杭州电子科技大学

硕 士 学 位 论 文

题　　目：汉语语义组块识别研究

研 究 生　常若愚

专　　业　计算机软件与理论

指导教师　吴铤　教授

完成日期　2015 年 3 月

杭州电子科技大学硕士学位论文

汉语语义组块识别研究

研 究 生： 常若愚

指导教师： 吴 铤 教 授

2015 年 3 月

Dissertation Submitted to Hangzhou Dianzi University

for the Degree of Master

Research of Chinese Semantic Chunk Recoginition

Candidate: Chang Ruoyu

Supervisor: Prof. Wu Ting

March, 2015

杭州电子科技大学
学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘要

随着当今社会信息化进程的加快以及互联网的飞速发展，自然语言处理技术被广泛应用于机器翻译、信息检索和人机交互等信息处理领域。经过多年发展，自然语言处理逐渐由基于规则的方法过渡到基于统计的方法。基于规则的方法以深层分析和理解自然语言为目的，在实现中复杂且困难；基于统计的方法以浅层处理自然语言为目的，便于利用计算机实现。

语义组块分析技术是自然语言处理中浅层语义分析和句法分析的代表，旨在解释自然语言中语法和语义之间的关联。组块的长度介于句子和单词之间，在各种自然语言中有着不同的划分，本文主要在汉语上展开相关的研究工作。

汉语的组块分析并没有统一的描述体系，因研究目的不同，研究者们各自提出了不同的组块分析体系。组块分析是浅层句法分析技术，基于对汉语句子语法和语义综合考虑进行分析的目的，本文在组块分析的相关任务语义角色标注问题上，沿用前人对语义组块的定义，对语义组块识别阶段的关键技术进行了深入的研究。

语义组块分析是自然语言处理中浅层语义分析和句法分析的重要内容，本文针对汉语语义组块识别中普遍存在的召回率不高这一问题，提出了一种新的标注方式：IO 标注法，并利用支持向量机（SVM）模型二类分类的特性充分地发挥了该标注法只有两种标识的优势，在语义组块识别阶段极大地提高了召回率进而提升了 F1 值。同时，本文也使用条件随机场（CRF）模型对语义组块按 I、O 标识进行了序列标注的研究。实验结果表明，在汉语的宾州命题库上，结合 IO 标注法的基于支持向量机的语义组块识别系统可以取得最好的性能，将 F1 值提高到了 80.30%，高于采取其它标注法的系统，实验还进一步表明不同标注法对语义组块识别系统性能的影响。

本文具体的组织结构如下：首先，介绍了语义组块识别的流程及评价方法，从中可知，经过语义组块识别后，句子中的各成分被标注了不同的标识，表征该成分是否是语义组块，本文以标注方式作为切入点，提出了一种全新的标注法，将其应用到语义组块识别阶段，并与传统的标注方式做出比较；其次，结合 IO 标注法，使用统计机器学习方法 CRF 和 SVM 建立统计模型，分别将语义组块识别作为序列标注问题和二元分类问题进行研究，实验结果与对比系统进行了比较，验证了在语义组块识别这一问题上，基于 SVM 模型的语义组块识别方法在

IO 标注法下可以取得最好的性能；最后，本文将新的语义资源加入现有系统，以期从新的角度研究语义组块。

关键词：语义组块分析，条件随机场，机器学习，支持向量机，语义角色标注

ABSTRACT

With the acceleration of informatization progress and the rapid development of Internet nowadays, natural language processing technology is widely used in many information processing fields such as machine translation, information retrieval, human-computer interaction etc. After years of development, natural language processing gradually achieved transition from rule-based approach to statistic-based approach. The rule-based approach aims at deeply analyzing and understanding natural language, which is complicated and difficult to achieve; while statistic-based approach aims at superficially processing natural language, which is easy to achieve with the computer.

Semantic chunk analysis technology represents shallow semantic analysis and syntactic analysis, which aims at explaining the relevance between syntax and semantics. The length of chunk lies between sentence and word, and the chunk is divided differently in various natural languages. The thesis mainly researches Chinese semantic chunk recognition.

There isn't any unified description system on Chinese chunk analysis. For different purposes, different researchers proposed different chunk analysis system respectively. Chunk analysis is a kind of shallow parsing technology. Aiming at analyzing Chinese sentence syntax and semantics synthetically, the thesis applies the definition of semantic chunk in the matter of remarking the related task semantic role to chunk analysis, and deeply researches the key technology during semantic chunk recognition.

The semantic chunk analysis research is an important direction in shallow semantic parsing domain. To improve the precision of semantic chunk recognition, a new IO labeling method was proposed in this paper. The IO labeling has only two tags, which was combined with the advantages of SVM for improving the recall rate and F1 value in Chinese semantic chunk recognition. At the same time, this paper presented the research results on semantic chunks' serial labeling using I or O tags by applying CRF model. The experimental on Chinese Proposition Bank results show that our proposed combined IO labeling and SVM method can obtain best

performance and reach 80.3% F1 value in Chinese semantic chunk recognition. The experiments also show that the different labeling methods affect the performance of semantic chunk recognition.

In the thesis, the specific research contents are as follows: First, the process and evaluation method of recognizing semantic chunk is introduced, from which we can see that the various components in the sentence are labeled differently after semantic chunk recognition, which represent whether these components are semantic chunk or not. The thesis proposes a new labeling method which is applied to semantic chunk recognization and compared with traditional labeling method. Second, the statistical machine learning methods such as CRF (Conditional Random Field) and SVM (Support Vector Machine) are applied to build statistical models combing with IO labeling method. Then semantic chunk recognition is researched as sequence labeling and binary classification respectively. Comparing the results of the above two methods, we find that, in the matter of semantic recognition, SVM-based semantic chunk recognition can get the best result with IO labeling method. Finally, the new semantic resources join the current system, and the semantic chunk is researched from a new perspective.

Keywords： Semantic Chunk Analysis, Machine Learning，CRF, SVM, Semantic Role Labeling

目录

摘要 I
ABSTRACT III
第 1 章 绪论 1
1.1 课题背景及研究意义 1
1.1.1 课题背景 1
1.1.2 研究意义 3
1.2 课题研究现状 4
1.3 本文研究内容 5
1.4 本文组织结构 6
第 2 章 语义组块分析概述 8
2.1 汉语语义组块定义 8
2.2 语义组块分析的应用 9
2.3 语义组块识别系统 10
2.3.1 语义组块识别流程 10
2.3.2 语义组块识别语料库 11
2.3.3 语义组块识别验证方法 13
2.3.4 语义组块识别评测标准 14
2.4 本章小结 14
第 3 章 语义组块识别 16
3.1 语义组块标注方式 16
3.1.1 传统标注方式 16
3.1.2 IO 标注法 17
3.2 基于 CRF 的语义组块识别方法 19
3.2.1 CRF 模型简介 19
3.2.2 特征抽取与特征模板 19
3.3 基于 SVM 的语义组块识别方法 20
3.3.1 SVM 模型简介 20
3.3.2 标注模型的构造 21
3.4 对比系统 22

3.5 实验结果及讨论 23
3.6 本章小结 25
第 4 章 语义分析技术与语义组块识别 26
4.1 语义信息在语义组块识别中的应用 26
4.1.1 语义分析 26
4.1.2 语义组块识别的语义信息选取 26
4.2 基于语义信息的语义组块识别原理 27
4.2.1 同义词词林 27
4.2.2 语义特征对语义组块识别的影响 28
4.3 实验及讨论 29
4.3.1 实验设计 29
4.3.2 实验结果及分析 30
4.4 本章小结 32
第 5 章 总结与展望 33
5.1 本文总结 33
5.2 工作展望 34
致 谢 36
参考文献 37
附 录 41

第1章 绪论

自然语言处理（natural language processing，NLP）作为探索计算机与自然语言通信的方法和理论，在当今信息社会有着重要的作用。近年来，随着计算机技术特别是互联网的蓬勃发展，使自然语言处理得到了空前的重视并逐步发展成为相对独立的学科。通过对自然语言完全的句法和语义分析，可以实现计算机与自然语言无障碍的交流。然而，人们在多年研究之后，依然无法解决这一难题，因此出现了浅层的句法和语义分析技术。本文研究的语义组块识别就是该技术的一种实现。

1.1 课题背景及研究意义

1.1.1 课题背景

自然语言是信息的重要载体，计算机的普及和互联网的迅速发展带给人们海量的信息，这些庞大的信息需要一些自动化的计算机技术辅助处理，自然语言处理在机器翻译（machine translation，MT）、人机交互（human-computer interaction，HCI）、信息检索（information retrieval，IR）等信息处理领域的实用价值使其成为这一技术的合适选择。

自然语言处理研究一般涉及语法、语义和语用三个层次。语法层面侧重于研究语言中组成句子序列的规则和句子结构成分相互之间的联系，通俗来讲就是研究句子到底如何说；语义层面侧重于研究语言的含义，覆盖语言中的各级单元：词、词组、句子和段落等，通俗来讲就是研究这个语言单元说得是什么；语用层面是从语言的使用者角度来研究他们对语言使用和传递过程中互相之间的影响，这方面的研究大多是从语言学角度考虑的[1]。以上三个层面中，对语法和语义层面的研究是目前利用计算机进行自然语言处理所关注的问题，语用层面更多是被语言学家关注。

对自然语言进行语法和语义层面的研究一般有两种方法：理性主义（rationalist）和经验主义（empiricist）[2]，也是自然语言处理中两种基本的研究策略。理性主义着力于规则，经验主义着力于统计[3]。在 20 世纪 60 年代，研究者们对理解自然语言的探索主要是依靠规则分析句子的语法和获取语义，这实际上是受传统语言学研究方法影响的惯性思维，而自然语言的语法规则又可以很方便地利用计算机的算法来描述，因此，到 20 世纪 70 年代为止，基于规则的研究方法是自然语言处理的主流。

基于规则的自然语言处理将语言本身作为一种语法，认为自然语言是由文法规则产生的，这就带来了两个问题：首先，对于现实应用中的真实文本，文法规则哪怕只覆盖一小部分，也会产生非常复杂的大量规则，而且文法规则到最后会出现很多互相矛盾的情况；其次，假设能写出覆盖所有自然语言现象的文法规则，由于其错综复杂的结构，用计算机进行解析也是非常困难的。可以看到，基于规则的自然语言处理方法复杂度太高，特别是在语义分析层面，词语的多义性非常依赖上下文，依靠规则很难分辨词语的歧义[4]。

在 20 世纪 80 年代，自然语言处理的研究者们开始更多的关注实用化的解决方法，研究角度从单纯的句子语法分析和语义分析，转向了更贴近应用的机器翻译、语音识别和知识获取等，出现了基于统计的自然语言处理研究方法。

基于统计的自然语言处理方法认为自然语言的知识系统来源于大规模的真实语料，因此这种研究方法又称为语料库语言学，它具备几种特性：首先，计算机可以存储大规模的文本语料，并对其进行一致性和可靠性的分析，具有实际的可实验性；其次，大规模的语料库在不断更新中也体现了自然语言的演变，特别是在目前网络词语层出不穷的情况下，使用真实语料进行研究比添加新的自然语言规则在可行性上要好的多；最后，基于语料库的方法不仅可以统计语言特征的数量，还能够对真实文本的定量模式进行定性的功能解释[5]。

在基于统计的自然语言处理中，句子的语法分析一向是重点和难点。经过不断的研究发展，在分析大规模真实语料时，完全的句法分析在实现中依然不理想。在此前提下，一些研究人员转变思路，将“分治”的思想运用到句法分析中，以浅层分析（shallow parsing）的策略来降低完全句法分析的难度。浅层句法分析又称之为组块分析（chunk parsing），由 Steven Abney[6]率先提出，他把句法分析问题分成三个步骤：（1）组块识别：用基于有限状态分析的块识别器进行快速识别；（2）组块内部结构分析：给组块内部的成分赋以合适的结构；（3）组块间关系分析：通过块连接器将不同组块连接成为完整的句法树。在此基础上，CoNLL（Conference on Computational Natural Language Learning）国际会议连续推出了几项组块分析相关的共享任务，包括子句识别、语义角色标注（Semantic Role Labeling，SRL）等[7]。

CoNLL-2000 的共享任务在 Abney 的基础上，将英语组块定义为：句子由一些组块构成，每个组块由句法相关的词语所组成，组块具有不重叠、无交集和非嵌套的特性。在汉语的组块分析中，国内研究人员大多借鉴英语的组块分析理论和技术来对汉语进行相关的研究工作，特别是对汉语组块的定义主要参考了英语的组块定义体系。然而汉语的句法体系迄今为止并没有一个公开的训练语料资

