分析方法。最后在结论部分对本文研究内容进行了整体的总结。本文第二章、第三章和第四章之间是顺承关系，第二章共指消解是作为浅层语义分析的预处理阶段，因为共指消解在文景转换中有着十分重要作用，如果没有该模块将无法确定文中共出现了多少个实体，也无法判断动作的发出者究竟是哪个实体；然后，第三章先从特征选择方面对浅层语义分析问题进行了研究；第四章在第三章基础上又从机器学习方法的角度对浅层语义分析进行了进一步的优化；第五章相对较为独立，以前面相关研究为启发，提出了一种基于认知的浅层语义分析的方法，基于认知的方法未来自然语言处理相关研究的重要趋势，因此本文对该方法进行了探索。

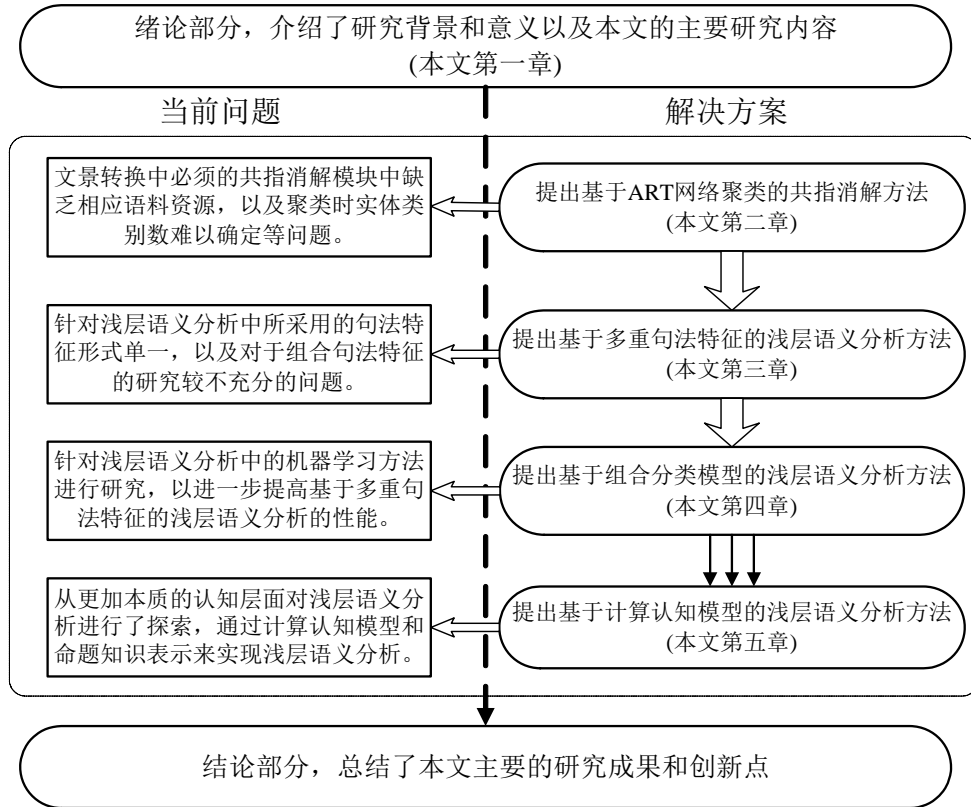


图 1-4 本文各章节之间的关系示意图

Fig.1-4 Illustration of relationship between thesis chapters

第2章 基于 ART 网络的聚类共指消解方法

2.1 引言

共指消解(Coreference Resolution)是一种通用的自然语言理解技术,在文本处理相关任务中扮演着重要角色。共指是自然语言中一种很普遍的现象,是指自然语言上下文中两个语言单位指向现实世界中同一实体的一种特殊语义关联,能够简洁高效地表述语义、衔接上下文。按照互指单元的成分可具体分为人称代词指代、指示代词指代、名词短语指代、部分整体指代等。共指消解的目的就是确定篇章中各个名词短语所指向的实体,其结果可以表示为对篇章中名词短语集合的一个划分,称为共指划分。因此,在实际应用中共指消解的过程通常形式化为一个求解共指划分的过程^[83]。

本文研究主要面向文景转换任务,最终目标是将文本转换为场景或动画,而生成动画所需最重要的语义信息是动作(或关系)以及动作(或关系)相关的人物及实体,而在自然语言描述中这些人物或实体经常会通过代词来表达,为了明确这些代词现实中的指代对象,因此只有先对篇章中的代词进行共指消解预处理,才能在其基础上进行浅层语义分析。例如,下面这段故事文本:“一头狮子在海边游荡,看见鲨鱼跃出水面,便劝他出来与自己结为同盟。”文中包括名词或代词短语‘一头狮子’、‘海’、‘水’、‘鲨鱼’、‘他’、和‘自己’(由于故事采用拟人手法,故用代词‘他’),只有先通过共指消解得出篇章中名词短语的共指划分,才能通过共指划分识别出篇章中共出现了多少个人物或实体(本例中的共指划分为: { ‘一头狮子’, ‘自己’ }、{ ‘鲨鱼’, ‘他’ } 和 { ‘海’、‘水’ }),进而识别动作的发出者或承受者究竟是哪个人物或实体。此外,代词所包含的语义信息也不如名词本身丰富,因此,标记出代词的指代对象也为后续的语义角色标注工作提供了更加丰富的语义信息。

共指消解是一项具有挑战性的研究课题,目前的主要方法有基于规则的方法和基于统计的方法,早期基于规则的方法主要利用手工建立的领域知识库,根据词性标记和句法分析等知识设计规则进行共指消解;基于统计的方法是目前的主流方法,取得了较好效果,其中还可细分为分类和聚类两种方法^[84]。有指导的分类方法主要用于标注语料较充分的情况下,效果也相对较好;但是目前中文共指消解领域中标注语料缺乏,人工标注语料库的代价很大,因此采用无指导的聚类方法是中文共指消解中更为合理的选择。本文基于这样的考虑采用了一种基于 ART 网络的聚类共指消解方法,该方法在具有相对较高的识别准确性同时,受领域和标注语料库的限制较少。

基于聚类的共指消解研究目前还处于一个探索阶段,从文献方面看主要有以下几项代表性工作:1999年,Cardie 首先提出无指导的名词短语共指消解方法^[85],采用了短语本身、短语中心词、短语序号、代词类型、冠词类型、同位语、专有名词、单复数、语义类别 9 个特征和基于启发式规则的聚类方法;2003年,Bergler 提出基于模糊集理论的共指消解方法,采用语义距离、短语重合程度、缩写词、简单代词消解和短语中心词 5 条模糊规则进行处理,然后经过模糊集合并和解模糊化过程形成共指链,其结果并不理想^[86]。2004年 Bean 和 Riloff 利用信息抽取方法获取上下文信息,然后根据这些信息判断指代语与先行语的相容性^[87]。在中文方面,2006 年香港理工大学的 Wang 和 Ngai 提出基于改进的 K 均值聚类算法的中文无指导共指消解^[88],选定 12 个适于中文的聚类特征,采用 Cardie 的距离度量,在人工标注的 30 篇语料上获得了较好效果;2007 年南京大学周俊生等采用类似 Cardie 的特征空间和距离度量,提出一种基于图划分的无指导共指消解方法^[89]。

目前聚类共指消解方法面临的主要问题是聚类结束条件难以判断。其原因是共指消解问题中篇章内的真实实体数量无法从原文中直接获得,而且难以准确估计。因此导致聚类数目也无法预知,聚类数目参数对于聚类的效果具有重要影响^[90]。目前的方法大都以从实验数据中获得的经验收敛阈值作为结束的条件,无法做出准确的判断,从而影响了聚类共指消解的性能。其次还有一个关键问题,现有聚类共指消解研究中所采用的聚类特征大多是直接从基于有指导的共指消解特征中移植过来的,并未充分考虑聚类的特性,因此一些区分度较弱特征会给聚类结果带来很大的干扰。

本研究提出将自适应谐振理论(Adaptive Resonance Theory,简称 ART)引入到中文共指消解任务中,提出了基于 ART 网络的中文聚类共指消解方法。与其他聚类方法不同,该方法能够隐式地利用名词短语自身特征进行聚类,有效地克服了聚类过程中输出类别数目难以确定这个主要问题。此外能够根据数据集来动态调节网络参数,控制聚类算法的输出类别,同时解决了距离度量难以定义和样本点波动大的问题。另外,我们还采用一种基于信息增益率的特征选择方法过滤掉区分度较弱的聚类特征,以减少这些特征对聚类算法的干扰。整体上看该方法具有适于共指消解问题,不依赖领域标注语料库,具有高效性、鲁棒性和可移植性的特点,可直接应用于真实文本。

本章首先介绍了基于信息增益率的聚类特征选择方法,以减少从区分度较弱的特征给名词短语聚类所带来的干扰。然后,详细介绍基于自适应谐振理论网络的聚类中文共指消解方法的基本框架和算法描述,包括如何利用名词短语自身特征,如何通过改变网络参数动态调节聚类数量等一系列问题。最后,我

们在自动内容抽取(Automatic Content Extraction, 简称 ACE)语料库上对该方法进行了相关实验, 实验结果表明该方法在同类研究中取得了较好的效果。

2.2 基于信息增益率的特征选择

特征集合的构造是基于机器学习方法的共指消解中的首要环节, 特别是对于聚类方法, 选择特征的质量是聚类效果的决定因素之一。传统分类方法通常能够根据训练语料中的分布情况, 对于区分度较弱的特征, 赋予其较小的权重。但是对于聚类方法, 这些特征可能会带来较大噪声, 因此本文采用基于信息增益率特征选择方法过滤掉区分度较弱的特征。信息增益是信息论中的重要概念, 它能够有效衡量给定特征区分训练样本的选择能力, 决策树 ID3 算法中增长树的每一个步骤就是使用信息增益作为特征选择的标准^[91]。其基本思想可概括为, 一个特征的信息增益就是数据集在被这个特征分割前后期望熵的差值。在数据集 D 上, 特征 A 的信息增益 $Gain(D, A)$ 可表示为公式(2-1)。

$$Gain(D, A) = E(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} E(D_v) \quad (2-1)$$

式中 D ——数据集;

A ——特征;

v ——特征的值;

D_v —— D 中特征 A 取值为 v 的子集。

公式(2-1)中 $Values(A)$ 代表特征 A 所有可能值的集合, D_v 代表 D 中特征 A 取值为 v 的子集, 即 $D_v = \{d \in D \mid A(d) = v\}$; $E(D)$ 表示数据集 D 的熵, 如果 D 中存在 c 个类别, 则 $E(D)$ 可由公式(2-2)计算得到。

$$E(D) = - \sum_{i=1}^c p_i \log_2(p_i) \quad (2-2)$$

式中 p_i ——数据集 D 中第 i 个类别的概率。

一般认为信息增益值越大的特征区分度就越高, 但该方法存在一种倾向性偏差, 即取值较多的特征信息增益值往往较大, 尽管其区分度能力可能并不好, Quinlan 在文献[92]中证实了信息增益率可有效减少这种倾向性偏差。因此本文采用信息增益率代替信息增益来作为特征的选择标准, 特征 A 在数据集 D 上的信息增益率 $GainRatio(D, A)$ 如公式(2-3)和公式(2-4)所示。

$$GainRatio(D, A) = \frac{Gain(D, A)}{SplitInfor(D, A)} \quad (2-3)$$

$$SplitInfor(D, A) = - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \log_2 \left(\frac{|D_v|}{|D|} \right) \quad (2-4)$$

首先本文根据目前领域内的相关文献^[85,88,89]，总结了具有代表性的聚类共指消解特征作为候选特征，如表 2-1 所示。我们用这些特征来验证本文提出的基于 ART 网络的聚类共指消解方法。根据上述方法，采用 ACE 标准语料库中的广播新闻(Broadcast News，简称 BN)部分(详见本文 2.4.1 节)作为特征选择数据集，分别计算表 2-1 中 11 个候选聚类特征在该数据集上的信息增益率，最后计算信息增益率均值，结果如表 2-2 所示。其中少数 $SplitInfor(D, A)$ 值为零的实例将被忽略，这种情况表示该篇章中所有名词短语中特征 A 的取值都相同，因此不具有区分度。本文忽略表 2-2 中信息增益率不足 50%的三个特征，即词汇级特征专有名词、短语级特征中心词词频和上下文特征短语序号，为避免这些区分能力较弱特征对聚类的干扰，我们将上述三个特征从特征集合中去除，选择表 2-1 中其余的八个特征——性别、单复数、短语本身、中心词、中心词词性、短语句序号、生物性和语义类别作为共指消解的聚类特征。

表 2-1 候选聚类特征及其描述

Table 2-1 The clustering feature candidates and their descriptions

特征类别	特征名称	特征概述
词汇级	性别	包括：雄性、雌性、未知
	单复数	包括：单数、复数、未知
	专有名词	该短语是否为专有名词或缩写词
短语级	短语本身	名词短语本身词串
	中心词	短语的中心词词串
	中心词词性	短语的中心词词性：名词或代词
	中心词词频	中心词在篇章中出现的频次
上下文	短语序号	短语按照出现先后顺序序号
	短语句序号	短语所在句子在文章中的序号
语义级	语义类别	按照 ACE 实体类别标准划分
	生物性	短语所指实体是否具有生物性

在构造好的聚类特征集合中，对于短语词串、中心词串、中心词词性、性别、单复数、短语句序号、生物性这 7 个特征取值的获取，可采用词典或规则的方法直接从语料库中获得。唯有名词短语的语义类特征无法直接获得，语义类别特征较为复杂，该特征的取值根据 ACE 实体标注规范^[93]，可将实体划分成 7 个大类 45 个子类。为避免过多噪声，本文将这 7 个大类作为实体语义类别特征的值域，即 PER(人)、ORG(组织机构)、GPE(地理、政治实体)、LOC(处所)、FAC(人造建筑)、WEA(武器)、VEH(传输设备)。统计表明，这 7

个类别的分布很不均匀，在 ACE 中文语料中 PER、ORG、GPE 三大类实体所占比重较大，占样本总数的 86%，其余四类所占比例均不足 5%。从表 2-2 中可见该特征的信息增益率较高(0.717)，能够为共指消解提供了的重要信息。因此本文采用一种基于支持向量机(Support Vector Machine，简称 SVM)模型的分类方法对名词短语的语义类特征进行识别。SVM 分类方法凭借其在实际应用中的高准确率，适于高维特征空间以及小样本集合等特点，广泛应用在自然语言处理领域很多分类问题中，并表现出了良好的效果。

表 2-2 各候选聚类特征的信息增益率

特征类别	特征名称	信息增益率
词汇级	性别	0.656
	单复数	0.580
	专有名词	0.456
短语级	短语本身	0.865
	中心词	0.830
	中心词词性	0.501
	中心词词频	0.425
上下文	短语序号	0.397
	短语句序号	0.533
语义级	语义类别	0.717
	生物性	0.775

此处的分类目标是将实体分为 7 个语义类别，而 SVM 分类方法是一种典型的二值分类方法。与其他二值分类器面临的问题类似，在面对多值分类问题时可以采用“一对多”(One-versus-All Classification)和“一对一”(Pairwise Classification)两种策略^[94]，构造多个二值分类器，然后将其组合成为一个多值分类器。“一对多”策略是指为每个类别构造一个分类器，来区分该类模式与其余的所有类别模式，因此须建立与待分类别数目相同的 SVM 分类器，最后通过各分类器的分类置信度判断样本所属类别；“一对一”策略则要为任意两类之间都要训练一个 SVM 分类器，针对 N 元分类问题需要构建 C_N^2 个分类器，最后通过分类器大数投票的方式选出样本所属类别。文献[94]证明了对于 SVM 多值分类问题，“一对一”策略的识别正确率要高于“一对多”策略。因此本文采用“一对一”方法，将多值分类转换为多个二值分类的组合来处理。选择四个分类特征为：词本身、上下文词(窗口大小取 1)、中心词以及中心词在知网(又称 HowNet)中的概念定义。HowNet 是一部类似于 WordNet 的汉语语义知识辞典^[50]，其功能强于普通的对应词表，此处将 HowNet 中对词汇的概念定义，即 DEF 域的值作为一个重要特征。例如，中文句子“台北桃园中正机场

昨天晚上发生意外”中 ORG 型实体“台北桃园中正机场”的四个特征：本身，上下文窗口词，中心词以及中心词在 HowNet 中概念定义的取值分别是：‘台北桃园中正机场’、‘Null, 昨天’、‘机场’和‘facilities|设施’。

在获得特征集合中的八个特征：性别、单复数、短语本身、中心词、中心词词性、短语序号、生物性和语义类别的取值之后，就可以用特征对文本中的名词短语进行描述，形成的特征向量形式，以句子“台北桃园中正机场昨天晚上发生意外”中的名词短语“台北桃园中正机场”为例，用本文选择的八个特征可描述为向量(‘无’，‘单数’，‘台北桃园中正机场’，‘中正机场’，‘NN’，‘2’，‘无’，‘FAC’)，将上述特征值进行数值化后即可作为基于 ART 网络的聚类共指消解算法的输入模式。

2.3 基于 ART 网络的中文共指消解

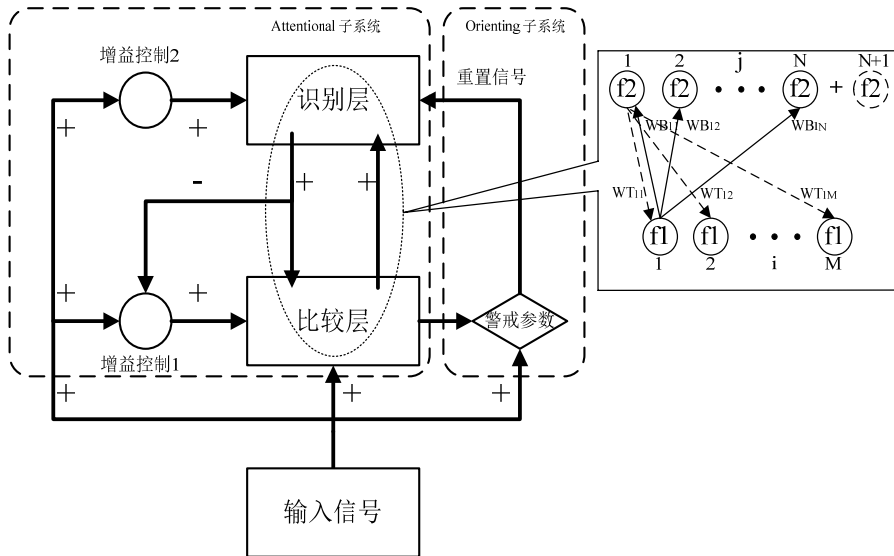


图 2-1 ART1 模型的结构框架

Fig.2-1 The architecture of ART1 model

本文提出了一种新颖的基于 ART 网络模型的共指消解聚类方法，针对在聚类指代消解中面临的输出类别数量难以确定的问题，该方法充分地利用了名词短语自身特征，通过实验来调节网络参数动态控制聚类算法的输出类别数目，同时该方法还能够有效解决特征空间维数较高、样本点波动大等聚类消解相关问题。ART 网络是一种基于自适应谐振理论，具有自组织、自稳定能力的竞争学习式人工神经网络，该理论体系内包括 ART1、ART2 和 ART3 三个基本模型^[95]。ART1 模型主要针对处理离散的输入模式，ART2 模型在其基础上扩展到连续的输入模式，ART3 模型模拟了化学神经键中的可计算属性，由于结构复杂，目前还未见相关应用。

本文中所处理的名词短语特征值均为离散型，因此采用 ART1 模型(为简化描述，后文中出现的 ART 全部是指 ART1)，其结构如图 2-1 所示。ART 网络模型由关注(Attentional)和取向(Orienting)两个子系统构成，关注子系统基本结构是一个由比较层和识别层所组成的双层神经网络，是整个体系的核心部分，用来处理对于模型来说较为“熟悉”的输入模式，主要包含三个部分：

(1) 基于短期记忆(Short-Term Memory, 简称 LTM)的比较层神经元和识别层神经元。其中比较层神经元是计算层，该层神经元节点的输入于三个部分：输入模式本身、识别层的反馈信号和增益控制单元的偏置信号。识别层神经元即对应输入模式的所属类别，由于网络中聚类数量是动态增加的，因此识别层单元的数量也随之逐渐增加；

(2) 基于长期记忆(Long-Term Memory, 简称 LTM)的两层神经元之间的突触连接网络。该网络中的突触连接是双向的，分为自顶向下的连接和自底向上的连接两部分，自底向上的权向量 WB 代表学习到的聚类中心，自顶向下的权向量 WT 代表目前网络熟悉的输入模式。

(3) 增益控制单元，一种保持网络活跃的机制。

当网络处理一个不“熟悉”的输入模式时，取向子系统将发挥作用，它根据一个新的输入模式能否被识别层神经元所识别来判断重置信号的激发。如果输入模式被某个识别层节点识别，曾将其归为识别层节点所对应的类别；如果输入模式无法被任何识别层节点识别，则为该输入模式创建一个新的识别层节点，对应以这个输入模式为中心的新类别。

警戒参数是 ART 网络模型中最重要的参数之一，本质上是控制新类别产生的阈值，该参数取值对聚类效果的影响很大。如果警戒参数设置过低，则即使输入模式和学习到与其最接近的聚类中心相似度很低，该聚类中心对应的识别层节点也会接受该模式。反之，如果警戒参数设置过高，即使输入模式和学习到与其最接近的聚类中心相似度很高，该聚类中心对应的识别层节点也无法接受该模式，这样网络就会频繁产生新的类别中心。总体来说，警戒参数取值的大小与最终的类别数目正相关。本文中警戒参数的取值 θ 通过实验方法调节得到，详见本章 2.4.3 小节。

本文将 ART 网络模型引入到共指消解问题中，将数值化后的名词短语特征向量作为输入信号，用识别层神经元来识别名词短语的共指划分。根据自适应谐振理论，不断对关注子系统中网络连接权重进行更新学习。当一个输入模式 X 首次出现时，比较层的输出就是 X ，经过一段时间的迭代和增益控制的调整，如果比较层神经元的输出越接近某个识别层单元，也就是它所对应的聚类中心，则对这个识别层节点的刺激就越强烈，然后识别层会再将这种刺激反馈

给比较层。通过这种激励反馈过程的反复迭代，相应节点之间的权值被不断强化，最终形成一种“谐振”状态。当输入模式引起网络谐振时，识别层中最活跃神经元所表示的类别即代表该输入模式所属的共指划分。如果输入模式未能引起网络谐振，则说明识别层神经元对于该输入模式不敏感，则为该模式建立新的识别层神经元节点，将其作为一个新的共指划分。下面给出基于 ART 网络的名词短语聚类共指消解的具体算法描述：

(1) 首先初始化参数，包括警戒参数、失效神经元集合、识别层神经元集合，以及网络突触权向量，包括两部分：自顶向下权向量 WT 和自底向上突触权向量 WB 两部分。失效神经元集合表示在样本的一次学习过程中已经判定失效的神经元节点集合，这些神经元将不再对该样本起作用。假设模型比较层含有 M 个节点，识别层含有 N 个节点，那么参数初始化可表示为公式(2-5)。

$$\begin{aligned} V = \theta, \quad Invalid = \emptyset, \quad R_j = 0, \quad j \in [1, N] \\ WT_{ij} = 1/N, \quad WB_{ij} = 1/M, \quad i \in [1, M], \quad j \in [1, N] \end{aligned} \quad (2-5)$$

式中 V ——警戒参数；

$Invalid$ ——失效神经元集合；

WT ——自顶向下突触权向量；

WB ——自底向上突触权向量。

(2) 将名词短语特征向量 $X: (x_1, x_2, \dots, x_M)$ 作为网络输入，经比较层激活函数 $f_1(x)$ 作用后得到向量 $X': (x'_1, x'_2, \dots, x'_M)$ ，见公式(2-6)和公式(2-7)。其中 G_1 为增益控制变量， $P: (p_1, p_2, \dots, p_M)$ 代表识别层的反馈信号，初值为全部零。

$$x'_i = f_1(x_i) = \left[\frac{x_i + G_1 + p_i}{2} \right], \quad i \in [1, M] \quad (2-6)$$

$$G_1 = (x_1 \vee x_2 \vee \dots \vee x_M) \wedge (\overline{R_1 \vee R_2 \vee \dots \vee R_M}) \quad (2-7)$$

式中 x_i ——输入特征向量的第 i 个分量；

G_1 ——增益控制变量；

p_i ——第 i 个识别层的反馈信号。

(3) 然后通过转换后特征向量 C 与自底向上权向量 WB 的内积，计算出识别层节点的净激活，得到识别层中最活跃的有效神经元节点 R_{Active} ，该节点最有可能代表该样本所属类别，具体方法见公式(2-8)和(2-9)。

$$Net_j = Net_j + \sum_{i=1}^M WB_{ij} \cdot x'_i, \quad j \in [1, N] \quad (2-8)$$

$$R_{Active} = \arg \max_{R_j} Net_j, R_j \in \overline{Invalid} \quad (2-9)$$

(4) 再将此最活跃的认识层神经元反馈回比较层，经过自顶向下权向量 WT 作用后得 $X'' : (x_1'', x_2'', \dots, x_M'')$ ，见公式(2-10)。最后通过计算 X 与 X'' 的相似度 $Sim(X, X'')$ 判断网络是否到达谐振状态，其计算方法见公式(2-11)：

$$x_i'' = \sum_{j=1}^M WT_{ij} \cdot x_j', i \in [1, M] \quad (2-10)$$

$$Sim(X, X') = \frac{\sum_{i=1}^M x_i''}{\sum_{i=1}^M x_i} \quad (2-11)$$

(5) 如果 $Sim(X, X'')$ 小于警戒参数 V ，则将该认识层节点的激活值重置为 0，并将该节点加入失效神经元集合，如公式(2-12)所示。然后返回第 4 步，继续在认识层中寻找有效神经元。

$$R_{Active} = 0, Invalid = Invalid \cup \{R_{Active}\} \quad (2-12)$$

如果 $Sim(X, X'')$ 大于或等于警戒参数 V ，则判断网络已达谐振状态，将输入样本 X 的类别判定为 R_{Active} 所表示的类别，然后根据公式(2-8)和公式(2-9)更新与神经元 R_{Active} 相关的自顶向下突触权值 WT 和自底向上突触权值 WB ，更新方法如公式(2-13)。

$$WT_{Active_i} = x_i \wedge WT_{Active_i}, \quad (2-13)$$

$$WB_{Active_i} = \frac{|x_i \wedge WT_{Active_i}|}{\beta + |WT_{Active_i}|}, i \in [1, M] \quad (2-14)$$

(6) 若所有认识层神经元均无法满足 $Sim(X, X'') \geq V$ 的条件，则判断网络未达到谐振状态，为输入样本 X 创建一个新的聚类中心及其对应的认识层神经元节点，并根据公式(2-5)初始化所有与其相连的权值。待篇章中所有名词短语样本聚类完成后，得到 C_1, C_2, \dots, C_k 共 k 个类别，然后再把具有相同名词短语字符串元素的类别进行归并，至类别数目稳定为止，归并后的类别即为该篇章中实体的共指划分，归并策略形式化描述见公式(2-15)。

$$\begin{aligned} & \text{If } \exists e_1 \in C_a, \exists e_2 \in C_b, a, b \in [1, k] \\ & \quad Prototype(e_1) = Prototype(e_2), \\ & \text{Then } Combine(C_a, C_b), k \leftarrow k - 1 \end{aligned} \quad (2-15)$$

下面举例说明该算法的聚类流程：假设 ART 网络模型的参数已经初始化完毕，并且模型已经对数据集中的前 $i-1$ 个数据 x_1, x_2, \dots, x_{i-1} 完成了聚类，当前模型的输入模式为一个名词短语对应的二进制特征向量 $x_i: (x_{i1}, x_{i2}, 6, x_{iM})$ 。首先，计算出 x_i 通过比较层激活函数 $f_1(x)$ 作用后的值 $x'_i: (x'_{i1}, x'_{i2}, 6, x'_{iM})$ ，根据公式(2-6)。然后， x'_i 经过自底向上连接权值 WB 的内积作用后到达每个识别层节点，其中第 j 个识别层节点 R_j 的激活量为 x'_i 与自底向上权向量 WB_j 的内积，参见公式(2-8)。接着，选出激活量最大的识别层节点 R_{Active} ，并将该节点的激活量通过自顶向上权向量 WT 反馈回比较层 $x''_i: (x''_{i1}, x''_{i2}, 6, x''_{iM})$ ，根据公式(2-10)。然后，计算 x_i 和 x''_i 的相似度，如果相似度值小于模型的警戒参数，则将这个识别层节点 R_{Active} 设为失效，在样本 x_i 的后续聚类过程中将不再起作用；如果该值大于或等于模型的警戒参数，则将该输入样本划分到识别层节点 R_{Active} 所对应的名词短语类别，并更新与 R_{Active} 相关的自顶向下权值 WT 和自底向上权值 WB ，根据公式(2-13)和公式(2-14)；若所有识别层神经元均不能使 x_i 和 x''_i 的相似度大于或等于模型的警戒参数，为将输入样本 x_i 划分为新的类别，并为该类别创建新的识别层神经元节点，至此模型对输入模式 x_i 的聚类过程结束。最后，当所有样本 x_1, x_2, \dots, x_M 的聚类过程结束后，把具有相同名词短语的类别根据公式(2-15)进行合并后，就得到了对样本集的共指划分结果。

