工学硕士学位论文

中文语义角色标注的方法研究

刘怀军

哈尔滨工业大学

2007 年 7 月

国内图书分类号：TP391.2
国际图书分类号：681.37

工学硕士学位论文

中文语义角色标注的方法研究

硕士研究生：刘怀军
导　　师：刘挺教授
申请学位：工学硕士
学科、专业：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2007年7月
授予学位单位：哈尔滨工业大学

Classified Index：TP391.2
U.D.C.: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON THE METHODS OF CHINESE SEMANTIC ROLE LABELING

Candidate： **Liu Huaijun**
Supervisor： **Prof. Liu Ting**
Academic Degree Applied for： **Master of Engineering**
Specialty： **Computer Science and Technology**
Affiliation： **School of Computer Science and Technology**
Date of Defence： **July, 2007**
Degree-Conferring-Institution： **Harbin Institute of Technology**

摘要

中文语义角色标注是近年来中文信息处理的一个热点，它能够广泛应用到信息检索、问答系统、信息抽取等领域中。句法分析对语义角色标注的影响很大，使用不同句法分析方法进行语义角色标注是目前该领域研究的主流。因此本文主要针对中文语义角色标注中特征的构造和选择，使用依存句法进行中文语义角色标注，以及多句法分析结果相结合进行了深入研究。

特征的构造和选择一直是中文语义角色标注的一个难点。本文针对中文短语结构句法和依存句法的特点，提出了许多有效的新特征和组合特征，并通过 χ^2 显著性检验、贪心算法的特征选择方法找到最优特征集，有效提高了系统性能。在最优特征集上，基于高质量短语结构句法和依存句法的系统，性能分别达到了 91.31%和 85.22%。

基于依存句法进行中文语义角色标注是一种新方法。本文首先通过 Penn2Malt 转化 Chinese Proposition Bank（CPB）得到高质量语料进行实验，并使用丰富的特征和有效的后处理规则，使得系统性能达到 85.22%。然后采用哈工大信息检索研究室的语言技术平台，实现了基于完全自动依存句法的中文语义角色标注系统。本文利用依存句法丰富的依存关系类型，系统精确率提高到 77.22%，最终超过了基于自动短语结构句法的系统精确率。

多句法结合能够充分利用句法层次的信息指导语义角色标注。利用多种句法分析器输出结果的不同，找到更多的句法结构参与语义角色的标注，以期提高语义分析的召回率，从而提高语义分析的性能。本文采用自动短语结构句法和自动依存句法相结合的语义分析方法，综合利用短语结构句法标注的较高召回率和依存句法标注的较高精确率，并给出对角色候选分类处理的有效结合策略，最终提高了系统的性能。从而证实了这种多句法相结合的中文语义角色标注方法是有效的。

关键词 中文语义角色标注；短语结构句法；依存句法；特征选择；多句法相结合

Abstract

Chinese Semantic Role Labeling (SRL) is a hot topic of Chinese language processing, which can be applied in the field of Information Retrieval, Question Answering System, Information Extraction and others. Syntactic parsing is necessary for SRL. Different syntactic views are used for SRL, which is the mainstream recently. The paper made a deep research on three aspects: feature construction and selection on Chinese SRL, Chinese SRL using dependency parsing, and Chinese SRL based combination of multi-syntactic parsing results.

Feature construction and selection are the difficult tasks in Chinese SRL. The paper proposes many new features and combined features, according to the characteristics of Chinese phrase-based syntactic parsing and dependency parsing. Best features are selected by χ^2 significant test and Greedy Algorithm in order to improve the performance. On the optimal selection of the feature set, the performances of the systems based on high-quality phrase-based syntactic and based on dependency parsing increased to 91.31% and 85.22% separately.

Dependency parsing used for Chinese SRL is a new approach. First, Chinese Proposition Bank (CPB) is converted to Chinese Dependency Bank by Penn2Malt. After rich features and post-process rules used, the performance achieved 85.22%. Then, the paper constructs a Chinese SRL system based automatic dependency parsing, using the Language Technology Platform of HIT-IR (HIT-IR LTP). The precision of the system increased to 77.22% by using rich dependency types, and surpassed the one based automatic phrase-based syntactic.