源，针对不同的组块分析策略无法提供统一的评价方法。从现有的研究来看，研究人员针对不同的研究目的给出了各自不同的组块定义体系[8]。

组块分析在句法分析范畴内，分析和识别句子的主要框架，给自然语言处理提供了基础的研究技术。针对汉语意合性语言的特性，对汉语进行组块分析应该有语义因素的考虑，这是因为，组块可以揭示句子的结构，而语义能够反映句子的意义。语义分析和句子的语法分析一样，是理解自然语言的重要手段，例如，句子“我打扫干净了房间”与“房间被我打扫干净了”表达形式虽不同，但在语义层面可由“我”、“打扫”、“房间”统一表示。可见，结合语义信息对组块进行分析可以综合考虑句子的语法分析和语义分析两方面，有助于自然语言处理的发展，目前，这方面的研究比较少见。

1.1.2 研究意义

自然语言处理是个多学科交叉的研究方向，其发展依赖于计算机技术、数学、语言学、心理学和人工智能等相关学科的同步发展。当今社会是一个信息社会，信息的表示、存储和传播主要是通过语言来实现，语言是人类行为的一个基本方面，是我们生活中的重要组成部分，书面语言记录了人类社会长久以来积累的知识，而口语则是人与人之间交流的基本方式。

不同的科学领域对于自然语言研究的侧重点不同，语言学家研究语言本身的体系结构，探究特定词语的组合可以形成句子的成因、句子具有某种意义的缘由等问题；哲学家研究人类的目标、意图等认知能力是如何和语言构成联系的；心理语言学家研究人类是如何识别自然语言里面句子的合理结构、如何确定词语的合理意义等问题；计算语言学家研究怎么用计算机技术构建一个围绕自然语言的计算方法[9]。当然，自然语言处理技术的进步离不开上述学科的共同发展。

在社会活动中，自然语言处理技术处处影响着人们的生活。目前，随着互联网的发展，社会信息化的程度越来越高，而且趋向于智能化的发展。自然语言处理本身就是人工智能领域的重要课题，在与人类生活息息相关的各方面发挥着作用。如，在社交网络方面，自然语言处理技术用来计算用户的情感趋向，方便人们的交流；在电子商务方面，自然语言处理技术用来进行商品推荐，方便人们的生活。人们在互联网上使用的社交工具、搜索引擎、电子邮箱乃至计算机的语言输入法，都离不开自然语言处理技术的发展。自然语言处理研究的重要性由此可见一斑。

目前，随着统计方法在自然语言处理实践和应用中的迅速发展，从大规模真实文本里面获取自然语言规律的语料库语言学也得到了长足的发展。在应用领域，自然语言处理在过去的 20 几年中发生了巨大的变化，例如，对自动问答任

务的需求基本上被网页搜索与数据挖掘所代替。新的应用愈加依靠语料的作用和浅层的自然语言处理方法，客观的说明在语料库的基础上对自然语言进行浅层句法和语义分析的重要性，组块分析正是这一技术的具体体现。

组块分析在文本分类、信息抽取、信息检索、问答系统和机器翻译系统有着广泛的应用。其中，在机器翻译领域中的作用较突出，本文在组块分析基础上进行语义组块分析研究的目的就是为了将其应用到机器翻译领域。语义组块具有语义独立性、结构可靠性的特点[10]，对于机器要翻译的两种自然语言来说，这些特性提高了翻译的准确度。同时，语义组块比词语的颗粒度大，而且语义组块带有句法意义和语义信息，词语不带任何句法意义。因此，在机器翻译中，将语义组块作为双语对齐的基本单位，可以提高系统的性能。

1.2 课题研究现状

目前，组块分析的技术主要有基于统计和基于规则两种。前面已经介绍过，自然语言处理领域的主流技术大多偏向基于统计的机器学习方法，组块分析作为其中的重要问题也不例外。主流的组块分析方法所使用的统计模型主要有隐马尔科夫模型（Hidden Markov Model，HMM）[11]、条件随机场（Conditional Random Fields，CRF）模型[12]、支持向量机（Support Vector Machine，SVM）模型[13]、最大熵（Maximum Entropy，ME）模型[14]和神经网络（Neural Network，NN）模型[15]等。

隐马尔科夫模型预测状态转移的优势使其在组块分析上得到很好的应用，该模型依然有着一定的缺点，比如计算得出的都是局部的最优值、无法利用上下文信息。条件随机场模型可以利用任意的上下文信息，在序列标注上具有独特的优点，可以结合多种特征，在组块分析中应用较广，缺点是性能过于依赖对特征的选择和优化。支持向量机模型是一种优异的分类器，在很多领域都有出色的表现，在组块分析中，可以有效地提高泛化性能，缺点是在非线性问题的应用中没有通用的解决办法。最大熵模型在组块分析中的应用主要在于其特征选择灵活这一特点，但有着较严重的数据稀疏问题。神经网络模型分类的准确度高，适合对组块进行分类，缺点是需要大量的参数、训练时间过长。

李珩等[16]使用隐马尔科夫模型，针对该模型的缺点，提出了一种导入上下文信息的策略，在公开的语料库哈工大汉语树库上，设计了一组转换函数来获取训练模型的信息，分别对汉语和英语进行组块识别的研究，称之为基于增益的隐马尔科夫模型组块分析方法。周雅倩等[17]借鉴最大熵模型在英语组块识别上的应用，实现了基于该模型的组块识别器，在汉语的宾州树库上，选取有效的特征对汉语组块识别进行了研究和分析。黄德根等[18]采用分布式策略，在北大汉语

语料库上，将前期的组块识别标记作为新的特征，使用条件随机场模型进行二次识别，并针对不同的组块“颗粒度”进行了研究。孙广路等[19]将条件随机场模型运用到组块识别中，在哈工大汉语树库上，取得了较好的实验结果。R.Collobert[20]等结合神经网络结构和学习算法，将卷积神经网络应用到词性标注、命名实体识别、语义角色标注和组块分析四个任务中，提供了一种新的研究思路，基于卷积神经网络模型的组块识别虽然性能略低于其它几种，但是该方法不依赖之前的自然语言处理体系，研究角度较新颖，是开创性的工作。

在上述机器学习的研究方法中，又分为有监督学习方法、半监督学习方法和无监督学习方法。目前，依靠大规模的语料库，有监督学习方法可以取得更好的效果，但是需要耗费大量的人力物力，而无监督和半监督学习方法也取得了一定的研究成果。

汉语的组块分析经过多年的研究发展，提供了多种的研究角度。不同研究者们定义的汉语组块体系虽然互不相同，但是研究的目的是一样的，都旨在揭示汉语句子浅层的语法和语义关系。目前，有大量的单纯基于句法的组块分析研究，这些研究大多偏重于使用具体策略对组块分析的改进，而将组块分析加入语义因素考虑进行基于语义的组块分析研究还比较少见。

在汉语的语义组块分析方面，丁伟伟等[21]发现语义组块可为语义角色标注服务，并将语义角色标注问题转化为语义组块的识别问题，对汉语的语义组块进行识别和分类，得到了与基于句法分析的方法可比较的实验结果。刘海霞等[22]结合语义信息，赋予组块语义标记，进行了汉语功能组块自动识别的工作，通过在组块识别中引入词语的语义信息，改善了功能块识别的性能。王鑫等[23]将组块分析加入构词法信息，采用分步标注法和直接标注法两种策略，提高了汉语语义标注的性能。何塞克等[24]在组块分析与语义角色标注的基础上，提出一种基于多任务学习方法（multi-task learning，MTL）中交互结构优化（alternating structure optimization，ASO）算法的策略，以有监督的方式提升了有语义标注的汉语组块分析系统性能。以上几种研究方法使用的语料库都是中文宾州树库，有着不同的研究目的，侧重点各不相同，在语义组块分析相关研究较少的情况下提供了一些研究思路。

1.3 本文研究内容

本文主要研究汉语方面的语义组块。汉语的组块体系很大程度上受到了完全句法分析方法的影响，分析效果也因此受到了限制，本文试图从语义分析方向着手来寻求汉语句子句法分析的突破，把语义分析的思想与句子的组块分析相结合，进行汉语语义组块的研究，有助于推动组块分析相关任务的发展，也有利于

句子浅层句法分析理论和语义分析理论的发展。

结合相关的研究进展，本文综合考虑汉语句子的语法和语义两个方面，以对句子的浅层分析为目的，在包含大规模真实文本的语料库上，使用统计机器学习理论，将不同的机器学习算法应用到语义组块识别中，在识别过程中提出了新的方法，并运用了与算法结合的新策略，深入研究了汉语的语义组块分析技术。

本文使用汉语的宾州命题库（Chinese Proposition Bank，CPB）作为语料库，在研究前期，主要做了以下准备工作：首先，对语料的文本进行分析，研究语料的文本结构，使其更好的应用于语义组块识别；其次，分析不同的机器学习算法优劣，经过反复实验，确定适合本文研究方法的统计学习模型；最后，在研究的过程中，探索对语义组块识别阶段的改进方式。

在对语义组块分析做了相关的前期研究后，本文主要在以下几个方面展开工作：

（1）解析语料库。同样的语料库可用于不同的自然语言处理研究方向，对于不同的研究任务，从语料库中抽取候选信息的方式各异，语料中与语义组块分析相关的信息需要解析出来作为候选，以便后续处理。

（2）从语料库中提取有效的特征。经过自然语言处理多年的发展，研究者们发现算法采用的特征是决定系统性能的主要因素。因此，对于语义组块分析来说，构造特征是选取合适的机器学习算法的必要前提。

（3）提出新的语义组块标注方式。语义组块分析一般分为语义组块识别和语义组块分类两个阶段，在汉语语义组块识别方面，大多采用传统的组块标注方式。本文针对研究的目的，提出了一种新的标注方式，采用不同的机器学习算法来验证此标注方式的有效性。

（4）将语义组块与语义资源相结合以实现语义组块识别系统的提升。语义组块是对句子语义信息的有效描述，语义资源在单词的层面上按语义给词语归类，本文将语义资源《同义词词林》与语义组块识别结合起来，用更丰富的语义信息来提升语义组块识别的性能，并分析了在不同算法下，这种研究方式得出差异化结果的原因。

在没有特别指明的情况下，本文中出现的组块分析及语义组块分析都代表针对汉语进行的研究。

1.4 本文组织结构

本文共分 5 章，具体的组织结构如下：

第 1 章，绪论。首先阐述了组块分析在自然语言处理中的意义，介绍了组块分析及其相关研究；其次剖析了语义分析运用在汉语组块分析上的可行性，并说

明了语义组块分析在实际应用领域的价值；最后给出了本文的研究内容以及各章节相关信息。

第 2 章，语义组块分析概述。这一章主要介绍语义组块分析的整个系统，前半部分将语义组块在实际应用领域机器翻译中的具体作用进行了分析，随后，给出了汉语语义组块的定义。后半部分以图解的方式展示了语义组块识别的流程，比较了相关研究常用语料库的异同，确定了对语义组块识别系统的评价方法。

第 3 章，语义组块识别研究。在这一章，本文在语义组块标注方式上提出了有别于传统标注方式的 IO 标注法，详尽的分析了该标注法对语义组块识别的影响。在 IO 标注法下，分别将语义组块识别作为序列标注问题和二元分类问题，使用统计机器学习的方法，利用 CRF 和 SVM 建立统计语言模型，进行语义组块识别的实验。在最后，将本系统最好的结果与对比系统的实验结果做出了比较和剖析。

第 4 章，语义分析技术与语义组块识别。针对语义组块识别研究句子中语法和语义信息的目的，将实验语料库 CPB 外部的语义资源引入系统，经过对汉语语义资源的分析，选择了合适的语义词典，从中抽取词语的语义信息，实现了在不破坏语料库句子结构的情况，给语义组块识别训练模型提供了新的特征，并分析了新的语义组块识别系统在 CRF 和 SVM 下的不同表现，阐述了这种影响的原因。

第 5 章，总结与展望。将本文的研究工作和成果进行了总结，探究了语义组块识别在未来值得研究的方向。

第2章 语义组块分析概述

2.1 汉语语义组块定义

组块分析在自然语言处理领域是基础性的问题，英语组块有一个公认的权威解释，相比之下，因研究目的不同，研究者们对汉语组块的体系描述各有所异。

李建素等[25]将汉语的组块定义为一种非递归短语结构，该结构符合一定句法功能，结构内部有一个核心词，组块内的其它成分围绕该核心词进行扩展。该组块体系将汉语的组块类型分为五种，规定了不同类型的组块之间不能互相包含，是按照严格的语法规则进行定义的。

周强等[26]针对汉语的语言特点，将组块的边界判别作为独立于语法描述形式的句子拓扑结构，通过引入了词界块和成分组的概念，意在形成一个独立于各种语法体系的组块描述体系。该汉语组块体系在后续研究中不断改进，依次提出了基本块[27, 28, 29]、功能块[30, 31]和事件描述小句[32]等定义，并在此基础上遵循穷尽性和线性的原则标记组块，将汉语分成八种组块类型，构建了大规模的组块语料库。