该方法改进了一般共指消解聚类算法中采用人工设定收敛阈值的方法控制聚类过程结束，而是通过实验来调节警戒参数 V 的值，对聚类类别数量进行控制。与传统的神经网络模型不同，本文采用的 ART 模型是一种高速、高效的不含隐含层的双层神经网络聚类模型，能够根据输入模式快速自动学习，网络参数的更新与类别的划分同步完成。该方法能够较好识别未知样本并为其动态创建新的类别，符合共指消解问题的特性。此外，该算法还具有很好的稳定性，对样本的波动并不敏感，共指消解中样本分布的不规律性对聚类结果造成的影响也比较小。为了验证该方法的有效性，我们利用上述方法在标准的 ACE 关系抽取语料库上进行了相关实验，下面将对实验部分进行详细描述。

2.4 实验及结果分析

2.4.1 实验数据

中文共指消解方面的语料库资源较少，ACE 关系抽取语料是目前该领域内为数不多的含有共指关系的语料库，该语料库由美国语言数据联盟 (Linguistic Data Consortium, 简称 LDC) 发布，主要有英语、汉语和阿拉伯语

三个部分，语料库的规模及一致性能够满足一般实验的需求。因此，本文选用 ACE-2005 汉语部分共 633 篇文档作为数据集，其内容主要是新闻题材，来源于三种媒体：广播新闻(Broadcast News, 简称 BN)、新闻专线(Newswire, 简称 NW)和网络日志(Weblog, 简称 WB)，其详细统计信息见表 2-3。

表 2-3 ACE-2005 汉语语料库相关统计信息

Table 2-3 Statistics of the ACE-2005 Chinese corpus

语料库	提及 mentions	代词 pronouns	实体 entities	文档 documents
广播新闻(BN)	13501	1195	6248	298
新闻专线(NW)	14341	1173	6552	238
网络日志(WB)	6479	978	2614	97
全部	34321	3346	15414	633

为验证共指消解方法的有效性，减少上述错误情况对本实验产生的影响，本文在已标注名词短语的层级上进行相关实验。直接采用 ACE 中 mention 级标注结果作为候选名词短语集合，在其基础上抽取特征向量。

2.4.2 评测指标和方法

共指消解中的评测指标比较特殊，不同于一般自然语言处理领域所采用的精确率、召回率以及 F 值。我们采用共指消解领域中特有的评测指标，即 MUC-6 中定义的基于链接的精确率(P)、召回率(R)和 F 值(F)，其计算方法如公式(2-16)、公式(2-17)和公式(2-18)所示，详见文献^[96]。

$$P = \frac{\sum (|S'_i| - |p(S'_i)|)}{\sum (|S'_i| - 1)} \quad (2-16)$$

$$R = \frac{\sum (|S_i| - |p(S_i)|)}{\sum (|S_i| - 1)} \quad (2-17)$$

$$F = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R}, \quad (\text{本文 } \beta \text{ 取 } 1) \quad (2-18)$$

式中 S —— 共指消解的参考答案；

S' —— 共指消解系统输出的答案；

S_i —— S 中的第 i 个共指集合；

p(S_i) —— 集合 S_i 根据 S' 中集合分布形成的划分；

p(S'_i) —— 集合 S'_i 根据 S 中集合分布形成的划分；

接着本文在第 2.2 节所述的聚类特征集合基础上，选择两种经典聚类方法

作为基准方法(Baseline):

(1) 启发式规则聚类: 即文献[85]中采用的共指消解方法, 先将篇章中的每个名词短语单独归为一类, 按照逆序逐个扫描名词短语, 计算它与前面名词短语之间的距离, 当该距离小于设定阈值时, 将这两个名词短语所属类别合并, 扫描至文中首个名词短语时算法结束。

(2) 基于划分的 K 均值聚类(K-means): 即文献[88]中采用的共指消解方法, 首先将篇章中的所有名词短语通过常规的 k-means 方法进行聚类, 然后进一步拆分名词短语簇直到聚类停止, 具体过程如下: 先计算所有聚类的中心, 然后选出类内距离最大的一类, 将该聚类中心删除, 添加该类中距离最远的两个实例作为新的聚类中心。直到所有聚类的类内距离均小于设定阈值时结束。

2.4.3 实验结果和分析

首先, 我们给出了本章第 2.2 节中所述的基于 SVM 分类器的名词短语语义类型特征识别的结果。本文采用 SVM-light 工具作为分类器^[97], 利用线性核函数用 ACE-2005 中文语料库的 BN 部分对分类器进行训练, 然后从其余语料中随机选取了 5000 个名词短语进行了测试, 这些名词短语的语义类别分布情况见表 2-4, 各语义类型的识别结果见表 2-5。按照各语义类型比例加权后, 取得了 88.5%的整体平均 F 值。在随后的测试过程中, 我们将由该 SVM 分类所识别的名词短语语义类型作为共指消解输入模式中对应的语义类型特征值。

表 2-4 5000 条测试名词短语的语义类别分布信息

Table 2-4 Distribution of semantic type of the 5000 testing noun phrases

	PER	ORG	GPE	LOC	FAC	WEA	VEH
名词短语数	2240	1015	1029	137	196	160	223
所占比例 (%)	44.8	20.3	20.58	2.74	3.92	3.2	4.46

表 2-5 SVM 实体语义类型识别结果

Table 2-5 Results of semantic type recognition using SVM

	PER	ORG	GPE	LOC	FAC	WEA	VEH
P	0.846	0.914	0.956	0.919	0.857	0.981	0.97
R	0.992	0.796	0.82	0.738	0.725	0.849	0.854
F	0.913	0.851	0.882	0.818	0.785	0.91	0.908

接着, 我们在 ACE-2005 语料库 BN、NW 和 WB 三部分上对本文提出的基于 ART 网络的聚类共指消解方法, 并与两种 Baseline 方法进行了比较, 结果如表 2-6 所示。基准方法中基于启发式规则和 K 均值聚类方法分别取得 55.2%和 63.3%的平均 F 值。经分析发现, 启发式规则方法适合于表层特征较

为明显的名词短语，如‘单复数特征’和‘短语本身’，在获得较高的精确率同时也会形成许多零散类别，导致召回率偏低；K 均值聚类方法采用了有效的名词短语距离度量和聚类中心分裂方法，因此性能上有所提升，但在最终聚类算法终止条件判断上影响了整体性能。

表 2-6 ACE 语料上的聚类共指消解结果

Table 2-6 Results of cluster-based coreference resolution on ACE corpus

Method		Precision	Recall	F-score
BN	Heuristic	0.823	0.415	0.552
	K-means	0.735	0.546	0.627
	A R T	0.815	0.621	0.705
NW	Heuristic	0.791	0.430	0.558
	K-means	0.748	0.549	0.633
	A R T	0.793	0.617	0.694
WB	Heuristic	0.812	0.406	0.541
	K-means	0.739	0.573	0.645
	A R T	0.810	0.636	0.712

本文中的基于 ART 网络的聚类共指消解算法有效地克服了这些问题，从整体上看，根据表 2-6 中可以算出在整个语料库和完全相同特征空间条件下获得了 70.2% 的整体平均 F 值，结果明显地优于两种基准方法；从 BN、NW 和 WB 三个部分看，本文方法取得的 F 值均优于两种基准方法，验证了本文基于 ART 网络的聚类共指消解方法的有效性。

表 2-7 名词性短语的共指消解实验结果

Table 2-7 Results of coreference resolution after ignoring pronoun phrases

Dataset		Precision	Recall	F-score
原结果	BN	0.815	0.621	0.705
	NW	0.793	0.617	0.694
	WB	0.810	0.636	0.712
无代词	BN	0.839	0.620	0.713
	NW	0.832	0.635	0.720
	WB	0.847	0.671	0.749

经过对聚类错误样本的分析，我们发现本文提出的方法能够较好地识别对名词性短语之间的共指关系，对于代词性短语与一般名词短语的聚类效果稍差，分析其原因主要在于指代性短语自身的信息量较少。比如在句子“普京没

有说谁是贝加尔金融集团的老板，但他表示该集团的背后是一群拥有石油行业经验的个人”中，名词短语‘他’与‘普京’之间存在共指关系，但在对其聚类时短语级特征所起到的作用很小，在实验中‘他’被分成一个单独类别，而类似的指代性短语比重占名词短语总数的 9.75%。表 2-7 给出了名词短语的聚类共指消解实验结果。

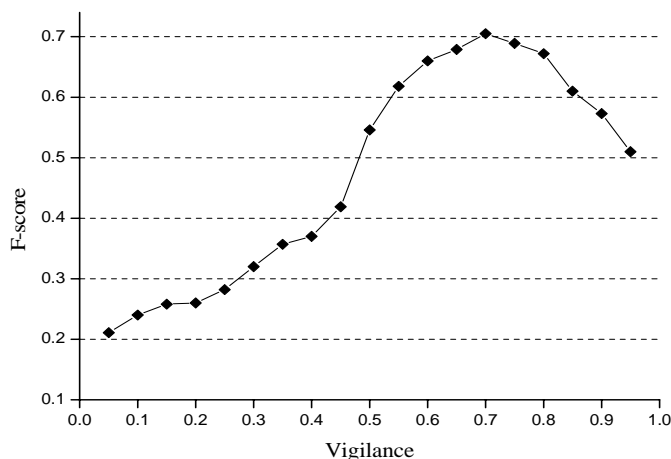


图 2-2 共指消解 F 值随警戒参数变化示意图

Fig.2-2 F-score of coreference resolution versus vigilance parameter

可见，在处理一般名词短语时效果提升明显，因此对于这部分代词性短语我们拟采用一些针对代词的后处理规则单独对其进行处理，以提高其识别精度。另外，警戒参数作为网络的重要参数，对于聚类效果影响很大。因此，我们调查了警戒参数随 F 值的变化曲线，采用上述 ART 网络的聚类方法，以 0.05 为步长调整警戒参数，观察该参数改变对于整个系统 F 值的影响，从图 2-2 中可见当警戒参数取值为 0.70 时达到最优平均 F 值。因此我们将该值作为 ART 网络警戒参数的初始值进行聚类。

表 2-8 移除各个特征对共指消解系统整体性能的影响

Table 2-8 The effect of removing each feature at a time to the coreference resolution system

移除特征	Precision	Recall	F-score
性别	0.779	0.615	0.687
单复数	0.783	0.607	0.684
短语本身	0.742	0.593	0.659
中心词串	0.736	0.602	0.662
中心词词性	0.807	0.610	0.695
短语句序号	0.794	0.611	0.691
语义类别	0.787	0.582	0.669
生物性	0.785	0.596	0.678
无(全部特征)	0.804	0.623	0.702

另外共指消解的评测指标：基于链接的准确率和召回率本身也存在一种偏

差，它对于独自组成一个实体类别的名词短语分数较低，因此这会造成一种倾向于产生类别较少系统的不公正现象^[98]，上述因素都对本方法的 F 值产生一定影响。最后，为了验证本文所采用聚类特征的有效性以及每个特征对聚类共指消解算法的贡献程度，我们采用移除的策略，通过观察缺少某特征时系统 F 值的下降幅度来评价该特征对于聚类共指消解的贡献，实验结果见表 2-8。

表 2-9 增加聚类特征对共指消解系统整体性能的影响

Table 2-9 The effect of adding clustering feature to the coreference resolution system

增加特征	Precision	Recall	F-score
专有名词	0.801	0.628	0.704
中心词词频	0.795	0.627	0.701
上短语序号	0.787	0.623	0.695

从表 2-8 中可见，单个特征移除后 F 值下降最明显的，也就是对共指消解问题贡献度最大的三个特征分别是名词短语本身、中心词串及其语义类别，与本章第 2.2 节中所述的信息增益率方法特征区分度所得结果相符，说明了本文中基于信息增益率的聚类特征选择方法的有效性。接下来我们又评价了第 2.2 节中被移除的信息增益率最低的三个特征，即词汇级特征专有名词、短语级特征中心词词频和上下文级特征短语序号对于聚类系统的影响。在原系统的基础上分别加入这三个特征后，系统 F 值的变化如表 2-9 所示。从表 2-9 中可见，增加被信息增益率方法移除的聚类特征后系统的 F 值影响很小，进一步验证了本文基于信息增益的聚类特征选择方法的有效性。

表 2-10 与其它聚类共指消解系统的对比结果

Table 2-10 The comparison results with other cluster-based coreference resolution systems

	Precision	Recall	F-score
H.K.PolyU.(2006)	0.771	0.629	0.693
NanJing U.(2007)	0.547	0.716	0.621
ART	0.781	0.607	0.683

最后，本文与另外两个效果较好的基于聚类的中文共指消解方法研究工作进行了比较，分别是香港理工大学和南京大学的相关工作^[88,89]。前者采用的语料库规模较小，是手工标注的 30 篇 TDT3 语料，仅含 410 个实体和 1640 次提及；后者的语料库与本文相同。本文与上述系统比较的结果见表 2-10，可见，在相同的评价指标下，本文方法的结果在中间位置，取得的 F 值较前者略差，约低 1 个百分点，但比后者高出后者近 6 个百分点。较前者低的主要原因是由于评价语料上有所不同。综上所述，本文提出的基于 ART 网络的聚类共指消解方法是有效的，其整体性能达到了领域内相关研究中较高的水平。

2.5 本章小结

共指消解是面向文景转换的浅层语义分析中必要的预处理步骤，它本身是一个语义层级的复杂问题，另外缺乏领域内的相关语料等因素又给该问题增加了难度。出于上述考虑，本文提出基于 ART 网络的聚类共指消解方法，有针对性地解决了聚类共指消解中输出类别数目无法预知这一难题。从现有文献上来看，本文是将基于神经网络的聚类方法应用于中文共指消解问题中，并取得了较好的实验效果。该方法充分利用了名词短语自身特征，能够通过实验来调节网络参数动态控制聚类的输出类别数目。在聚类特征选择方面，为了减少了区分度较弱特征给聚类所带来的干扰，本文采用了一种基于信息增益率的特征选择方法，从语法、句法、上下文和语义四个层面选取了适于聚类算法的八个特征。其中采用了基于 SVM 分类器的语义类别特征抽取模块，能够高效地对语义类这一重要特征进行识别。最后，我们在 ACE-2005 中文语料库上进行了一系列实验，实验结果验证了本文提出的基于 ART 网络的聚类共指消解较一般聚类共指消解方法性能上有明显的提高。

第3章 浅层语义分析中的特征选择方法研究

3.1 引言

浅层语义分析是近年来自然语言处理方面的研究热点，该技术广泛应用于自然语言领域的各项任务中，提供结构化的浅层语义信息。目前浅层语义分析采用的形式主要是语义角色标注(Semantic Role Labeling, 简称 SRL)，中文语义角色标注是指从汉语句子里识别出与目标谓词相关的语义角色(或称论元)并判断其对应的功能类型，根据目标谓词和语义角色之间的约束关系，可以按照功能把语义角色分为若干类型，如施事、受事、与事、时间、地点、工具等等。本文研究的中文浅层语义分析主要面向文景转换任务，其研究内容、目标和方法与通用的浅层语义分析研究基本一致，不同之处在于本文对与动画生成相关度较高的语义角色类型更加关注，如施事、受事、地点、时间；相反对那些与动画生成相关度较小的语义角色则关注度较低，如条件、目的、原因、方式、程度、范围等。

文献[11]和[12]中已证实了句法分析对于目前的语义角色标注是必要的。根据语义角色标注中采用的句法分析形式，语义角色标注的基本标注单元可分为词、短语或句法成分三种。词标注单元主要用于基于依存句法分析语义角色标注系统；短语主要用于基于 Chunk 的语义角色标注系统；句法成分主要用于基于短语结构句法分析的语义角色标注系统。而从整体效果上看，以词或句法成分为标注单元的语义角色标注的效果较好^[99]；从语义完整性上看，语义角色应该是在语义上较为完整的实体或描述，句法成分在粒度上与其最为相近。因此，本文选择句法成分，也就是短语结构句法分析树中的节点，作为语义角色基本标注单元。

基于特征的有指导机器学习方法是目前语义角色标注中的最主流方法这种语义角色标注方法的可分为两个阶段：一是语义角色识别，目标是从句子中抽取所有充当语义角色的句法成分；二是语义角色分类，也就是判断语义角色识别阶段所得的语义角色的类型。另外，在进行语义角色识别之前，通常还要进行一个重要的剪枝步骤，用简单的方法过滤句法分析树中很多不可能成为语义角色的句法成分，以缩小候选范围，提高准确率^[21]。另外，语义角色标注通常要先将句法成分与其对应的谓词相结合，组成“谓词-论元”二元组，然后再将二元组转换成为特征向量作为学习和预测的样本，以获得句法成分中与谓词相关的一些语义信息。

基于特征和机器学习的浅层语义分析方法中有三个关键因素：一是特征选

择，二是机器学习方法，三是训练语料库。本章首先从特征选择角度对浅层语义分析方法进行了探索，先发掘出足够的对于浅层语义分析问题区分能力较强的句法特征，然后再将其用于机器学习方法中。本章提出一种基于多重句法特征的中文浅层语义分析方法，该方法首次融合了短语结构和依存结构两种类型的句法特征，在此基础上我们又构造了多个二元组合特征，通过一种基于统计的组合特征选择方法筛选出其中有效的组合句法特征。最后，我们利用短语结构特征、依存结构特征和组合句法特征等多重句法特征进行后续的分类过程。

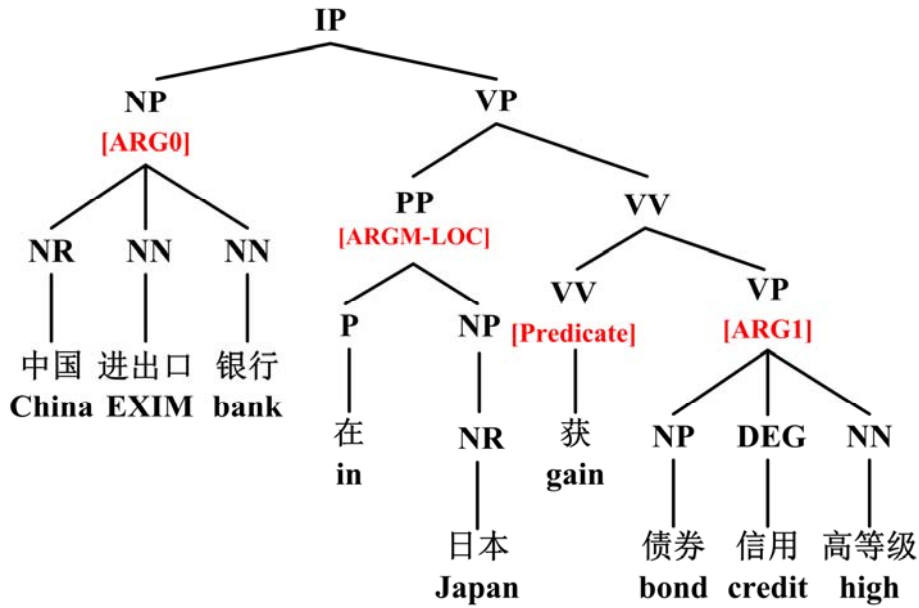


图 3-1 基于短语结构句法分析的语义角色标注

Fig.3-1 SRL based on phrase structure syntactic parsing

在标注形式方面，本文采用了 PropBank 语料库的语义角色标记体系^[15]，具体形式如图 3-1 所示，带有方括号标注的节点即为语义角色句法成分，方括号中的内容是其语义角色类型。相比 FrameNet^[13]和 VerbNet 等^[14]标记形式，这种形式大幅地减少了语义角色类型的数量，能够有效地减轻数据稀疏现象，更有利于统计机器学习方法。该体系中把语义角色分成两类：一类是核心语义角色，标记为 Arg0-Arg5，Arg0 通常表示动作的施事，Arg1 通常表示动作的受事，Arg2-Arg5 根据不同的谓词具有不同的语义含义；另一类是修饰性语义角色，包括副词、时间、地点、方式、条件、目的等 13 个子类型，标记为 ARGM，另外再附加上其子类型标记。例如，图 3-1 中‘PP’节点对应的句法成分‘在日本’是一个地点类型的修饰性语义角色，因此应标记为类型 ARGM-LOC。

从基于特征的语义角色标注研究现状来看，还未见对于多种句法形式以及在其基础上进行特征组合的相关研究报道，目前只有少数几个关于采用多种类

型句法特征的浅层语义分析的研究^[31,70,100]，但其中都未涉及对组合句法特征的研究。目前还有很多能够提高系统性能的组合句法特征尚未发掘，缺乏快速高效的组合句法特征选择方法是该问题的主要障碍。因此本文在采用多种句法形式的基础上，又提出了一种基于统计的组合句法特征构造方法，该方法能够以基本特征集合和语料库为基础，构造并筛选出有助于语义角色识别和分类的组合句法特征，从特征选择的角度提高语义角色标注整体性能。

本章首先总结了现有的语义角色标注系统中证实有效的短语结构特征和依存特征作为基本特征集合，以此为基础构造组合句法特征。然后提出一种基于语料库的统计方法，用来估计每个组合句法特征划分类别的能力，筛选出效果较好的组合句法特征用于语义角色识别和分类步骤。最后，我们采用支持向量机(SVM)分类模型在 CPB 标准语料库上对上述方法进行了语义角色识别和分类实验，筛选出对于语义角色识别和分类过程有效的组合句法特征，并在由基本特征和组合特征构成的特征集合上用 SVM 分类器进行了相关实验，验证了本文提出方法的有效性，本章 3.4 小节中对这部分内容进行了详细描述。

3.2 基本特征集合的构建

3.2.1 分类模型的选择和句法树的剪枝

本文选择支持向量机(Support Vector Machine, 简称 SVM)模型作为语义角色标注的分类模型。SVM 模型是目前基于有指导机器学习的语义角色标注领域中性能最好的判别模型之一^[101]，而且在自然语言处理其他任务上也有成功的应用。本文 4.3.5 节中对 SVM 模型进行了详细的介绍，该模型具有适于高维特征空间、适于小规模样本、适于非线性问题、推广能力强等优点。语义角色标注包含语义角色识别和语义角色分类两个子任务，因此在用 SVM 分类方法进行语义角色标注时，需要针对这两个子任务分别进行学习。语义角色识别是一个二值分类任务，语义角色分类是一个多值分类任务。SVM 是一种典型的二值分类方法，因此在语义角色识别阶段，可直接采用 SVM 模型进行学习和预测；而在语义角色分类阶段，我们采用“一对多”的策略处理该多值分类问题，训练与待分语义角色类别数相同的 SVM 分类器，分别针对各语义角色类别进行判断，该策略的详细描述见本文 2.2 节。接下来我们先对剪枝方法进行了后处理，然后分别构建了基本特征集合和组合特征集合，最后在由基本特征和组合特征组成的特征集合上用 SVM 模型进行学习和预测。

基于短语结构句法分析的语义角色标注首先需要一个的剪枝预处理过程以过滤掉短语结构分析树中一些不可能成为语义角色的句法成分，保留尽量少的候选句法成分以提高准确性。目前最常用的剪枝方法是 Xue 在文献[21]中提出

的基于启发式规则的方法，该方法可描述为以下三个步骤：

- (1) 将目标谓词在句法分析树中所处的节点设置为当前节点。
- (2) 抽取当前节点的所有兄弟节点放入语义角色候选集合。如果其中某个兄弟节点为介词短语，则将该节点的子节点也放入语义角色候选集合。
- (3) 然后将当前节点的父节点设置成当前节点，重复上述抽取过程，直至达到子句根节点为止。

在实际应用中我们发现该剪枝方法抽取出的候选语义角色中，仍包含较多冗余句法成分。本文在上述剪枝方法的基础上，通过在语料库上的统计分析，引入一种后处理方法能够进一步减轻冗余，缩小候选语义角色的范围。我们在 CPB-1.0 语料上统计了语料库中语义角色与其对应的短语类型的共现情况，通过对训练语料中正反例的对比发现正确语义角色对应的短语类型在全部语义角色对应短语类型中所占比率为 45.3% (24/53)，且正例中 24 种正确短语类型对应的语义角色数量在全部语义角色中所占比率为 84.16%；相应地，正确语义角色的父节点短语类型在全部语义角色父节点短语类型中所占比率为 63.6% (14/22)，且正例中 14 种正确父节点短语类型对应的语义角色数量在全部语义角色中所占比率为 93.65%。因此，根据上述信息，本文提取了正例中出现的 24 个短语类型和 14 个父节点短语类型作为判断条件，详见表 3-1，将从上述基于启发式的剪枝方法获得的句法成分中短语类型及其父节点短语类型不在类型集合中的句法成分过滤掉，从而进一步减少候选语义角色的数量，提高后续分类过程的效率。

表 3-1 正例中出现的短语类型和父节点短语类型

Table 3-1 Phrase types of constituents and their parent nodes in the positive instances

短语类型	父节点短语类型
NP,ADVP,PP,IP,QP, LCP,VP,DP,DVP,CP, PRN,UCP,VV,LST,NN, DNP,CLP,AD,VA,NR, NT,CD,PN,VCD,	VP,IP,NP,PP,CP, QP,VS,VRD,LCP,DVP, UCP,DNP,ADVP,VPT

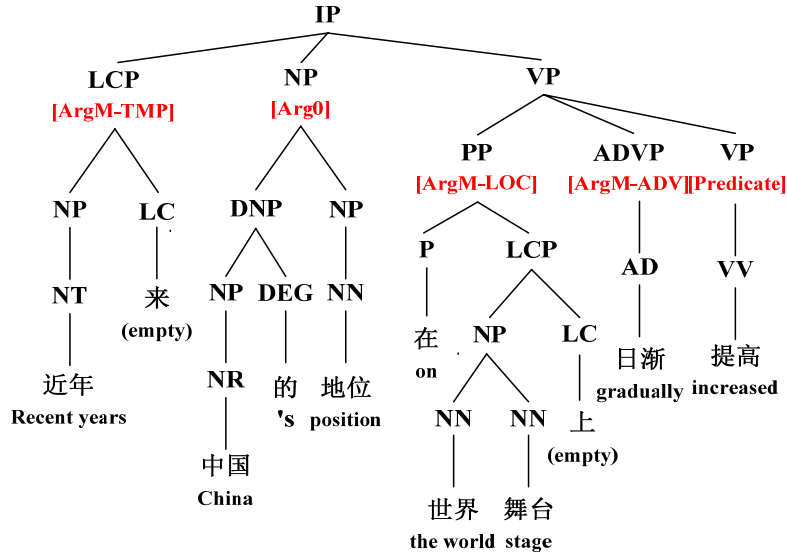
实质上本文句法树剪枝后处理方法是基于这样一个前提，即语料规模能够涵盖所有可能作为语义角色的短语类型。在实际中，本文对采用的 CPB-1.0 语料库进行了测试，其中共有近十万个语义角色正例，而句法成分的短语类型仅有 50 多种(包括词性标注类型)，在这种情况下任何一种合理的语义角色短语类型在近十万个正例中一次都不出现的概率极小。因此该前提在本文所采用的

语料库上是合理的，实验结果也证实了该方法的可行性。

3.2.2 短语结构句法特征集合的构建

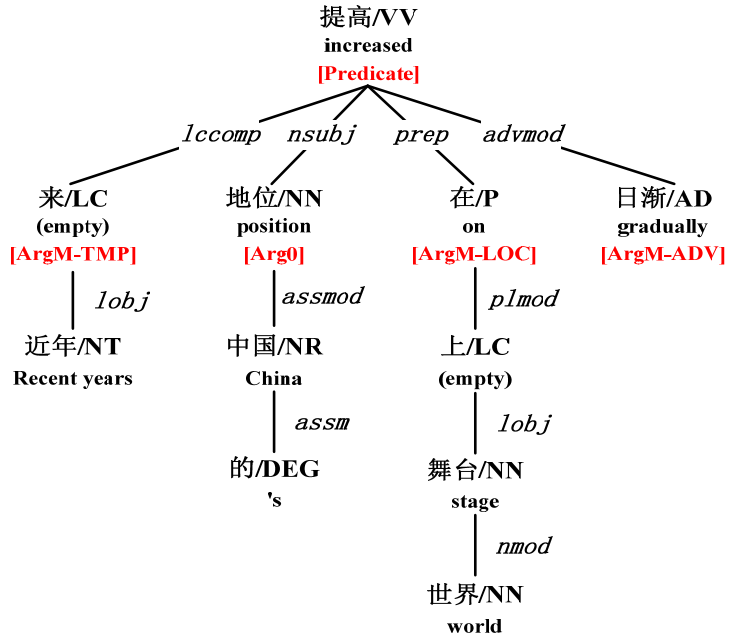
与一般浅层语义分析系统所采用的特征集合不同，本文中采用的特征集合可以分成两个子特征集合：基本特征集合和由基本特征所组成的组合特征集合。另外，本文的基本特征集合还包含了两个层次的句法特征：一是短语结构句法特征，二是依存结构句法特征。首先，我们调研了现有文献中出现的有效短语结构特征和依存结构句法特征，在此基础上又引入一些新定义的句法特征构成基本特征集合。然后，为了深度发掘这两种句法形式中的特征，我们将基本特征集合中的不同特征进行两两组合，构造出大量候选组合句法特征。然后采用一种基于统计的组合特征选择方法，根据在语料库上的统计信息，从候选组合句法特征中选择出最有效的一部分构造组合特征集合。最后，将基本特征集合和组合特征集合的并集作为分类特征集合，用这些分类特征将样本表示成特征向量，然后再用机器学习方法进行训练和预测。