Combination of multi-syntactic parsing results can help SRL by more syntactic information. More argument candidates can be found from the label results of systems based diffrent syntactic parsing, then the recall is expected to improve, and the performace is improved final. The paper combines the automatic phrase-based parsing and automatic dependency parsing for Chinese SRL. It makes use of high recall of system based phrase-based syntactic，and high precision of sytem based dependency parsing. The classification of the argument candidates proves to be efficient in the synthesis strategies. The

performance is improved final. The experiment shows that combination of multi-syntactic parsing results is effective for Chinese SRL.

Keywords Chinese Semantic Role Labeling; Phrase-based Syntactic Parsing; Dependency Parsing; Feature Selecting; Combination of Multi-Syntactic Parsing

目录

第1章 绪论

1.1 课题研究的学术背景及意义

1.1.1 课题研究背景

随着知识经济的兴起和计算机技术的迅猛发展，人类正在步入一个信息化时代。在信息化时代提高汉语的功能与地位，加速汉语的传播，首要任务之一是加快汉语言文字信息处理的研究。计算机的出现以及其在自然语言研究领域的应用，大幅度提高了语言信息的处理速度和质量。为了使计算机具有理解、处理和生成自然语言的能力，必须使计算机能够分析自然语言语句的含义，也就是进行语义分析。

所谓语义分析，指的是根据句子句法结构和句中每个实词的词义推导出能够反映这个句子意义的某种形式化表示。例如句子“他打开了纸箱”和“纸箱被他打开了”在语义上都可以统一表示为“打开（他，纸箱）”。

20 世纪 70 年代以来，语义分析越来越受到从事自然语言处理的学者们的重视，这一时期的研究主要集中于语义理解、知识表示和推理等复杂问题上，开发出来的系统在特殊句子或小故事理解方面[1]，性能上有很大提高。然而，这种方法需要获取大量的知识，受限于知识工程进展的瓶颈。因而，为了避开深层困难的语义理解，自然语言学者们将注意力集中到浅层的、简单并且实用的语义分析任务上。到了 90 年代，随着统计学习方法的发展，人们在一些简单的应用上取得了很大的进展，例如分词，词性标注，句法分析等等[2,3]。同时由于语义角色标注在问答系统、信息抽取、机器翻译等领域有着广泛的应用。因此目前语义角色标注引起了越来越多从事自然语言理解研究和应用的学者们的重视。在许多著名的国际会议以及国际期刊上，语义分析的文章也越来越多。

基于机器学习的浅层自动句法取得了一定的进展，并且许多的研究浅层语义的学者也构建了许多相关的语料库，比如英文的语料库PropBank，NomBank[1]和中文的Chinese Proposition Bank[2]（CPB）。目前基于手工标注句

[1] http://nlp.cs.nyu.edu/meyers/NomBank.html

法的语义角色标注已经达到了很高的性能，但是基于自动句法的语义角色标注系统，汉语系统的性能较之英文系统还有相当的差距。所以，中文的语义角色标注角色分析还有很多的挑战，有很多需要挖掘、值得研究的内容。

1.1.2 语义角色标注的应用

语义角色标注作为许多学者努力的目标，也是实现深层语义分析的一种途径。较之深层语义分析，语义角色标注分析有许多优点：有明确的任务，有相当规模的语料资源，分析结果便于评测，便于应用到其他领域等。

在自然语言处理领域，通过统计的机器学习方法进行语言分析一直是一个热点问题，并且也取得了相当的成就。语义角色标注综合利用了底层的分词、词性标注、句法分析、命名实体识别等的信息，一方面对这些底层的技术是一个考验，再一方面也是测试机器学习技术的一个平台。作为自然语言理解的底层研究，语义角色标注的最终目标是应用到高层研究中。在信息抽取、自动问答、机器翻译等领域中，语义角色标注有着广泛的应用。

信息抽取是从文本中选择出信息，然后创建一个结构化表示的过程。20世纪 80 年代开始,基于消息理解会议(MUC) 、自动内容抽取会议(Automatic Content Extraction, ACE)等几个因素的推动，信息抽取系统研究迅速开展起来。信息抽取不仅能自动获取人们需要的信息，还能帮助提高信息检索等技术的性能。把语义角色标注应用信息抽取，能有效提高信息抽取的性能[4]。同时语义角色标注也可以看作一种通用的信息抽取，能指导信息抽取技术的发展。