孙广路等[33]将汉语组块定义如下：组块是由带句法功能标记的词序列组成，是一种特殊的短语形式，组块内部包含一个核心词以及其前置修饰短语，不包含后置附属短语，各组块之间非递归、非嵌套、不重叠。该组块体系对组块的划分仅仅根据表面信息，如词语、词性标注等，使划分的颗粒度尽可能的大，不考虑句子的整体句法结构和组块之间的距离约束，同样是按照严格的语法规则进行定义的。

上述汉语组块描述体系各不相同，与之类似，汉语的语义组块定义也是独立其它描述体系的。周强等[34]指出，汉语的功能组块分析对应于英语的语义角色标注，是在句法层面对语义角色标注进行的模拟。这一定程度上表示，在浅层的句子分析层面，组块分析可为语义角色标注服务。袁彩霞等[35]对汉语功能组块进行了深入的研究，将汉语句子中具有不同语法功能或语义角色的成分统一用“功能标记”来标识。

由这些研究可以看出，功能组块分析的目的是为了将句子分析从语法分析层面扩展到语义分析层面，而语义角色标注是组块分析的相关任务。从研究目的来说，语义角色标注与功能组块分析并无二致，都是为了建立语法形式和语义信息在句法分析中的联系。

语义角色标注研究旨在对句子进行浅层的语义分析，不对句子包含的语义信息进行深入的研究。具体的说，语义角色标注针对句子中的谓词-论元结构进行分析，标记出句子中单个谓词（动词、部分名词以及形容词）的所有论元，识别出目标谓词语义角色的边界，通过标注论元的语义角色来显示汉语句子句法结构骨架。简单来讲，语义角色标注就是对句子中给定的谓词，以其为中心，分析各成分与它的关系，并用语义角色来描述这些关系。

Hacioglu 等[36]在英语上进行过类似的工作，称之为语义论元组块分析。丁伟伟等[14]借鉴语法组块分析的定义，在汉语上将论元的识别、分类定义为语义组块的识别与分类。无论称之为语义论元组块，还是语义组块，都旨在揭示句子结构的语义关系。

在这些研究的基础上，本文将组块分析应用于汉语的语义角色标注，将论元的识别统一称之为语义组块识别进行研究。

2.2 语义组块分析的应用

对自然语言进行组块分析离不开对句子的分析处理，组块分析描述句子结构的能力介于完全句法分析和单词表达之间[37]。结合语义分析技术的语义组块是一个句子表述完整语义的核心部分，这种特性使其在机器翻译领域有着实际的应用价值。

不同的自然语言在表现形式上各不相同，语言的传递和交流是人类社会的基本活动行为。通俗来讲，人类运用语言主要是为了表达自己的“意思”，每一个句子代表着不同的意义，在处理不同语言间的交流障碍时，如何准确无误的传递语言的“意思”是首先需要考虑的方面。语义组块正是句子中表达语义的单位，因此，语义组块分析技术在用计算机处理自然语言翻译时起着重要的作用。

机器翻译研究涉及到计算机科学、语言学、人工智能和数学等学科，分为基于规则的机器翻译系统（rule-based machine translation system，RBMTS）、基于实例的机器翻译方法（example-based machine translation method，EBMT）和基于统计的机器翻译方法（statistical machine translation，SMT）。

基于规则的方法是早期机器翻译的主流方法，依靠双语词典的编纂和研究者总结的翻译规则，使用计算机将自然语言句子依次按词典和规则进行解码。这和基于规则的其它自然语言处理任务一样：缺陷较大、实现太复杂。自然语言本身具有复杂性、歧义性和灵活性等特点，很难用规则或是其它形式化手段来覆盖所有句子。当规则达到一定规模时，语言的歧义性使得再增加新的规则几乎是不可能的，基于规则的机器翻译方法的瓶颈也在于此。

八十年代中期产生的基于实例的机器翻译方法[38]从已有的翻译经验知识出

发，将语言句子切分为翻译经验中见过的片段，将这些片段与翻译知识进行匹配，得出翻译结果[39]。主要是通过对已有翻译资源的总结，得出实例库，设计规则处理实例库中的歧义性等问题。

基于统计的机器翻译方法由 Brown 等[40]提出，与其它基于统计的机器学习方法原理是一样的，使用统计学习的理论、在大规模的双语语料库上进行训练以建立翻译模型，将模型进行解码得到翻译结果，这也是现在主流的机器翻译研究方法。

现有的机器翻译系统结果还不够理想，主要表现在从源语言句子到目标语言句子的翻译过程中，句子片段译文的错误选择及译文顺序的错误组织。这是由于在对句子进行分析之前对句子进行了不合理的切分，把一些只有作为整体才具有正确意义的语言片段切分开了，导致切分后的各语言片段意义不完整，使各片段的译文选择和译文顺序不正确而影响了翻译结果的质量。

因此，要提高翻译质量就有必要对句子按照一定的“颗粒度”划分，把语义相对独立的语言片段划分成一个单位，结合语义信息进行组块切分的研究意义正在于此。句子中包含主要语义信息的组块被标识出来，在翻译过程中，双语之间有相对应的语义组块，翻译结果保证了句子意思的完整性，因此，对汉语和英语句子进行语义组块分析对改善汉英机器翻译系统有着重要的研究意义。

除此之外，语义组块分析在自动问答、信息抽取和信息检索等领域也有着诸多应用前景，由于本文对于语义组块的研究目的是服务于机器翻译的，其它应用方向不再赘述。

2.3 语义组块识别系统

2.3.1 语义组块识别流程

在英语句子中，单词之间是分隔开来的，而汉语的词语之间并没有这种语言自身的界限。因此，在汉语的自然语言处理的一些问题中，正确的分词以及词性标注是必要的前提条件，汉语语义组块识别也是如此，语义组块识别系统的起点是分词后且带有词性标注的句子。

汉语句子的组成是复杂的，同一个句子可能包含不同的谓词，针对不同的谓词，有不同的谓词-论元结构，在这种情况下，需要将句子中每一个谓词都对应这个句子的一个副本，分别对其进行语义组块识别。

由以上分析，将语义组块识别的整体流程简单罗列如下：首先，对汉语句子进行分词和词性标注；其次，确定句子的目标谓词，如果一个句子包含多个谓词，可将句子进行同等数量的拷贝，对于不同的谓词分析各自的语义组块；最后，对句子进行语义组块识别，标注出目标谓词的语义组块。

完整的语义组块识别流程如图 2.1 所示。

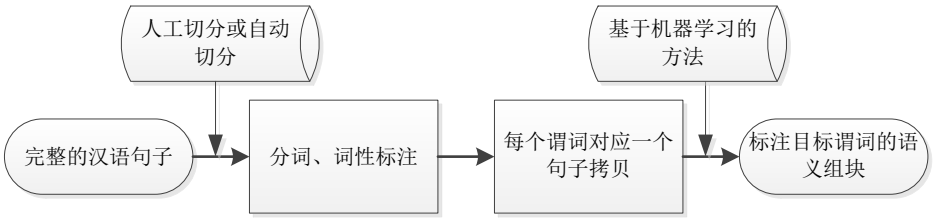

图 2.1 语义组块识别流程

在第一个步骤中，假设句子的分词和词性标注都是正确的。本文实验所使用的语料库 CPB 的文本内容已经是正确切分好的汉语句子，分词是自然语言处理领域中的另一个基础性问题，不属于本文的研究范围，也不是本文研究的着力点，这里不做详细论述。在第二个步骤中，每个谓词对应一个句子拷贝是针对未知实例来说的，本文实验语料的测试集是 CPB 中的文本，不存在这个问题。

以上两个步骤具有普适性：对于所有的汉语句子来说，在经过这两个步骤后，都可以使用语义组块识别系统建立的模型来进行语义组块的标记，也就是说适合任何属于或不属于语料库的句子。

当然，科学的研究方法是需要验证和评测的，在最后一个步骤中，需要进行严格的实验设置，利用有效的评价方法来评估整个系统的性能。利用机器学习的方法对目标谓词语义组块进行标注识别是整个语义组块识别系统的核心部分，也本文后续章节要分析的内容。

2.3.2 语义组块识别语料库

任何信息处理系统都需要数据和知识库的支撑，自然语言处理系统也是如此，统计学习方法在自然语言处理中的广泛应用使得语料库语言学得到了不断发展，并且对语言研究的相关领域产生越来越大的影响[41]。

语料库包含的是在实际生活中真实应用的语言语料，语料库的信息往往需要经过分析和处理才能运用到特定的研究领域中。与自然语言处理领域其它基于统计机器学习的研究方法一样，语义组块分析需要在经过加工处理的大规模真实文本的语料库上进行研究。

英语方面标注了语义信息的语料库比较成熟，主要有 FrameNet[42]、PropBank[43]和 NomBank[44]。U.C.Berkeley 开发的 FrameNet 是以框架语义作为标注理论的语义型词典，描述了谓词的语义框架以及各框架元素间的联系。宾夕法尼亚大学对宾州树库 Penn TreeBank 语料中的动词做出语义标记，建立了命题库 Proposition Bank，也就是 PropBank。PropBank 基于 Penn TreeBank 手工标注的句法分析结果标注了句子中各谓语动词的谓词-论元结构，准确率高，在语义标

注评测中被广泛使用，CoNLL 共享任务（2004、2005）就以 PropBank 作为语料库。PropBank 中的谓词仅包含动词，作为补充，纽约大学标注了 NomBank，其与 PropBank 标注的是同一批树库，区别是 NomBank 将句子中的名词性谓词及其论元结构也标注了出来，并且 NomBank 里的语义角色可以互相覆盖。

对应于英语中这三种语料库，汉语有 Chinese FrameNet[45]，Chinese Proposition Bank，和 Chinese NomBank[46]。Chinese FrameNet 以 FrameNet 为参照，形式化的描述了汉语的语义。Chinese PropBank 与英语的 Propbank 类似，将汉语的宾州树库 Chinese Penn TreeBank 加入语义信息的标注，Chinese NomBank 也是如此，对应于英语的 NomBank。

鉴于宾州树库 PropBank 应用的广泛性，考虑与相似系统的对比性，本文针对汉语语义组块分析使用汉语的宾州命题库 Chinese PropBank，所用版本是 CPB1.0，对应的汉语宾州树库 Chinese Penn TreeBank 的版本是 CTB5.0。CTB 的语料内容起初标注的是新华社的 10 万词次的新闻文本[47]，后又增加了人民日报、香港新闻电讯和从其它自然语言翻译过来的中文语料，规模扩大到 40 万词次[48]。

Chinese PropBank 语料库中例子（chtb_207.fid 中 824 行—835 行）的句法树如图 2.2 所示。

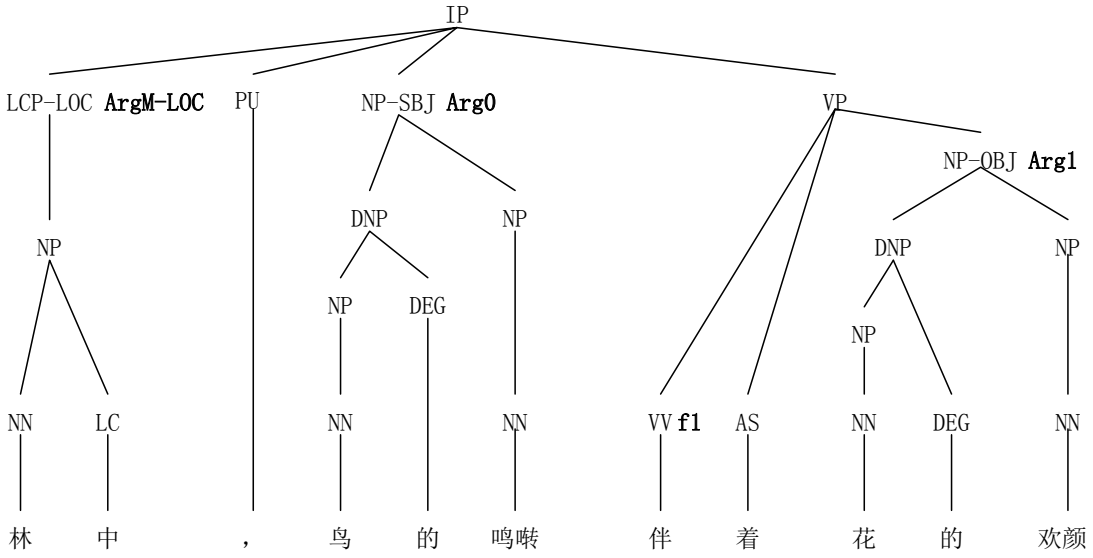

图 2.2 PropBank 例子

图 2.2 中的“ArgM-LOC”、“Arg0”和“Arg1”是论元标记，语料库 CPB 和 CTB 内容是一样的，具有对应关系，区别只在于 CPB 多了这些语义标记，本文统一用 CPB 来描述实验语料库。