首先介绍短语结构句法分析与依存句法分析的基本形式和特点。短语结构句法树主要表达的是句法成分之间的层次关系，由终结符、非终结符以及短语标记构成。自底向上短语结构句法分析过程就是利用产生式规则把句子的终结符归约为根节点的过程，如图 3-2 a)所示；依存句法树主要表达的是单个词之间的依存关系，由词、依存弧和依存关系构成。依存分析过程中就是识别句子中词语之间的支配或从属关系，支配与被支配成分之间由一条有向的依存弧连接，依存弧上标记着依存关系，如图 3-2 b)所示。依存分析可以由短语结构分析结果通过中心规则转换生成，转换过程中会丢掉跨度和跨度上的句法标识信息。短语结构句法树表达的是句法成分之间的层次关系和句法功能，但其形式较为复杂，易导致特征取值的稀疏现象；依存句法树表达的是词与词之间的依存关系，形式简洁，更加贴近语义层，而对短语之间的层次关系句法功能缺乏描述。两种句法形式的特征能够在浅层语义分析中互相补充，提供更加充分的句法信息。例如，图 3-2 a)中短语结构分析树中的 NP 型句法成分‘中国的地位’被标记为‘Arg0’，该句法成分到目标谓词‘提高’之间的路径特征为‘NP-IP-VP-VP’，比较复杂；而在图 3-2 b)中依存结构分析树上的词节点‘地位’到目标谓词‘提高’之间仅通过一条类型为‘nsubj’的依存弧连接，在表达上更加的显式和直接，对于谓词及其对应句法成分之间关系的识别要更加容易。除此之外，其它与句法路径相关的特征也都会得到简化。下面将详细介绍本文所采用的短语结构句法特征和依存结构句法特征。



a) 基于短语结构句法分析的语义角色标注

a) SRL based on phrase structure syntactic parsing



b) 基于依存句法分析的语义角色标注

b) SRL based on dependency structure syntactic parsing

图 3-2 基于依存句法分析的语义角色标注

Fig.3-2 The architecture of ART1 model

本文调研了现有文献，从中选出了 26 个效果较好的短语结构句法特征 [21,25,28,102]。然后我们定义了 8 个新的短语结构句法特征，为了补充一些新的句法信息和减少原特征值的稀疏性，这两部分特征联合构成了短语结构句法特征集合。表 3-2 和表 3-3 分别描述了 26 个选择的和 8 个自定义的短语结构句法特征，包括它们的名称、标号、含义描述以及实例。表格中的说明实例均以图 3-2 a) 中标注为 ‘Predicate’ 的节点和 ‘Arg0’ 的节点作为谓词和论元。

表 3-2 选择的短语结构特征及其描述

Table 3-2 The selected phrase structure features and their descriptions

标号	特征	描述
C1	谓词	谓词的原形, e.g. ‘提高’
C2	句法路径	句法成分到谓词之间的单向句法路径, e.g. ‘NP ↑ IP ↓ VP ↓ VP’
C3	短语类型	句法成分短语根节点的类型, e.g. ‘NP’
C4	位置	句法成分位于谓词的前或后(二值), e.g. ‘前’
C5	语态	谓词是主动还是被动形式(二值), e.g. ‘主动’
C6	短语中心词	句法成分短语的中心词, e.g. ‘地位’
C7	次范畴框架	谓词所在动词短语的展开形式, e.g. ‘VP-> PP+ADVP+VP’
C8	句法框架	句法成分周围 NP 的分布形式, e.g. ‘Cur_NP_v’, Cur 是当前句法成分, v 表示当前谓词
C9	短语中心词词性	句法成分的中心词的词性, e.g. ‘NN’
C10	部分句法路径	句法成分与谓词最低公共祖先(LCA)路径上, LCA 子节点之间的路径, e.g. ‘NP ↑ IP ↓ VP’
C11	短语首词	短语的第一个词, e.g. ‘中国’
C12	短语末词	短语的最后一个词, e.g. ‘地位’
C13	短语首词词性	短语第一个词的词性, e.g. ‘NR’
C14	短语末词词性	短语最后一个词的词性, e.g. ‘NN’
C15	左兄弟类型	句法成分左兄弟短语的类型, e.g. ‘LCP’
C16	右兄弟类型	句法成分右兄弟短语的类型, e.g. ‘VP’
C17	左兄弟中心词	句法成分左兄弟短语的中心词, e.g. ‘近年’
C18	右兄弟中心词	句法成分右兄弟短语的中心词, e.g. ‘提高’
C19	左兄弟中心词性	句法成分左兄弟短语中心词的词性, e.g. ‘NT’
C20	右兄弟中心词性	句法成分右兄弟短语中心词的词性, e.g. ‘VV’
C21	部分路径层数	句法成分和 LCA 节点之间的层数之差, e.g. ‘1’
C22	时间指示词	短语中是否有时间指示词(二值), e.g. ‘否’
C23	谓词词性	谓词的词性, e.g. ‘VV’
C24	父节点类型	句法成分父节点短语的类型, e.g. ‘IP’
C25	父节点中心词	句法成分父节点短语的中心词, e.g. ‘提高’
C26	父节点中心词性	句法成分父节点短语中心词的词性, e.g. ‘VV’

在表 3-3 中所述新特征中, 特征‘地点指示词’(C27)与特征‘时间指示词’(C22)类似, 该特征主要为识别 ArgM-LOC 型语义角色设计, 判断一个句法成分中是否含有地点指示词; 特征‘短语词性框架’(C28)和‘短语子类型框架’(C29), 参照谓词相关特征‘次范畴框架’(C7)提出, 而这两个特征主要针对句法成分, 主要目的是为了补充句法成分的内部模式信息。其余 5 个新定义的特征都是为了减少特征值稀疏的问题, 特别是句法路径相关的特征, 其中: 特征‘LCA 左分支类型’(C30)和‘LCA 右分支类型’(C31)是针对特征‘部分句法路径’(C10); 特征‘句法路径词袋’(C32)、‘短语的词性词袋’(C33)和‘短语子类型词袋’(C34)分别针对特征‘句法路径’(C2)、‘短语词性框架’(C28)和‘短语子类型框架’(C29)。

表 3-3 定义的短语结构句法特征及其描述

Table 3-3 The defined phrase structure features and their descriptions

标号	特征	描述
C27	地点指示词	句法成分中是否有地点指示词(二值), e.g. ‘否’
C28	短语词性框架	句法成分中所有词的词性序列, e.g. ‘NR-DEG-NN’
C29	短语子类型框架	句法成分的子节点类型的序列, e.g. ‘DNP-NP’
C30	LCA 左分支类型	LCA 节点类型, 以及其左侧分支子节点的类型, e.g. ‘IP-NP’
C31	LCA 右分支类型	LCA 节点类型, 以及其右侧分支子节点的类型, e.g. ‘IP-VP’
C32	句法路径词袋	特征句法路径(C2)的词袋形式, e.g. ‘{IP, NP, VP}’
C33	短语的词性词袋	特征短语词性序列(C28)的词袋形式, e.g. ‘{DEG, NN, NR}’
C34	短语子类型词袋	特征短语子类型序列(C29)的词袋形式, e.g. ‘{DNP, NP}’

3.2.3 依存结构句法特征集合的构建

依存结构句法分析能够有效地表达词与词之间的依存关系, 但是依存分析中不包含句法成分的信息。因此如果要使用依存特征来标记句法成分的语义角色, 需要先将每个句法成分映射到依存分析树中的一个或多个节点上。本文将每个句法成分映射到其语义中心词节点上, 用该语义中心词的依存结构句法特征作为句法成分的依存结构句法特征。接着采用类似的方法, 本文调研了现有文献, 从中选出了 35 个效果较好的依存结构句法特征, 作为依存结构句法特征集合^[30,31,70,73,80]。表 3-4 中详述了这 35 个依存结构句法特征的名称、标号、含义描述以及实例。表格中的说明实例均以图 3-2 b)中标注为‘Predicate’的节点和‘Arg0’的节点作为谓词和论元。

表 3-4 选择的依存结构特征及其描述

Table 3-4 The selected dependency structure features and their descriptions

标号	特征	描述
D1	谓词关系类型	谓词与其父节点直接的依存关系, e.g. ‘root’
D2	论元关系类型	论元与其父节点直接的依存关系, e.g. ‘nsubj’
D3	关系路径	论元到谓词之间的依存关系路径, e.g. ‘nsubj’
D4	谓词子节点词性框架	谓词所有子节点的词性序列, e.g. ‘LC-NN-P-AD’
D5	谓词子节点关系框架	谓词所有子节点的关系序列, e.g. ‘NR’
D6	谓词子节点关系集合	特征谓词子节点词性框架(D4)的词袋形式, e.g. ‘{ AD, LC, NN, P}’
D7	谓词子节点词性集合	特征谓词子节点词性框架(D5)的词袋形式, e.g. ‘{NR}’
D8	谓词父节点词	谓词父节点对应的词, e.g. ‘Root’
D9	论元父节点词	论元父节点对应的词, e.g. ‘提高’
D10	谓词父节点词性	谓词父节点词的词性, e.g. ‘Root’
D11	论元父节点词性	论元父节点词的词性, e.g. ‘VV’
D12	论元左侧词	论元左侧的词, e.g. ‘的’
D13	论元右侧词	论元右侧的词, e.g. ‘在’
D14	论元左侧词词性	论元左侧词的词性, e.g. ‘DEG’
D15	论元右侧词词性	论元右侧词的词性, e.g. ‘P’
D16	论元左侧词关系	论元左侧词与其父节点的关系, e.g. ‘assm’
D17	论元右侧词关系	论元右侧词与其父节点的关系, e.g. ‘prep’
D18	论元左兄弟词	论元左兄弟节点对应的词, e.g. ‘来’
D19	论元右兄弟词	论元右兄弟节点对应的词, e.g. ‘在’
D20	论元左兄弟词性	论元左兄弟节点词的词性, e.g. ‘LC’
D21	论元右兄弟词性	论元右兄弟节点词的词性, e.g. ‘P’
D22	论元左兄弟关系	论元左兄弟节点与其父节点的关系, e.g. ‘lccomp’
D23	论元右兄弟关系	论元右兄弟节点与其父节点的关系, e.g. ‘prep’
D24	存在依存关系	谓词和论元之间是否存在依存关系, e.g. ‘Yes’
D25	存在依存类型	谓词和论元之间存在依存关系的类型, e.g. ‘nsubj’
D26	词性路径	论元到谓词之间的词性路径, e.g. ‘NN-VV’
D27	词性路径长度	论元到谓词之间的词性路径的长度, e.g. ‘2’
D28	关系路径长度	论元到谓词之间的关系路径的长度, e.g. ‘1’
D29	最高支持动词	从 LCA 到论元的路径上的首个动词, e.g. ‘提高’
D30	最低支持动词	从 LCA 到论元的路径上的最末动词, e.g. ‘提高’
D31	最高支持名词	从 LCA 到谓词的路径上的首个名词, e.g. ‘None’
D32	最低支持名词	从 LCA 到谓词的路径上的末个名词, e.g. ‘None’
D33	LCA 节点词	谓词和论元的 LCA 节点对应的词, e.g. ‘提高’
D34	LCA 节点词性	谓词和论元的 LCA 节点词的词性, e.g. ‘VV’
D35	LCA 节点关系	谓词和论元的 LCA 与其父节点的关系, e.g. ‘Root’

为了保持短语结构句法分析和依存结构句法分析之间的一致性, 本文采用一种基于中心规则的方法将短语结构句法树中直接转换成依存结构句法树

^[103]，这种中心规则能够获得句法成分中的中心。我们将表 3-2 和表 3-3 中所述的短语结构句法特征和表 3-4 中所述的依存结构句法特征联合起来构成本文的基本特征集合，然后在此基础上构造和选择信息更为丰富的组合句法特征。

3.3 基于统计的组合特征集合构建

文献[21]和[28]已证实了加入由基本特征组成的组合特征能够有效提高语义角色标注的性能，目前大多数语义角色标注系统都会采用若干组合特征。但是由于特征组合的方式较多，采用了不当的组合特征非但不会提高系统的性能，反而会大大增加特征空间的维数，提高运算的复杂性。目前对于组合特征选择方法研究还不充分，大多数都是通过观察到某种特定现象后人工定义得到。本文提出一种基于统计的方法，定义了一种新的统计量对组合特征进行筛选，根据各组合特征在语料中相应的分布情况，快速高效地发掘出对于分类帮助较大的有效组合特征，构造出组合特征集合。

本文中训练集合是从 CPB 语料库中获得的，正例可以直接根据人工标记从语料中抽出，反例的构造方法在语义角色识别和语义角色分类过程中有所不同，分为两种情况：在语义角色识别阶段，随机地从候选句法成分中选择不是语义角色的成分作为反例，使正反例数量较为均衡，避免由于反例数量过多造成分类器的偏差^[47]；在语义角色分类阶段，对于每个待识别的语义角色类型都要构造一组正反例训练语料，对于某一语义角色类型，其正例同样可直接从标注预料中获得，反例则是由除该类型之外其余语义角色类型的正例所组成，因此反例的数量通常要比正例的数量多。采用这中反例构造方法的目的是为了训练反例更接近预测时反例出现的真实情况，其根本原因是在语义角色识别和语义角色分类两个阶段候选实例生成机制的不同。

首先将基本特征集合中的全部 N 个特征放入特征集合，每次针对两个不同基本特征 f_a 和 f_b ($a \neq b, a, b \in [0, 1, \dots, N]$)，将这两个基本特征构造一个新的组合特征 f_{ab} ，并将其放入特征集合中作为第 $N+1$ 个元素，即 f_{N+1} 。假设训练集合 D 由正例和反例两部分组成 $D = \{D_{pos}, D_{neg}\}$ 。在特征向量化的过程中，分别将正例和反例中的特征进行数值化，这样能够保证在正反例样本中特征的维数之间具有统计性差异。在特征向量化之后的正反例训练数据集可表示为 $D_{pos} = \{(x, y) | x \in D, y = 1\}$ ， $D_{neg} = \{(x, y) | x \in D, y = -1\}$ ，其中 x 代表待分类句法成分所对应的特征向量，采用向量的稀疏表示方法，即仅使用值为 1 的特征维数来表示特征向量，值为 0 的维数则不进行表示， y 代表该候选句法成分所属的语义角色类型。在训练数据集 D 中，首先我们计算特征集合中每个特征 f_i 在正例和反例中的样本均值 $Mean_{f_i}(D)$ 和样本标注差 $S_{f_i}(D)$ ，如公式(3-1)和公式

(3-2)所示。

$$Mean_{f_i}(D) = \frac{1}{|D|} \sum_{x \in D} x(i), \quad i \in [1, N+1] \quad (3-1)$$

$$S_{f_i}(D) = \sqrt{\frac{1}{N} \sum_{x \in D} (Mean_{f_i}(D) - x(i))^2}, \quad i \in [1, N+1] \quad (3-2)$$

式中 D ——数据集；

$x(i)$ —— x 中特征 f_i 的值在特征空间中的维数；

f_i ——特征集合中的第 i 个特征；

N ——特征集合中的特征数量。

然后将特征 f_i 在正反例集合中样本均值之差的平方，作为该特征正反例之间的类间距离 $InterDist_{f_i}(D_{pos}, D_{neg})$ ，如公式(3-3)所示；将该特征在正反例样本中的样本方差之和，作为该特征正反例之间的类内距离 $IntraDist_{f_i}(D_{pos}, D_{neg})$ ，如公式(3-4)所示。

$$InterDist_{f_i}(D_{pos}, D_{neg}) = (Mean_{f_i}(D_{pos}) - Mean_{f_i}(D_{neg}))^2 \quad (3-3)$$

$$IntraDist_{f_i}(D_{pos}, D_{neg}) = S_{f_i}^2(D_{pos}) + S_{f_i}^2(D_{neg}) \quad (3-4)$$

公式(3-4)中 $S_{f_i}^2(D_{pos})$ 和 $S_{f_i}^2(D_{neg})$ 分别代表特征 f_i 的取值在正例和反例中的样本方差。实质上类间距离 $InterDist_{f_i}(D_{pos}, D_{neg})$ 描述了特征 f_i 正例样本中心和反例样本中心之间的距离，如果其值较小，则说明特征 f_i 的正例样本中心和反例样本中心相距较近，反之则相距较远；类内距离 $IntraDist_{f_i}(D_{pos}, D_{neg})$ 是特征 f_i 正例样本方差和反例样本方差之和，如果其值较小，则说明该特征 f_i 的所有正例样本距正例样本中心与所有反例样本距反例样本中心的总和较小，也就是说正反例样本较其对应的中心较为集中，反之则较为发散。受 Fisher 线性判别模型的启发^[104]，我们采用类间距离与类内距离的比值作为判断组合特征区分能力强弱的依据，比值越大则说明特征 f_i 对于类别区分的作用就越明显。因此，为每个组合特征 f_i 定义了一个统计量 $g(f_i)$ ，其计算方法见公式(3-5)。

$$g(f_i) = \frac{InterDist_{f_i}(D_{pos}, D_{neg})}{IntraDist_{f_i}(D_{pos}, D_{neg})} \quad (3-5)$$

然后，为了比较各个组合特征之间的差别，我们还需要对统计量 $g(f_i)$ 进行标准化，通过计算其标准化后的值 $Z\text{-score}$ ，作为该特征的最终判断指标。

该值能够有效衡量某个样本与样本均值之间的相差多少个标准差，但是该值依赖于总体分布的均值和方差，本文中采取一种简化方式，用样本均值和样本方差来代替总体分布的均值和方差，那么 $Z\text{-score}(g(f_i))$ (下文中简记为 $Z(f_i)$) 的计算方法如公式(3-6)、(3-7)和(3-8)所示。

$$Z(f_i) = \frac{g(f_i) - \overline{g(f_i)}}{S_G} \quad (3-6)$$

$$\overline{g(f_i)} = \frac{1}{N+1} \sum_{i=1}^{N+1} g(f_i), \quad i \in [1, N+1] \quad (3-7)$$

$$S_G = \sqrt{\frac{1}{N} \sum_{i=1}^{N+1} (g(f_i) - \overline{g(f_i)})^2}, \quad i \in [1, N+1] \quad (3-8)$$

式中 $\overline{g(f_i)}$ ——统计量 $g(f_i)$ 的样本均值；

S_G —— $g(f_i)$ 的样本标准差；

这样，计算出两个基本特征 f_a 、 f_b ，以及组合特征 f_{ab} 所对应的统计量 $Z(f_a)$ 、 $Z(f_b)$ 和 $Z(f_{ab})$ 之后，我们又定义了一个统计量 $I(f_{ab})$ 来衡量引入某组合特征之后，整体区分度的提高幅度，该统计量表示组合特征 f_{ab} 的 $Z(f_{ab})$ 值与两个基本特征之中得分较高的 Z 值之差，其计算方法如公式(3-9)所示。

$$I(f_{ab}) = Z(f_{ab}) - \text{Max}(Z(f_a), Z(f_b)) \quad (3-9)$$

我们先过滤掉对应 $I(f_{ab})$ 值为负的组合特征，由于这些组合特征不能够带来区分性能的提高。最后，根据 $Z(f_{ab})$ 的值对所有未被过滤的组合特征进行排序，选择其中的前 N 个(Top- N)作为组合特征集合， N 的具体取值通过后续在开发集上进行实验调整获得。该方法不但能够有效过滤对于正反例样本间的均值无明显差异的组合特征，保留适于语义角色识别和分类的特征，同时运算量较小，能够保证在处理大量组合特征时的速度，为接下来基于 SVM 模型的学习和分类提供较为有效的组合特征。需要注意的是本文在利用 SVM 分类器处理语义角色识别和语义角色分类两个子任务时，以及在处理语义角色分类中各个语义类别的分类任务时，所采用的基本特征集合是相同的，但采用的组合特征集合并不相同，原因是这些任务的正反例训练语料截然不同，因此在各自训练语料上根据上述方法获得的组合特征也不尽相同。

3.4 实验及结果分析

3.4.1 实验数据及评测指标

本文采用 CPB-1.0 语料库进行相关实验,该语料库在形式上与英文中的 PropBank 类似,是目前中文语义角色标注研究中的通用标准语料库。该语料库是在宾州中文树库基础上手工标注了其中的语义角色信息,共包含 760 篇文档,10364 个句子,4854 个谓词,以及 92959 个语义角色。为了便于与其他系统进行比较,我们采用一种较为通用的分割方法将其中的前 100 篇文档(从 chtb_001.fid 到 chtb_100.fid)作为测试语料,后 32 篇文档(从 chtb_900.fid 到 chtb_931.fid)作为开发语料,其余的 628 篇文档(从 chtb_101.fid 到 chtb_899.fid)作为训练语料。然后采美国康奈尔大学开发的 SVM-light 工具(version 6.02)作为 SVM 分类器^[97],由于本文采用的词汇化特征较多,导致特征空间维数较高,因此分类过程采用线性核函数。采用美国斯坦福大学开发的 Stanford Parser (version 1.6)作为自动短语结构句法分析器^[105]。为了验证和分析本文提出的基于多重句法特征的浅层语义分析方法在语义角色识别和语义角色分类这两个 SRL 子任务上的不同效果,我们将对这两个子任务分别进行评价。另外,本文还分别在 CPB 语料库中标注的正确句法分析结果上以及含有一些错误的自动句法分析结果上进行了语义角色识别和语义角色分类实验。实验过程中,语义角色标注仅以短语结构句法分析作为输入,依存句法分析的结果根据短语结构句法分析结果转换获得。语义角色识别阶段的输入实例为经过规则方法剪枝后剩余的全部候选句法成分,语义角色分类阶段的输入实例为语义角色识别阶段自动识别所得的句法成分,两个步骤之间衔接紧密,中间不经过任何人工干预。本文的评价指标采用 SRL 领域中广泛使用的准确率(A)、精确率(P)、召回率(R)以及 F 值(F)四个指标,具体计算方法见本文 1.2.4.1 小节。

3.4.2 组合特征选择的实验结果及分析

首先,我们根据本文 3.3 节中所述基于统计的组合特征选择方法,利用带有正确句法标注的训练语料作为数据集,选择出区分度较高的组合特征。表 3-5 中给出了在语义角色识别阶段以及语义角色分类各个阶段中根据 z-score 选择出的前十名组合特征,每个组合特征由两个基本特征中间通过一个加号连接的方式来表示。从表 3-5 中可见,目前较为常用的组合特征,如文献[21]中提出的组合特征‘谓词+短语中心词’(C1+C6)和‘位置+语态’(C4+C5)都出现在表中,特别是‘谓词+短语中心词’(C1+C6)组合特征在除 LOC 外其他的语义类别中结果都排在第一位,可见该特征的区分程度很高。在 LOC 语义类别

中组合特征‘语态+地点指示词’(C5+C27)取得了最好结果,在一定程度上也反映了新定义特征 C27 的作用。此外,我们还发现了一些目前未被发现的有效组合特征,包括‘短语中心词+次范畴框架’(C6+C7)、“短语中心词+谓词词性”(C6+C23)和‘谓词+短语类型’(C1+C3),在后文中将对这些组合特征的效果做出进一步验证。

表 3-5 根据 z 值排名前十位的组合特征

Table 3-5 The top-10 combined features ranked by z-score

Rank	SRI	ARG0	ARG1	ARG2	ADV	LOC	TMP
1	C1+C6	C1+C6	C1+C6	C1+C6	C1+C6	C5+C27	C1+C6
2	C1+D3	C32+C30	C30+D31	C1+D1	C30+D27	C9+D17	C22+C27
3	D25+D14	C6+C7	C30+D32	C1+C7	C30+D28	C9+D13	C6+C7
4	C4+D25	C1+C2	C5+C30	C7+C6	C1+C11	C2+C9	D26+D27
5	D25+D22	C1+C12	C30+D24	C1+C5	C24+D33	C23+C27	D26+D28
6	D25+D20	C23+C6	C30+C21	C1+C23	C30+D25	C9+C20	C23+D26
7	D25+D21	C1+C3	C4+C5	C23+C6	C24+D9	C14+C32	C5+D26
8	D25+D18	C10+D35	C1+C10	C1+C3	C27+C2	C10+C14	D26+D31
9	D25+D19	C10+D1	C30+D10	C5+C6	C22+C2	C9+C26	D26+D32
10	D25+D35	C10+D28	C4+C6	C1+D5	C24+D13	C14+C2	C23+C6

另外,我们还发现了在组合特征集合中出现频率最高的基本特征有:‘谓词’(C1)、“短语中心词”(C6)、“LCA 左分支类型”(C30)、“存在依存类型”(D25)和“词性路径”(D26)。这些基本特征在与其他特征组合后表现出了较好的区分能力。然后,我们在开发集合上估计了采用前 N 个组合特征中参数 N 的取值,结果表明当 N 的值达到 20 之后分类结果几乎不再发生变化。因此在实验中,我们将 N 的值分别设为 5, 10 和 20,这样产生的组合特征集合的大小分别为 28, 60 和 114 个组合特征,然后我们分别采用这些集合来进行后续的分类实验,根据整体的精度和效率决定参数 N 的取值。

3.4.3 正确句法分析基础上的实验结果及分析

首先,我们在带有正确句法标记的测试语料上对本文 3.2.1 节所述的剪枝后处理方法进行了评价,采用两个指标对剪枝效果进行评价,一是剪枝召回率,为剪枝后保留的正确语义角色数量与总正确语义角色数量之比;二是剪枝效率,表示正确剪掉的短语数与总短语数之比。测试语料中共含有 12270 个语义角色和 380623 个短语,测试结果详见表 3-6。从表 3-6 中可以看出,本文中

的剪枝后处理方法剪枝召回率未有明显下降，但剪枝效率却有明显的提高，提高幅度达 2%。剪枝效率的提高能够减少后续分类的样本数量，进而降低语义角色分类的错误率。

表 3-6 剪枝后处理方法实验结果

Table 3-6 Results of the post-processing method for pruning

方法	召回语义角色	正确剪枝短语	剪枝召回率(%)	剪枝效率(%)
原剪枝方法	12017	344447	97.94	90.50
加入后处理	12012	352421	97.90	92.59

然后，为了详细了解本文提出的方法各个阶段的有效性，我们采用基于线性核函数的 SVM 分类器和 6 个不同的分类特征集合，在训练语料上构造了 6 个浅层语义分析系统，分别如下：

- (1) 仅采用基于短语结构句法特征作为分类特征集合，记作 CFO。
- (2) 仅采用基于依存结构句法特征作为分类特征集合，记作 DFO。
- (3) 同时采用基于短语结构句法特征和基于依存结构句法特征作为分类特征集合，记作 CDF。
- (4) 同时采用基于短语结构句法特征和基于依存结构句法特征，以及 Top-5 的组合特征作为分类特征集合，记作 CDF+Top5。
- (5) 同时采用基于短语结构句法特征和基于依存结构句法特征，以及 Top-10 的组合特征作为分类特征集合，记作 CDF+Top10。
- (6) 同时采用基于短语结构句法特征和基于依存结构句法特征，以及 Top-20 的组合特征作为分类特征集合，记作 CDF+Top20。

首先，我们在带有正确短语结构句法标记的测试语料上评价了语义角色识别结果，依存结构句法分析的结果直接采用 Stanford Parser 中集成的句法成分到依存的中心规则转换器获得，在正确句法分析基础上测试 6 个 SRL 系统的语义角色识别结果如表 3-7 所示。

表 3-7 在正确句法分析上的语义角色识别结果

Table 3-7 Results of semantic role identification using gold parses

System	A (%)	P (%)	R (%)	F (%)
CFO	97.87	97.04	97.30	97.17
DFO	92.76	92.90	84.19	88.33
CDF	97.98	97.44	97.25	97.34
CDF+Top5	98.12	97.56	97.58	97.57
CDF+Top10	98.15	97.61	97.62	97.61
CDF+Top20	98.18	97.68	97.64	97.66

从表 3-7 中可见, ‘CDF’ 和 ‘CDF+Top20’ 系统之间仅比 ‘CFO’ 有很小的提高, 幅度小于 1%。也就是说, 依存特征的加入和组合特征的方法对于语义角色识别阶段分类的影响较小, 没有产生明显的效果提升。接下来我们仍在带有正确句法标记的测试语料上评价了语义角色分类结果, 表 3-8 给出了 6 个系统对于各个语义角色类型分类结果的 F 值。