机器翻译（MT）是利用计算机把一种自然语言转变成另一种自然语言的过程。目前，统计机器翻译受到人们的高度重视，而自然语言理解的也被广泛应用在这个领域。尤其语义角色标注等技术的应用[5]，使得机器翻译系统达到了较好的效果。

复述（Paraphrasing）[6]在国内也有学者称为改写，与问答、文摘和翻译并列被美国认知心理学家G.M.Olson认为是判别计算机是否理解自然语言的四条标准。其含义是让计算机自动判断两个自然语言语句是否表达相同的含义。我们知道，理解句子的含义正是语义分析的目的，因此，利用语义角色标注的结果，必然能够较好的完成复述的任务。

[2] http://www.cis.upenn.edu/~chinese/cpb/

1.2 中文语义角色标注

1.2.1 中文语义角色标注的定义

语义角色标注是浅层语义分析的一种可行方案[7]，中文语义角色标注（Semantic Role Labeling，SRL）是目前中文语义分析的一种主要实现方式。主要目的是识别所有与动词有关的句法成分的语义角色，并且给他们赋予一定的角色类型，比如核心角色施事、受事、工具等，还有附加角色时间、地点、方式等。

它采用“谓语动词-角色”的结构形式，不对整个句子进行详细的语义分析，而只是标注句法成分为给定谓语动词的语义角色，每个语义角色被赋予一定的语义含义。例如“[委员会 施事][明天 时间]将要[通过 谓语动词][此议案 受事]。”其中，“施事”表示动作的发出者，“受事”表示动作的接受者等。在中文语义角色的标注结果中，句子 1、2、3 虽然表达方式不同，但是标注的结果都相同。

句子 1：委员会明天将要通过此议案

句子 2：此议案明天将要被委员会通过

句子 3：明天，委员会将要通过此议案

1.2.2 语料资源及主要机器学习方法

进行中文语义角色标注，和其他基于有指导机器学习的自然语言处理任务一样，也需要语料资源的支持。目前，中文语义角色标注的研究主要使用三种资源：Chinese Proposition Bank(CPB)[8]，Chinese Nombank[9]，Chinese FrameNet[10]。

CPB是Upenn基于Penn Chinese Treebank(PCT)标注的汉语语义角色标注资源，在Penn Chinese Treebank句法分析树的对应句法成分中加入了语义信息。Penn Chinese Treebank的标注数据主要来自新华新闻专线、Sinorama新闻杂志和香港新闻。图1-1是CPB中一个句子的标注实例。CPB包含 20 多个语义角色，相同的语义角色对于不同目标动词有不同的语义含义。其中核心的语义角色为ARG0-5 六种，ARG0 通常表示动作的施事，ARG1 通常表示动作的影响等等。其余的语义角色为附加语义角色，用前缀ARGM表示，后面跟一些附加标记（Secondary Tags）来表示这些参数的语义类别，如

ARGM-LOC表示地点，ARGM-TMP表示时间等。

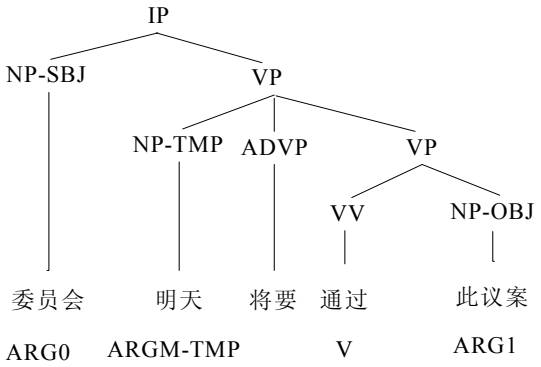

图 1-1 Chinese Proposition Bank 中的一个标注实例

Figure 1-1 A instance illustrating in Chinese Proposition Bank

Chinese Nombank 把传统 English Proposition Bank[11] 和 English Nombank[12]的标注框架，扩展到对中文名词性谓语动词的标注。Chinese Nombank在PCT数据上加入了语义层的标注信息，像CPB一样，也标注了两类语义角色：核心语义角色和附加语义角色。Chinese NomBank标注了名词性谓语动词的框架集（Framesets），不过规模只是CPB中对应动词性谓语动词标注框架集的一少部分。Chinese NomBank中的角色位置有两类情况。第一类，角色在以名词性谓语动词为核心词(Head Word)的名词短语中。第二类，当以名词性谓语动词为核心词的名词短语作支持动词(Support Verb)的主语时，允许语义角色在名词短语外。图1-2是Chinese NomBank的一个标注实例。