PropBank 中的论元标注根据的是 Dowty[49]的原型理论，包含 20 多个论元标

记，分为核心论元和非核心论元两大类，核心论元又分为施事、受事、与事等多种论元，共有 Arg0~Arg5 六种，包括的词语范围较广，这些核心论元可以组成一个基础的汉语句子。这六种类型之外的论元是非核心论元，按照功能分为不同的小类，如方式、时间、地点、材料等，通常在句子中起修饰性的作用。

图 2.2 的句子里，目标谓词是“伴”这个动词，标识“f1”表示该句子中的“伴”属于谓词“伴”的第 1 个子语类框架（frameset），子语类框架的划分是因为同一个谓词可能有多重语义，不同的语义下的句法结构有可能不同，考虑句法的因素以额外的标识进行区分有利于标注的一致性。

该例子标记出了“伴”的三个论元：“林中”、“鸟的鸣啭”、“花的欢颜”。其中，“鸟的鸣啭”是“伴”的核心论元，“Arg0”表示它是原型施事论元，在句子中有自立性和使动性的特点[50]，描述了谓词所代表事项的独立存在，并且实施了动作。同样的，“花的欢颜”的标识“Arg1”表示它是核心论元且是原型受事论元，描述在谓词所代表事项中受到的影响。“林中”是“伴”的非核心论元，由“ArgM-LOC”标识，当中的“ArgM”是非核心论元的标记，“LOC”是二级功能标记，表示谓词“伴”的地点信息，此类二级功能标记有多种类型，比如“EXT”表示程度、“DIR”表示方向、“TMP”表示时间信息等。

这三个词语块表征了整个句子的主要语义成分，即为该句子的语义组块。

2.3.3 语义组块识别验证方法

不同的统计学习方法会给出不同的模型，在建立统计语言模型的过程中要进行验证，在自然语言处理领域，一般采取正则化验证（regularization validation）和交叉验证（cross validation）两种方法。正则化验证是一种结构风险最小化策略的实现，而交叉验证在组块分析的研究中使用更为广泛。

语义组块识别采取的是交叉验证方法，该方法又分为简单交叉验证、S 折交叉验证和留一交叉验证[51]。简单交叉验证将数据划分为训练集（training set）和测试集（test set）两部分，于各种条件下使用训练集来训练模型，得出不同的统计模型，在测试集上评价这些模型之间的测试误差，选择误差最小的模型。S 折交叉验证将数据平均地划分为 S 个部分，使用其中 S-1 个部分的数据作为训练集来训练模型，剩下的 1 个部分作为测试集来测试模型，这个过程可以重复进行 S 次（将 S 个部分依次作为测试集），然后选出误差最小的模型。留一交叉验证是 S 折交叉验证的特殊情况，在数据量很少的时候使用。

本文在对语义组块识别系统的具体实验中，将语料库中划分出训练集、测试集和开发集（development set）。其中开发集又称之为验证集（validation set），和测试集的功能一样，也是用于选择模型的。这种划分方法是在语料充足情况下的

交叉验证法，在选择模型阶段开发集代替了测试集，最终用测试集来评估开发集选出的模型。更重要的是，这种划分方法是为了和相似系统保持一致，有比较的验证本文语义组块识别系统的有效性。

2.3.4 语义组块识别评测标准

在语义组块识别阶段，通常采用准确率 Precision、召回率 Recall 和 F-Measure（又称为 F-Score）来评测系统的性能。本文用标识“I”来标注句子中的语义组块，标识“O”来标注句子中的非语义组块，这种标注方式是本文首先提出的，有其独特的优势，将在后续实验中详细分析。

设 TrueI 为实验系统将句子中的语义组块正确标注为“I”的个数，FalseI 为将非语义组块错误的标注为“I”的个数，FalseO 为将语义组块错误的标注为“O”的个数。则 Precision 和 Recall 的计算公式如下：

$$\mathrm{Pr}ecision = \frac{TrueI}{TrueI + FalseI} \tag{2.1}$$

$$\mathrm{Re}call = \frac{TrueI}{TrueI + FalseO} \tag{2.2}$$

Precision 和 Recall 指标有时候会出现矛盾的情况，因此需要 F-Measure 来综合考虑，F-Measure 是 Presicion 和 Recall 的加权调和均值，计算公式如下：

$$F - Measure = \frac{(\alpha^2 + 1) \bullet \mathrm{Pr}ecision \bullet \mathrm{Re}call}{\alpha^2(\mathrm{Pr}ecision + \mathrm{Re}call)} \tag{2.3}$$

根据不同的研究目的，可设置 α 为不同的值，当 α =1 时，即为最常见的 F1 值。F1 值综合考虑了 Precision 和 Recall 的结果，是有效的检验方式，由于相似系统都使用 F1 值评测，本文也使用 F1 值对语义组块识别进行评价，计算公式如下：

$$F1 = \frac{2(\mathrm{Pr}ecision \bullet \mathrm{Re}call)}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{2.4}$$

2.4 本章小结

语义组块分析技术具有实际应用价值，本章分析了语义组块分析在自然语言处理领域的应用，举例说明其在实际应用中所起的作用，表明了语义组块研究的重要性。汉语的组块描述体系各不相同，本章在前人的研究基础上，沿用相似系统的语义组块定义，将汉语句子里对特定谓词的论元识别作为语义组块的识别，

通过分别描述普通汉语句子的识别情况和实验语料中的识别情况，详尽的剖析了语义组块识别的整个流程，为后续的研究做好铺垫。

在对本文所研究内容的有效性验证方面，本章给出了领域内广泛应用的评测标准，详细介绍了语义组块识别的语料资源，阐述了本文实验所使用语料库的背景，展示了语料库中的文本结构，对汉语句子中各语义成分的功能标记在语义组块识别中的应用做出了说明。本章是进行语义组块分析的必要前提。

第3章 语义组块识别

3.1 语义组块标注方式

3.1.1 传统标注方式

在自然语言处理中，将句子进行划分，依研究目的不同进行标记，是文本分析的必要步骤。这是一种有效且相对容易处理的预处理方式，在此基础上可以进一步的对句子进行分析，使用详细的标记来描述切分出来的句子成分。

对于组块切分来说，就是在组块分析体系下，将句子分成不重叠的部分，给各部分标注不同的标识，这些标识的标注形式一般分为人工标注、半自动标注和自动标注三种。人工标注耗费大量人力物力，标注的结果不断校正，得到接近完全正确的标注结果；自动标注是使用标注工具，这些工具一般是基于统计学习理论开发的，并且随着理论的发展，标注工具也随之改进；半自动标注介于两者之间，先进行自动标注，再对结果加以人工纠偏，是一种有指导的自动标注方式。

基于统计的自然语言处理问题一般都是在人工标注的语料库上进行研究，这是因为人工标注的语料库经过多年的完善，具有极高的精确度，在建立统计语言模型的时候不会受语料来源的影响，从而产生除算法之外的误差。在具体的研究中，将语料库的训练集中人工标注的语料进行处理，采用不同的策略建立统计语言模型，然后在测试集中把去除真实标注结果外的部分作为输入，统计模型会预测出测试集数据的标注结果，这是一个自动标注的过程。最后将自动标注的结果与测试集中人工标注的真实结果做比较，计算出系统各性能指标。

无论是自动标注还是人工标注，在同一个问题中使用的标注表示方法是相同的。不同的标注表示方法在不同的研究中对研究结果会产生有差别的影响，同时标注的结果也是在实际实验中的最终指标。

在语义组块识别中，词语的组块标识代表着它是否属于语义组块，可见，词语的标记过程就是组块的识别过程，因此，采用何种标注表示方法，是语义组块识别阶段需要着重考虑的问题。

语义标注是句子中词语序列到标注集合的一个映射问题，标注集合里面的标识是由标注方式决定的。传统的组块标注方式一般采取基于 Ramshaw[52]提出的 IOB 表示法，这种表示法的初衷是用来标记句子中的基本名词短语（base NP），base NP 的识别问题是组块识别问题的雏形。Sang 等[53]对 IOB 表示法进行了改造，提出了另外三种标注表示法。这四种标注方式分别为：IOB1、IOB2、IOE1

和 IOE2。

在语义组块识别的应用中，基于 IOB 的标注方式有一个共同点，即标识“I”表示标注的词语包含在一个语义组块中，标识“O”表示标注的词语不在任何语义组块中。具体到每种表示法，又有如下区别：

（1）IOB1 中的“B”表示当前词语是一个语义组块的起始词，与前一语义组块临界（前面的词语标识为“I”）。前面的词语不是一个语义组块的成分时，语义组块起始词仍用“I”表示（前面的词语标识为“O”）。

（2）IOB2 的“B”表示当前词语是一个语义组块的起始词，且不考虑前词的标识，组块以“B”开始，以“I”结尾。

（3）IOE1 与 IOB1 类似，只是界定标识由“B”标识语义组块开始改为由“E”标识语义组块末尾。

（4）IOE2 与 IOB2 类似，“E”标识语义组块末尾且不考虑后词的标识，组块以“I”开始，以“E”结尾。

在上述四种标注法之外，Uchimotoetal[54]提出了更详细的 start/end 标注法，共有五种标识：B、I、E、O 和 S。此标注法应用到语义组块识别上有如下解释：“B”表示当前词语是一个语义组块的起始词，“I”表示当前词语包含在一个语义组块中，“E”表示当前词语是一个语义组块的终结词，“O”表示当前词语不在任何语义组块中，“S”表示当前词语是一个语义组块且此组块仅有一个词。

自然语言的一个句子中词语的顺序是固定的，在同样的排序下，同一个词语在不同的语义组块标注法下得到的组块标记可能是不同的。在 IOB1 和 IOB2 标注法下，语义组块的起始词被标注了与同一组块其它词语不同的标记；在 IOE1 和 IOE2 标注法下，语义组块的终结词被标注了与同一组块其它词语不同的标记；在 start/end 标注法下，同一组块的各部分词语之间得到的标记差异更大。这种差异在统计模型的建立过程中，会随着统计学习方法的不同，得到不同的实验结果。

3.1.2 IO 标注法

经由对传统标注方式的分析和研究，本文提出了一种全新的标注方式：IO 标注法，将其运用到语义组块的识别中，并与相似系统在以上五种传统标注方式下实验效果最好的一种做出比较。

IO 标注法将语义组块的标识分为两类：字母“I”或字母“O”。“I”代表当前词语在一个语义组块内部，“O”代表当前词语不在任何语义组块内。

本文与其它语义组块分析系统一样，使用 CPB 作为实验语料库，CPB 作为被广泛使用的语义资源，有其独特的风格[55]。该语料库对于句子中论元的标注是非叠加的，即句子中任两个语义组块不存在重叠的部分。由此特性结合 IO 标

注法有如下分析：在语料库中，词语序列组成汉语句子，这些词语的语义组块标记由一串连续的“I”或“O”组成。

这种标注方式会带来一个问题：当句子中两个语义组块紧邻时，无法区分两个组块的边界，例如，一个语义组块的起始词标记为“I”，它的前一个词是另一个语义组块的末尾，组块标记也为“I”。而传统的五种标注法对于类似情况均采取了特殊处理方式，如引进“B”表示组块的开始、引进“E”表示组块的结尾等。

表面上看 IO 标注法无法识别语义组块的边界，然而，从另外一个角度分析，语义标注一般分为语义组块识别和语义组块分类两个阶段，在语义组块分类阶段，要对识别出来的语义组块按语义成分进行归类，不同的谓词-论元结构决定了不属于同一语义组块的词语很难归为同一类（汉语的句子成分结构顺序中，主语一般在谓语前面，谓语在宾语前面，句首状语在全句前面，不同成分词语的语义差别比较明显）。再结合汉语宾州命题库 CPB 论元标注非重叠的特性，可以将辨识组块边界的任务交给分类阶段。

IO 标注法在语义组块识别阶段的优势正在于此，采取了更少的标识，避免过多的标记对统计模型的影响，从而显著地提升了语义组块识别的召回率和 F1 值。五种传统标注法与本文标注法的对比示意图如图 3.1（句子为 CPB1.0 文件 0034-bai-chu.xml）。

	欧元	**诞生**	**在即**	**，**	**法国**	**许多**	**书店**	**、**	**书摊**	**和**	**报摊**	**纷纷**	**摆出**	**以**	**欧元**	**为**	**主题**	**的**	**书籍**	**。**
IOB1	I	I	I	O	I	I	I	I	I	I	I	B	O	I	I	I	I	I	I	O
IOB2	B	I	I	O	B	I	I	I	I	I	I	B	O	B	I	I	I	I	I	O
IOE1	I	I	I	O	I	I	I	I	I	I	E	I	O	I	I	I	I	I	I	O
IOE2	I	I	E	O	I	I	I	I	I	I	E	E	O	I	I	I	I	I	E	O
start/end	B	I	E	O	B	I	I	I	I	I	E	S	O	B	I	I	I	I	E	O
IO	I	I	I	O	I	I	I	I	I	I	I	I	O	I	I	I	I	I	I	O