表 3-8 在正确句法分析上的语义角色分类结果

Table 3-8 Results of semantic role classification using gold parses

System	Arg0 (F%)	Arg1 (F%)	Arg2 (F%)	ADV (F%)	LOC (F%)	TMP (F%)	ALL (F%)
CFO	92.40	90.57	59.98	96.25	86.80	98.14	91.23
DFO	90.70	88.22	56.95	94.54	81.23	97.37	89.14
CDF	92.85	91.29	63.35	96.55	87.55	98.32	91.86
CDF+Top5	93.96	92.79	73.48	97.13	88.63	98.31	93.22
CDF+Top10	94.15	93.23	74.18	97.42	87.17	98.57	93.41
CDF+Top20	94.10	93.19	75.13	97.23	88.05	98.48	93.46

表 3-8 中结果表明, 本文提出的基于多重句法特征的方法在语义角色分类阶段具有明显的效果。采用组合特征的系统 CDF+Top5 整体 F 值达到 93.22%, 与不含组合特征的 CFO、DFO、CDF 三个系统相比总体 F 值提高幅度均大于 2%。在每个具体语义类别的分类上, 本文方法也都取得了很大幅度的提升, 除了时间类型(ArgM-TMP)的语义角色原本已经达到了较高的水平。综合表 3-7 和表 3-8 的结果来看, 本文提出的组合特征选择方法能够有效地提升基于短语结构句法特征和基于依存结构句法特征的语义角色标注, 且这种提升主要发生在语义角色分类阶段。

此外, 我们还发现基于依存结构句法特征集合的 DFO 系统在识别和分类两个阶段均与其他系统有明显的差距, 在分类阶段差距略有减小。其原因主要是本实验中语义角色的标注单元为句法成分, 而依存结构中不含任何句法成分信息, 我们是通过将句法成分映射到中心词上的方法来提取句法成分的依存结构特征的, 因此依存特征中缺失了句法成分中的结构化信息, 导致其性能较差。而在分类阶段这种差距的减小说明了句法成分的中心词在语义角色识别阶段起到了重要的作用。

3.4.4 自动句法分析基础上的实验结果及分析

为了正确评价本文提出方法在实际应用中的真实有效性, 我们在经过 Stanford Parser 进行自动短语结构句法分析的测试语料上重复了上一小节中的

实验，依存分析的结果同样由自动短语结构句法分析的结果经中心规则转换得到。在自动句法分析基础上测试 6 个系统的语义角色识别结果如表 3-9 所示。

表 3-9 在自动句法分析上的语义角色识别结果

Table 3-9 Results of semantic role identification using automatic parses

System	A (%)	P (%)	R (%)	F (%)
CFO	71.54	68.72	70.62	69.66
DFO	68.86	65.06	60.68	62.79
CDF	73.53	70.63	72.75	71.67
CDF+Top5	73.62	70.69	72.98	71.82
CDF+Top10	73.65	70.71	73.08	71.88
CDF+Top20	73.67	70.70	73.16	71.91

表 3-9 中的结果表明，尽管相比基于正确句法分析的语义角色识别结果有很大程度的下降，降幅约为 26%，但本文提出的方法在基于自动句法分析的语义角色识别阶段表现出了较好的效果，较基于单一句法特征的系统 F 值提高了 2%。通过对错误结果的分析，我们发现相比基于正确句法分析结果大幅度下降的原因在于基于规则的剪枝策略将自动句法分析中的错误进一步放大。统计发现，在基于正确句法分析的语义角色识别阶段，语义角色中被错误剪枝掉的部分所占比率约为 2%，效果较好。而在基于正确句法分析的语义角色识别阶段，这部分比率升至 17%，是导致性能下降的主要原因。接下来，我们测试了在自动句法基础上的语义角色分类结果，表 3-10 给出了 6 个系统对于各个语义角色类型的分类结果 F 值。

表 3-10 在自动句法分析上的语义角色分类结果

Table 3-10 Results of semantic role classification using automatic parses

System	Arg0 (F%)	Arg1 (F%)	Arg2 (F%)	ADV (F%)	LOC (F%)	TMP (F%)	ALL (F%)
CFO	89.20	88.90	54.47	93.93	81.80	94.38	88.24
DFO	88.79	89.32	50.21	91.27	78.26	93.86	87.63
CDF	89.75	89.87	57.71	95.28	84.22	94.71	89.16
CDF+Top5	90.75	90.97	65.64	95.53	84.45	94.45	90.16
CDF+Top10	90.96	91.37	67.25	95.31	84.49	94.61	90.45
CDF+Top20	90.94	91.29	67.42	95.22	84.39	94.65	90.42

从表 3-10 中可见，相比基于正确句法分析的，语义角色分类阶段的结果也有一定程度下降，但降幅仅为 4%，远没有语义角色识别阶段那么明显，表

明了语义角色分类阶段受句法分析中错误的影响相对较小。表 3-10 中数据与表 3-8 中基本趋势类似,采用组合特征系统的提升幅度约为 2%。接着我们给出了在正确句法分析和自动句法分析基础上 6 个语义角色标注系统的整体结果,详见表 3-11。

表 3-11 语义角色标注整体结果

Table 3-11 Results of semantic role labeling using both gold and automatic parses

System	Gold Parse (F%)	Auto Parse (F%)
CFO	89.29	63.13
DFO	82.69	60.34
CDF	90.01	65.56
CDF+Top5	91.47	66.37
CDF+Top10	91.68	66.61
CDF+Top20	91.76	66.61

从表 3-11 中可见,在正确句法分析基础上采用多重句法特征的 CDF+Top20 系统的结果比 CFO 系统高出 2.5%,说明本文提出的基于多重句法特征方法在具有良好句法分析基础条件下的可行性。而在自动句法分析基础上,其结果更是高出了 3.5%,在自动句法分析上的效果更为明显进一步验证了本文方法在实际应用中的有效性。并且根据本文 1.2.4 节中介绍的卡方显著性检验方法,我们对系统之间性能的差异进行了统计显著性检验,发现在本文提出的多重句法特征系统与单一句法特征系统(CFO)在正确句法分析和自动句法分析条件下性能上的差异具有统计显著性($p < 0.05$),验证了本文方法在理论上的可行性和在基于正确句法分析结果的浅层语义分析上的有效性。

此外,从表 3-11 中的后三行结果中我们发现,加入 Top10 组合特征后相对于加入 Top5 组合特征有一定的效果提升,其代价是额外引入 32 个组合特征;而加入 Top20 组合特征后相对于加入 Top10 组合特征效果几乎没有提升,其代价是额外引入 54 个组合特征,因此在本文基于 CPB 语料库的浅层语义分析中,采用 Top10 组合特征是较为合理的选择。

3.4.5 组合句法特征的性能分析

为了验证每个组合句法特征在实际分类过程中的性能,并发掘出其中较为有效的组合句法特征,我们对 CDF+Top10 系统中采用的 60 个组合特征进行了排序,排序指标为在基本特征集合基础上引入该组合特征时 SRL 系统的 F 值提高幅度。表 3-12 中给出了排名在前 20 的组合特特征。

表 3-12 排名前 20 位的组合特征及各自性能

Table 3-12 Top-20 combined features and their performance improvement

Rank	Feature	ΔF (%)	Rank	Feature	ΔF (%)
1	C1+C6	0.611	11	C10+D1	0.413
2	C1+C10	0.593	12	C5+D26	0.404
3	C4+C6	0.557	13	C24+D9	0.395
4	C9+C20	0.503	14	D25+D35	0.395
5	C23+C6	0.494	15	C30+D24	0.377
6	C1+C3	0.458	16	C9+C26	0.377
7	C9+D13	0.449	17	C10+D28	0.368
8	C14+C10	0.431	18	C30+D29	0.365
9	C1+C5	0.422	19	C30+D30	0.361
10	C24+D33	0.413	20	C6+C7	0.361

表 3-12 中排名的 20 个组合特征中有一半含有依存结构句法特征, 说明了在浅层语义分析中依存结构句法信息能够为短语结构句法分析提供有效的补充。此外, 我们发掘出一些新的有效的组合特征, 例如: ‘谓词+部分句法路径’(C1+C10)、‘位置+短语中心词’(C4+C6)、‘短语中心词词性+右兄弟中心词性’(C9+C20)和‘短语中心词词性+论元右侧词’(C9+D13)。通过观察我们发现这些组合特征并不都是由重要的基本特征所组成, 一些效果普通的基本特征也能够构造出有效的组合特征, 例如: ‘部分句法路径’(C10), ‘短语中心词词性’(C9), ‘右兄弟中心词性’(C20)和‘论元右侧词’(D13)。另外, 我们还发现表 3-12 中给出的实际分类过程中组合特征的性能排名与表 3-5 中的排名并不完全一致, 这说明了本文中提出的组合特征估计方法还不能完全精确地判断组合特征的性能。本文提出的组合特征选择方法, 主要是为了提高处理大量组合特征时的效率, 组合特征的性能是通过计算语料中正反例样本之间统计信息得到; 而在本节实验中, 组合特征的性能是通过对语料中的样本进行逐个测试得到的, 因此产生二者在精确度上的差异。在实际应用中, 由于组合特征的巨大数量和语料库规模等条件制约, 采用逐个测试方法是不现实的。

3.4.6 整体性能对比

最后, 我们将本文提出的方法与目前效果最好的四个中文浅层语义分析系统进行了比较。第一个是文献 Xue 和 Palmer 在文献[47]中所提出的系统, 该系统采用 9 个基本特征和 2 个组合特征, 采用最大熵模型作为分类器, 记为 ‘XP’; 第二个是刘怀军等在文献[58]中所提出的系统, 该系统采用 19 个基本特征和 10 个组合特征, 也采用最大熵模型作为分类器, 记为 ‘Liu’; 第三个

是车万翔在文献[55]中所提出的系统，采用一种基于混合卷积树核方法直接比较两个句法成分之间的相似度，记为‘Che’；第四个是 Xue 在文献[29]中所提出的系统，该系统是在‘XP’系统基础上建立的，只是又引入了若干新的特征，记为‘Xue’；本文采用对比系统是前面取得效果较好的‘CDF+Top10’系统，记为‘Ours’。上述系统都是以句法成分作为标注单元，也都以 CPB 语料库作为数据集，与本文系统的设置一致，表 3-13 给出了在正确和自动句法分析条件下各系统性能之间的比较结果。

表 3-13 与其他中文 SRL 系统过的比较结果

Table 3-13 The comparison results with other Chinese SRL systems

System	Gold Parses (F%)	Auto Parses (F%)
XP	90.3	61.3
Liu	91.31	—
Che	91.67	65.42
Ours	91.76	66.61
Xue	92.0	66.8

从表 3-13 中可见本文提出的基于多重句法特征的方法的在正确和自动句法分析条件下的结果都仅次于‘Xue’系统，但是比其他三个系统‘XP’、‘Liu’和‘Che’的结果都要好，说明本文提出方法的性能基本达到了目前中文浅层语义分析领域内的先进水平，进一步证明了本文提出的多重句法特征集合在中文浅层语义分析问题上的有效性。

3.5 本章小结

本章从特征选择角度对中文浅层语义分析进行了深入研究，提出一种基于多重句法特征的中文语义角色标注方法。该方法将短语结构句法和依存句法两种类型的句法特征进行融合，构成基本特征集合，为浅层语义分析提供了更为丰富的句法信息。然后在基本特征集合的基础上，提出一种基于统计的组合特征选择方法，根据各个特征在语料库中的分布状况，利用类间距离和类内距离之比标准化后的值作为统计量，快速有效地筛选出适于语义角色识别和语义角色分类的组合特征，然后将这些组合特征和基本特征整合为分类特征集合，根据分类特征集合构造相应的输入特征向量，再利用 SVM 分类器进行学习和预测。最后在 CPB 标准语料库上的实验结果证明了该方法的有效性，在基于正确句法分析和自动句法分析的基础上，分别取得了 91.76% 和 66.61% 的整体 F

值，达到了目前中文浅层语义分析领域内较高的水平。并且我们对每个组合句法特征的性能进行了分析，挖掘出区分度较高的组合句法特征，包括：‘谓词+部分句法路径’ (C1+C10)，‘位置+短语中心词’ (C4+C6)，‘短语中心词词性+右兄弟中心词性’ (C9+C20)和‘短语中心词词性+论元右侧词’ (C9+D13)等，填补了目前中文浅层语义分析领域中组合特征研究方面的空白。

第4章 基于组合分类模型的浅层语义分析方法

4.1 引言

通过对现有浅层语义分析方法的总结,发现统计机器学习方法是目前解决浅层语义分析问题最为有效方法。统计机器学习方法简单来说就是自动地从数据或经验中获取知识并以此提高自身性能的一种算法,是继专家系统之后人工智能领域应用中所广泛采用的方法。这种方法避免了专家系统依赖大规模知识库、缺乏学习和发现的能力、扩展和移植代价较大等缺点。机器学习方法仅需要一些某种语言现象的已知数据作为训练语料,通过概率统计模型对训练语料进行学习,就可以实现对该语言现象的识别和预测。基于有指导机器学习方法的浅层语义分析有三个关键点,一是选择的语言学特征,二是采用的机器学习方法,三是训练机器学习方法所采用的语料库。目前,中文浅层语义分析语料库方面的改进较为复杂,需要一个长时间的积累过程和大量的人力物力。因此,本文主要从语言学特征和机器学习方法这两方面对浅层语义分析问题进行了研究,在上一章中我们对浅层语义分析中语言学特征的部分进行了深入分析和讨论,本章将从机器学习方法角度研究如何提高中文浅层语义分析的性能。

基于有指导的分类方法目前是浅层语义分析领域中最主流的方法,该方法基本流程如图 4-1 所示。其中常用的、效果较好的分类模型主要有决策树、感知器、最大熵、支持向量机等。这些模型以“谓词-论元”作为分类样本,每个“谓词-论元”对应目标谓词和该谓词的一个候选句法成分组成。候选句法成分是利用句法树剪枝算法,从剪枝后的句法分析树中提取得到。然后,根据已经定义好的语言学特征集合,把“谓词-论元”进行数值化,形成一种特征向量的形式作为分类器的输入模式。在分类过程中,先通过一个基于二值分类算法的语义角色识别步骤来判断一个“谓词-论元”对中的候选论元是否是目标谓词的真实论元;再通过一个基于多值分类算法的语义角色分类步骤来判断在上一步骤中所识别论元的语义角色类型,如施事、受事、时间、地点等。

目前,从机器学习方法方面提高浅层语义分析的效果相对比较困难,因为对于各种分类算法的研究都已经比较充分,各种常用分类模型几乎都被应用过,对于这方面的研究几乎到达了一种瓶颈状态,也使得目前大多数工作都集中在特征选择方面。针对这种情况,本文提出一种基于组合分类模型的浅层语义分析方法,该方法采用一种输入相关的选通机制(Input-dependent Gating Mechanism)将多个单一分类算法整合在一起,通过调整选通系统中的参数使在对每个样本的分类过程中能够充分发挥各分类算法的优势,减少单一分类算法

所产生的错误，该方法目前已经在很多领域中取得了较好的效果^[106,107]。本文将这种基于 EM 算法的组合分类模型引入到中文浅层语义分析问题上，探索通过基本分类方法的有效组合，能否给基于特征向量和分类方法的浅层语义分析方法带来性能上的提升，提升幅度究竟有多少。

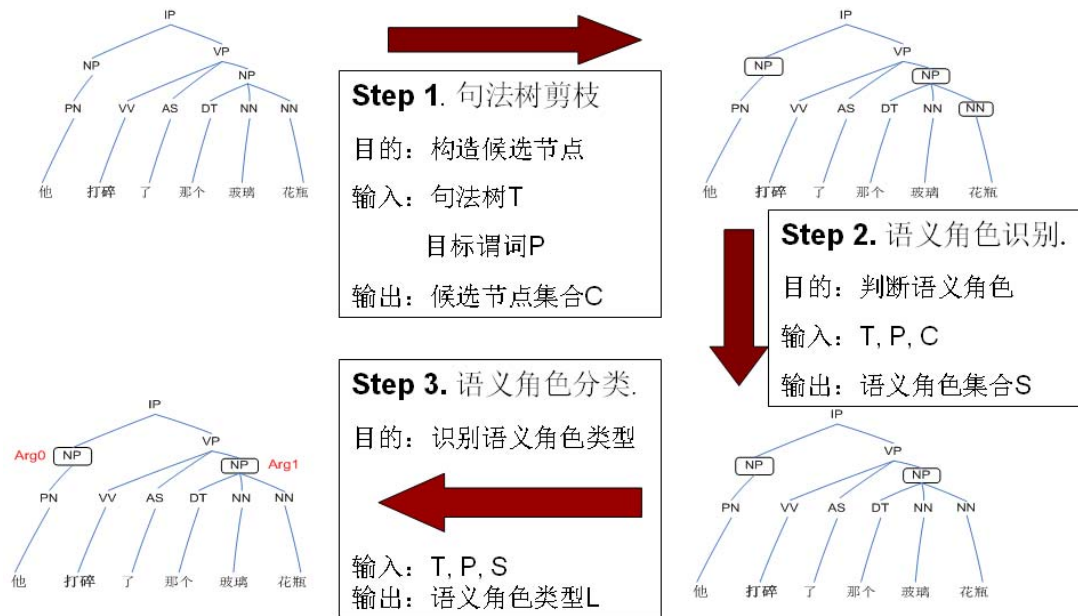


图 4-1 基于机器学习的语义角色标注方法框架^[25]

Fig.4-1 The framework of machine learning-based semantic role labeling method^[25]

本章中第 4.2 小节介绍了基于组合分类模型的浅层语义分析方法的基本结构以及组合分类方法所采用的特征集合。第 4.3 小节讨论了组合分类模型中采用的各个基本分类模型。第 4.4 小节着重介绍了一种基于 EM 算法的组合分类模型方法中参数学习的过程。第 4.5 小节我们在 CPB 语料库上对基于该方法的中文语义分析方法进行了实验，给出了详细实验结果及分析。最后在第 4.6 小节中对该方法进行了总结。

4.2 基于组合分类模型的浅层语义分析方法

本文提出的基于组合分类模型的浅层语义分析模型的基本结构如图 4-2 所示。我们采用一种组合机器学习方法，由多个预先训练好的浅层语义分析基本分类模型组成，然后通过一种输入相关的选通机制将这些基本分类模型集成到一起。选通机制在组合模型中扮演着十分重要的角色，可以通过调整选通系统中的参数协调各个基本分类模型，控制组合模型的输出。最后采用期望最大化 (Expectation Maximization, 简称 EM) 算法在训练语料上对选通系统中的参数进行学习，获得最终的浅层语义分析组合模型。

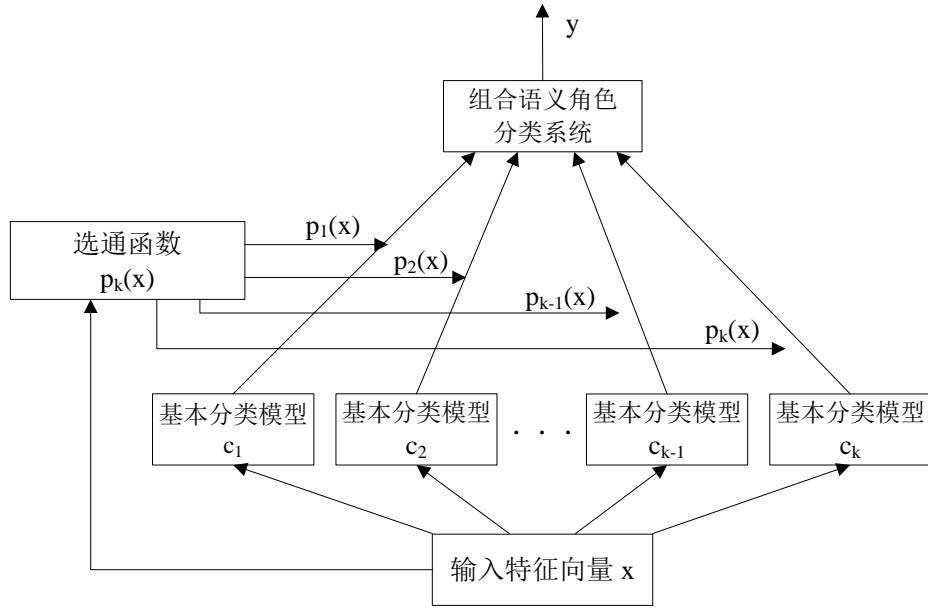


图 4-2 基于组合分类模型的语义角色标注方法

Fig.4-2 The basic framework of classifier combination-based SRL method

如图 4-2 所示, 假设组合模型中由 k 个预先训练好的语义角色基本分类模型组成: (c_1, c_2, \dots, c_k) 。每个分类器输出一个类别号, 分别代表待分语义类别。组合模型的输入包含三个部分: 剪枝后句法树 t 、目标谓词 p 和输入样本 x 。输入样本 x 代表候选语义角色所对应的特征向量; y 代表目标变量, 即组合模型输出 x 的语义角色类别; y_i 是各个基本模型的目标变量, 代表对应基本模型输出 x 的语义角色类别, 那么目标变量 y 的条件概率可表示为公式(4-1)。

$$p(y|t, p, x) = \sum_{j=1}^k \pi_j(x) \cdot p_j(y|t, p, x) \quad (4-1)$$

式中 x ——输入模式;

t ——剪枝后的句法树;

p ——目标谓词;

p_j ——第 j 个基本模型的输出概率。

公式(4-1)中, $\pi_j(x)$ 是输入相关的选通系数, 也可视为一个输入相关的组合系数, 取值范围在 0 到 1 之间, 且满足总和为 1 的约束条件; 本文中采用 softmax 函数作为选通函数, 其基本形式如公式(4-2)所示。

$$\pi_j(x) = \frac{\exp(\mathbf{w}_j^T x)}{\sum_i \exp(\mathbf{w}_i^T x)} \quad (4-2)$$

式中 \mathbf{w}_j ——第 j 个基本分类模型的权重。

另外，在公式(4-1)中 $p_j(y|t, p, x)$ 代表每个基本分类模型独立的概率，也就是第 j 个分类器的目标变量 y_j 的条件概率，可以表示成公式(4-3)。

$$p_j(y|t, p, x) = p(x \in S_y | c_j) \quad (4-3)$$

式中 c_j ——第 j 个基本分类模型；

S_j ——类别语义类别为 y 的样本集合。

假设训练数据集 D 中包含 n 个训练样本 $X = (x_1 \sim x_n)$ 。则根据公式(4-1)，目标变量条件概率的对数似然函数可以表示成公式(4-4)。

$$L(D|\theta) = \sum_{i=1}^n \ln p(y_i | t, p, x_i) \quad (4-4)$$

其中 θ 表示组合模型的参数向量，我们将采用一种基于 EM 算法的参数训练方法在训练数据集上估计出公式(4-2)中的选通函数 $\pi_j(x)$ ，然后就可以得到完整的基于组合分类模型的浅层语义分析模型。

综上所述，本文提出的基于组合模型的浅层语义分析可以分成如下四个步骤：第一步是特征选择，先给出构造各个基本分类模型必须的分类型特征集合，这里我们采用本文第3章中获得效果最好的‘CDF+Top10’特征集合，详见本文第3.2和3.4两节；第二步是构造多个独立的浅层语义分析模型，首先我们介绍各个分类器的基本模型以及特点，然后利用分类特征分别对这些基本模型进行训练；第三步是组合分类模型的参数训练，采用一种基于 EM 算法的参数训练方法，通过最大化目标变量条件概率似然函数的数学期望，迭代地求出组合模型中选通系统的参数 $\pi_j(x)$ ；最后一步，我们在 CPB 语料库上，对基于该组合分类模型的中文浅层语义分析进行了系统的评测。

4.3 基本浅层语义分析模型的构造

在训练组合模型之前，我们必须构造多个独立的基本分类模型作为组合模型中的基本元素，这些分类模型的性能将会对整个组合模型的分类型性能产生重要影响。因此，本文选择了五个较为有效的语义角色分类算法构造了五个分类模型作为组合模型的基本元素，这五个模型分别是 K 近邻、决策树、感知器、最大熵和支持向量机模型，下面我们将对这五种模型的基本原理以及各分

类模型的构造过程进行必要的介绍。

4.3.1 K 近邻(K-Nearest Neighbor, KNN)模型

K 近邻模型是一种理论上较为成熟的基于实例的分类模型^[109,110]。该模型基本思想是：一个样本的类别是由训练语料中与该样本距离最近的 K 个相邻样本所决定的，通过这 K 个最近邻样本之间多数投票的方法即可获得待分类样本的类别。在特征空间中样本之间距离是衡量样本相似性的度量，两个样本之间的距离越近，它们的相似度就越高。因此，如果与一个样本在特征空间中最相似的 K 个样本中的大多数都属于某一个类别，则判断该样本也属于这一类别是合理的。以一个简单的例子来说明，当 K=1 时，则将一个样本的类别判断为与它最近样本的类别。

KNN 算法的优点在于简单有效，不需要复杂的参数训练过程，该模型仅有一个参数 K，通常是在开发集上采用交叉验证的方法估计得到的。在分类过程中，输入模式的标记是通过与其距离最近的 K 个近邻的标记进行投票产生的。该方法主要靠周围有限个邻近的样本，而不是靠划分类别区域的方法来确定所属类别的，因此对于类别区域交叉或重叠情况较多的数据集来说，KNN 方法更为适合。KNN 算法的距离度量通常采用两个特征向量在特征空间中的欧氏(Euclidean)或马氏(Mahalanobis)距离，欧氏距离 D_E 和马氏距离 D_M 的计算方法如公式(4-5)和公式(4-6)所示。

$$D_E(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})(\vec{x} - \vec{y})^T} \quad (4-5)$$

$$D_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \quad (4-6)$$

式中 \vec{x} 、 \vec{y} ——特征空间中的两个样本向量；

Σ ——特征空间中样本集合的协方差矩阵。

KNN 算法本质上是一个局部模型，因为它在类别决策上只依据最邻近的若干个样本的类别来决定待分样本所属的类别。该算法的主要不足之处在于，对于数据集的局部结构较为敏感。当样本数量不平衡，例如某一类别的样本数量远多余其他类样本数量时，很可能导致在分类一个新样本时，该样本的 K 个近邻中大数量类的样本占多数。因此对于这种情况，我们采用增加权值的方法来改进，定义与样本距离越小的近邻权值越大。此外，该方法还有一个缺点就是它对语料库规模的依赖性较强，样本相似度计算量较大。样本的特征空间通常有很高的维数，训练样本也必须达到一定的数量，否则将导致类别信息不足。因此在确定 K 个最近邻时需要计算待分文本和所有训练样本的相似度，计算量较大，分类速度也比相对较慢。

本文中构造 KNN 浅层语义分析基本模型的基本流程是：

- (1) 计算要进行语义角色分类的句法成分 x_i 与训练集合 $D: \{x_1, x_2, \dots, x_N\}$ 中其它样本之间的欧氏距离，方法参见公式(4-5)。
- (2) 根据上一步中计算出的距离对训练集合 D 中所有样本进行排序。
- (3) 通过在开发数据上进行交叉验证的方法，以最小均方误差为准则，估计出模型中参数 K 的最优取值。
- (4) 为排序后距离最大的前 K 个样本中的每一个样本分配一个权值，权值大小为每个样本到输入样本 x_i 的距离的倒数。
- (5) 将这前 K 个样本按照标注的语义角色类别划分成若干个集合，将每个类别集合中所有样本的权值相加作为该类别的概率输出。

4.3.2 决策树(Decision Tree, DT)模型

决策树，或称判定树，是一种树型结构分类模型^[90,111]。决策树的根节点是整个数据集合空间，每个非叶节点是对样本某个属性的测试，该测试将数据集合空间分割成两个或更多块，每个叶节点则是属于类别的记录。其分类原理是将待分类的样本从决策树的根部输入，在通过树的非叶节点时对样本的某个属性进行判断，根据结果选择不同的路径逐层下降，到达树的叶节点时即可判断该样本的所属类别。其本质上是通过一系列规则对数据进行分类的过程。经典决策树算法主要有 CART 算法^[113]、ID3 算法^[114]和 C4.5 算法^[92]。

决策树分为分类树和回归树两种，分类树主要针对离散变量，回归树主要针对连续变量，浅层语义分析中所采用的特征大多为离散特征，因此本文主要采用分类树。构造分类决策树的过程为：首先寻找初始划分，整个训练数据集作为产生决策树的集合，训练集中每个记录都带有类别标记。然后决定特征集合中哪个特征属性域是目前最好的分类指标。一般的做法是穷尽所有的属性域，对每个属性域划分的好坏做出度量，选出最好的那个划分对应的特征属性域。决策树的创建就是根据样本特征的不同取值建立树的分支，然后在每个分支子集中重复建立下层结点和分支。建决策树的关键在于建立分支时对样本特征不同取值的选择。选择不同的特征会使划分出来的样本子集不同，影响决策树生长的快慢以及决策树结构的好坏，从而影响产生规则信息的质量，可见，决策树算法的技术难点也就是选择一个好的特征进行分支。