山西大学构建的 Chinese FrameNet 是基于框架语义的，是一种 FrameNet 风格的中文词典。它描述了词汇单元以及参与者框架元素之间的关系，也包含了框架元素的详细句法信息。Chinese FrameNet 的架构和 English FrameNet 相似，并且有许多来自 English FrameNet 的翻译，但是作了一些相应的修改和创新，增加了相应语义角色的汉语名称。目前 Chinese FrameNet 已经有 130 多个汉语框架，并且还在不断增加，Chinese FrameNet 还没有在网络上共享。

语义角色标注通常被看作分类问题，目前的研究大多基于有指导的机器学习方法。在语义角色标注中，主要使用的机器学习方法有支持向量机(SVM)[13]，最大熵(Maximum Entropy)[14]，SNoW(Sparse Network of Winnows)[15]等。

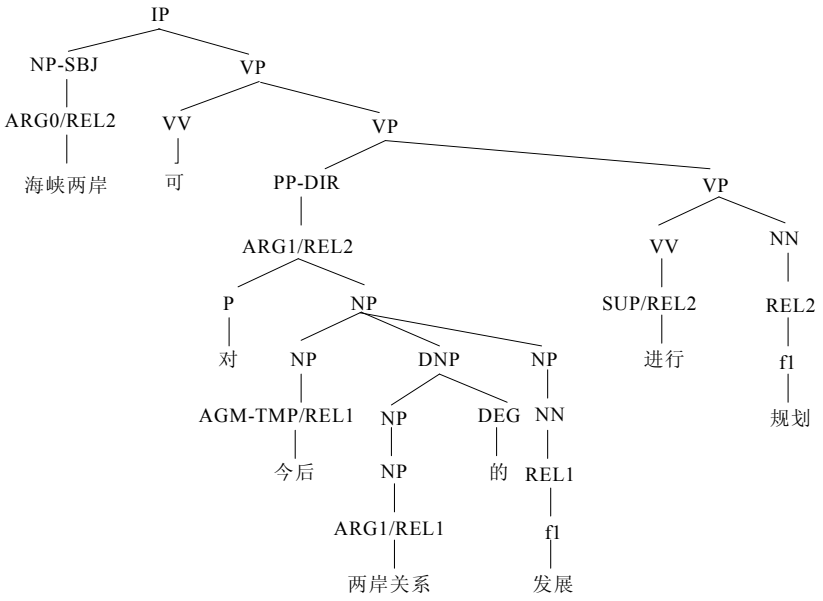

图 1-2 Chinese NomBank 中的一个标注实例

Figure 1-2 A instance illustrating in Chinese NomBank

支持向量机(Support Vector Machine, SVM) 是支持向量机是基于Vapnik[16]提出的统计学习原理构建的一种线形分类器，后成功的应用于文本分类[17]等自然语言处理领域。它是Cortes&Vapnik1995 年首先提出来的[18]，近年来机器学习研究的一项重大成果。根据Vapnik&Chervonenkis 的统计学习理论，如果数据服从某个(固定但未知的) 分布，要使机器的实际输出与理想输出之间的偏差尽可能小，则机器应当遵循结构风险最小化原理，而不是经验风险最小化原理，通俗地说就是应当使错误概率的上限最小化。支持向量机正是这一理论的具体实现。

最大熵(Maimum Entropy)的基本思想是为所有已知的因素建立模型，而把所有未知的因素排除在外[19]。也就是说要找到这样一个概率分布，它满足所有已知的事实，且不受任何未知的因素的影响。最大熵分类器已经成功应用于信息抽取，句法分析等多个自然语言处理领域。此方法通过计算特征函数的线性组合$\sum \partial \lambda_i f_i(c,h)$来估计一个概率模型，如公式(1-1)所示。

$$P(c \mid h,\lambda) = \frac{\exp \sum_i \lambda_i f_i(c,h)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',h)} \tag{1-1}$$