图 3.1 不同标注法的比较

组块的标注方式和其它一些自然语言处理技术是一样的，例如命名实体识别（named entity recognition）、语义角色标注等，与传统的五种标注法一样，IO 标注法也可以运用到这些问题上。因此，本文提出的 IO 标注法不但有助于语义组块分析的发展，对上述自然语言处理的相关任务来说也不无裨益。

在语义组块识别的应用中，IO 标注法经过实验证明是目前最好的标注方式，当然，目前并没有研究表明哪种标注法更好，上述几种标注方法在研究中都有其适用领域。

3.2 基于 CRF 的语义组块识别方法

3.2.1 CRF 模型简介

统计语言模型的建立在自然语言处理系统中是非常重要的，模型在特定问题上表现的优劣直接影响了整个系统的性能。得益于统计机器学习新的理论和方法的不断涌现，自然语言处理在建立模型时有了更多的选择，基于条件概率无向图的 CRF 模型在自然语言处理中使用较为广泛。

CRF 模型是一个用于预测的统计模型，主要用来标注有序列性结构的数据，常用于词性标注、句法分析等问题。本文经过研究和尝试，考虑 CRF 模型在序列标注上的优势，将其应用到语义组块识别阶段。

具体来说，对于给定的输入序列 X 和输出序列 Y，CRF 通过定义条件概率 P（$Y \mid X$）来描述模型，这个概率定义如下：

$$\exp(\sum_{j} \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_{k} \mu_k s_k(y_i, X, i)) \qquad (3.1)$$

公式 3.1 中 $t_j(y_{i-1}, y_i, X, i)$ 是转移函数，表示对于输入序列 X，输出序列在 i 及 $i-1$ 位置上标记的转移概率。$s_k(y_i, X, i))$ 是状态函数，表示对于输入序列 X，输出序列在 i 位置的标记概率。两个函数前面的参数 λ 和 μ 分别是函数的权重，是在训练集中估计出来的。出于描述方便的考虑，转移函数和状态函数统一表示如下：

$$F_j(Y, X) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, X, i) \qquad (3.2)$$

每个特征函数 $f_j(y_{i-1}, y_i, X, i)$ 表示状态特征或转移特征。综合以上，CRF 的条件概率定义如下：

$$P(Y \mid X, \lambda) = \frac{1}{Z(X)} \exp(\lambda_j \bullet F_j(Y, X)) \qquad (3.3)$$

式（3.3）中，$Z(X)$ 是规范化因子：

$$Z(X) = \sum_{Y} \exp(\lambda_j \bullet F_j(Y, X)) \qquad (3.4)$$

对于语义组块识别来说，这里的变量 Y 表示输出的标注序列，变量 x 表示需要标注的观测序列。在训练模型时，利用训练集通过极大似然估计得到条件概率的模型 P（$Y \mid X$），在对模型进行测试时，对于测试集中给定的输入序列，求出条件概率最大的输出序列。

3.2.2 特征抽取与特征模板

建立的基于 CRF 的统计语言模型能否有效的对语义组块进行预测和标注，

特征的选取很重要。一般来说，特征的种类越多，模型从训练集中学习到的“知识”越多。然而，特征太多有时候会导致系统太复杂，而且一些相关性小的特征反而会降低系统的性能。

因此，在现有研究的基础上，结合本文所使用的语料库 CPB 的特点，经过反复实验，从 CPB 中抽取以下特征用于基于 CRF 模型的语义组块识别：词、词性、是否目标谓词、词性序列的组合。其中，词的特征分为-1、0、+1，分别表示前一个词、当前词语和后一个词。词性的特征分为-2、-1、0、+1、+2，依次表示前两个词的词性、当前词语的词性和后两个词的词性。是否目标谓词分为-1、0、+1，分别表示前一个词、当前词语和后一个词是否为目标谓词，并标为 Y\N（是\否）。词性序列的组合选取的稍复杂一些，从前三个词的词性到后三个词的词性都有涉及。

本文使用开源工具包 CRF++作为 CRF 模型的实现，该工具包使用 LBFGS（Limited-Memory Quasi-Newton Method）算法对 CRF 模型进行了优化。CRF++有两类特征模板：Unigram 和 Bigram，分别描述一元特征和二元特征。其中 Unigram 可模拟 Bigram，因此，实验中使用 Unigram 来实现抽取的特征组合。模板文件中每一行是一个模板，每个模板由形如%X[row，col]的样式来指定语料的特征，在按行处理输入数据时，row 和 col 分别控制行偏移量和列位置，例如，%X[-2，2]代表取的是当前行再前 2 行、2 列的特征。形如%X[row，col]/%X[row，col]之类的模板是用 Unigram 一元特征来模拟二元特征的。

具体的实验参数设置如下：考虑 Unigram 可模拟 Bigram，统一使用 Unigram 特征模板；为了过滤掉只出现了一次的特征，频率参数 f 设置为 2；为了防止统计机器学习方法中常出现的过拟合问题，拟合参数设置为 5。在 CRF++可选的规范化算法中，使用缺省的 CRFL2 规范化算法。

具体实验结果在本章的实验分析部分给出。

3.3 基于 SVM 的语义组块识别方法

3.3.1 SVM 模型简介

语义组块的识别不仅可以作为对词语进行 IO 标记的序列标注问题，也可以作为将词语按 IO 标记进行二元分类的问题。在现有的统计学习理论中，SVM 模型在分类问题中表现优异，可作为合适的分类器进行语义组块识别。

SVM 模型是在高维特征空间使用线性函数假设的学习系统[56]，广泛应用于文本分类、信息过滤、短语识别和词义消歧等自然语言领域。SVM 模型本质上是一种二类分类模型，在分类方面具有良好的性能，基本思想是在特征空间中找一个决策平面，该平面可以最优地分隔两个分类中的数据。SVM 模型可构建从

简单到复杂的模型：线性可分 SVM 模型、线性 SVM 模型以及非线性 SVM 模型，简单模型是复杂模型的特殊情况。

拿线性可分 SVM 模型举例来说，假设给定特征向量空间上的训练集 $T=\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_i, Y_i)\}$，其中 $X \in R^n$，$Y_i \in \{+1, -1\}$，$i=1, 2, \cdots, n$，X_i 是第 i 个特征向量，Y_i 是 X_i 的类标记，当 $Y_i=+1$ 时，称 X_i 是正类，当 $Y_i=-1$ 时称 X_i 是负类，（X_i，Y_i）称为样本点。SVM 模型在特征空间找到决策平面，将特征向量分到不同类中。决策平面可由方程 $w \bullet x+b=0$ 表示，由法向量 w 和截距 b 决定，它将特征向量空间分为两部分，一部分为正类，一部分为负类。

如图 3.2 所示，上方虚线上的两个样本点 X 和下方虚线上的样本点 O 组成了一个决策平面，将数据样本分成了两类。当然，在实际应用中并不会如此完美的将两类间隔开，而是利用各种策略寻找一个最优化的决策平面。虚线上的三个样本点就是 SVM 模型的支持向量（support vector，SV），起着“支撑”决策平面的作用。

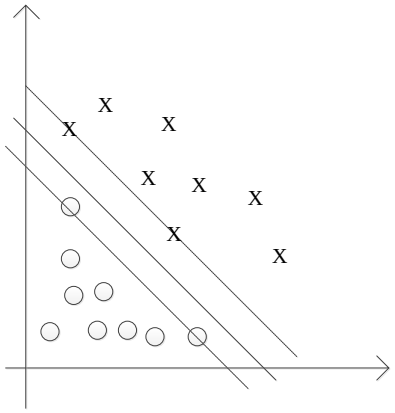

图 3.2 二元分类

关于 SVM 模型有很多相关的研究，本文使用 SVM 模型主要考虑其二类分类的特性，结合本文提出的采用两个标识来标记语义组块的 IO 标注法，在语义组块识别阶段可发挥较大的优势。

3.3.2 标注模型的构造

在基于 CRF 模型的语义组块识别方法中，将语义组块标记的识别作为序列标注问题来处理，而基于 SVM 模型的语义组块识别方法则将语义组块标记的识别作为分类问题进行处理，思路的转变提供了不同角度的研究方法。在 SVM 模型中，特征的选取依然是非常重要的。

基于 SVM 模型的语义组块识别需要解决三个问题：特征选取、参数训练和测试。在统计语言模型的训练阶段，选取词、词性标注、当前词语之前 n 个词的组块标记来对句子中第 i 个词进行组块标记，这是一个模型的“学习”过程，为

了将这个过程简单化的描述，可参见表 3.1。

其中，w_i 是句子中第 i 个词，t_i 是 w_i 的词性标注，c_i 是 w_i 的组块标记。在这里，采取前向分析法，即对于当前词语 w_i，用 w_{i-2}、w_{i-1} 的标记 c_{i-2}、c_{i-1} 来分析，对应的，有后向分析法。这些候选标记的选择就是需要解决的第一个问题：特征选取。

表 3.1 特征列表

词	词性标注	语义组块标记
w_{i-2}	t_{i-2}	c_{i-2}
w_{i-1}	t_{i-1}	c_{i-1}
w_i	t_i	$\boldsymbol{c_i}$
w_{i+1}	t_{i+1}	
w_{i+2}	t_{i+2}	

在表 3.1 中，对于当前词语 w_i 来说，表格里选取的这些特征可以组成一个窗口，如图 3.3 所示：

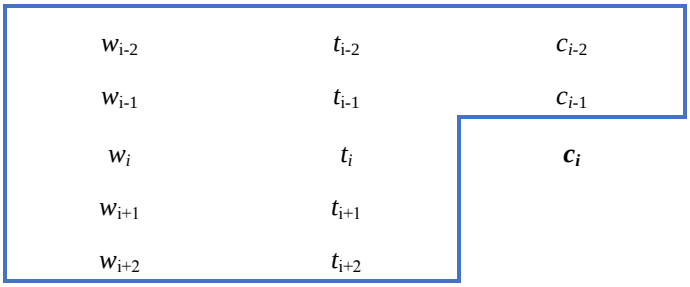

图 3.3 特征窗口

在实验阶段，需要进行反复的实验并对训练结果进行测试，选择最优的窗口大小。窗口大小的调整是基于 SVM 模型的语义组块识别问题中参数选择的一部分，这个过程需要不断对训练出来的模型进行测试，也就是需要解决的后两个问题：参数选择和测试。

以上是对 SVM 模型在语义组块识别问题中具体工作流程的模拟，在具体的实验中，本文使用 TinySVM 作为 SVM 模型的具体实现，选用 Kudo 等[57]开发的基于 TinySVM 的 YamCha 开源工具包进行实验，经过实验调优，上述窗口参数 window-size 设置为 6，核函数参数 d 设置为 2.