本文采用的 ID3 算法是 Quilan 提出的一种经典决策树算法，该算法的核心是在决策树各级结点上选择特征时，通过信息增益这一指标来选择特征，使得在每一个非叶结点进行测试时，能够获得关于测试样本最大的类别信息。其具体方法是通过检测所有的特征，选择信息增益最大的特征产生决策树结点

(关于信息增益的定义和计算方法的描述见本文第 2.2 节), 由该特征的不同取值建立分支, 再对各分支的子集递归调用该方法建立决策树结点的分支, 直到所有子集仅包含同一类别的数据为止, 最后得到的决策树即可用来对新样本进行分类。该算法的优点主要在于: 分类准确率高且速度最快; 模型产生的分类规则易于理解; 可同时接受离散型和连续型数据。其缺点是: 待分类别较多时, 决策树就会过于复杂, 导致预测效果较差; 属性相互关系协调不够, 难以表达复杂概念; 还有该模型抗噪声能力较差。

本文采用 ID3 算法建立决策树分类模型, 具体步骤如下:

- (1) 初始化决策树 T , 构造一个根节点 (D, F) , 其中 D 是全体训练样本集合: $\{x_1, x_2, \dots, x_N\}$, F 为全体特征集合: $\{f_1, f_2, \dots, f_M\}$ 。
- (2) 如果 T 中所有的叶节点 (D', F') 都满足, D' 中所有样本均属于同一语义角色类别或者 F' 为空, 则决策树 T 的构造完成, 算法结束。
- (3) 任取一个不具备(2)中所述条件的叶节点 (D'', F'') , 对于 F'' 中的每个特征 f_i , 计算其信息增益率 $\text{GainRatio}(f_i, D'')$ 。
- (4) 选择具有最高信息增益的属性 f_{\max} 作为节点 (D'', F'') 的测试特征, 并根据该特征的不同取值, 构造分支节点。
- (5) 返回第(2)步, 重复上述过程。

测试时, 将输入的候选句法成分从决策树的根节点送入, 在每次通过树的非叶节点时对输入样本的某个属性进行判断, 根据结果选择不同的分支路径逐层下降, 到达树的叶节点时即可判断出该输入句法成分所属的语义角色类别。

4.3.3 感知器(Perceptron)模型

感知器模型是一个具有单层计算单元的前馈式神经网络, 其神经元为线性阈值单元, 是一种典型的线性判别模型^[114, 115]。感知器模型由输入层和输出层两部分构成, 输出层是计算单元层。由于浅层语义分析任务是个多值分类问题, 因此本文采用一种多输出感知器模型, 输入层和输出层都由多个神经元构成, 输入部分的神经元与输出层的各种神经元间均有连接。当输入部分将输入数据传送给连接的处理单元时, 输出层就会对所有输入数据进行加权求和, 再由阈值型作用函数产生一组输出类别。

本文中 $X:(x_1, x_2, \dots, x_n)$ 代表输入模式, 即某个候选句法成分对应的特征向量, $Y:(y_1, y_2, \dots, y_m)$ 代表输入模式 X 对应的输出语义角色类别。 w_{ij} 表示输入层神经元 x_i 与输出层神经元 y_j 之间的连接权值。在每个输出层节点 y_j 处, 先将各个输入层分量 x_i 通过连接边加权 w_{ij} 之后累加, 然后通过一个阈值激活函

数来判断当前输入数据是否属于该类别，如公式(4-7)所示。最后把全部输出层节点整合即可得到样本的输出类别。感知器模型的优点是理论基础牢固、分类精度较高、通用性较好；缺点主要是对线性不可分问题的处理能力较差，抗噪声能力不足，收敛速度较慢。

$$y_j = \begin{cases} 1, & \text{if } \sum_{i=1}^n w_{ij}x_i - \theta_j \geq 0, i \in [1, n], j \in [1, m] \\ 0, & \text{if } \sum_{i=1}^n w_{ij}x_i - \theta_j < 0, i \in [1, n], j \in [1, m] \end{cases} \quad (4-7)$$

式中 x_i ——第 i 个输入层神经元；
 y_j ——第 j 个输出层神经元；
 w_{ij} —— x_i 与 y_j 之间的连接权值；
 θ_j ——激活函数阈值。

不同于前面所述的 K 近邻和决策树模型，感知器模型需要一个显式的训练过程，通过带有标记的训练语料对模型中的连接权值 w_{ij} 和激活函数阈值 θ_j 进行学习。感知器模型学习的基本思想是：逐步地将训练集中的样本输入到网络中，根据输出类别与标注类别之间的差别来调整网络中的权值 w_{ij} 和阈值 θ_j 。另外，参数学习过程对于每个输出层单元 y_j 都是独立的，因此本文以一个固定输出层单元 y 为例来说明参数学习的过程。对于一个固定输出层单元 y ，待学习的参数有 w_1, w_2, \dots, w_n 和权值 θ 。首先，我们通过向原模型增加一个输入层单元 $x_0 = -1$ ，并将 θ 作为该输入单元的权值 w_0 ，构造出一个新的感知器模型，并将新模型的阈值 θ_{new} 设为 0。根据公式(4-7)可知，新模型与原模型等价。这样便将问题简化为学习新模型的权向量 $\mathbf{w}:(w_0, w_1, w_2, \dots, w_n)$ ，具体学习过程如下：

- (1) 初始化权向量 $\mathbf{w}:(w_0, w_1, w_2, \dots, w_n)$ ，给权向量中的每个分量一个较小的非零数作为它的初始值。
- (2) 循环对每个输入的训练样本 $X:(x_1, x_2, \dots, x_n)$ 进行训练，样本标注语义角色类别为 c_i 。在第 t 次迭代时，由公式(4-7)计算出该样本的实际输出类别 $y(t)$ 。
- (3) 比较实际输出类别与标注的语义角色类别，如果两者一致，则不对权值进行修改，返回(2)步对下一个训练样本进行训练；如果两者不一致，则采用一下公式(4-8)对输入层和输出层之间的权向量进行修正。

$$w_i(t+1) = w_i(t) + \eta[c_i - y(t)]x_i, \quad i = 0, 1, \dots, n \quad (4-8)$$

式中 η ——学习率，用来控制模型收敛的速度；

c_i ——样本 x_i 对应的标注语义角色类别。

- (4) 如果，权向量不再变化或者迭代次数 t 超过设定阈值，则模型训练过程结束。否则，将 t 增加 1，继续对训练样本进行迭代学习。

4.3.4 最大熵(Maximum Entropy, ME)模型

最大熵模型是一种灵活的指数概率模型，已经广泛地应用在自然语言处理领域的各项任务中，包括浅层语义分析等^[116,117]。熵原本是一个物理学概念，它是描述事物无序性的参数，熵越大则无序性越强。信息论的开创者香农认为，信息的作用是消除人们对事物了解的不确定性，他把不确定的程度称为信息^[119]。随机事件的信息熵可以定义为：假设离散随机变量 x ，有 w_1, \dots, w_n 共 n 种可能的取值，每个取值 w_i 出现的概率用 $p(w_i)$ 来表示，则随机事件的信息熵是各项取值 w_i 的概率与其概率对数乘积的相反数的总和，如公式(4-9)所示。

$$H(p) = -\sum_{i=1}^N p(w_i) \log p(w_i) \quad (4-9)$$

最大熵模型的基本原理是：在对一个随机事件的概率分布进行预测时，预测分布应当满足全部已知的条件，而对未知的情况不要做任何主观假设，即遇到不确定情况时，就要保留所有可能性。这种情况下概率分布的信息熵最大，概率分布最均匀，预测的风险最小。简言之，就是要保留全部不确定性，将风险降低到最小。如果没有任何约束条件，那么当所有概率相等，也就是随机变量满足平均分布时信息熵最大。最大熵模型是以最大熵理论为基础建立的一种选择模型方法，即从符合约束条件的分布中选择条件熵最大的最优后验概率分布 $p(y|x)$ ，如公式(4-10)所示。

$$p^*(y|x) = \arg \max_{p \in C} H(y|x) \quad (4-10)$$

式中 C ——约束条件集合。

假设训练样本集合表示为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，其中 x 表示训练样本， y 表示 x 的标记类别，则给定 x 时 y 的条件熵 $H(y|x)$ 如公式(4-11)所示。

$$H(y|x) = -\sum_{x,y} p(x,y) \log p(y|x) \quad (4-11)$$

公式(4-10)中的约束条件通常由特征函数来表达。特征函数是最大熵模型

中的重要部分，其作用主要是对概率分布模型加以限制，使模型能够利用上下文的信息。特征函数一般情况下是一个二值函数，用来表达 x 与 y 之间存在的某种关系，通过这些关系对概率分布进行约束，如公式(4-12)所示。通过求解在限制条件下具有最大熵值的分布即可获得最优模型。

$$f_i(x, y) = \begin{cases} 1, & \text{if } x \text{ and } y \text{ are in some relationship} \\ 0, & \text{otherwise} \end{cases} \quad (4-12)$$

对于给定的训练样本集合 $\{(x, y)\}$ 和选定的 k 个特征函数 f_i ，约束条件 C 为：每个特征 f 在经验分布 $\hat{p}(x, y)$ 下的期望值 $\hat{p}(f)$ 应与其在实际分布 $p(y|x)$ 下的期望值 $p(f)$ 相符，形式化描述见公式(4-13)。

$$C = \{p \in P \mid p(f_i) = \hat{p}(f_i), i = 1, \dots, k\} \quad (4-13)$$

综上所述，最大熵模型可以简单描述成如下形式：

$$p^*(y|x) = \arg \max_{p \in C} H(y|x) \quad (4-14)$$

$$\text{s.t. } C = \{p \in P \mid p(f_i) = \hat{p}(f_i), i = 1, \dots, k\} \quad (4-15)$$

由公式(4-14)，通过拉格朗日乘数法，即可求出最优概率分布 $p^*(y|x)$ ，求解过程详见文献^[120]。最终最大熵模型的形式化描述可表示为公式(4-16)。

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x, y)\right) \quad (4-16)$$

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x, y)\right) \quad (4-17)$$

式中 λ_i ——拉格朗日乘子；

f_i ——特征函数；

$Z_\lambda(x)$ ——归一化因子。

最大熵模型的优点是理论基础牢固，特征选择较为灵活，分类精度较高，通用性很强。缺点是数据稀疏问题严重，且对语料库依赖性较强。本文采用 Zhang 开发的最大熵工具包^[121]，作为分类器进行浅层语义分析相关实验。测试过程中，将候选句法成分转换为特征向量后作为最大熵模型的输入，最大熵模型可直接输出该输入句法成分属于各个语义角色类别的概率。

4.3.5 支持向量机(Support Vector Machines, SVM)模型

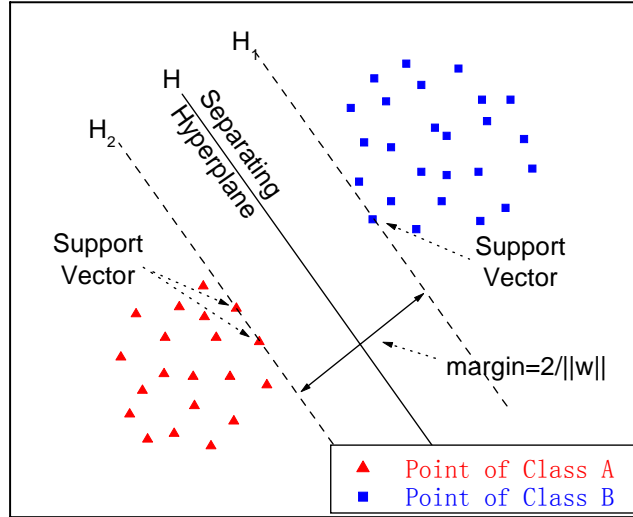


图 4-4 SVM 模型分类示意图

Fig.4-4 Demonstration of classification using SVM model

支持向量机的基本思想是先通过某种非线性映射(核函数)把输入向量映射到一个高维特征空间, 然后在高维空间中构造最优的分类超平面实现对类别的划分, 通过高维空间中计算超平面的方法来代替在原特征空间中计算复杂非线性曲面^[122]。图 4-4 给出了该模型的示意图, 最优超平面是指不但能将两类训练样本正确分开, 而且使每一类样本中与超平面距离最近的样本到超平面的距离最大的那个超平面, 形象地说就是使图 4-4 中分类边缘(margin)最大的超平面。这些与超平面距离最近的训练样本, 也就是图中超平面 H_1 和 H_2 上的点被称为支持向量, 最终 SVM 模型由这些支持向量所表达, 故称为支持向量机, 该模型目前已成功应用在 SRL 等领域中^[102]。

假设训练样本集合表示为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 x_i 是一个 n 维向量, 表示一个训练样本, y_i 表示 x_i 对应的类别标记, 取值为 1 或 -1。这样支持向量机模型的判别函数 $g(x)$ 可以用分类超平面的法向量 \mathbf{w} 和偏移量 b 来表示, 如公式(4-18)所示。

$$g(x) = \text{sign}(\mathbf{w}^T x + b) \quad (4-18)$$

式中 \mathbf{w} ——分类超平面的法向量;

b ——偏移量。

当 x_i 是正例时, 判别函数 $g(x)$ 值为 1; 当 x_i 是反例时, 其值为 -1。 \mathbf{w} 也可视为 x_i 对应的 n 维权重向量。求解最优超平面的过程也就是求解使图 4-4 中两

个超平面 H_1 和 H_2 之间的距离(即 margin)最大的超平面 H 的法向量 \mathbf{w} 和偏移量 b ，且该超平面须满足能够将所有训练样本正确分类这一条件。因此支持向量机模型的问题求解过程可以简要描述为公式(4-19)和公式(4-20)。

$$\arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (4-19)$$

$$\text{s.t. } y_i(\mathbf{w}^T x_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, n \quad (4-20)$$

经论证后可知该问题是一个线性约束下的凸二次优化问题，可采用拉格朗日乘数法对其进行求解，先将上述问题转换成对偶形式，把不等式约束条件下的优化问题转化为等式约束条件下的优化问题，简化求解的过程。最终得到支持向量机的判别函数见公式(4-21)和公式(4-22)。

$$g(x) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i (x_i^T x) + b \right) \quad (4-21)$$

$$b = y_k - \sum_{i=1}^n \alpha_i y_i (x_i^T x_k) \quad (4-22)$$

其中， x_k 为支持向量， α_i 是训练样本 x_i 所对应的拉格朗日乘子，大多数 α_i 的值都为 0，值不为 0 的 α_i 所对应训练样本 x_i 即为支持向量。从公式(4-21)中可见，支持向量机模型的判别主要是通过输入空间中计算测试向量 x 与所有支持向量 x_i 的内积，再乘以一定的权重后累加得到的。然而，当输入空间中的样本线性不可分时，需要通过满足一定条件的核函数把样本映射到另一个线性可分的高维特征空间，在高维空间中构造分类超平面对样本进行划分，采用核函数后的支持向量机判别函数见公式(4-23)和公式(4-24)。

$$g(x) = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right) \quad (4-23)$$

$$K(x_i, x) = \phi(x_i)^T \phi(x) \quad (4-24)$$

公式(4-24)中， $\phi(x)$ 表示到高维特征空间的映射函数，实际上这种映射是隐式实现的，支持向量机中采用核方法能够直接在输入空间中计算出映射到高维空间后的内积，提高了计算效率。常用的核函数主要有：线性核、多项式核、高斯核、Sigmoid 核等等。

支持向量机分类模型的优点在于：分类准确度很高、泛化能力强、善于处理小样本、非线性以及高维模式识别问题。其主要缺点是难以对大规模训练样

本进行学习和难以直接应用于多分类问题。本文中以美国康奈尔大学开发的 SVM-light 工具包^[97]作为 SVM 分类器进行浅层语义分析相关实验。浅层语义分析是一个多值分类问题。不同于最大熵模型能够直接处理多值分类问题，SVM 是一个二值分类模型，因此本文为每一个待分语义角色类别构造一个 SVM 分类器。在测试过程中，将候选句法成分转换为特征向量后作为输入样本，分别送入训练好的各个语义角色类别对应的 SVM 分类模型中。然后将输入样本到各个 SVM 模型的分超平面的距离进行归一化后，作为该输入样本属于各个语义角色类别的输出概率。

4.4 基于 EM 算法的组合模型参数训练方法

根据 4.2 节中所述的组合分类模型基本结构，本文提出的基于组合模型的浅层语义分析方法包含以下四步：1) 特征选择；2) 构造多个独立的基本分类模型；3) 组合基本分类模型并训练参数；4) 组合模型的评价。上一小节中我们对第 2 步构造基本分类模型进行了详细描述，而特征选择方法在本文第 3 章中也已有很详细介绍，此处不再赘述。下面我们着重介绍第 3 个步骤，组合分类模型的构造及基于 EM 算法的参数训练方法。假设组合模型中包含一组 k 个训练好的语义角色基本分类模型 (c_1, c_2, \dots, c_k) ，每个基本模型都会输出一个语义角色类别。首先为每个样本 x_n 引入一组由 k 个二值指示变量组成的隐变量 z_{nk} ，其值为 0 或 1，用于指示组合分类器中的哪个成员模型对生成该样本起作用。例如， z_{nk} 的值为 1 时表示第 k 个基本分类模型对样本 x_n 的分类起作用。 z_{nk} 的分布满足伯努利分布，其概率可以表示为： $p(z_{nk}=1) = \pi_k(x_n)$ 。则 z_n 的概率密度函数可以表示为公式(4-25)。

$$p(z_n) = \prod_{j=1}^k \pi_j(x_n)^{z_{nj}} \quad (4-25)$$

假设 x_n 代表输入样本，即用特征向量表示的潜在语义角色； y_n 代表包含分类结果的目标变量，则给定样本 x_n 、句法分析树 t 和目标谓词 p 情况下 y_n 的条件概率可表示为：

$$p(y_n | z_{nk}=1) = p_k(y_n | t, p, x_n) \quad (4-26)$$

其中 $p_k(y | t, p, x)$ 是第 k 个分类器的输出概率，表示给定分类器 c_k 条件下 x 属于 y 类的条件概率。将公式(4-26)中的 z_{nk} 整合，可得 y_n 关于 z_n 的条件概率：

$$p(y_n | z_n) = \prod_{j=1}^k p_j(y_n | t, p, x_n)^{z_{nj}} \quad (4-27)$$

根据公式(4-25)和(4-27), y_n 与 z_n 的联合概率可以表示为公式(4-28)。

$$p(y_n, z_n) = p(z_n) \cdot p(y_n | z_n) = \prod_{j=1}^k \pi_j(x_n)^{z_{nj}} \cdot p_j(y_n | t, p, x_n)^{z_{nj}} \quad (4-28)$$

其中 $\pi_j(x)$ 是选通系统的函数, 是一个根据输入样本 x 的变化而变化的组合系数, 其取值范围从 0 到 1 并且约束所有 $\pi_j(x)$ 的总和为 1, 其形式如公式(4-2)所示。接下来我们采用 EM 算法, 通过最大化对数似然函数来训练组合模型的参数 $\pi_j(x)$ 。假设训练数据集 D 中含有 n 个样本 $X: x_1, x_2, \dots, x_n$, 对应的语义类别标记为 $Y: y_1, y_2, \dots, y_n$, y_n 和 z_n 联合概率分布的对数似然函数可表示为公式(4-29)。

$$L(D | \theta) = \sum_{i=1}^n \ln p(y_i, z_i) \quad (4-29)$$

将公式(4-28)代入公式(4-39)右部的对数函数中, 即可得到引入隐变量后的对数似然函数, 可表示为公式(4-30)。

$$L(D | \theta) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \ln(\pi_j(x_i) \cdot p_j(y_i | t, p, x_i)) \quad (4-30)$$

接下来我们用 EM 算法迭代地估计选通函数 $\pi_j(x)$ 中的权重参数 w_j , 算法分为四个步骤:

- (1) 首先模型的初始化, 为每个参数赋一个初值, 本文取 $1/k$ 。
- (2) E 步, 计算对数似然函数的数学期望, 如公式(4-31)所示。

$$Q(\theta, \theta^{(t)}) = E[L(D | \theta)] = \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} (\ln \pi_j(x_i) + \ln p_j(y_i | t, p, x_i)) \quad (4-31)$$

其中 θ 表示模型的参数, 在第 t 次迭代后 θ 的估计值记为 $\theta^{(t)}$ 。 γ_{ij} 是隐变量 z_{ij} 的期望值, 可表示为公式(4-32)。

$$\gamma_{ij} = \frac{\pi_j(x) p_j(y | t, p, x)}{\sum_i \pi_i(x) p_i(y | t, p, x)} \quad (4-32)$$

(3) M 步: 利用拉格朗日乘数法在 $\sum_j \pi_j(x) = 1$ 的约束下, 即可得到使 $Q(\theta, \theta^{(t)})$ 达到最大化的参数 θ , 如公式(4-33)所示。

$$\theta = \arg \max_{\theta} Q(\theta, \theta^{(t)}) \quad \text{s.t.} \quad \sum_j \pi_j(x) = 1 \quad (4-33)$$

交替进行 (2)、(3) 两个步骤, 当参数 θ 值稳定不变时算法结束。

通过逐步调整模型的参数，可以使参数和训练样本的似然概率逐渐增大，最后使模型收敛于一个极大值。下面对本文算法的收敛性进行简要说明，首先第（2）步中的对数似然函数的数学期望函数可以表示成：

$$Q(\theta, \theta^{(t)}) = E[L(D|\theta)] = \sum_y L(D|\theta) p(y|X, \theta^{(t)}) \quad (4-34)$$

其中，对数似然函数 $L(D|\theta)$ 还可以表示成公式(4-35)。

$$L(D|\theta) = \log(p(X, Y|\theta)) = \log p(Y|X, \theta) + \log p(X|\theta) \quad (4-35)$$

将公式(4-35)带入公式(4-34)中可得：

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \sum_y L(D|\theta) p(y|X, \theta^{(t)}) \\ &= \sum_y \log p(y|X, \theta) p(y|X, \theta^{(t)}) + \sum_y \log p(X|\theta) p(y|X, \theta^{(t)}) \\ &= \sum_y \log p(y|X, \theta) p(y|X, \theta^{(t)}) + \log p(X|\theta) \cdot \sum_y p(y|X, \theta^{(t)}) \\ &= \sum_y \log p(y|X, \theta) p(y|X, \theta^{(t)}) + L(\theta|X) \end{aligned} \quad (4-36)$$

根据公式(4-36)，将 $Q(\theta^{(t+1)}, \theta^{(t)})$ 与 $Q(\theta^{(t)}, \theta^{(t)})$ 相减可得：

$$\begin{aligned} &Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) \\ &= L(\theta^{(t+1)}|X) + \sum_y \log p(y|X, \theta^{(t+1)}) p(y|X, \theta^{(t)}) \\ &\quad - L(\theta^{(t)}|X) - \sum_y \log p(y|X, \theta^{(t)}) p(y|X, \theta^{(t)}) \\ &= L(\theta^{(t+1)}|X) - L(\theta^{(t)}|X) - \sum_y \log \frac{p(y|X, \theta^{(t)})}{p(y|X, \theta^{(t+1)})} p(y|X, \theta^{(t)}) \end{aligned} \quad (4-37)$$

公式(4-37)中的最后一项是 KL 散度，其值不小于零。因此似然函数 L 的值随着其数学期望 Q 值的增大而增大，当参数 θ 使 Q 取得极大值时，观测数据的似然函数也会在相同的 θ 处取得极大值，也就是说 EM 算法会收敛到对数似然函数的局部最优值。通过上述方法，在训练数据集上迭代计算出参数 $\pi_j(x)$ 的值后，我们就可以采用组合分类模型来对新样本进行浅层语义分析。根据图 4-2 中描述的模型基本框架，已知输入样本 x ，浅层语义分析的最终结果可以通过计算的每个成员分类模型 j 的输出结果与该成员模型对应的选通系数 $\pi_j(x)$ 的乘积，再将所有成员分类器得到的乘机相加，所有成员模型的累加和就是组合分类器系统的输出目标值，即代表这输入样本对应语义角色类型。

4.5 实验结果及分析

4.5.1 实验数据及评测指标

为了验证本文提出的基于组合分类模型的浅层语义分析方法的有效性，我

们仍以本文第3章中采用的语料库，即中文浅层语义分析领域中最常用的CPB语料库为数据集，在上面进行了相关实验。该语料库中包含760篇文章，10364个句子，4854个谓词，以及92959个语义角色。仍采用第3章中数据集的划分方法，将其中的前100篇文档(从chtb_001.fid到chtb_100.fid)作为测试语料，后32篇文档(从chtb_900.fid到chtb_931.fid)作为开发语料，其余的628篇文档(从chtb_101.fid到chtb_899.fid)作为训练语料。评价指标仍采用准确率(A)、精确率(P)、召回率(R)以及F值(F)，详细计算方法见本文1.2.4小节。

首先，我们利用628篇训练语料分别来训练K近邻(KNN)、决策树(DT)、感知器(PER)、最大熵(ME)以及支持向量机(SVM)这五个基本分类模型。接着我们采用第4.2节中所述的方法，将这五个基本分类模型构造成为一个组合分类模型，然后利用EM算法对组合分类模型中的参数进行学习。我们分别在正确和自动短语结构句法分析两个阶段对各基本分类模型以及本文提出的组合分类模型(记为CM)在浅层语义分析性能进行了评价。其中，正确句法分析结果可直接从CPB语料中获得，自动短语结构句法分析器采用斯坦福大学开发Stanford Parser (version 1.6)句法分析器^[105]。

4.5.2 正确句法分析基础上的实验结果及分析

首先我们根据测试语料中的句法标注信息，也就是在没有错误句法分析干扰的条件下，测试了本文提出的方法取得的结果。另外我们还分别对核心语义角色和修饰性语义角色进行了测试。核心语义角色的标注结果见表4-1，修饰性语义角色的标注结果见表4-2。

表4-1 核心语义角色的标注结果

Table 4-1 The labeling results of core semantic roles

	KNN	DT	PER	ME	SVM	CM
A (%)	85.48	86.33	89.04	92.13	92.76	93.45
P (%)	86.16	86.24	89.36	92.17	93.07	93.72
R (%)	84.39	85.45	87.90	90.16	91.50	93.06
F (%)	85.27	85.90	88.63	91.26	92.28	93.39

表4-2 修饰性语义角色的标注结果

Table 4-2 The labeling results of adjunctive semantic roles

	KNN	DT	PER	ME	SVM	CM
A (%)	84.63	85.67	87.85	90.17	91.08	91.65
P (%)	85.19	85.40	88.37	90.02	91.79	91.31
R (%)	83.90	84.71	86.52	88.89	90.14	91.14
F (%)	84.54	85.06	87.44	89.45	90.96	91.23

根据表 4-1 和表 4-2 可以看出, 本文提出的组合分类模型方法取的结果(表中的 CM 列), 比其他 5 个基本模型有明显的效果提升。在 5 个基本模型中, SVM 模型取得了最好的效果, 在核心语义角色和修饰性语义角色上分别取得了 92.28%和 90.96%的 F 值, ME 模型的效果次之。而本文提出的 CM 模型比最好的 SVM 基本模型还有所提高, F 值达到 93.39%和 91.23%。另外, 核心语义角色的整体效果比修饰性语义角色的效果要高, 主要是由于修饰性语义角色较为稀疏并且在语言描述上也较为灵活和复杂。下面给出各模型在正确句法分析上的整体浅层语义分析结果, 详见表 4-3。

表 4-3 正确句法分析上浅层语义分析的整体结果

Table 4-3 The shallow semantic parsing results using gold parses

	A (%)	F (%)
KNN	85.09	84.94
DT	86.03	85.52
PER	88.50	88.09
ME	91.24	90.44
SVM	92.0	91.68
CM	92.63	92.41

从表 4-3 中能够更加清楚地看出, 本文提出的 CM 模型取得了表中 6 个模型中的最好结果, F 值达到 92.41%。相比基本模型中分类性能最好 SVM 模型 F 值的提高 0.7%。通过卡方显著性检验(方法详见本文 1.2.4 节)发现, 本文提出的组合分类模型(CM)相比其他基本模型在性能上的提高均具有统计显著性($p < 0.05$), 验证了本文方法在理论上的可行性和在基于正确句法分析结果的浅层语义分析上的有效性。

4.5.3 自动句法分析基础上的实验结果及分析

表 4-4 核心语义角色的标注结果

Table 4-4 The labeling results of core semantic roles

	KNN	DT	PER	ME	SVM	CM
A (%)	60.12	62.19	64.74	66.94	67.82	69.07
P (%)	60.18	62.16	64.65	66.89	68.07	68.60
R (%)	58.47	61.01	63.08	66.84	66.44	68.18
F (%)	59.31	61.58	63.86	66.87	67.25	68.39