其中， f 是特征函数， c 表示类别， h 表示上下文历史信息。

目前最大熵模型是语义角色标注中应用最为广泛的模型之一[20,21]，究其原因，主要是因为最大熵模型能够较为准确地给出每个输出角色的概率值，并且方便的处理多类问题，另外一个不可忽视的原因就是最大熵模型较之支持向量机有更快的训练速度。然而，最大商模型也并非完美，首先是其不便于使用复杂的结构特征，另外其不支持特征自动组合，需要手工组合特征的缺点也限制了其性能的进一步提高。

SNoW(Sparse Network of Winnows)[22]是建立在预定义或者增量变化的特征空间上的线性函数的稀疏网络。这种稀疏网络是以数据驱动的方式来分配和连接特征与目标单元的，并且它的计算是基于活性特征而非整个特征空间的。它采用多种更新规则，包括：感知器，Winnow，梯度变化回归算法和Bayes算法等。SNoW作为一种学习结构框架，使得使用者可以通过设置结构参数、更新规则、训练策略等来建立自己的分类器结构。

Boosting[23]是基于判别式的分类方法，其基本思想是组合多个弱分类器Schapire等人证明组合这些弱分类器可以形成一个强分类器。M`arquez等人以及Surdeanu等人[24]使用AdaBoost（Boosting思想的一种实现）算法进行语义角色标注，取得了不错的效果。

1.2.3 标注单元及谓语动词

语义标注的基本单元可以是句法成分（Constituent）、短语（Phrase）、词（Word）或者依存关系（Dependency Relation）等等。

在图1-1的短语结构句法分析树中，每一个非终结节点，如NP-SBJ，NP-TMP，VP等，都是句法成分。一般认为每个语义角色是与某一句法成分相对应的。基于句法成分的语义角色标注，就是以句法成分为单元， 从句法树中提取合适的结点作为语义角色的候选。然后，对候选集中的每个句法成分，把它分到对应的语义角色类别。现在多数的语义角色标注系统通常都是以句法成分为基本的标注单元。

以短语和词为单元[25,26,27]的语义角色标注，可以表示成组块分析（Chunking Task）问题。从基本的短语和词中抽取特征，并且给句子中每个短语和词一个标记。通常采用IOB标记方法，短语和词在语义组块内部被标为I（Inside），在语义组块外部被标为O（Outside），是语义组块的开始被标为B（Begin）。

依存语法通过分析语言单位内成分之间的依存关系揭示其句法结构。以

依存关系为单元[28]的语义角色标注，充分利用了节点间的依存关系对语义角色结构的暗示，以不同于短语结构句法的方式进行语义角色的分析。

中文语义角色标注中，谓语动词是核心，语义角色是围绕谓语动词进行标注的。谓语动词有两类：一类是动词性质，另一类是名词性质。

CPB 的语料是以动词性质的谓语动词为核心，标注其语义角色的。其中，谓语动词有四类：状态动词（VA），系动词（VC），{有，没有，无}作动词（VE）以及其他动词（VV）。其他动词包括情态动词、可能性动词、行为动词等。不同类型的动词，其框架结构（FrameSet）不同，所带的语义角色个数和类型也不同。

Chinese NomBank 的语料则是以名次性质的谓语动词为核心的，标注语义角色的。和动词性质的谓语动词相比，名次性质的谓语动词有自己的一些特点。首先，后者对角色类别的预测能力不如前者。前者的不同句法位置的成分会充当对应的语义角色类别，比如主语（Subject）一般作谓词的施事（ARG0），而宾语一般作受事（ARG1）。其次，后者有很少的语义角色，尤其附加语义角色（ARGM-）。并且，后者一般是单义词，框架结构比较简单。

1.2.4 标注步骤及特征构造

语义角色标注系统一般通过三个阶段实现[29]：第一是过滤阶段（Pruning），使用一些启发式规则把大部分不可能成为语义角色的句法成分从句法分析树中过滤掉；第二阶段进行语义角色识别（Identification），用二元分类器把角色候选分为正例和反例，正例是语义角色，反例非语义角色；第三是语义角色分类阶段（Classification），使用多类分类器把第二阶段识别的语义角色分到对应的角色类别。也有系统会加入一个基于启发式规则的后处理阶段（Post-Process）。