SVM 模型下的语义组块识别实验结果在本章的后续小节给出。

3.4 对比系统

基于句法分析的论元标注研究一般要经过从句法树剪出候选论元的步骤，而语义组块分析的方法并不需要这个阶段。在针对论元标注的语义组块分析方面，对比系统将语义组块的识别作为一个序列标注问题，选择基于 CRF 模型的方法，分别研究了语义组块识别和分类阶段所需要的特征，并给出了在不同的标注方式下语义组块识别的系统性能。

出于对实验结果进行比较的目的，本文使用 CPB1.0 作为数据，该数据标记的是 CTB5.0（Chinese Treebank）从标号 chtb_001 到 chtb_091 的文件，实验数据与上述文献相同。

在对比系统的语料设置里，这些数据被划分为三部分：训练集语料为 chtb_081—chtb_899，测试集语料为 chtb_001—chtb_040 以及 chtb_900—chtb931，开发集语料为 chtb_041—chtb_080。由于 CPB1.0 的格式问题，本文数据设置与这个语料设置相似，但不是尽然相同，在训练集和测试集的比例上是相同的（9 比 1），并且有部分是完全相同的。从数据的随机性来看，这样的数据设置是合理且在某种程度上更有效的。

对比系统的实验证明了不同的标注方式对于语义组块识别的影响，并且在 start/end 标注方式下取得了在传统的标注方式中最好的实验结果，本文将新的标注方式 IO 标注法应用到语义组块识别阶段，从而得到了更高的召回率和 F1 值。其中，本系统的召回率提升非常显著，从而拉高了衡量系统整体性能的 F1 值。

3.5 实验结果及讨论

语义组块识别使用准确率 Precision、召回率 Recall 和 F-Score 来评价系统的性能，在后续的语义组块分类阶段，由于语义组块已被正确的识别，仅使用分类的精确率 Accuracy 就可以评测分类的效果[58]，因此，语义组块识别阶段在语义组块研究中起着重要的作用。

对于语义组块的识别，本文使用基于 CRF 模型的序列标注策略和基于 SVM 模型的分类策略，结合 IO 标注法，得到的实验结果如表 3.2 所示：

表 3.2 基于 I0 标注法的实验结果

	Precision	Recall	F-Score
CRF模型	73.79%	71.93%	72.85%
SVM模型	78.78%	81.87%	80.30%

可以看到基于 CRF 模型的方法准确率高于召回率，表现出了其在序列标注上的优势，同时，实验结果也揭示了在 IO 标注法下基于 SVM 模型的识别方法远远优于基于 CRF 模型的识别方法，具体表现为在召回率上基于 SVM 模型的方

法比基于 CRF 模型的方法高了 9.94%，显著地提升了系统的性能。

究其原因， SVM 本身是一种二类分类模型，它的基本原理是在特征空间上将正负类数据样本最优化分隔，在分类特别是二元分类问题上表现优异，IO 标注法仅使用两种标识，最大限度利用了 SVM 模型的这个特性。而 CRF 作为一种序列标注模型，在组块标识种类较多的应用中更能发挥优势，对比系统的实验验证了这一特性，该系统在 start/end 标注法下，使用 CRF 模型，得到了比基于 IOB 的四种标注方法更好的实验结果（这四种标注法只有三个类型的组块标识，start/end 标注法有五个）。表 3.3 给出了在传统的 start/end 标注法下本文与对比系统的实验结果。

表 3.3 start/end 标注法下实验结果比较

	Precision	Recall	F-Score
文献[21]	81.00%	73.58%	77.11%
本文	65.73%	63.97%	64.84%

在具有五种组块标识的 start/end 标注法下，基于 SVM 模型的识别效果明显不如对比系统。而在具有三种组块标识的 IOB 标注法下，实验结果差距就缩小了，以广泛使用的 IOB2 标注法为例，实验结果对比如表 3.4 所示：

表 3.4 IOB2 标注法下实验结果比较

	Precision	Recall	F-Score
文献[21]	78.45%	71.48%	74.80%
本文	70.75%	71.23%	70.99%

表 3.3 和表 3.4 证实了前面的推断，当组块标识种类减少时，基于 SVM 模型的组块识别方法更有效。由前面的分析可知，IO 标注法的引入可以合理利用 SVM 模型二类分类的本质，引入后的实验结果也证实了本文基于 SVM 模型的语义组块识别结合 IO 标注法可以取得更好的识别效果，本文方法与对比系统最好实验结果的比较如表 3.5 所示：

表 3.5 本文方法与对比系统的实验结果比较

	Precision	Recall	F-Score
文献[21]	81.00%	73.58%	77.11%
本文	78.78%	**81.87%**	**80.30%**

表 3.5 中文献[21]的实验结果是在 start/end 标注法下取得的，基于其它四种传统标注法的实验结果在三个指标中都低于该结果，可以清晰地看出，对比系统在准确率上维持在 80%以上的水准，比本系统略高 2.22%。而对于本系统来说，IO 标注法的优势又一次体现在召回率方面，召回率显著地提升了 8.29%，从而进一步地提高了 F1 值，使 F1 值提升了 3.19%，使其达到了 80%以上，验证了 IO 标注法在语义组块识别阶段的有效性。

值得注意的是，本文合理地利用了 SVM 模型的特点和 IO 标注法的优点，在提取的特征中仅使用了词、词性、位置信息和是否目标谓词，使得本文的方法在处理汉语文本语料中具有较强的泛化能力，相较之下，对比系统采取了更多的特征。在统计机器学习中，如果对语料的特征选取过于复杂，可能出现训练模型对已知语料的预测能力很好而对未知语料的预测能力很差的情况，该情况称之为过拟合。简单来讲，就是只在和训练语料结构一致的数据上才能得到较好的预测结果。本文所使用的四个特征在汉语文本中都是普遍存在的，因此，本文所用方法对特征的合理提取不仅具备更好的普适性，而且可以有效地避免了训练模型过拟合的问题。

本文的语义组块识别系统仍然有改进的空间。从对比系统的性能可以看出，基于 SVM 模型的语义组块识别方法在分类中“牺牲”了一定的准确率，这个代价是值得的，换来的是 F1 值高达 3 个百分点以上的提升，而 F1 值是衡量系统整体性能的最重要的指标。但是，这说明了本文在准确率这个指标上还有进一步提高的可能性。

3.6 本章小结

本章详细介绍了语义组块识别的研究方法，提出了一种全新的标注方式，举例说明了新的 IO 标注法与传统的 IOB1、IOB2、IOE1、IOE2 和 start/end 标注方式的异同处，阐述了 IO 标注法的优势并将其运用到语义组块识别中。

结合新的标注方式，本章使用统计机器学习方法 CRF 和 SVM 建立统计语言模型，对模型的原理做了简要的介绍，分析了两种模型在语义组块识别中所需要的特征。基于 CRF 模型的识别方法将语义组块识别作为序列标注的问题，用条件概率来预测句子中词语的组块标记；基于 SVM 模型的识别方法将语义组块识别作为分类问题，通过决策平面来对句子中词语进行组块标记的分类。最终的实验结果表明，在 IO 标注法下，基于 SVM 模型的语义组块识别方法同时发挥了标注法和模型的优势，取得的系统性能比基于 CRF 模型的方法更好。随后，将本文的系统与相似系统做出对比，进一步验证了本系统的性能。在本章的最后，分析了实验中可以改进的地方，为未来的工作提供了研究方向。

第4章 语义分析技术与语义组块识别

4.1 语义信息在语义组块识别中的应用

4.1.1 语义分析

利用计算机对人类通过自然语言所传达的信息进行不同形式的处理，使计算机能够理解人类的自然语言，一直是人工智能领域的分支自然语言处理追求的目标[59]。人类之间进行信息传递一般依靠自然语言的口头或者书面形式，这些信息经由词语组成单个的句子，这些完整的句子包含丰富的信息，也是理解自然语言的基本单元。

在自然语言的句子中，词语之间的联系主要被句法和语义所约束，而针对语义方面的研究无疑是“理解”句子意义的重要手段[60]。语义分析可以对句子进行深入的知识获取推理，揭示句子的含义，理解句子的意思。语义分析技术在多个领域都有相关应用背景，和组块分析也有着紧密的关联。

在结合语义分析的组块研究中，语义标注是需要研究的内容，语义分析有别于语法分析，是一种推导出能够反映句子意义表现形式的研究技术，是自然语言处理领域研究的基本问题之一，而语义组块分析的目的是为了更准确的标注出句子中的语义成分，最终是为了将自然语言从机器可读提升到机器可理解的程度。

在以往的很多语义标注研究中，大多从语料库中抽取特征进行研究，构造的特征集合经由处理语料库中文本生成。抛开研究手段，单从目的来看，语义组块的识别和其它语义标注任务一样，是为了更有效的标注出句子的语义成分。而对于每个词来说，有其固定的语义信息。因此，本文考虑将语义组块的识别加入词语的语义特征，以期得到更好的识别效果。

4.1.2 语义组块识别的语义信息选取

语义组块识别所使用的语料库本身已含有语义信息，由谓词-论元结构所表示，因此，新引入的语义信息必须以某种合适的形式出现在特征构造阶段，不然有破坏原来语义信息的可能性。所以，有必要对各种汉语语义资源做出分析，找出最适合语义组块识别的语义资源。

从以上的分析不难看出，在已标注了语义关系的语料库中，新引入的语义资源应该具有表征单个词语语义信息的功能，才不至于破坏已有的语义结构。从这个角度来看，将语料库中最小的单位（这里是已切分好的词）的语义信息加入特征提取阶段，可以做到在特征里面“添加”上语义信息而不影响已有的特征。具

体来说，就是使用某种语义资源，将其中词语的语义信息合理的运用到训练集和测试集中，以实现新的语义特征构造。

目前的汉语语义资源有 HowNet、《同义词词林》等。HowNet 揭示了概念的共性和个性以及概念之间的各种联系，描述了词汇之间的各个层面的语义关系，使用多种的词汇关系和语义关系来表示词汇知识，包含了 17000 多个词义概念，这些概念是由若干与概念有关的义原组成的，是一个按语义关系网络组织的知识系统。《同义词词林》是自然语言处理领域中著名的语义资源，由梅家驹等[61]编制而成，不仅包括了词语的同义词，而且包含了这个词语广义的相关词，当中的每个词语都有对应的语义编码，可以看作是一个语义词典。

对上述语义资源的分析得知，《同义词词林》针对每个特定的词语，有唯一的语义编码，在本文实验的语料库 CPB 中，每个最小的单位恰恰是切分过的词语。在本文第 2 章中，介绍过 CPB 中句子的文本结构，这些句子是一种句法树库的结构，而切分好的词语在句法树中是叶子结点。将语义编码的信息添加到叶子结点上并不会改变整个句法树，因此可考虑将其应用到语义组块识别中，从而实现有效的加入新语义特征而不改变原有的特征结构。

4.2 基于语义信息的语义组块识别原理

4.2.1 同义词词林

《同义词词林》里的每个词语对应唯一的语义编码形式，是一个具有词语和语义映射关系的语义词典，主要通过同义词的集合来表示词语的语义信息。由于原版《同义词词林》著作时间久远，有很多新词没有加入，而一些旧的词语随着汉语的演变又成为了生僻字，在此情况下，哈工大信息检索研究室将其改进为《同义词词林扩展版》，对原版的词语进行了删减。

《同义词词林》收录的词语为 53859 条，《同义词词林扩展版》参考现有的词典资源，对比人民日报语料库中词语出现的频次，去除掉了 14706 个在实际中几乎不再应用的词语，保留了原有的 39099 条词语。同时，《同义词词林扩展版》也进行了扩充的工作，投入了大量的人力和物力，最终的词典中收录了 77343 条词语。规模的扩大意味着在对汉语进行基于统计的自然语言处理研究时，从中得的语义信息更加丰富。

《同义词词林扩展版》除了针对词语本身进行了改进外，整个词典的文本结构与《同义词词林》是相同的，为了描述方便，本文将两者统一称之为《同义词词林》。

在《同义词词林》中，按照树状的层次结构将汉语词语归为大、中、小三个类别。其中，大类有 12 个，中类有 97 个，小类有 1400 个。小类的下面有很多

词语，这些词语又根据词语的语义远近和关联性划分成若干词群，词群进一步的细分为若干项，属于同一项的词语有两种关系：词义相同、有比较强的关联性。这样，在大中小三类下面又分出了两个级别，可以看成一共有五个级别的分类，表征五层结构。

对于单个的词语来说，在《同义词词林》中有对应的 8 位编码，包含了该词语在五个级别的从属信息。举例说明如图 4.1：

Aa01B02= 群众 大众 公众 民众 万众 众生 千夫
Ba02B15# 合格品 一级品 真品 正品 佳品
Ca12C04@ 善后

图 4.1 《同义词词林》举例

在上述例子中，8 位的编码对应的是同义或者语义相关的词语，这些词语都有唯一的一个 8 位编码，按照顺序依次表示的含义为：第 1 位表示词语所属的第一级别的分类标识；第 2 位表示词语所属的第二级别的分类标识；第 3 位和第 4 位用数字组成的两位数来表示词语所属的第三级别的分类标识；第 5 位表示第四级别的分类标识；第 6 位和第 7 位是两位整数，表示第五级别的分类标识，是五层结构里面的最小单元。8 位编码的最后一位是五个级别之外的标识，这是因为第五个级别的分类结果需要额外的表明，正如例子中所示，有的是同义词的集合，有的是关联词的集合，有的只有单独的一个词。基于使用的考虑，需要给予这三种情况有区别的标识，分别是“=”、“#”和“@”。其中，“=”表示“相等”、“同义”；“#”表示“不等”、“同类”；“=”表示“独立”，在《同义词词林》中没有与其在语义上有任何关联的词语。

词语编码规则如表 4.1 所示：

表 4.1 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	A	a	0	1	B	0	2	=(#\@)
符号性质	大类	中类	小类		词群	原子类		
级别	一	二	三		四	五		

4.2.2 语义特征对语义组块识别的影响

在本文前面的研究中，基于 CRF 模型的语义组块识别方法将语义标注作为序列标注的方式来处理，主要的优点在于其按条件概率转移状态的特性。在一些统计模型中，需要将观测对象与其它对象进行独立性的假设，而 CRF 模型计算

的是整体的联合概率分布分解成的若干子集的条件概率分布，不但避免了这一缺点，还可以有效的利用序列的上下文信息。

在训练过程中，CRF 模型只需要考虑当下已经出现的观测序列的状态即可，并没有严格分类的独立性要求，可以有效地利用整个序列的特征。加入的语义信息给整个序列加入了新的观测点，而这些信息是从《同义词词林》中抽取出来的，句子中的同义词和相关词的在这些观测点的语义编码是一致的，对于从句子的语义分析是有帮助的，因此可知如果在输入序列中加入新的语义特征，能够提高语义组块识别的效果。