表 4-5 修饰性语义角色的标注结果

Table 4-5 The labeling results of adjunctive semantic roles

	KNN	DT	PER	ME	SVM	CM
A (%)	58.28	61.03	62.91	66.43	66.03	67.59
P (%)	58.02	60.77	62.60	66.03	66.36	67.39
R (%)	57.52	60.23	62.51	64.96	65.33	67.01
F (%)	57.77	60.50	62.56	65.49	65.84	67.20

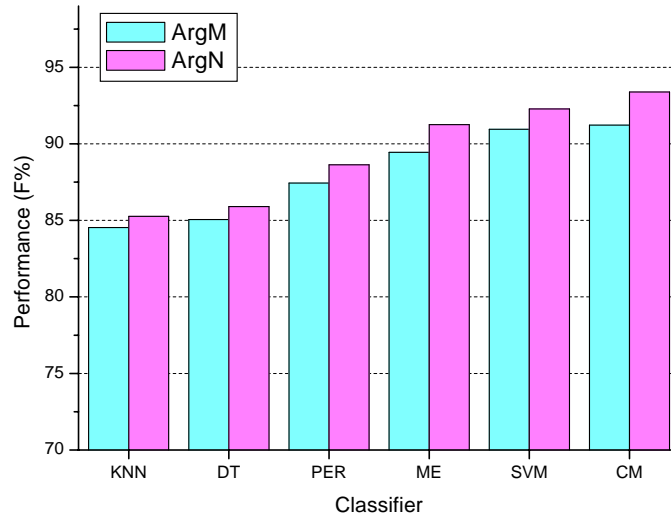
为了验证本文提出的方法在带有噪声的实际任务中的性能，我们在 Stanford Parser 产生的自动句法分析结果基础上，重复进行了上述实验。核心语义角色的标注结果见表 4-4，修饰性语义角色的标注结果见表 4-5。从表 4-4 和表 4-5 中可见，本文中的 CM 模型在自动句法分析基础上仍取得了最好的效果，在核心语义角色和修饰性语义角色上分别取得了 68.39% 和 67.2% 的 F 值，比基本模型中性能最好的 SVM 模型高出 1.14% 和 1.36%。整体提高幅度跟在正确句法分析基础上的提高幅度相当，可见本文方法对数据的依赖性较弱，在带有噪声的自动句法分析数据上仍能保持较好的效果。通过上述实验得到各分类模型在自动句法分析上的整体浅层语义分析结果，见表 4-6。

表 4-6 自动句法分析上浅层语义分析的整体结果

Table 4-6 The shallow semantic parsing results using automatic parses

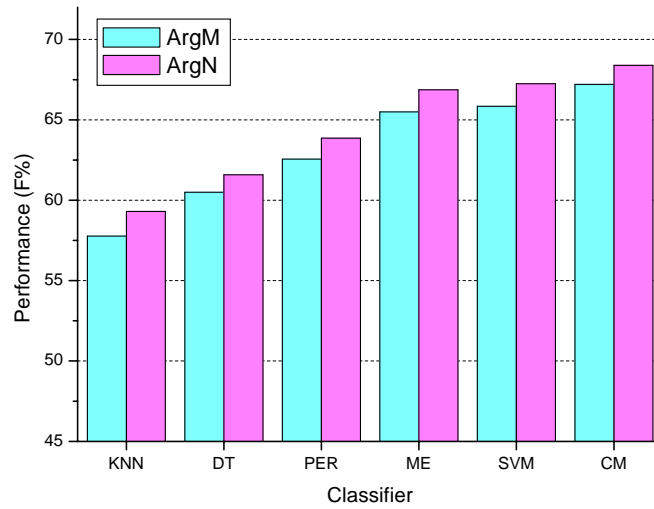
	A (%)	F (%)
KNN	59.28	58.61
DT	61.66	61.09
PER	63.91	63.27
ME	66.71	66.24
SVM	67.01	66.61
CM	68.4	67.85

表 4-6 中的结果表明，本文提出的 CM 模型取得了表中 6 个模型中的最好的 F 值 67.85%。相比基本模型中分类性能最好 SVM 模型 F 值的提高 1.24%。同样通过卡方显著性检验我们发现本文提出的组合分类模型比其他基本模型在性能上的提高均具有统计显著性($p < 0.05$)，进一步验证了本文方法在实际问题中，也就是在基于自动句法分析的浅层语义分析上的有效性。



a) 基于正确句法分析结果

a) Results on gold syntactic parses



b) 基于自动句法分析结果

b) Results on automatic syntactic parses

图 4-5 浅层语义分析系统比较的柱状图

Fig.4-5 The comparison results of shallow semantic parsing systems in histogram form

最后我们通过一种直方图的形式，直观地描述了各个分类模型在正确句法分析和自动句法分析上的分类性能，参见图 4-5，其中核心语义角色标记为 ArgN，修饰性语义角色标记为 ArgM。从图中可以清楚的看到本文提出的 CM 模型的整体效果要高出其他基本机器学习模型，无论是在正确句法分析还是自动句法分析的基础上，而且修饰性语义角色的标注效果低于核心语义角色。综上所述，本文提出的基于组合分类模型的浅层语义分析方法有效地提高了基于特征和有指导分类方法的浅层语义分析的整体准确性，且对核心语义角色还是修饰性语义角色的分类都有所帮助。

4.6 本章小结

本章在上一章提出的特征集合基础上，提出了一种基于组合分类模型的中文浅层语义分析方法，从机器学习的角度进一步来完善中文浅层语义分析方法。首先该组合模型由多个预先训练好的浅层语义分析基本分类模型组成，其中包括 K 近邻、决策树、感知器、最大熵以及支持向量机五个分类模型，作为组合模型中的基本元素。然后通过一种输入相关的选通系统将基本分类模型集成到一起。可以通过调整选通函数中的参数协调各个基本分类模型，控制组合模型的输出。然后采用 EM 算法在训练语料上对选通函数中的参数进行学习。在 CPB 语料库上的相关实验结果表明，该方法能够明显改进中文语义角色分析的效果，在正确句法分析以及自动句法分析基础上的整体 F 值分别达到了 92.41% 和 67.85%，超过了本文第 3 章中提出的多重句法特征方法所取得的 91.76% 和 66.61% 的整体 F 值，同时优于 Xue 在文献[28]中提出系统的 92% 和 66.8% 的整体 F 值，取得了目前中文浅层语义分析领域中比较好的结果。而且该模型在核心语义角色和修饰性语义角色方面都获得了具有统计显著性的效果提升，验证了这种基于组合分类模型的中文浅层语义分析方法在整体和局部上的有效性。

第5章 基于计算认知模型的浅层语义分析方法

5.1 引言

在对浅层语义分析方法中的特征选择以及机器学习方法两个关键方面进行了研究和分析后, 本文从更本质的认知层面对中文浅层语义分析进行探索, 提出一种基于认知模型的中文浅层语义分析方法。认知科学是上世纪具有标志性的一门新兴研究学科, 是神经科学、心理学、信息科学、数学、语言学、人类学乃至自然哲学等学科交叉发展的结果, 已经引起了全世界科学家们的广泛关注。上世纪七十年代, 神经学、心理学、语言学和计算机科学等领域研究者们发现他们共同关注着一些重要问题, 如信息的获取、加工、存储和传递, 思维的产生和变化, 知识的表征以及理解的过程等。这些问题本质上都是对人类认知过程的研究, 只是层次和角度有所不同, 在对这些问题研究的基础上形成了认知科学这门跨领域的研究学科。认知科学主要以记忆、知觉、语言、表征、思维、推理、决策、规划等人类的认知行为作为研究对象来探索人类的认知过程, 涉及领域包括认知神经学、认知心理学、认知语言学、人工智能等。

国际上对于认知科学的研究较早。二十世纪七十年代之后, 大脑成像技术的发展突飞猛进。人们在研究认知功能与大脑的关系时, 通过事件相关电位(ERP)、功能核磁共振(fMRI)、正电子发射断层扫描术(PET)等成像技术能够更直接地观测到人在进行各种认知活动时, 大脑出现的相应结构与功能的变化。这些成像技术促进了认知神经学的迅速发展, 产生了认知神经科学的新趋势, 同时也迅速带动了整个认知科学理论的发展。目前认知科学研究者们在提认知理论的同时, 越来越重视可计算认知模型的建立, 通过将认知理论转换成为相应的可计算认知模型, 拟合一些已有的认知数据, 验证认知理论的正确性和有效性。而且在一些应用相关领域, 人们已经开始尝试通过构造可计算的认知模型来实现认知功能的物化。国内对于认知科学的研究基本上刚刚处于起步阶段, 中科院神经所、北京师范大学、复旦大学等机构均已开展了此方向的研究。2008年英国剑桥大学出版的《The Cambridge Handbook of Computational Psychology》一书中, 详述了近年来计算认知模型领域所取得的研究成果, 但其中大部分模型都是为了验证认知过程的某个层面, 很少有涵盖整个语言理解过程, 具备解释或预测能力, 并能应用于实际系统中的可计算认知模型^[123]。

从整体趋势上来看, 对计算认知模型的研究是人工智能和自然语言处理领域的发展方向之一, 也是从根本上解决通过计算机来实现智能及语言处理等相关问题的途径和方法。目前国内外对于面向实际应用问题的计算认知模型的研

究还很不充分, 本文提出了一种基于计算认知模型的浅层语义分析方法, 目前未见相关研究的文献报道, 该方法探索了认知模型在实际应用问题中的基本方法及其产生的效果, 对与基于认知模型的自然语言处理研究有重要意义。下面首先对自然语言处理领域相关的主要计算认知模型进行简要的介绍。

5.2 主要计算认知模型概述

1987年, Laird 等人^[124]提出的 SOAR(State Operator And Result)体系是早期代表性计算认知模型之一。SOAR 提供了一个整合的关于问题解决和学习的体系结构, 该体系包含三个层次: 知识层、问题空间层和符号层, 通过一个产生式系统来实现。知识层用来提供产生式形式的外部知识, 问题空间层是核心部分, 由一些状态和操作组成, 而问题的求解被抽象成为在问题空间中搜索目标状态的过程, 搜索过程是通过一种子目标机制对问题空间的整合实现的。该模型可应用于机器人控制、自然语言理解、学习和规划等实际任务^[124,125]。

1995年, Anderson 等人^[127]提出 ACT(Adaptive Control of Thought)认知模型, 该模型依据认知理论以人类记忆系统中的两个体系为基础: 陈述性记忆(Declarative Memory)和程序性记忆(Procedural Memory), 陈述性记忆用来存储知识、事实和信息, 在应用中具体包括任何能够用词语、图形或符号进行回忆和描述的内容; 程序性记忆是用来存储与动作或系列动作相关的信息。模型的核心与 SOAR 相似也是一个产生式系统, 通过模式匹配、产生式的选择和执行来实现, 该模型较适用于行为预测和智能训练系统^[128]。

连接主义模型(Connectionist Model)在人工智能领域被称作神经网络模型(Neural Network)^[129,130], 是在模拟人脑神经处理机制基础上建立的, 但远不及自然神经网络的复杂高效, 是人们对其的一种简单理解和抽象。该模型以有向图为拓扑结构的动态系统, 以多层感知器为组织形式, 通过对离散或连续的输入作出状态响应来实现信息处理。神经网络模型的优点在于能够充分逼近任意复杂度的非线性关系, 能够学习与适应严重不确定性系统的动态特性, 且具有较强的鲁棒性和容错性。但是该模型缺乏统一的网络模型和通用学习算法, 网络的层数、隐含层神经元数和作用函数类的选择缺少指导性原则^[131,134]。

1998年, Kinstch 提出了一种较为完善的构造整合(Construction-Integration Model, 简称 CI) 模型^[135], 在计算认知模型研究中取得了突破性的成果。CI 模型的形式是一个命题关联网, 由命题所组成的节点用来模拟神经元, 节点间的连接边模拟突触, 输入外界刺激会给整个网络带来一个初始激活, 然后激活会在网络的各个节点中扩散传播, 即扩散激活理论^[136]。通过在命题网络中模拟扩散激活的过程, 最后根据各节点的激活量来判断他们与输入刺激的语义相关程度。CI 模型是一种混合模型, 它集成了符号系统与联结主义模型各自

的特点，为自然语言语义计算提供了一种有效的计算手段，在词义消歧、共指消解、故事理解等实际应用中都取得了较好的成果^[137,138]。

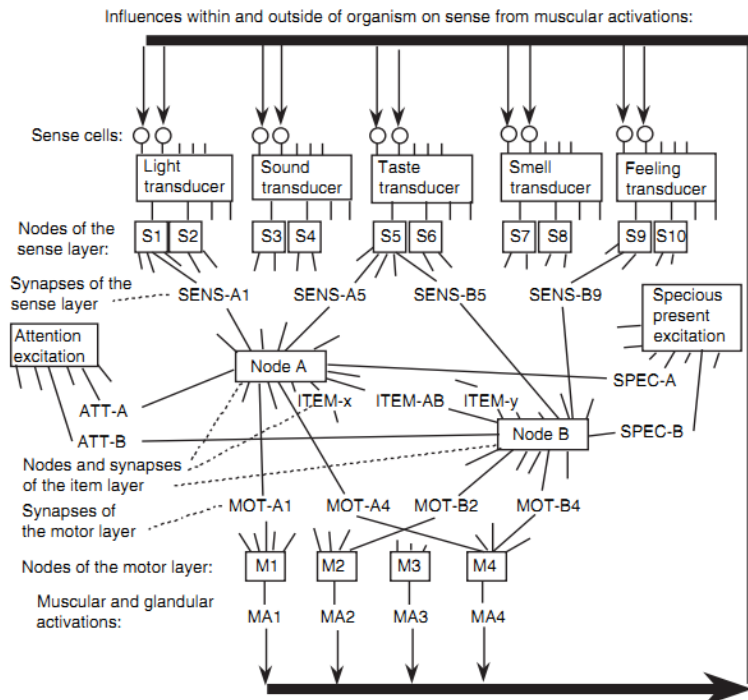


图 5-1 突触状态理论的基本框架^[140]

Fig.5-1 The basic framework of the synapse-state theory^[140]

2007 年图灵奖获得者 Peter Naur 在心理学研究基础上提出突触状态理论 (Synapse-State Theory)，用来描述问题解决的认知过程，他认为认知的方法才是实现让计算机具有思考能力的真正途径，其基本结构如图 5-1 所示^[140]。该理论认为理解的基本过程首先由感知细胞将其接收到的外界刺激传递给感知层神经元，感知层神经元的兴奋经过连通性突触将激活传递给大脑中相应的熟悉概念节点，这些概念节点用同样的方式将兴奋扩散传播给周围的节点，最后被激活的节点即为理解的结果。该理论是一种更完备的非迭代性扩散激活理论，遗憾的是它仅描述该理论的体系框架，却未实现相应的计算认知模型。

目前浅层语义分析中的主流方法是统计机器学习方法，如决策树、最大熵、支持向量机等。这种方法以‘谓词-论元’二元组作为分类样本，利用定义好的语言学特征将其转换成特征向量形式。然后再利用机器学习方法在数据集上对这些特征向量进行学习和预测。该方法以大规模语料库为基础，从已知语言现象的语言单位中抽取出一个相关的表层特征结构作为基本分析单元，通过计算特征结构之间的相似程度来识别新遇到的语言现象。基于统计机器学习方法主要存在以下问题：1) 基于统计机器学习方法以表层特征结构为基础，通过特征结构之间相似度的计算来判断样本中的语义关系，而并未考虑到上下

文信息，忽略了语言单位之间的内在联系。2) 基于统计机器学习方法主要句子为基本处理单元，因此无法体现句子之间的关系和篇章语义的连贯性，难以实现篇章级语义分析。3) 基于统计机器学习方法的系统性能依赖于训练语料库的质量和规模，因此通常受限于领域，系统移植性差。

本课题提出一种基于认知模型的中文浅层语义分析方法，能够弥补传统机器学习方法在这些方面的缺陷。该方法以语言理解的认知理论为基础，能够抽象地描述和表现人类认知的表征和过程，包括语义处理单元、信息加工的阶段、各个要素之间的相互关系以及阶段过渡过程中的行为等等。它从理解的本质出发来研究语义信息处理技术，能够有效地克服上述机器学习方法的缺点，是一种新型的、深层次的、跨领域的语义信息处理方法。该方法主要思想是从语言认知的本质出发，建立能够完整描述理解过程的计算认知模型，涵盖语言的表征、内部结构和处理以及结果的生成与描述等阶段，最终通过计算认知模型来实现中文语义分析。另外还提出一种适于认知模型的命题语义形式，作为语义计算的基本单元。与机器学习方法中的‘谓词-论元’二元组结构相比，这种命题形式中含有的语义信息更加丰富，不仅包含了谓词和论元之间的关系，而且还包含了同一谓词的多个论元之间的关系。另外，本文还充分利用自然语言处理技术为认知模型提供了有效的支持，借助目前较为成熟的一些自然语言处理中的技术和资源，对计算认知模型的应用起到必要的辅助作用。与基于机器学习的中文浅层语义分析方法相比，基于计算认知模型方法的主要缺点在于其对语言学特征的融合能力较差，而在认知模型的命题网络中这些语言学特征难以有效地表达，而机器学习方法能够利用丰富的语言学特征来指导浅层语义分析。

本章结构如下，第 5.3 小节定义了命题形式的语义知识表示形式，第 5.4 节中描述了基于计算认知模型的浅层语义分析方法的基本结构，第 5.5 节介绍了该方法的实现方法和具体细节，最后在第 5.6 节中对提出的基于计算认知模型的中文浅层语义分析方法进行了系统的评价。

5.3 命题语义表示形式的定义

首先我们提出一种基于命题逻辑的语义表示形式，用于描述本文所要分析的语义信息，根据每个句子的短语结构分析树构造潜在命题。然后我们利用改进的 CI 认知模型^[141-143]，将该命题语义表示形式引入到 CI 认知模型中，作为模型基本单元，即神经元。两个神经元之间通过突触连接起来，如果一个神经元被激活，那么这种激活状态会通过突触进行传播，使与该神经元通过突触相连接的其他神经元也被相应地激活，这种现象也被称做扩散激活。认知模型通过模拟这种扩散激活机制，使与上下文信息相符的候选命题的激活量逐渐加

强；相反，不相符的候选命题则会不断被抑制。当模型达到稳定状态时，根据各个命题节点的激活量来判断命题是否保留，最终获胜的命题所表达的语义信息即为认知模型语义分析的结果。首先我们给出命题形式语义表示的具体定义，该形式适合作为认知模型中的基本单元，符合计算认知模型的输入条件，并且其中蕴涵任务所需的语义信息，能实现对文本进行语义分析的目标。表 5-1 采用 BNF 范式对这种命题语义表示形式进行了定义。

表 5-1 命题形式语义表示的定义

Table 5-1 Definition of propositional semantic representation

No.	Derivation Rules
1.	<code><text> ::= {<proposition>}<proposition></code>
2.	<code><proposition> ::= <index> <content></code>
3.	<code><index> ::= <class><No></code>
4.	<code><class> ::= H O P T S E</code>
5.	<code><No> ::= {<digit>}<digit></code>
6.	<code><content> ::= <predicate>#<anchor>(<argument list>);</code>
7.	<code><predicate> ::= <verb></code>
8.	<code><anchor> ::= <position>-<offset></code>
9.	<code><argument list> ::= {<argument>,<argument>}</code>
10.	<code><argument> ::= <human> <object> <place> <time> <proposition></code>

从表 5-1 中可见，每个命题(proposition)由四个元素组成：索引(index)、锚点(anchor)、谓词(predicate)及论元序列(argument list)。索引是指为每个命题分配的一个标号，包括命题的类型、命题所在句子以及命题序号三个信息；锚点的作用是指谓词在自然语言句子中的位置，由位置和偏移两部分组成，分别代表谓词在句中的其实位置和谓词自身的长度；谓词就是指每个命题的核心动词；论元序列是指由命题谓词所辖的全部论元所组成的序列，命题共有五个类型，包括人物、实体、地点、状态和事件，其具体形式如表 5-2 所示。本文中允许命题嵌套，而且为了命题嵌套在表达上更为简洁，当一个命题作为另一个命题的论元时，我们直接采用被嵌套命题的索引来表示该命题论元。下面分别对各个命题类型的组成和特点进行介绍。

从表 5-2 中可见，人物和实体两个类型的命题形式较为固定，一个命题仅包含一个固定的谓词以及文本中一个表示人物或实体的名词短语作为论元；地点类型的命题略复杂一些，其谓词由一个方位词充当，论元为表示地点的名词短语。方位词在语言学中是指表示方向或位置的词，通常以“介词+名词+方位词”的形式出现在介词短语中，例如，“在椅子上”、“在电脑前”等。在汉语

中介词和方位词较为有限(详见附录 2); 状态和事件类型的命题比较复杂, 这两类命题在形式和内容上都十分灵活, 而且每个命题的内容都代表了句子中一组语义角色的分配形式。这两种命题中的谓词是文本中表达状态或者行为的动词, 论元的形式由谓词在上下文中的语义功能, 也就是动词的语义框架所决定。语义框架是一个描述性的框架, 由很多语义角色槽所组成^[13]。

表 5-2 命题的类型及示例

Table 5-2 Proposition categories and examples

Categories	Examples
Human	Hum(Tom/NNP);
Object	Obj(tree/NN);
Place	On(Obj(tree/NN));
State	Be(Hum(Tom/NNP), On(Obj(tree/NN)));
Event	Run(Hum(Tom/NNP), To(Obj(tree/NN)));

我们以中文句子“胡锦涛在东京参观了川崎环保城”为例来说明命题语义表示形式。该句子中仅包含一个动词‘参观’, 采用本文提出的知识表示方法, 该句子可以表示成表 5-3 中的命题形式。其中, H1、O2、O3 三个命题分别代表一个人物型命题和两个实体型命题, 这两种形式的命题是语义表示的基础; P4 是一个地点型命题, 说明实体 O2 所处的地点; 事件类型的命题 E5 是核心命题, 其中蕴涵了谓词‘参观’相关的语义角色信息, 包括 H1、O3 和 P4, 根据该谓词的语义框架, 可以从命题中还原出其所表达的语义: “H1(visitor) visit O3(thing_visited) at P4(place)”。可见该命题形式能够有效表达语义角色信息, 本文将浅层语义分析目标转化为构造这样的命题形式。

表 5-3 命题语义表示的示例

Table 5-3 Example of propositional semantic representation

Prop Index	Predicate-Anchor-Arguments
H1	Hum@1-3(胡锦涛/NNP);
O2	Obj@5-2(东京/NNP);
O3	Obj@10-5(川崎环保城/NNP);
P4	In@4-1(O2);
E5	Visit@7-2(H1, O3, P4);

5.4 基于认知模型的浅层语义分析基本方法

下面将介绍基于计算认知模型的浅层语义分析的整体思路和方法。上一小

节中提出了一种命题语义表示形式，这种命题形式表面上看是用一串字符(命题的谓词和论元)来表达一个概念。我们从另外一个角度出发，将命题视为一种函数，这个函数的参数可以是文本中的词，也可以是其他命题，函数的值表示该命题自身置信度，也就是命题的谓词在上下文中辖有该命题中全部论元的可能性。这样问题就简化为如何实现一个能够准确地反映命题的置信度函数。本文中我们利用改进的 CI 认知模型，即构造整合模型来实现这种函数映射。通过这种方式，浅层语义分析的问题可通过两个步骤实现，一是构造动词所有可能的命题形式，二是计算这些命题的函数值，找出其中函数值最大的。CI 模型是一种可计算认知模型，它结合了符号系统和连接主义的特点，已经应用在很多实际应用问题中，比如共指消解^[135]、半指导学习^[144]和故事理解^[139]中，并被证明是有效的。

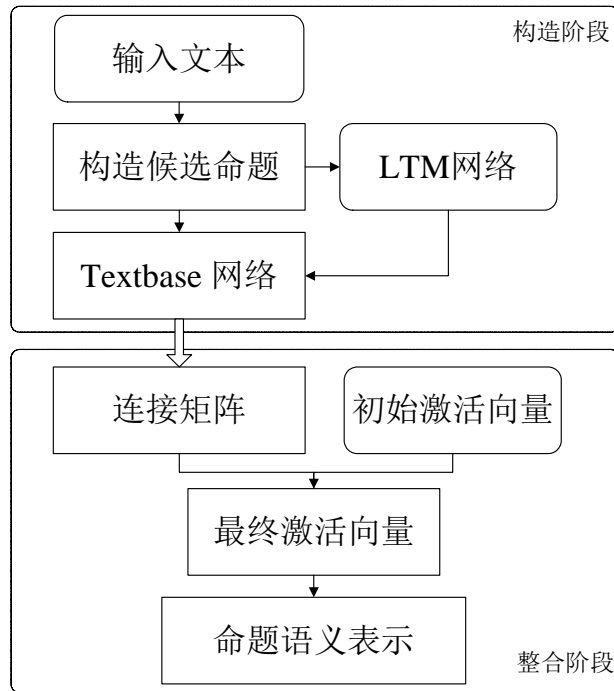


图 5-2 基于 CI 模型的浅层语义分析方法框架

Fig.5-2 The framework of the shallow semantic parsing method based on CI model.