SVM 模型的基本原理是求出可以正确划分训练集且使几何间隔最大化的决策平面。对于训练集中的样本语料来说，决策平面可以有无穷多个，但确定最大几何间隔的决策平面只有一个。这意味着，决策平面以最优解的方式将训练集分类时，不仅要能够分出明显是正类或负类的样本点，也要分出距离决策平面最近的样本点。当唯一的决策平面确定后，平面边界作为支持向量的数据点也就确定了。

SVM 模型的这种特点决定了在通过训练确定了决策平面后只有支持向量起作用，其它的数据并不起作用，而这些支持向量是由一些很“重要的”训练样本确定的。因此，基于 SVM 模型的语义组块识别方法将语义标注作为分类问题时，再加入新的语义特征，如果对支持向量的影响不大，那么对整个识别系统性能的影响可能是很细微的。

下一节的实验证明了本节的分析。

4.3 实验及讨论

4.3.1 实验设计

本实验的目的是为了验证语义特征的有效性，基于可比性的考量，实验语料与没有加入语义特征的语义组块识别系统是一样的，除语义特征外的其它特征也是相同的，其它诸如参数选择等也是如此。《同义词词林》并没有覆盖所有的汉语词汇，例如汉语句子中有些成分如人名、地点等并不含有语义信息，这些成分的语义特征在实验中统一设置为 0。

将基于语义信息的语义组块识别系统区别于上一章的语义组块识别系统的原因是，前人的研究中大多都在实验语料库中提取信息、构造特征，可以视为在语料库内部利用句法结构、语义标记和语法信息进行研究，是一个“封闭”的体系。本文利用语义词典《同义词词林》，将外部“开放”的资源加入原体系中，并且做到了只引入新特征而没有破坏原体系的整体结构，不失为一种创新的做法。

表 4.2 是在语料中引入语义特征后的例子：

表 4.2 带语义特征的语义组块标注举例

词	其它特征	语义特征	语义组块标记
麦克马拉曼	…	0	I
打扮	…	Dc03D01=	0
得	…	Gc02C01=	0
像	…	Dk32A01=	I
个	…	Dn08A20=	I
牧师	…	Am03C02#	I

在上表中，“麦克马拉曼”是人名，语义特征为 0；“打扮”的最后一位语义编码为“=”，表示该词在同义词词林中有同义词，查《同义词词林》可知，“打扮”的同义词有“装束”、“妆饰”和“扮相”；“牧师”的最后一位语义编码为“#”，表示该词在同义词词林中有同类的相关词，分别是“传教士”、“教士”和“使徒”。

可以看到，同义词之间的语义和词性关联性很强，相关词之间也有较强的联系。在训练模型的过程中，这些扩展到训练语料中的语义信息，能够有效的描述词语在句子中扮演的语义角色，从而改进统计语言模型，使其在语义组块识别阶段发挥更好的作用。

4.3.2 实验结果及分析

表 4.3 中展示了加入语义特征后的语义组块识别系统在不同模型下的变化，这个实验结果符合前面对 CRF 模型和 SVM 模型在新系统下表现的预期。

表 4.3 基于语义信息的语义组块识别结果对比

模型	语义特征	Precision	Recall	F-Score
CRF	未加入	73.79%	71.93%	72.85%
CRF	加入	**74.68%**	**72.47%**	**73.56%**
SVM	未加入	78.78%	81.87%	80.30%
SVM	加入	78.88%	81.49%	80.17%

可以看到，基于 CRF 模型的语义组块识别方法，在引入了外部的语义资源后，准确率提高了 0.89%，召回率提高了 0.54%，F1 值提高了 0.71%。这种提升在于新加入的语义特征使数据样本的序列特征更丰富，而 CRF 模型在训练序列性数据样本上具有一定的优势，在两者的有机结合下，基于 CRF 模型的语义组块识别系统充分利用了加入新的特征后的数据整体的序列特征，使得系统的三个

指标都有了一定程度的提升。

与之对应的是，SVM 模型在加入语义特征后，改变微乎其微，甚至将系统的性能降低了一点点。这说明了，基于 SVM 模型的语义组块识别系统对句子中的成分按 IO 标记进行二元的语义组块分类时，加入新的语义特征没有使通过训练得到的决策平面有太大的区别，也就是说这些语义特征在决策平面的支持向量中并不起决定性因素。

实验结果表明了，虽然基于 SVM 模型的语义组块识别方法在“封闭”的体系中取得了比基于 CRF 模型的方法更加优异的结果，然而，在体系“开放”添加新的资源方面，适应能力不如 CRF 模型。当然，这种结论是基于 IO 标注法下的语义组块识别系统得出的。

基于 CRF 模型的方法在加入语义特征后系统三个指标都得到了提升，虽然最终结果依然没有基于 SVM 模型的方法好，但是，在未来的研究中，如果基于 CRF 模型的方法在引进外部资源后系统性能继续提升，而基于 SVM 模型的方法依然保持不变的话，这种结果有改变的可能性。

本文的对比系统使用 CRF 模型在 start/end 标注法下取得了最好的实验结果，而本文使用 CRF 模型在 IO 标注法下得到的系统性能是低于对比系统的。这表示，基于 CRF 模型的语义组块识别方法不仅可以加入语义特征加以改进，还可以从其它标注方式的角度研究以提升系统性能。当然，实验的最终目的是为了更好地识别句子中的语义组块，虽然基于 CRF 模型的方法在加入语义特征后识别效果全面提升，但是，本文将 SVM 模型结合 IO 标注法发挥了二类分类的优势，该方法依然是目前最好的语义组块识别策略。为了进一步的比较说明，将本文的方法结合传统的五种标注法与对比系统做出比较，如表 4.4 所示：

表 4.4 五种传统标注法下的实验对比及 I0 标注法下本文实验结果

标注法	文献[21]			本文		
	Precision	Recall	F-Score	Precision	Recall	F-Score
IOB1	78.09%	69.49%	73.54%	74.14%	74.02%	74.08%
IOB2	78.45%	71.48%	74.80%	70.75%	71.23%	70.99%
IOE1	77.53%	70.42%	73.81%	74.14%	71.50%	72.80%
IOE2	79.53%	72.30%	75.74%	70.23%	71.26%	70.74%
start/end	81.00%	73.58%	77.11%	65.73%	63.97%	64.84%
IO				78.78%	**81.87%**	**80.30%**

表 4.4 揭示了本文方法在不同标注法下的规律。由于 IOB1 和 IOE1 标注法对组块的标记位置是固定的，即组块开头固定使用标识“B”或组块结尾固定使

用标识“E”，使得在这两种标注法下，本文方法与对比系统各有优劣。这说明了在组块标识从 start/end 标注法的五种减少为 IOB 式标注法的三种时，本文所使用方法性能开始得到提升。当组块标识减少为 IO 标注法的两种时，本文的方法得到了最好的组块识别效果，与对比系统在五种标识的 start/end 标注法下取得的最好结果对比，F1 值提升了 3.19%。

值得一提的是，在同一计算机上，两种模型的训练时间差异很大，加入语义特征后，基于 CRF 模型的语义组块识别方法仅需要 5 分钟 10 秒，而基于 SVM 模型的语义组块识别方法则长达 2 小时 38 分钟 15 秒。在加入语义特征前，这两个时间分别是 5 分钟 4 秒和 2 小时 34 分 10 秒。

在 CPB 这样规模的语料库中，基于 SVM 模型的语义组块识别方法耗费小时量级的时间代价，取得较好的系统性能是可以接受的。但可以预见的是，随着研究的发展，语料库资源会越来越丰富，规模也会越来越大，特别是近年来与计算机相关的应用领域都强调大数据的重要性，基于 CRF 模型的语义组块识别方法在适应能力以及训练的时间开销上的优势或许未来可以在语义组块识别阶段得到更好的发挥。

4.4 本章小结

已有的语义组块分析研究是在实验的语料库中“封闭”的处理，也就是说特征信息的抽取、模型的构造都是基于所使用的语料库，这样做一方面是保证研究方法的可对比性，另一方面是因为在已有的文本语料中添加外部的特征信息有可能导致中文本结构遭到破坏。

在此情况下，本章首先将《同义词词林》作为语义词典抽取语义特征，由于词典的映射特性，相当于给每个词语添加了类似词性标注的一个标识，加入单个词语的语义信息并不破坏实验语料的句子结构。同时，词典中语义编码和词语一对多的特性决定了语义相同或者相关的词语具有相同的语义标记，从而实现了引入外部语义资源而又不影响原有体系的效果。其次，本章结合所使用的机器学习算法，分析了在不同算法下引入外部语义信息可能对实验效果产生的影响，并在实验中得到了验证。在本章的最后，对于实验中发现的一些值得探究的点做出了剖析，为语义组块分析提供了新的研究角度。

第5章 总结与展望

5.1 本文总结

近年来，随着互联网的不断发展，人类生活愈来愈趋向信息化和智能化，自然语言处理是人工智能领域重要的信息处理技术，在搜索引擎、社交网络和智能系统等应用中随处可以见自然语言处理技术的身影。

统计学习理论的不断发展、计算机能力的提高和信息量的不断增加，使基于统计的自然语言处理研究得到了长足的发展。在基于统计的自然语言处理中，浅层的句法和语义分析是研究的重点，语义组块分析是其中的研究方向之一。

语义组块分析以句子为单位，通过对句子的语法结构进行研究，标记句子中的语义成分，分析句子的浅层信息。句子的语法结构揭示了句子中词语之间的关联方式，这种结构指明了词语如何组成短语、组块等颗粒度更大的句子成分。不过，句法的语法结构并不能反映出句子的意思，需要对句子中词语之间的语义关系进行分析。这也是语义组块分析的意义所在。

就本文具体的研究成果来说，主要包含以下几个方面：

（1）对汉语组块分析的各种描述体系进行研究和对比，考虑自然语言处理中的机器翻译问题，结合前人的研究基础，对汉语句子进行语义组块分析的研究。将句子按语义组块进行切分有助于表达句子内部语义相对独立的各成分，在机器翻译中，源语言的语义组块可对应目标语言语义相同的组块，是语义组块分析技术在实际领域的真实应用。

（2）提出了一种全新的标注方式：IO 标注法。本文对传统的五种标注法对组块识别的影响进行了研究，发现组块的标记方法在不同的统计学习理论下，会对系统的性能产生一定的作用。以此为切入点，本文提出了 IO 标注法，分析了 IO 标注法带来的问题，并提出了解决方案，从而在语义组块识别这一阶段，取得了比传统标注方式更好的效果。

（3）结合 IO 标注法和统计机器学习方法，分别将语义组块识别作为序列标注问题和二元分类问题进行研究。本文利用 CRF 模型在序列标注上的优势和 SVM 模型二类分类的本质来对语义组块识别建立统计语言模型，在语义组块分析方面研究者们常用的语料库 CPB 上，有效地利用了 CPB 的文本结构、IO 标注法和统计模型三者的契合点。经过实验，基于 SVM 模型的语义组块识别方法在 IO 标注法下，取得了 80.30%的 F1 值。目前在语义组块识别阶段的相关研究中，

使用相似的实验设置，还没有看到 80%以上的 F1 值，这也验证了本系统的有效性。

（4）基于进一步研究浅层句法和语义分析的目的，本文对语义组块识别系统进行了改造。原有语义组块识别系统以及其它对语义组块识别的研究，都是仅从语料库内部提取特征来建立统计语言模型的，考虑到语义组块识别对谓词-论元结构的分析本身就是一种语义分析的技术，本文将外部的语义资源加入现有系统进行研究。研究的结果表明，基于 CRF 模型的语义组块识别方法可以很好的融合进新的语义信息，虽然整体结果不如未加入语义信息的基于 SVM 模型的识别方法，但是为语义组块分析提供了新的研究方法和角度，具有一定的研究价值。

5.2 工作展望

本文的研究工作取得了一些阶段性的成果，但对于语义组块分析来说，还需要开展进一步的研究工作。

近年来自然语言处理相关技术快速发展，涌现了许多新的研究思路和方法，而自然语言处理自从研究伊始，就与人工智能的发展密切相关。随着互联网特别是移动互联网当下对人类生活的全面覆盖，人类社会产生信息量呈爆炸性增长，可以说，当今社会已处在一个大数据时代，海量的数据对人工智能中基于统计的机器学习方法的重要性不言而喻。

在大数据的影响下，语义分析、情感计算和文本蕴含等有着很强实际应用价值的技术成为当下自然语言处理领域的研究热点，受到学术界和工业界的广泛关注。语义组块分析作为自然语言处理中语义分析的关键技术，有着重要的研究价值，本文认为有以下几点是在未来的研究工作中需要关注的：