CI 模型是目前为数不多的较为成熟和完整的计算认知模型，其计算模型以关联网(Associated Network)作为基本结构，网络中的节点表示一个神经元(Neuron)，边表示神经元之间传递激活的突触(Synapse)。基于该模型的浅层语义分析基本结构如图 5-2 所示。在模型初始阶段，模型中的一些节点会从文本这个外界刺激(External Stimulus)中获得初始激活量，然后激活通过网络以扩散激活的形式进行传播，直到网络中到达一个相对稳定的状态时过程结束。最终网络中每个节点的激活量将反映该节点对应命题的置信度。可见 CI 模型提供

了一种有效的命题函数计算方法，命题由网络中的节点表示，然后用节点最终的激活量表示该命题的函数值。

除了命题节点和初始激活量之外，基于 CI 模型的浅层语义分析还需要一个长时记忆 (Long-Term Memory, 简称 LTM) 网络，LTM 网络是一个预先定义好的关联网络，其作用相当于人脑中的记忆库，具有知识存储的功能。该网络对于构建 Textbase 网络有着至关重要的作用，Textbase 网络是扩散激活算法的载体，是 CI 模型的核心。尽管 CI 模型不同于一般机器学习方法，通常需要具有一定规模的训练语料，但是在应用到具体任务时，仍需要一些带有标记的已知数据作为指导，预先构造一个 LTM 网络。

CI 模型的计算过程可以分成两个阶段：构造阶段和整合阶段。构造阶段的目的是构造一个 Textbase 网络。首先从输入文本中提取出所有可能命题，并将这些命题作为 Textbase 网络中的部分节点，称之为上下文节点；Textbase 网络中的另外一部分节点来自于 LTM 网络，称之为 LTM 节点，这部分节点是通过以上下文节点命题作为种子，采取启发式的方法从 LTM 网络中抽取得到的。在整合阶段，通过迭代扩散激活方法不断调整 Textbase 网络中各节点的激活量以及节点之间的连接权值，最终使网络达到一致。其间符合上下文和知识库的命题节点将会逐渐被加强，而不相关的节点会不断被削弱。CI 模型收敛时即可获得 Textbase 网络上各个节点的激活量。根据网络中最终获胜的命题节点，结合该命题中谓词的语义框架，即可还原命题表达的语义信息，也就是浅层语义分析的结果。本文将在第 5.5 节中给出 CI 模型构造和整合两个阶段的具体方法以及扩散激活算法的详细描述。

5.5 认知模型的构造和整合

5.5.1 构造候选命题

首先我们根据输入文本构造 CI 模型中的上下文命题节点。与机器学习方法类似，我们首先利用本文第 3.2.1 节中所述的剪枝方法对句法树进行剪枝预处理。通过剪枝方法过滤掉不满足句法约束的名词短语，根据剪枝后句法树中的短语和词性等信息，构造相应的候选命题。首先对于人物和实体类型的命题，来源于句法树中的名词短语，其谓词是固定的，只要关注其中的论元即可。我们根据句法树中名词短语的中心词词性，结合 HowNet 中文语义词典对人物和实体两个类型的词进行识别，中心词的词性和长度作为该命题的锚点。

其次，对于地点类型的候选命题，其中通常会蕴涵 ArgM-LOC 类型的语义角色，通过分析发现中文中地点类型命题主要来源于介词短语。根据汉语语言学中的解释，一个中文介词短语由三个部分所组成：一个介词，一个名词短

语和一个方位词，例如：“在沙发上”，“在路两旁”。其中介词和名词短语通常是必要的，方位词部分有时可以省略，特别是名词短语是地名时，例如：“在北京”、“在美国”等等。汉语中的介词和方位词都较为有限，因此我们通过定义规则模板的方法从句子中提取出“介词+名词短语+方位词”和“介词+名词短语”形式的部分，根据抽取的部分来构造地点类型命题。

最后，对于状态和事件这两个复杂命题类型的构造是关键部分。这两类命题中的谓词是文本中的目标动词，其论元的形式由动词在上下文中的语义框架所决定，动词的语义框架可在 CPB 语料库中已有标注，该语料库在前文中已经多次介绍过，是目前中文语义分析领域中最常用的语料库，其中标注了 4854 个中文动词的 5796 个语义框架，其中有接近 85% 的动词(共 4090 个)仅有一个语义框架，对于这部分动词，其相应的命题构造较为容易，而对于包含多个语义框架的动词，其相应的命题的构造则要复杂一些。

首先，对于仅有一个语义框架的动词，构造它们对应的候选命题的方法就是在现有的命题中进行搜索，找出满足其语义框架中各个槽语义约束的命题。以表 5-3 中的示例来说明，如果句子“胡锦涛在东京参观了川崎环保城”中动词参观仅有一个框架：“<arg0:visitor> visit <arg1:place visited>”，那么这个动词相关的候选论元有“visit(H1(胡锦涛), O2(东京))”和“visit(H1(胡锦涛), O3(川崎环保城))”；对于有多个语义框架的动词，我们为每个语义框架定义了一个动词价(valence)，先利用上下文信息过滤掉一些不合适的语义框架。该动词价定义为一个动词管辖核心语义角色的个数，其值代表了一个动词的语义框架所能包含核心语义角色的基本能力。在本文中，核心论元主要对应人物和实体两类命题。因此，我们通过比较在句子中人物和实体命题的个数之和与该语义框架的动词价来过滤掉不满足上下文的语义框架。如果一个语义框架的动词价大于这个总和，则将该语义框架过滤掉。然后，对未被过滤掉的每个动词框架分别构造上下文命题，方法与构造仅有一个语义框架动词的候选命题相同。

5.5.2 构造 LTM 网络

LTM 网络作为一个含有语义信息的长时记忆网络，在基于 CI 模型的构造阶段中有着重要作用。LTM 网络的基本结构跟上文中所述 CI 模型构造阶段所要建立的 Textbase 网络结构相同，都是采用命题关联网络的形式，其主要功能是为 Textbase 网络隐式地提供一些相关的语义信息，因为其中表达了一个命题与其他命题之间的几种重要关系。本文中 LTM 网络的构造是采用一种启发式规则的方法在已有语义角色标注语料库基础上自动创建的。

首先，根据上一小节中所述方法将标注语料库中的文本转换成命题形式，作为 LTM 网络中的节点。不同之处在于在标注语料库中我们能够根据语料库

中的标记直接获得各种类型的命题；而在没有标记的测试语料上，只能构造出可能的候选命题。获得了 LTM 网络中的命题节点后，我们将构造网络中节点之间的连接权重，这里该权重是可以累加的，范围在-1 到 1 之间。假设是 A 、 B 是 LTM 网络中的两个命题节点， $w(A,B)$ 代表节点之间的连接权重，我们利用以下四条启发式规则，来分配两个节点之间的权重：

(1) 重叠，如果命题 A 和命题 B 中包含相同的论元命题，那么将命题 A 和 B 之间的权重 $w(A,B)$ 增加 λ_1 。

(2) 嵌套，如果命题 A 与命题 B 的某个论元相同，那么将 A 和 B 之间的权重 $w(A,B)$ 增加 λ_2 。

(3) 互斥，如果命题 A 与命题 B 具有相同的谓词和锚点，那么将 A 和 B 之间的权重 $w(A,B)$ 增加 λ_3 。

(4) 论元角色，如果 A 是命题 B 的核心语义角色，那么将 A 和 B 之间的权重 $w(A,B)$ 增加 λ_4 。

构造 LTM 网络的目的是建立能够反映命题之间语义关联强弱的模型。当 LTM 网络命题中的某一个被激活时，LTM 中节点之间保存的语义关系也将会被发掘，然后在新的上下文网络中发挥作用。

5.5.3 认知模型的构造阶段

构造阶段的目标是构造了一个 Textbase 网络，作为 CI 模型的主体，在该网络上进行扩散激活。Textbase 网络中的节点表示命题，命题节点之间连接边的权值表示命题之间的语义关联度，网络中的命题节点可以分成两类：上下文节点和 LTM 节点。上下文节点就是根据自然语言描述构造的候选命题，表示上下文语义信息；而 LTM 节点是从 LTM 网络中抽取出的，表示记忆网络中的语义信息。每个上下文命题节点代表一个候选命题，其构造方法见本文 5.5.1 小节。LTM 节点是以上下文命题节点作为种子，通过一种基于概率的选择方法从 LTM 网络中抽取出来的。在基于概率的节点选择过程中，与种子相关程度越大的 LTM 节点被选中的概率越大。假设在 LTM 网络中的种子节点 i 与 n 个节点相邻，则在这 n 个节点中节点 j 被选择的概率可表示为公式(5-1)。

$$p(j|i) = \frac{w(i,j)}{\sum_{k=1}^n w(i,k)}, \quad 1 \leq j \leq n \quad (5-1)$$

式中 $w(i,j)$ ——种子节点 i 和 LTM 节点 j 之间的连接权重。

各 LTM 节点的抽取是相互独立的，一个种子节点可以获得多个 LTM 节点，如果抽取到的 LTM 节点与某个上下文节点一致，则将这两个节点合并成

一个，并且将种子节点与该节点连接边的权值叠加。抽取出全部上下文节点与 LTM 节点之后，我们又定义了三条规则，与之前定义的 LTM 网络中的四条规则：重叠、嵌套、互斥和论元角色相结合，来为 Textbase 网络分配权重。

(1) 继承，如果命题 B 是从 LTM 网络中提取的节点，命题 A 是提取 B 所使用的种子，那么将在 LTM 网络中命题 A 和 B 之间的权重 $w_{LTM}(A, B)$ 叠加到 Textbase 网络中命题 A 和 B 之间的权重 $w(A, B)$ 上。

(2) 路径，如果命题 A 和 B 都是上下文节点， A 是 LTM 节点 C 的种子， AC 和 AB 之间的路径相同，则 $w(A, B)$ 和 $w(A, C)$ 都增加 λ_5 。

(3) 无关联，如果命题 A 和 B 之间上述 6 条规则均不适用，则将 A 和 B 之间的权重 $w(A, B)$ 的值分配为 0。

综上所述，本文提出的基于 CI 模型的浅层语义分析中共有五个参数： λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 ，后文中将通过实验的方法在独立开发集合上调整这些参数的取值。整体上讲，Textbase 网络的构造过程由上下文节点、LTM 节点以及它们之间的连接权值构成。在构造阶段，网络中包含冗余甚至可能含有冲突；在后续的整合阶段，网络将被收敛成较为一致的结构，然后从中获得包含正确语义信息的上下文命题节点。

5.5.4 认知模型的整合阶段

整合阶段的目的是通过扩算激活算法使 Textbase 网络达到一致的收敛状态。首先需要对 Textbase 网络进行数值化，本文采用无向图的连接矩阵 (Connection Matrix, 简称 CM) 表示方法来将 Textbase 网络抽象成为一个 n 阶的矩阵。另外网络中的每个命题节点都附带一个实数变量，表达该命题节点的当前激活量。在原始的 CI 模型中，扩散激活是通过不断地用激活向量后乘以 CM 矩阵的方式，直到各个节点的激活量不再发生变化，即网络达到稳定状态为止，本文中对其进行了如下扩展。

假设 Textbase 网络中共有 n 个节点，也就是文本中共有 n 个候选语义命题；在欧几里德空间 R^n 中，实数 a_i 代表第 i 个节点的激活量，实数 w_{ij} 代表节点 i 和节点 j 之间的连接权重。那么，整个网络的激活量可以表示成一个激活向量 $\mathbf{A} = (a_1, a_2, \dots, a_n)$ 。网络中所有连接边的权重可以表示成一个 $n \times n$ 的连接矩阵 $\mathbf{CM}_{n \times n}$ ，见公式(5-2)。

$$\mathbf{CM}_{n \times n} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nn} \end{bmatrix} \quad (5-2)$$

当 i 等于 j 时, w_{ij} 代表节点与其自身的连接权重, 其值设定为 1。换句话说, 也就是公式(5-2)中连接矩阵 $\mathbf{CM}_{n \times n}$ 主对角线上的值全部为 1。然后我们根据人脑神经元的激活方式, 在 CI 模型中引入了一种激活量衰减机制用来调整网络的平衡点^[142]。我们定义了一个衰减函数 $D(t)$ 来描述激活量随着迭代次数的增加而逐渐衰减的情况, 如公式(5-3)所示。

$$D(t) = 1 - e^{-dt}, \quad d > 0, \quad t \geq 1 \quad (5-3)$$

式中 t —— 迭代次数;

d —— 激活量衰减速率。

衰减函数 $D(t)$ 的值域为 $[0,1)$, 随着迭代次数的增加, 其值逐渐趋向于 1。函数 $D(t)$ 的一阶导数为 de^{-dt} , 其值大于 0, 这使得了随着迭代次数的增加, 激活量的衰减也越来越大; 而其二阶导数的值小于 0, 这使得了随着迭代次数的增加, 激活量衰减的幅度越来越小。参数 d 用来控制激活量随着迭代次数增加而衰减的速率。在每次迭代过程中, 新的激活量可通过公式(5-4)计算出。

$$a_i(t+1) = \sum_{j=1}^n a_j(t) \cdot \mathbf{CM}_{ij} - a_i(t) \cdot D(t) \quad 0 \leq i, j \leq n \quad (5-4)$$

式中 t —— 迭代次数;

$a_i(t)$ —— 第 i 个节点在第 t 次迭代时的激活量。

根据公式(5-3)和公式(5-4)可见, 初始迭代时激活量的衰减较快, 随着迭代次数的增加衰减越来越慢, 衰减逐渐趋于平稳。从整个网络上来看, 网络中节点激活量的迭代方法可以描述成公式(5-5)。

$$\mathbf{A}(t+1) = \mathbf{A}(t) \cdot (\mathbf{CM}_{n \times n} - D(t) \cdot \mathbf{E}_n) \quad (5-5)$$

式中 $\mathbf{A}(t)$ —— 第 t 次迭代时的网络整体激活向量。

然后, 我们定义了三个函数用来描述 Textbase 网络扩散激活过程, 见公式(5-6)、公式(5-7)和公式(5-8)。其中函数 f_1 代表网络中激活量的迭代函数, 函数 f_2 是为了确保当前迭代时的激活量非负, 函数 f_3 用来归一化所有激活量的总和为 1。函数 f_2 和 f_3 的作用主要是确保扩散激活算法的收敛性。

$$f_1(\mathbf{x}) = \mathbf{x} \cdot (\mathbf{CM}_{n \times n} - D(t) \cdot \mathbf{E}_n) \quad (5-6)$$

$$f_2(\mathbf{x}) = \begin{pmatrix} \max(x_1, 0) \\ \max(x_2, 0) \\ 7 \\ \max(x_n, 0) \end{pmatrix} \quad (5-7)$$

$$f_3(\mathbf{x}) = \frac{\mathbf{x}}{\sum_{i=1}^n |x_i|} \quad (5-8)$$

为了简化描述，我们定义了一个复合函数 $F(x) = f_3(f_2(f_1(x)))$ ，其中变量 x 表示激活向量，其取值范围如公式(5-9)所示。

$$D = \left\{ x \in R^n \mid \sum_{i=1}^n x_i = 1 \right\} \quad (5-9)$$

完成上述定义后，我们给出扩散激活算法的详细描述，如图 5-3 所示。

```

输入：初始激活向量A(0)
输出：最终激活向量A(last)
1: BEGIN
2:     CT = 10-3, //convergence threshold
3:     t = 1; //iteration numbers
4:     do
5:         A(t) = F(A(t-1));
6:         t = t + 1;
7:     while ||A(t-1) - A(t-2)|| > CT
8:     return A(t);
9: END
    
```

图 5-3 扩散激活算法描述

Fig.5-3 The description of the spreading activation algorithm

从图 5-3 中可见，在 Textbase 网络建立好之后，扩散激活算法有两部分输入：初始激活向量 $A(0)$ 和收敛阈值 CT ，整个算法就是不断通过迭代地计算网络中每个节点的激活量，使得激活量向与初始激活节点连接密切的节点不断积累，使得符合上下文信息的节点最终获胜。本文 Textbase 网络中，初始激活量主要是根据输入文本产生，所有 LTM 节点的初始激活量均为 0，所有上下文节点的初始激活均设为 $1/M$ ，其中 M 为网络中上下文节点的总数。最后，当网络中所有节点的激活量的变化幅度小于设定收敛阈值时，网络达到稳定状态，各个节点的激活量即为最终激活量。我们将所有命题化根据其锚点划

分成 N 个集合，并且保留每个集合中最终激活量最大的候选命题作为获胜命题。最后根据获胜命题并结合该命题中谓词的语义框架就可以还原出命题对应的语义角色信息，以表 5-3 中的例子来说明，句中动词“参观”的框架为“<arg0:visitor> visit <arg1:place visited>”，如果网络最终获胜的命题节点为“visit(H1(胡锦涛), O3(川崎环保城))”，那么句中动词“参观”对应的 Arg0 类型语义角色即为“胡锦涛”，Arg1 类型语义角色即为“川崎环保城”。

5.6 实验结果及分析

为验证本文所提出的基于 CI 认知模型的浅层语义分析方法的有效性，我们在中文浅层语义分析领域的通用标准语料库 CPB 上对该方法进行了评价。首先我们在带有正确句法标注 CPB 语料库上进行了相关评价，在评价过程中以目前主流的基于特征向量的浅层语义分析方法作为的基准方法(Baseline)，采用朴素贝叶斯(NB)模型和最大熵模型(ME)作为基准方法中的分类器，采用第 1.2.3.1 节中所述的标准特征：谓词、句法类型、次范畴框架、路径、位置、语态和中心词作为分类特征集合。我们将 CPB 语料库划分为三个部分，将其中的前 100 篇文档(从 chtb_001.fid 到 chtb_100.fid)作为测试语料，后 32 篇文档(从 chtb_900.fid 到 chtb_931.fid)作为开发语料，其余的 628 篇文档(从 chtb_101.fid 到 chtb_899.fid)作为训练语料用于 NB 和 ME 分类器的训练，以及 LTM 网络的构造。通过在开发集合上面按照步长参数调整，选取拟合数据效果较好的参数值。根据调整结果本文中 CI 认知模型的 5 个参数 λ_1 到 λ_5 取值分别为 0.2、0.4、-1.9、0.5 和 0.7，衰减函数中控制衰减速度的参数 d 取值为 0.5。我们采用的评价指标仍是精确率(P)，召回率(R)和 F 值(F)。实验结果如表 5-4 所示，其中 ArgN 代表谓词相关的核心论元，ArgM 代表修饰性论元。

表 5-4 在 CPB 数据集上的浅层语义分析结果

Table 5-4 Results of shallow semantic parsing on the CPB dataset

System		P	R	F
NB	ArgN	0.801	0.855	0.827
	ArgM	0.826	0.647	0.726
	Total	0.802	0.731	0.765
ME	ArgN	0.859	0.836	0.847
	ArgM	0.763	0.625	0.687
	Total	0.853	0.824	0.838
CI	ArgN	0.803	0.817	0.810
	ArgM	0.834	0.725	0.776
	Total	0.804	0.813	0.808

表 5-5 三个语料库的统计信息

Table 5-5 Statistics of the three datasets

Data Set	Doc Num	Sent Num	Predicate Num	ArgN Num	ArgM Num	Word Num	AVG Length
CPB	760	10,364	37,183	63,963	2,757	552,389	55.3
D1	100	705	2,181	2,739	82	13,049	18.5
D2	100	3,052	7,449	8,402	322	47,286	15.5

从实验结果中可以看出, 本文提出基于命题网络和 CI 模型的中文浅层语义分析方法在标注了句法分析结果的 CPB 数据集上取得了 80.8% 的 F 值, 比采用 NB 模型基准方法取得的 76.5% 的 F 值高出了 4.3%, 但是比最大熵模型的方法低了 3%, 可见该方法能够在噪声较小的数据集上取得与机器学习模型相当的实验结果, 但是与浅层语义分析领域中效果最好的机器学习模型还有一定差距。此外, 为了进一步验证该方法在实际应用中的效果, 我们在另外两个文景转换数据集上对该方法进行了测试, 分别是 100 篇伊索寓言中的文档和 100 篇网络下载的儿童故事, 标记为 D1 和 D2, 数据集的统计信息见表 5-5。

我们使用美国斯坦福大学开发的 Stanford Parser^[105]作为自动句法分析器来构造 D1 和 D2 的短语结构句法分析树, 并且人工标注了其中的浅层语义分析结果。考虑到这两个数据集的规模较小, 我们增加了 80 篇文档到之前的 CPB 训练集中用来训练 NB 和 ME 机器学习模型以及构造 LTM 网络, 剩余的 20 篇文档用于测试, 在 D1 和 D2 两个新数据集上的实验结果详见表 5-6。

表 5-6 在另外两个数据集上的浅层语义分析结果

Table 5-6 Results of shallow semantic parsing on D1 and D2 dataset

System		D1			D2		
		P	R	F	P	R	F
NB	ArgN	0.726	0.663	0.693	0.653	0.629	0.641
	ArgM	0.698	0.578	0.632	0.708	0.683	0.695
	Total	0.730	0.656	0.691	0.658	0.630	0.644
ME	ArgN	0.742	0.723	0.732	0.681	0.627	0.655
	ArgM	0.711	0.645	0.676	0.734	0.521	0.609
	Total	0.740	0.718	0.729	0.684	0.621	0.652
CI	ArgN	0.709	0.705	0.707	0.670	0.621	0.645
	ArgM	0.758	0.720	0.739	0.723	0.517	0.603
	Total	0.714	0.708	0.711	0.676	0.617	0.645

从表 5-6 中可见, 在自动标注句法分析带来较大噪声的数据集合 D1 和 D2 上, 本文提出的 CI 模型整体性能达到 0.711 和 0.645, 较 NB 基准方法分别高

出 2% 和 0.1%，但仍较 ME 方法仍低 1.8% 和 0.7%。该结果还反映了本文方法在 D1 数据集上提高幅度略高于 D2 数据集上的提高幅度，其原因是基于特征的方法对于句法分析的依赖性很强，对与其中的错误较为敏感，而本文方法对于句法错误的敏感度较低。D1 数据集中的句子的平均长度较 D2 数据集中的长，句式也更为复杂，因此 D1 数据集的自动句法分析准确率要低于 D2 数据集。句法分析是目前基于机器学习方法的中文浅层语义分析中的主要瓶颈，但在本文提出的方法中，句法信息仅作为修改连接权重的规则之一，对于整体效果的影响要低于基于机器学习方法。

接下来，我们观察了本文提出的认知方法随着 LTM 网络规模增加性能的变化情况，以及基于机器学习的基准方法中随着训练语料规模的增加结果性能的提升，来进一步验证 LTM 网络在实际应用中的效果，并发掘认知方法和传统机器学习方法的不同。图 5-4 给出了在正确句法分析条件下浅层语义分析系统 F 值随 LTM 网络规模变化的示意图。

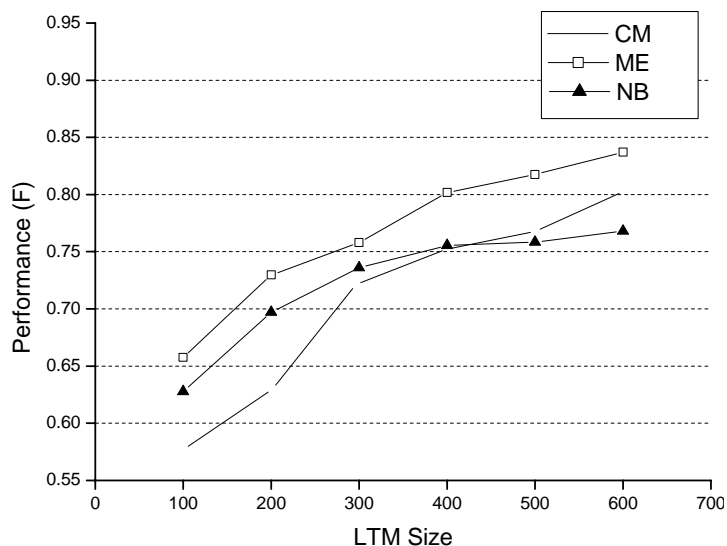


图 5-4 系统性能与 LTM 网络规模的关系示意图

Fig.5-4 Performance versus size of LTM network

从图中可见，随着 LTM 网络规模的增大整体 F 值呈上升趋势，然而当规模到达一定程度时，上升趋势逐渐减缓，基于机器学习的方法的性能，也随着训练语料规模的增加结果有明显的提升。但二者的不同之处在于，机器学习方法在语料较少的初始阶段效果较好，而认知方法中在初始阶段效果比较差；但随着训练语料到达一定规模，本实验中为 400 篇文档左右，机器学习方法效果提升变得缓慢，而认知方法效果的提升相对较快。可见本实验中训练语料规模的增大对于认知方法的帮助较大。总体上看，虽然基于认知模型的方法取得了与一般机器学习方法相当的效果，但仍与领域内较好的机器学习方法有差距。

与传统基于机器学习的中文浅层语义分析方法相比,本文方法的优点在于对于句法分析的依赖性较小,对其中的错误容忍度也比较高。由于在基本单元和模型上都与传统方法有很大差别,本文方法对句法框架较为单一的动词效果较好,原因在于对此类命题构造的候选命题,也就是 Textbase 网络中的节点是根据动词框架产生的,复杂形势和数量较多的命题都会明显地增加系统的复杂程度,在后续的工作中,可尝试采用对动词语义框架聚类的方法针对这一现象进行改进。

5.7 本章小结

本文提出一种新颖的基于命题网络和构造整合认知模型的中文浅层语义分析方法,该方法在语义的表达和计算方式上与目前现有方法均有很大不同。首先,我们定义了一种基于命题形式的语义知识表示形式用来支持本文所采用的扩展 CI 认知模型。然后,我们根据自然语言文本构造出候选命题和 LTM 网络,通过一种关联网络的形式来表达命题及命题之间的关系。与传统的基于特征向量的机器学习方法相比,它能够更加有效地在计算过程中引入丰富的上下文以及常识知识,通过模拟人脑中神经元之间信息传递的扩散激活算法,使符合上下文和常识的命题节点会被不断被加强;相反地,与上下文或常识信息相悖的命题不断被削弱,最后网络到达稳定状态时,分析过程结束。通过分析最终获胜的命题,结合该命题中谓词的语义框架,得出浅层语义分析的结果。实验结果表明基于认知模型的方法能够取得与一般机器学习方法相当的整体效果,但与目前领域内最好的机器学习方法仍有一定的差距。另外,认知模型还处于一个初级阶段,本文充分借助目前较为成熟的自然语言处理中的技术和资源为认知模型提供了有效支持,同时认知模型也弥补了自然语言处理技术在篇章语义信息处理上的不足。

本文提出的这种基于命题网络和认知模型的中文浅层语义分析方法中,首次将由目标谓词及其所有论元组成的命题形式作为语义分析和计算的基本单元,更加合理和有效地表达了多个论元相互之间的语义关联,为浅层语义分析提供了更丰富和有效的上下文信息。因此,本文提出的基于认知模型的中文浅层语义分析方法的另一个重要意义在于为浅层语义分析提供了一种新的计算单元,并通过实验验证了基于这种命题单元的浅层语义分析的可行性。

结 论

本文针对面向文景转换任务的中文浅层语义分析方法中存在的若干问题展开全面深入的研究。具体来说,本文取得的主要成果和创新点如下:

(1) 提出一种基于 ART 网络的无指导中文名词短语共指消解方法,有效地解决了目前聚类共指消解中输出类别数目难以确定的难题。该方法通过调整 ART 网络中的参数动态地控制聚类数量,在聚类算法中还采用了一种基于信息增益率的特征选择方法,减少了区分度较弱特征给聚类所带来的干扰。在 ACE 标准语料库上的结果表明,共指消解识别准确率高于一类聚类算法。

(2) 提出一种基于多重句法特征的中文浅层语义分析方法,从语言学特征方面对中文浅层语义分析进行改进。提出将短语结构句法特征和依存句法特征两种类型的句法特征进行融合,并且在这两个特征集合基础上,提出一种基于统计的组合特征选择方法,能够根据各个特征在语料库中的分布,发掘出适于各分类阶段的组合句法特征。实验表明,利用短语结构句法特征、依存句法特征以及在两者基础上构造的组合句法特征所组成的多重特征集合能显著地提高浅层语义分析的性能,在正确和自动句法分析的 CPB 语料库上分别取得了 91.76% 和 66.61% 的整体 F 值。

(3) 提出了一种基于组合分类模型的中文浅层语义分析方法,从机器学习方法的角度进一步对中文浅层语义分析进行完善。首先在前面提出的多重句法特征基础上,采用五种机器学习方法构造了五个基本分类模型作为组合模型中的元素。然后通过选通系统将五个基本分类模型有机地结合到一起,通过调整选通函数中的参数协调各个基本分类模型的结果,控制组合模型的输出。最后用 EM 算法对选通函数中的参数进行学习,在 CPB 语料库上进行了相关训练和测试,结果表明该方法能进一步提升中文语义角色分析的效果,在正确和自动句法分析的 CPB 语料库上分别达到 92.41% 和 67.85% 的整体 F 值。

(4) 提出了基于认知模型的中文浅层语义分析方法,以认知理论为基本依据,通过模拟人类的语言认知过程,从根本上研究浅层语义分析。首先我们设计了一种适于认知模型和文景转换的命题语义表示形式,这种命题形式能够简洁高效地自然语言中蕴涵表达空间语义信息,我们将该命题形式作为认知模型中的基本单元。通过在认知模型上模拟人脑中神经元的扩散激活机制,使得符合上下文约束的命题节点不断被加强,不符合的逐渐被削弱。最终当网络达到稳定状态时,根据被激活的节点即可还原出动词相关的浅层语义分析结果。

尽管取得了上述这些阶段性的研究成果，本文的浅层语义分析研究中仍存在许多有待进一步探索和改进的工作，主要包括一下几个方面：

（1）在共指消解方面，建立针对名词性短语和代词性短语之间的单独共指消解问题模块，结合机器学习方法与语言学规则，提升这部分的识别正确率；其次是发掘新的高效聚类特征，充分利用上下文信息，优化聚类模型并探索更加有效的聚类方法。

（2）在特征选择方面，引入更加丰富的结构化特征，挖掘更深层次的句法信息，结合基于核方法的浅层语义分析的优点，充分利用多种形式句法分析中句法树的结构化信息。另外还要研究特征模版的构造，即如何从在大规模的特征中高效精确地组合出最优的特征模版。

（3）在机器学习方面，研究基于无指导或半指导的中文浅层语义分析方法，在互联网时代，网络中拥有海量的非结果化或半结构化信息和数据，因此研究如何通过无指导或半指导的方法，充分地利用这些信息和数据来提高浅层语义分析的性能，是我们需要研究的课题。

（4）深入研究计算认知模型，探索不依赖于句法分析的中文浅层语义分析方法。目前浅层语义分析十分依赖句法分析，而句法分析以及基于句法分析结果的剪枝是造成中文浅层语义分析中错误的主要原因。认知理论中已有实验证明在人脑中句法分析和语义分析阶段是并行的。因此，深入研究认知理论和探索计算认知模型是解决语义分析问题的根本方法。

参考文献

- [1] Christensen J, Mausam, Soderland S, et al. Semantic Role Labeling for Open Information Extraction[C]. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL) 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, 2010:52-60
- [2] Shen D, Lapata M. Using Semantic Roles to Improve Question Answering[C]. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL), 2007:12-21
- [3] Bilotti M W, Ogilvie P, Callan J, et al. Structured Retrieval for Question Answering[C]. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 2007:351-358
- [4] Fillmore C J, Baker C F. Frame Semantics for Text Understanding[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics (NACCL) 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, 2001
- [5] Liu D, Gildea. Semantic Role Features for Machine Translation[C]. Proceedings of the 23rd International Conference on Computational Linguistics (COLING), 2010:716-724
- [6] Adorni G, Manzo M D, Giunchiglia F. Natural Language Driven Image Generation[C]. Proceedings of the International Conference on Computational Linguistics(COLING). Stanford, California, 1984:495-500
- [7] Coyne B, Sproat R. WordsEye: An Automatic Text-to-Scene Conversion System[C]. Proceedings of the Annual Conference on Computer Graphics, Los Angeles, USA, 2001:487-496
- [8] Gruber J S. Studies in Lexical Relations[D]. Ph.D. Dissertation, Massachusetts Institute of Technology (MIT), 1965
- [9] Fillmore C J. The Case for Case. In Bach and Harms (eds.): Universals in Linguistic Theory[M]. New York:Holt, Rinehart and Winston, 1968:1-88
- [10] Chomsky N. Lectures on Government and Binding[M]. Dordrecht:Foris, 1981
- [11] Gildea D, Palmer M. The Necessity of Syntactic Parsing for Predicate Argument Recognition[C]. Proceedings of the Association for Computational Linguistics(ACL), 2002:239-246