（1）随着机器翻译领域的发展，在实际的机器翻译系统中，使用统计机器学习方法一般不采用单一的统计模型，而是使用多种统计模型对系统进行融合，这种系统融合技术在机器翻译领的各种评价方法中都取得了较好的结果。该技术通过一定的策略，将多种统计模型得到的相似结果进行融合，抽取其中有用的信息，将其归纳为最终结果。由于基于统计的机器学习方法本质都是一样的，语义组块分析也可使用模型融合的技术，将如 CRF、SVM 等在语义组块分析中表现较好的模型分别进行实验，采用一定的机制将各模型的实验结果融合，可能会取得比现有系统更好的性能。

（2）深度学习（deep learning，DL）是目前人工智能领域研究的热点，并在语音识别、图像识别等领域取得了很好的效果。在自然语言处理领域，深度学习取得的成效不如语音、图像领域，但依然是有发展空间的，R.Collobert 等[62]将深度学习用于自然语言处理，在词性标注、组块分析、命名实体识别和语义角

色标注四个自然语言处理的基础任务中取得了不错的效果，最终实验结果略低于已有的研究。Google 于 2013 年开源了软件工具 word2vec，该工具在深度学习理论的基础上，将词语转换为向量的形式，将对词语的处理转变为对向量的计算，通过计算词语向量空间的相似度来表示词语的语义相关性。这些研究都是语义组块分析中值得探究的方向。

（3）文本研究使用的语料库是宾州命题库 CPB，第 4 章使用的语义资源是《同义词词林》，在基于统计的自然语言处理中，语料库是很重要的。可以预见的是，随着相关研究的进展，语料资源也会不断发展，语义组块分析技术在统计学习理论和语料库的共同发展下，有着良好的研究前景。

致 谢

时光匆匆如白驹过隙，转眼间在杭州电子科技大学的研究生生活即将结束。回首过去，心里唯有感恩，感恩自己遇到那么多可敬可爱的人。在毕业离别之际，我在此向所有曾经关心、帮助和支持我的人表示由衷的感谢和祝福。

首先我要向我的导师吴铤教授表示由衷的敬意，吴老师严谨的治学态度和豁达的心态令我每每想起不禁心生高山仰止之感，非常感恩在我的人生路上能遇到这样一位导师。

同时还要感谢王荣波老师对我的耐心培养，王老师在科研工作上对我进行了无微不至的指导，细致而又深入地教我如何做科研。在生活中，王老师也给予了我亲切的关怀，很感恩研究生期间王老师对我的教导。

由衷感谢黄孝喜老师、谌志群老师、姚金良老师、王小华老师、周昌乐老师、杨冰老师、朱文华老师、陆蓓老师和吴海虹老师，正因为有了他们的指导，我的科研工作才得以进展。

感谢实验室的周建成、吴腾飞、李杰、罗兰、张华、董瑜同学，他们陪我走过研究生生活的风风雨雨，让我体会到友谊的可贵，感谢同学们对我无私的帮助。同时还要感谢实验室的师兄们、师弟师妹们，让我在实验室度过的这段时光无比温暖。

感谢杭州电子科技大学提供了那么好的学习和生活环境，学校的一切都令我永难忘记。

最后我要感谢我的父母，感谢他们多年的养育之恩，在此祝福爸爸妈妈身体健康。

参考文献

[1] 宗成庆.统计自然语言处理[M]. 北京：清华大学出版社，2008：8-9.

[2] Manning D，Schutze H. Foundation of statistical natural language precessing[M]. MIT Press，1999：1-5.

[3] 周惠巍.模糊限制信息检测中融合方法的研究[D]. 大连：大连理工大学，2012.

[4] 冯志伟.自然语言处理中理性主义和经验主义的利弊得失[J]. 长江学术，2007，2：79-85.

[5] 俞士汶.计算语言学浅介[J]. 术语标准化与信息技术，2009，3：34-39.

[6] Abney S. Parsing by chunks[C]. Dordrecht：Kluwer Academic Publishers，1991：257-278.

[7] 周强.汉语基本块描述体系[J]. 中文信息学报，2007，21（3）：21-27.

[8] 李业刚，黄河燕.汉语组块分析研究综述[J]. 中文信息学报，2013，27（3）：1-8.

[9] 江铭虎.自然语言处理[M]. 北京：高等教育出版社，2006：1-3.

[10] 程崴，赵军，徐波.一种面向汉英口语翻译的双语语块处理方法[J]. 中文信息学报，2003，17（2）：21-27.

[11] Ahmed A A，Aly S. Appearance-based Arabic Sign Language recognition using Hidden Markov Models[C]. Engineering and Technology (ICET) 2014 International Conference：IEEE，2014：1-6.

[12] Sharma P，Sharma U， Kalita J. Named entity recognition in Assamese using CRFS and rules[C]. Asian Language Processing (IALP) 2014 International Conference：IEEE，2014：15-18.

[13] Fan S X， Chen L D，Wang X，et al. Shallow parsing with Hidden Markov Support Vector Machines[C]. Machine Learning and Cybernetics (ICMLC) 2014 International Conference：IEEE，2014：827-830.

[14] Yang Z，Li M，Zhu Z，et al. A maximum entropy based reordering model for Mongolian-Chinese SMT with morphological information[C]. Asian Language Processing (IALP)，2014 International Conference：IEEE，2014：175-178.

[15] Gong C，Li X，Wu X. Recurrent neural network language model with part-of-speech for Mandarin speech recognition[C]. Chinese Spoken Language

Processing (ISCSLP) 2014 9th International Symposium：IEEE，2014：459-463.
[16] 李珩，杨峰，朱靖波，姚天顺.基于增益的隐马尔科夫模型的文本组块分析[J]. 计算机科学，2004，31（2）：152-154,192.
[17] 周雅倩，郭以昆，黄萱菁.基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展，2003，40（3）：440-446.
[18] 黄德根，王莹莹.基于 SVM 的组块识别及其错误驱动学习方法[J]. 中文信息学报，2006，20（6）：17-24.
[19] 孙广路，郎非，薛一波.基于条件随机域和语义类的中文组块分析方法[J]. 哈尔滨工业大学学报，2011，43（7）：135-139.
[20] Ronan C，Jason W. Natural language processing（almost）from scratch[J]. The Journal of Machine Learning Research，2011：1-34.
[21] 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）：53-61,74.
[22] 刘海霞，黄德根.语义信息与 CRF 结合的汉语功能块自动识别[J]. 中文信息学报，2011，25（5）：53-56.
[23] 王鑫，孙薇薇，穗志方.基于浅层句法分析的中文语义角色标注研究[J]. 中文信息学报，2011，25（1）：116-122.
[24] 何赛克.语义角色标注中的关键技术研究-多任务学习方法在组块分析中的应用[D]. 北京：北京邮电大学，2009.
[25] 李建素，刘群，白硕.统计和规则相结合的汉语组块分析[J]. 计算机研究与发展，2002，39（4）：385-391.
[26] 周强，孙贸松，黄昌宁.汉语句子的组块分析体系[J]. 计算机学报，1999，22（11）：1158-1165.
[27] 周强.汉语基本块描述体系[J]. 中文信息学报，2007，21（3）：21-27.
[28] 张昱琪，周强.汉语基本短语的自动识别[J]. 中文信息学报，2002，16（6）：1-8.
[29] 宇航，周强.汉语基本块的内部关系分析[J]. 清华大学学报（自然科学版），2009，49（10）：136-140.
[30] 周强，赵颖泽.汉语功能块自动分析[J]. 中文信息学报，2007，21（5）：18-24.
[31] 陈亿，周强. 分层次的汉语功能块描述库构建分析[J]. 中文信息学报，2008，22（3）：24-31,43.
[32] 王颖. 基于 CRF 的汉语语块分析和事件描述小句识别[C]. 北京：2009.
[33] 孙广路. 基于统计的中文组块分析技术研究[D]. 哈尔滨：哈尔滨工业大学，

2008.
[34] 周强，李玉梅. 汉语块分析评测任务设计[J]. 中文信息学报，2010，24（1）：123-128.
[35] 袁彩霞.中文功能组块分析及应用研究[D]. 北京：北京邮电大学，2009.
[36] Hacioglu K，Ward W. Target word detection and semantic role chunking using support vector machines[C]. Edmonton：Association for Computational Linguistics，2003：25-27.
[37] 李建素，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003，1722-1727.
[38] He X.Using word dependent transition models in HMM based word alignment for statistical machine translation[C]. Association for Computational Linguistics，2007：80-87.
[39] 杨南.基于神经网络学习的统计机器翻译研究[D]. 合肥：中国科学技术大学，2014.
[40] Liang P，Taskar B，Klein D. Alignment by agreement[C]. Association for Computational Linguistics，2006：104-111.
[41] 谭键.语料库及语料库语言学的发展与应用[J]. 西北工业大学学报（社会科学版），2005，25（1）：61-63.
[42] Baker C F，Fillmore C J，Lowe J B. The Berkley frameNet project[C]. Montreal：Association for Computational Linguistics. 1988：86-90.
[43] Palmer M，Gildea D，Kingsbury P. The proposition bank：an annotated corpus of semantic roles[J]. Computational Linguistics. 2005，31（1）：71-106.
[44] Meyers A，Reeves R，Macleod C. The nombank project: an interim report[C]. Boston：Association for Computational Linguistics，2004：24-31.
[45] You L P，Liu K Y. Building Chinese frameset database[C]. NewYork：IEEE，2005：301-306.
[46] Xue N. Annotating the predicate-argument structure of Chinese nominalizations[C]. Genoa：Association for Computational Linguistics，2006：1382-1387.
[47] Xue N. Xia F. The Bracketing Guildlines for the Penn Chinese Treebank. University of Pennsylvania：Technical Report IRCS，2000：00-08.
[48] 黄昌宁，靳光靳.从宾州中文树库观察三个汉语语法问题[J]. 语言科学，2013，12（2）：178-192.

[49] Dowty D. Themantic Proto-Role and Argument Selection[J]. Language，1999，67：3.
[50] 袁毓林.论元角色的层级关系和语义特征[J]. 世界汉语教学，2002，2：10-23.
[51] 李航.统计学习方法[M]. 北京：清华大学出版社，2012：13-15.
[52] Ramshaw L，Marcus M. Text chunking using transformation-based learning[C]. Springer Netherlands，1999：157-176.
[53] Sang E，Kim T. Representing text chunks[C]. Hong Kong：Association for Computational Linguistics，1999：173-179.
[54] Uchimoto K，Ma Q. Named Entity Extraction Based on a Maximum Entropy Model and Transformation Rules[C]. Hong Kong：Association for Computational Linguistics，2000：326-335.
[55] 冯文贺，姬东鸿.命题库：分析与展望[J]. 外语电化教学，2010，136：25-38.
[56] 汪海燕，黎建辉，杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究，2014，31(5)：1281-1286.
[57] Taku K，Yuji M. Chunking with support vector machines[C]. Morristown：North American Chapter of the Association for Computational Linguistics，2001：1-8.
[58] 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，2007，13（3）：565-573.
[59] James A，刘群等译. 自然语言理解[M]. 北京：电子工业出版社，2005，1-3.
[60] 李明琴，李涓子，王作英.语义分析和结构化语言模型[J]. 软件学报，2005，16（9）：1523-1533.
[61] 梅家驹. 同义词词林[M]. 上海：上海辞书出版社，1983.
[62] Ronan C. Deep Learning for Efficient Discriminative Parsing[C]. International Conference on Artificial Intelligence and Statisitcs，2011：No. EPFL-CONF-192374.

附 录

主要术语对照表:

语义组块	Semantic Chunk
组块分析	Chunk Parsing
语义角色标注	Semantic Role Labeling，SRL
机器学习	Machine Learning，ML
宾州汉语树库	Penn Chinese TreeBank，CTB
宾州命题库	Chinese PropBank，CPB
自动问答	Question Answering，QA
信息检索	Information Retrieval，IR
隐马尔科夫模型	Hidden Markov Model，HMM
最大熵	Maximum Entropy，ME
条件随机场	Conditional Random Fields，CRF
支持向量机	Support Vector Machine，SVM
神经网络	Neutral Network，NN
深度学习	Deep Learning，DL

作者在读期间发表的学术论文及参加的科研项目

学术论文：

常若愚, 吴铤, 王荣波. 基于 IO 标注和 SVM 的汉语语义组块识别研究[J]. 计算机应用研究. 已录用.

科研项目：

[1] 国家自然科学基金（61202281）
[2] 浙江省自然科学基金（LY12F02006）
[3] 教育部人文社会科学研究项目青年基金（12YJCZH201）

汉语语义组块识别研究

作者：常若愚
学位授予单位：杭州电子科技大学