- [12] Punyakanok V, Roth S, Yih W. The Necessity of Syntactic Parsing for Semantic Role Labeling[C]. Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI). 2005: 1117-1123
- [13] Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Project[J]. Proceedings of COLING-ACL'98 Joint Conference, 1998:86-90
- [14] Kipper K, Korhonen A, Ryant N, et al. A Large-scale Classification of English Verbs[J]. Language Resources and Evaluation, 2008, (42):21-40
- [15] Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles[J]. Computational Linguistics, 2005, 31(1):71-106
- [16] Meyers A, Reeves R, Macleod C, et al. Annotating Noun Argument Structure for NomBank[C]. Proceedings of the International Conference on Language Resources and Evaluation(LREC), 2004:803-806
- [17] 刘开瑛, 由丽萍. 汉语框架语义知识库构建工程[C]. 中文信息处理前沿进展——中国中文信息学会成立二十五周年学术会议论文集, 2006, (11):64-71
- [18] 周强. 汉语句法语义链接知识库开发[C]. 第八届汉语词汇语义学研讨会. 2007
- [19] Xue N, Palmer M. Adding Semantic Roles to the Chinese Treebank[J]. Natural Language Engineering, 2009, 15(1):143-172
- [20] Xue N, Palmer M. Annotating the Propositions in the Penn Chinese Treebank[C]. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, 2003:47-54
- [21] Xue N, Palmer M. Calibrating Features for Semantic Role Labeling[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2004:88-94
- [22] Toutanova K, Haghghi A, Manning C M. A Global Joint Model for Semantic Role Labeling[J]. Computational Linguistics, 2008, 34(2):161-191
- [23] Mitchell T M. Machine Learning[M]. New York: McGraw-Hill, 1997
- [24] 李军辉. 中文句法语义分析及其联合学习机制研究[D]. 苏州: 苏州大学, 2010, 27
- [25] Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles[J]. Computational Linguistics, 2002, 28(3):245-288
- [26] Chen J, Rambow O. Use of Deep Linguistic Features for the Recognition and Labeling of Semantic Arguments[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2004, 41-48
- [27] Surdeanu M, Harabagiu S, Williams J, et al. Using Predicate-argument Structures for Information Extraction[C]. In Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL), 2003:8-15

- [28] Pradhan S, Hacioglu K, Krugler V, et al. Support vector learning for semantic argument classification[J]. Machine Learning Journal, 2005, 60(3):11-39
- [29] Xue N. Labeling Chinese Predicates with Semantic Roles[J]. Computational Linguistics, 2008, 34(2):225-255
- [30] Ding W, Chang B. Improving Chinese Semantic Role Classification with Hierarchical Feature Selection Strategy[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2008:324-323
- [31] Zhao H, Chen W L, Kit C. Semantic Dependency Parsing of NomBank and PropBank: An Efficient Integrated Approach via a Large-scale Feature Selection[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), Singapore, 2009:30-39
- [32] Boxwell S A, Dennis M D, Brew C. Brutus: A Semantic Role Labeling System Incorporating CCG, CFG, and Dependency Features[C]. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Singapore, 2009:37-45
- [33] Swier R S, Stevenson S. Unsupervised Semantic Role Labelling[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2004:95-102
- [34] Fuerstenau H and Lapata M. Graph Alignment for Semi-Supervised Semantic Role Labeling[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2009:11-20
- [35] Abend, O, Rappoport, A. Fully Unsupervised Core-Adjunct Argument Classification[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics(ACL), 2010:226-236
- [36] Swier R S, Stevenson S. Exploiting a Verb Lexicon in Automatic Semantic Role Labelling[C]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005:883-890
- [37] Grenager T, Manning C D. Unsupervised Discovery of a Statistical Verb Lexicon[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2006:1-8
- [38] Poon H, Domingos P. Unsupervised Semantic Parsing[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2009:1-10
- [39] Abend O, Reichart R, Rappoport A. Unsupervised Argument Identification for Semantic Role Labeling[C]. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009:28-36

- [40] Deschacht K, Moens M F. Semi-supervised Semantic Role Labeling Using the Latent Words Language Model[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2009:21-29
- [41] Moschitti A. A Study on Convolution Kernels for Shallow Statistic Parsing[C]. Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics(ACL), 2004:335-342
- [42] Moschitti A, Basili R. Verb subcategorization kernels for automatic semantic labeling[C]. Proceedings of the ACL-05 Workshop on Deep Lexical Acquisition, 2005:10-17
- [43] Collins M, Duffy N. Convolution kernels for natural language[J]. In Advances in Neural Information Processing Systems(NIPS), 2001, 14(1):625--632.
- [44] Che W X, Zhang M, Aw A T, et al. Using a Hybrid Convolution Tree Kernel for Semantic Role Labeling[J]. ACM Transactions on Asian Language Information Processing, 2008, 7(4):1-23
- [45] Zhang M, Che W X, Aw A T, et al. A Grammar-driven Convolution Tree Kernel for Semantic Role Classification[C]. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), 2007:200-207
- [46] Sun H, Jurafsky D. Shallow Semantic Parsing of Chinese[C]. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL), 2004:249-256
- [47] Xue N, Palmer M. Automatic Semantic Role Labeling for Chinese Verbs[C]. In Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI), 2005:1160-1165.
- [48] Sun W. Improving Chinese Semantic Role Labeling with Rich Syntactic Features[C]. Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (ACL), 2010:168-172
- [49] 安强强, 张蕾. 基于依存树的中文语义角色标注[J]. 计算机工程. 2010, 4:161-164
- [50] 董振东, 董强, 郝长伶. 知网的理论发现[J]. 中文信息学报, 2007,(21)4:3-9
- [51] 王步康, 王红玲, 袁晓虹等. 基于依存句法分析的中文语义角色标注[J]. 中文信息学报, 2010, 24(1):25-30
- [52] 李军辉, 周国栋, 朱巧明等. 中文名词性谓词语义角色标注研究[J]. 软件学报, 2010
- [53] 丁伟伟, 常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报, 2009, 23(5):53-62
- [54] 邵艳秋, 穗志方, 吴云芳. 基于词汇语义特征的中文语义角色标注研究[J].

中文信息学报, 2009, 6:3-10

- [55] 车万翔. 基于核方法的语义角色标注研究[D]. 哈尔滨: 哈尔滨工业大学, 2008.
- [56] 李济洪. 汉语框架语义角色的自动标注技术研究[D]. 太原: 山西大学, 2010.
- [57] 王红玲. 基于特征向量的中英文语义角色标注研究[D]. 苏州: 苏州大学, 2009.
- [58] 刘怀军, 车万翔, 刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, 21(1):80-84
- [59] 丁金涛, 王红玲, 周国栋等. 语义角色标注中特征优化组合研究[J]. 计算机应用与软件, 2009, 5
- [60] 刘娜. 中文特殊句式的语义角色标注[D]. 北京: 北京邮电大学, 2009.
- [61] Johansson R, Williams D, Berglund A, et al. Carsim: A System to Visualize Written Road Accident Reports as Animated 3D Scenes[C]. In Hirst, G. and Nirenburg, S., editors, ACL2004: Second Workshop on Text Meaning and Interpretation, Barcelona, Spain, 2004, 57-64
- [62] 陆汝钤, 张松懋. 从故事到动画片——全过程计算机辅助动画自动生成[J]. 自动化学报, 2002, 28(3):322-348
- [63] Carreras X, Màrquez L. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2004
- [64] Carreras X, Màrquez L. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2005
- [65] Surdeanu M, Johansson R, Meyers A, et al. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2008
- [66] Hajic J, Ciaramita M, Johansson R, et al. The CoNLL-2009 Shared Task: Joint Parsing of Syntactic and Semantic Dependencies in Multiple Languages[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2009
- [67] Litkowski K C. Senseval-3 task: Automatic Labeling of Semantic Roles[C]. Proceedings of the 4th International Workshop on Semantic Evaluations (SensEval), 2004:99-104
- [68] Baker C, Ellsworth M, Erk K. SemEval-2007 Task 19: Frame semantic Structure Extraction[C]. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), ACL, 2007:99-104

- [69] Zhou Q. SemEval-2010 task 11: Event Detection in Chinese News Sentences[C]. Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), ACL, 2010:86
- [70] Hacioglu K. Semantic Role Labeling Using Dependency Trees[C]. Proceedings of the International Conference on Computational Linguistics (COLING), 2004:1273-1276
- [71] Koomen P, Punyakanok V, Roth D, et al. Generalized Inference with Multiple Semantic Role Labeling Systems[C]. Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL), 2005:181-184
- [72] Roth D. Learning to Resolve Natural Language Ambiguities: A Unified Approach[C]. Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI) and Tenth Conference on Innovative Applications of Artificial Intelligence (IAAI), Menlo Park, CA, USA, 1998:806-813
- [73] Johansson R, Nugues P. Dependency-based Syntactic-Semantic Analysis with PropBank and NomBank[C]. Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL), 2008:183-187
- [74] Crammer K, Dekel O, Keshet J, et al. Online Passive-aggressive Algorithms[J]. Journal of Machine Learning Research, 2006, (7):551-585
- [75] Lin C J, Weng R C, Keerthi S S. Trust Region Newton Method for Large-scale Logistic Regression[J]. Journal of Machine Learning Research, 2008, (9):627-650
- [76] Zhang Y, Wang R, Uszkoreit H. Hybrid Learning of Dependency Structures from Heterogeneous Linguistic Resources[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2008:198-202
- [77] McDonald R, Pereira F, Ribarov K, et al. Non-Projective Dependency Parsing using Spanning Tree Algorithms[C]. In Proceedings of HLT-EMNLP, Vancouver, Canada, 2005:523-530,
- [78] Nivre J, Nilsson J, Hall J, et al. Maltparser: A Language-independent System for Data-driven Dependency Parsing[J]. Natural Language Engineering, 2007, 13(1):1-41.
- [79] Zhao H, Chen W, Kazama J, et al. Multilingual Dependency Learning: Exploiting Rich Features for Tagging Syntactic and Semantic Dependencies[C]. Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL), 2009:61-66
- [80] Che W, Li Z, Li Y, et al. Multilingual Dependency-based Syntactic and Semantic Parsing[C]. Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL), 2009:49-54
- [81] Ngai G, Wu D, Carpuat M, et al. Semantic Role Labeling with Boosting, SVMs,

- Maximum Entropy, SNOW, and Decision Lists[C]. Proceedings of SensEval-3, ACL, 2004
- [82] Johansson R, Nugues P. Using WordNet to Extend FrameNet Coverage[C]. Proceedings of SemEval-2007, ACL, 2007
- [83] 王厚峰. 汉语篇章的指代消解浅论[J]. 语言文字应用, 2004, 4: 113-119
- [84] 孔芳, 周国栋, 朱巧明等. 指代消解综述[J]. 计算机工程, 2010, 36(8): 33-36
- [85] Cardie C, Wagstaff K. Noun Phrase Coreference as Clustering[C]. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Maryland, USA, 1999: 82-89
- [86] Bergler S, Witte R, Khalife M, et al. Using Knowledge-Poor Coreference Resolution for Text Summarization[C]. Proceedings of the HLT-NAACL Workshop on Text Summarization. Edmonton, Canada, 2003: 85-92
- [87] Bean D, Riloff E. Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution[C]. Proceedings of HLT-NAACL, Boston, USA. 2004: 297-304
- [88] Wang C S, Ngai G. A Clustering Approach for Unsupervised Chinese Coreference Resolution[C]. Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 2006: 40-46
- [89] 周俊生, 黄书剑. 一种基于图划分的无监督汉语指代消解算法[J]. 中文信息学报, 2007, 21(2): 77-82
- [90] Xu R, Wunsch D. Survey of Clustering Algorithm[J]. IEEE Transactions on Neural Networks, 2005: 645-678
- [91] Quinlan J R. Induction of Decision Trees[J]. Machine Learning, 1986, 1: 81-106
- [92] Quinlan J R. C4.5: Programs for Machine Learning[M]. San Mateo, CA: Morgan Kaufmann, 1993
- [93] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Entities Version 5.5[S]. University of Pennsylvania, Philadelphia, USA. 2005.
- [94] Hsu C W, Lin C J. A Comparison of Methods for Multi-class Support Vector Machines[J]. IEEE transactions on Neural Networks, 2002, 13(2): 415-425
- [95] Carpenter G A, Grossberg S. A Massively Parallel Architecture for a Self-organizing Neural Pattern Recognition Machine[J]. Computer Vision, Graphics, and Image Process, 1987, 37(1): 54-115
- [96] Vilain M, Burger J, Aberdeen J, et al. A Model Theoretic Coreference Scoring Scheme[C]. Proceedings of the 6th Message Understanding Conference, Morgan Kaufmann, San Francisco, 1995: 45-52
- [97] Joachims T. Making large-Scale SVM Learning Practical. Advances in Kernel

- Methods[J]. Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1998.
- [98] Luo X. On Coreference Resolution Performance Metrics[C]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, 2005:25-32
- [99] Carreras X, Marquez L. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2005:152-164
- [100] Pradhan S, Ward W, Martin J H. Towards Robust Semantic Role Labeling[J]. Computational Linguistics, 2008, 34(2):289-310
- [101] Pradhan S, Hacioglu K, Krugler V, et al. Support Vector Learning for Semantic Argument Classification[J]. Machine Learning Journal, 2005,60(3):11-39
- [102] Pradhan S, Waed W, Haciolgu K, et al. Shallow Semantic Parsing using Support Vector Machines[C]. Proceedings of HLT-NAACL, 2004:233-240
- [103] Marneffe M C, MacCartney B, Manning C D. Generating Typed Dependency Parses from Phrase Structure Parses[C]. Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2006:449-454
- [104] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. 2nd Edition. New York: John Wiley, 2001
- [105] Levy R, Manning C D. Is It Harder to Parse Chinese, or the Chinese Treebank[C]. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL), 2003:439-446
- [106] Florian R, Ittycheriah A, Jing H, et al. Named Entity Recognition through Classifier Combination[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2003:168-171
- [107] Lu X, Wang Y, Jain A K. Combining Classifiers for Face Recognition[C]. Proceedings of IEEE International Conference on Multimedia and Expo, 2003:13-16
- [108] Jordan M I, Xu L. Convergence Results for the EM Approach to Mixtures of Experts Architectures[J]. Neural Networks, 1995, (8):1409-1431
- [109] Xu L, Jordan M I. On Convergence Properties of the EM Algorithm for Guassian Mixtures[J]. Neural Computation, 1996, 8(1):129-151.
- [110] Cover T M, Hart P E. Nearest Neighbor Pattern Classification[J]. IEEE Transactions on Information Theory, 1967,13(1): 21-27
- [111] Hall P, Park B U, Samworth R J. Choice of Neighbor Order in Nearest-neighbor Classification[J]. Annals of Statistics, 2008,36(5):2135-2152
- [112] Breslow L A, Aha D W. Simplifying Decision Trees: a Survey[J]. Knowledge

- Engineering Review, 1997, 12(1):1-40
- [113] Breiman L, Friedman J H, Olshen R, et al. Classification and Regression Trees[M]. Monterey, CA:Wadsworth and Brooks, 1983
- [114] Quinlan J R. Learning Efficient Classification Procedures and Their Application to Chess End Games[J]. Machine Learning: An Artificial Intelligence Approach, Palo Alto, CA, 1983
- [115] Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain[J], Cornell Aeronautical Laboratory, Psychological Review, 1958, 65(6):386-408.
- [116] Gallant S I. Perceptron-based Learning Algorithms[J]. IEEE Transactions on Neural Networks, 1990,1(2):179-191.
- [117] Park K M, Rim H C. Maximum Entropy Based Semantic Role Labeling[C]. Proceedings of the Conference on Computational Natural Language Learning (CoNLL), 2005:209-212
- [118] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3):565-573
- [119] Shannon C E. A Mathematical Theory of Communication[J]. Bell System Technical Journal, 1948, (27):379-423,623-656
- [120] Guinasu S, Shenitzer, A. The Principle of Maximum Entropy[J]. The Mathematical Intelligencer, 1985, 7(1):42-48.
- [121] Zhang L. Maximum Entropy Modeling Toolkit for Python and C++[EB/OL]. http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html, 2006
- [122] Vapnik V, Kotz S. Estimation of Dependences Based on Empirical Data[M]. New York:Springer, 2006
- [123] Sun R. The Cambridge Handbook of Computational Psychology[M]. Cambridge University Press, 2008
- [124] Laird J E, Newwell A, Rosenbloom P S. Soar: An Architecture for General Intelligence[J]. Artificial Intelligence, 1987, 33(3):1-64
- [125] Laird J E, Hucka M, et al. An Analysis of Soar as an Integrated Architecture[J]. ACM SIGART Bulletin, 1991, 2(4):98-103
- [126] Benjamin D, Lonsdale D, Lyons D. A Cognitive Robotics Approach to Comprehending Human Language and Behaviors[C]. Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction, 2007:185-192
- [127] Anderson J R, Lebiere C. The Atomic Components of Thought[M]. Mahwah, NJ:Lawrence Erlbaum Associates, 1998
- [128] Anderson J R. Cognitive Psychology and Its Implications[M]. New York: WH Freeman and Company, 1999

- [129] McClelland J L. Connectionist Models and Psychological Evidence[J]. *Journal of Memory and Language*, 1988, 27:107-123
- [130] Rumelhart D E. The Architecture of Mind: A Connectionist Approach[J]. *Foundations of Cognitive Science*, Cambridge:Bradford Books, 1989:133-159
- [131] Miikkulainen R. Text and Discourse Understanding: The DISCERN System[J]. *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. New York: Marcel Dekker, 1999
- [132] Brown D P, Tepper J A, Powell H M. Connectionist Natural Language Parsing[J]. *Trends in Cognitive Sciences*, 2002
- [133] Scholtes J C. Neural Networks in Natural Language Processing and Information Retrieval[D]. Netherlands:Universiteit van Amsterdam, 1993
- [134] Moisl H. Artificial Neural Networks and Natural Language Processing[J]. *Encyclopedia of Library and Information Science*, 2003
- [135] Kintsch W. The Role of Knowledge in Discourse Comprehension: A Construction and Integration Model[J]. *Psychological Review*, 1988, 95:163-182
- [136] Collins A M, Loftus E F. A Spreading Activation Theory of Semantic Processing[J]. *Psychological Review*, 1975, 82(6):407-428
- [137] Kintsch W. Comprehension: A Paradigm for Cognition[M]. Cambridge University Press, 1998
- [138] Kapusuz E. Refining the Representational Basis of the Construction-Integration Model of Text Comprehension with Syntactic Cues[D]. Turkey: Middle East Technical University, 2001
- [139] Mueller E T. Story Understanding[J]. *Encyclopedia of Cognitive Science*, London: Macmillan Reference, 2002
- [140] Nuar P. Computing versus Human Thinking[J]. *Communications of the ACM*, 2007, 50(1):85-94
- [141] Kintsch W. The Role of Knowledge in Discourse Comprehension: a Construction and Integration Model[J]. *Psychological Review*, 1988, 95(2):163-182
- [142] Guha A, Rossi J P. Convergence of the Integration Dynamics of the Construction-Integration Model[J]. *Journal of Mathematical Psychology*, 2001, 45:355-369
- [143] Sanjose V, Abarca E V, Padilla O M. A Connectionist Extension to Kintsch's Construction-Integration Model[J]. *Discourse Processes*, 2006, 42(1):1-35
- [144] Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation[R]. Pittsburgh:Carnegie Mellon University, CMU-CALD-02-107, 2002.

附 录

附录 1: PropBank 语料库中的修饰性语义角色标记及其含义

语义角色标记	具体含义
ArgM-ADV	Adverbial, 状语
ArgM-BNF	Beneficiary, 受益者
ArgM-CND	Condition, 条件
ArgM-DIR	Direction, 方向
ArgM-DIS	Discourse Marker, 话语标记
ArgM-DGR	Degree, 程度
ArgM-EXT	Extent, 范围
ArgM-FRQ	Frequency, 频率
ArgM-LOC	Locative, 地点
ArgM-MNR	Manner, 方式
ArgM-PRP	Purpose or reason, 目的或原因
ArgM-TMP	Temporal, 时间
ArgM-TPC	Topic, 主题

附录 2: 汉语中的介词集合和方位词集合

介词集合: {从, 自, 于, 打, 到, 往, 在, 当, 朝, 向, 被, 叫, 让, 给, 比, 和, 同, 按, 照, 依, 以, 凭, 为, 因, 对, 把, 向, 跟, 与, 同, 给, 除, 自从, 顺着, 沿着, 随着, 按照, 依照, 本着, 经过, 通过, 根据, 为了, 为着, 由于, 因为, 对于, 关于, 除了, 除去, 除非}

方位词集合: {上, 下, 前, 后, 东, 西, 南, 北, 左, 右, 里, 外, 中, 内, 旁, 边, 头, 尾, 面, 顶, 侧, 际, 东面, 西面, 南面, 北面, 东头, 西头, 南头, 北头, 上头, 下头, 前头, 后头, 里头, 外头, 之上, 之下, 之前, 之后, 之东, 之西, 之南, 之北, 之外, 之内, 之中, 之间, 上边, 下边, 前边, 后边, 东边, 西边, 南边, 北边, 左边, 右边, 里边, 外边, 上面, 下面, 前面, 后面, 左面, 右面, 里面, 外面, 以东, 以西, 以南, 以北, 以内, 以外, 上下, 前后, 南北, 左右, 里外, 内外, 底下, 顶上, 头里, 当中, 当间, 中间, 内中, 内里, 面前, 跟前, 头前, 背后, 旁边, 边上, 面上, 对面, 这儿, 这里, 那儿, 那里, 以远, 以近, 附近, 周围, 上部, 下部, 前部, 后部, 东部, 西部, 南部, 北部, 中部, 内部, 东端, 西端, 南端, 北端, 左端, 右端, 上端, 下端, 前端, 后端, 东侧, 西侧, 南侧, 北侧, 左侧, 右侧, 外侧, 内侧, 东南边, 东北边, 西南边, 西北边, 东南面, 东北面, 西南面, 西北面, 东南部, 东北部, 西南部, 西北部, 东南方, 东北方, 西南方, 西北方, 东南角, 东北角, 西南角, 西北角, 左上头, 左下头, 右上头, 右下头, 左上部, 左下部, 右上部, 右下部, 左上方, 左下方, 右上方, 右下方, 前上方, 后上方, 左上角, 左下角, 右上角, 右下角, 一边, 一面, 一头, 两边, 两头, 两旁, 两旁边, 四面, 四边, 四外, 四周, 前后左右, 四围, 四周围, 中心, 东南, 东北, 西南, 西北, 外边}

攻读博士学位期间发表的论文及其它成果

(一) 发表的学术论文

- [1] 李世奇, 赵铁军, 李晗静, 刘鹏远, 刘水. 基于特征组合的中文语义角色标注[J]. 软件学报. 2011, 22(2):222-232 (**EI: 20111313882255**)
- [2] 李世奇, 赵铁军, 陈晨, 刘鹏远. 基于 ART 网络的无指导中文共指消解方法[J]. 高技术通讯. 2009, 19(9): 926-932 (**EI: 20094512433589**)
- [3] **Shiqi Li**, Qin Lu, Tiejun Zhao, Pengyuan Liu, and Hanjing Li. Combining Constituent and Dependency Syntactic Views for Chinese Semantic Role Labeling[C]. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING). 2010:665–673
- [4] **Shiqi Li**, Pengyuan Liu, Tiejun Zhao, Qin Lu, and Hanjing Li. PKU_HIT: An Event Detection System Based on Instances Expansion and Rich Syntactic Features[C]. In Proceedings of SemEval-2010 workshop of ACL. 2010:304-307
- [5] **Shiqi Li**, Tiejun Zhao, Hanjing Li, Shui Liu. A Feature Combination Method for Semantic Role Classification[J]. Journal of Information and Computational Science. 2010, 7(1):127-133 (**EI:20102112950416**)
- [6] **Shiqi Li**, Tiejun Zhao, Hanjing Li, Pengyuan Liu, and Shui Liu. Using Cognitive Model to Automatically Analyze Chinese Predicate[C]. In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLPKE). 2010:119-126 (**EI: 20104813421267**)
- [7] **Shiqi Li**, Tiejun Zhao, Hanjing Li. Improving Spatial Semantic Analysis by a Combining Model[C]. In Proceedings of International Conference on E-Business and E-Government (ICEE). 2009:1430-1433 (**EI: 20104913459030**)
- [8] **Shiqi Li**, Tiejun Zhao, Hanjing Li. Spatial Semantic Analysis Based on a Cognitive Approach[J]. Computer and Information Science, Studies in Computational Intelligence, 2009, 208:93-103
- [9] **Shiqi Li**, Tiejun Zhao, Hanjing Li. Text-to-Scene Conversion: An Introductory Survey[J]. International Journal of Computational Science. 2009, 3(3):309-324
- [10] Pengyuan Liu, **Shiqi Li**. A Corpus-based Method to Improve Feature-based Semantic Role Labeling[C]. In Proc. of the NLPOE workshop of 2011

IEEE/WIC/ACM International Conference on WI-IAT. 2011 (已录用)

[11]Pengyuan Liu, Shui Liu, **Shiqi Li**. One sense per N-gram[C]. In Proc. of the NLPOE workshop of 2010 IEEE/WIC/ACM International Conference on WI-IAT. 2010:195-198 (**EI: 105013474269**)

[12]Pengyuan Liu, Yongzeng Xue, **Shiqi Li**, Shui Liu. Minimum Normalized Google Distance for Unsupervised Multilingual Chinese-English Word Sense Disambiguation[C]. In Proceedings of 4th International Conference on Genetic and Evolutionary Computing(ICGEC), 2010:252-255 (**EI: 20104813421267**)

(二) 参与的科研项目及获奖情况

- [1] 参与科研项目：融合语言知识与统计模型的机器翻译方法研究. 国家自然科学基金重点项目. 课题编号：60736014.
- [2] 参与科研项目：典型事件过程建模的研究. 国家自然科学基金项目. 课题编号：60803094.
- [3] 参与科研项目：基于 Ontology 的自然语言描述的空间概念三维可视化研究. 国家自然科学基金项目. 课题编号：60575041.

哈尔滨工业大学学位论文原创性声明及使用授权说明

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向文景转换的中文浅层语义分析方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：李世奇 日期： 年 月 日

学位论文使用授权说明

本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

(1) 已获学位的研究生必须按学校规定提交学位论文；(2) 学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；(3) 为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；(4) 根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：李世奇 日期： 年 月 日

导师签名：赵铁军 日期： 年 月 日

致 谢

值此论文完成之际，心中感慨良多。这篇论文能够得以顺利完成，使我既体会到了辛勤劳动后的喜悦，又深深地感到它与实验室大家庭成员的热心帮助是分不开的。

首先，感谢我的导师赵铁军教授这六年来的言传身教，我的每一分工作与成绩都和他的辛勤培养分不开。赵老师严谨的治学态度、渊博的专业知识、深厚的学术素养、平易近人的工作作风、诚恳严格的处事原则都使我十分钦佩，是我一生学习的榜样。

感谢李生教授的指导和帮助。他开阔的视野、敏捷的思维、渊博的学识无不是学生学习的楷模。

感谢李晗静老师对我的帮助和指导。我的很多工作都是在她的指导下完成的，她教会了我科学的研究方法和实事求是的治学态度。

感谢杨沐昀、郑德权、于浩、徐冰、朱聪慧等老师的鼓励与帮助，感谢尚福华、薛永增、曹海龙、李卫疆、谭红叶、王博、蒋宏飞、杨宇航、陈鄞、王浩畅、张淑等师兄和师姐在学习和生活上对我的帮助和指导，感谢在实验室一起学习与奋斗过兄弟姐妹：陈晨、刘水、李壮、喻洪勇、姚超、朴星海、张迪、刘树杰、党可、刘立刚、王月颖、郑银、陈百方、王浩、丁伟利、赵纪元、刘艳、林日、韩延海、叶利军、李理、郭键、季伟等，这篇论文的顺利完成与他们的密切合作和热心帮助是分不开的。

感谢实验室的所有成员，是他们给了我一个团结友爱、积极向上的集体，身为其中的一员，我深感自豪。

感谢在百忙之中对本论文提出宝贵修改和评审意见的各位专家。

感谢培育我的母校哈尔滨工业大学，感谢计算机学院辛勤的老师。

最后，深深感谢我的父母这么多年对我的养育和关怀，你们是我的一切，只要有你们，不管面对什么样的困难，我都会坚持度过。

个人简历

李世奇，男，1984 年 8 月 20 日出生于黑龙江省哈尔滨市。

2001 年 9 月起就读于哈尔滨理工大学计算机科学与技术专业，2005 年 7 月本科毕业并获得工学学士学位。

2005 年 9 月——2011 年 6 月，在哈尔滨工业大学计算机科学与技术学院，计算机应用技术学科攻读硕士及博士学位。

获奖情况：获 2011 年度腾讯科技卓越奖学金。

面向文景转换的中文浅层语义分析方法研究

作者: [李世奇](#)
学位授予单位: [哈尔滨工业大学](#)

引用本文格式: [李世奇](#) [面向文景转换的中文浅层语义分析方法研究](#)[学位论文]博士 2011