特征一直是决定统计自然语言处理系统性能的重要因素。相比特征空间较小的底层自然语言处理任务，比如分词、词性标注和命名实体（NE）识别，语义角色标注任务的一个显著特性就是特征空间很大。由Gildea等人[29]在其语义角色标注系统中使用的语言学特征往往被当作各个系统的基本特征所使用。并且在Pradhan等人[30]以及Xue 等人[31]的中文语义角色标注工作中也使用了许多有效的特征。下面我们介绍这些基本特征并简要分析其有效性。

● 句法成分结构特征

1. 短语类型

2. 中心词及其词性：在中心词提取中，我们使用Sun等人[32]的中心词规则（Head rules for Chinese）。

3. 句法成分第一个词及其词性

4. 句法成分最后一个词及其词性

5. 句法分析树中左、右兄弟句法成分的短语类型

● 谓语动词结构特征

1. 谓语动词

2. 子类框架：谓语动词父节点及其子节点。如图1-1中，“通过”的子类框架是VP→VV-NP-OBJ。

3. 谓语动词的类别信息：目前的中文语义角色标注任务中还没有统一规范的动词分类，Xue等人[31]的语义角色标注工作中从如下三个方面来对动词分类：

① 动词的框架集（Framesets）数

② 动词每个框架集的角色（Arguments）数

③ 动词每个框架集的句法交替（Syntactic Alternations）

● 谓语动词-句法成分关系特征

1. 路径：句法分析树中从当前句法成分到谓语动词的句法路径。如图1-1中，NP-TMP的路径是NP-TMP↑VP↓VP↓VV。

2. 部分路径：对路径特征的一种泛化，指当前句法成分到它和当前谓语动词的最近共同父节点的句法路径。如图1-1，NP-TMP的部分路径是NP-TMP↑VP。

3. 路径长度：句法成分和它的谓语动词之间的长度。

4. 位置：句法成分在谓语动词前面还是后面，这是一个二值特征。

5. 距离：句法成分和谓语动词的相对距离，在谓语动词前的句法成分的距离特征为负值，谓语动词后的句法成分距离特征为正值。

6. 句法成分的句法框架：句法框架特征包含谓语动词和围绕谓语动词的名词短语。句法框架特征中，谓语动词和这些名词短语作为核心，当前句法成分和它们相关联。如图1-1中，句法成分NP-OBJ的句法框架是np_np_v_NP-OBJ。

通过统计，测试语料中总会有动词在训练语料中没有出现过，从训练数

据中学习的最大熵模型就不能很好的对这些动词进行预测。CPB 中许多动词有相似的语义结构，比如动词“显现”和“显示”都带两个核心语义角色，主语指描述的实体，宾语指所描述实体的特性。这样，动词类别信息就可以在动词稀疏的情况下正确预测角色类别。

1.2.5 国际评测

对于语义角色标注系统的性能评测，我们使用信息检索中的评测方法，采用F-Score对最终系统的性能进行评价。其定义如公式(1-2)。

$$F\text{-}Score=\frac{2*Precision*Recall}{Precision+Recall} \tag{1-2}$$

其中准确率（Precision）和召回率（Recall）的定义公式(1-3)和(1-4)：

$$Precision=\frac{\text{正确标注为语义角色的个数}}{\text{分类器预测为语义角色总数}} \tag{1-3}$$

$$Recall=\frac{\text{正确标注为语义角色的个数}}{\text{测试数据中语义角色总数}} \tag{1-4}$$

2004 年和 2005 年CoNLL(Conference on Computational Linguistics Learning)举行对语义角色标注的评测SRL Shared Task[20,33]。CoNLL是一个面向各家学术机构和组织的评测会议，2005 年有 19 个单位参赛，评测数据除了Penn TreeBank中抽取外，也有其他领域数据。评测包括封闭测试（只使用主办方提供的数据），还包括开放测试（可以使用任何外部数据，如WordNet ，VerbNet等等）。此次评测更面向实际的语义角色标注系统。

1.3 中文语义角色标注的现状分析

目前的中文语义角色标注，Xue等人[31]主要依赖手工标注的高准确度语料CPB，才取得了 91.3%的性能。而CPB语料规模很小，只是对PCT中 760 多个文档进行了语义角色的标注。Chinese NomBank也只是在Chinese TreeBank中，对名词性谓语动词进行了语义角色标注，相对CPB而言，规模更小。并且，已有的语义角色标注的语料只是标注了有限几种语言的语料。所以，缺乏高质量大规模的语料一直是语义角色标注的一个瓶颈。并且根据Gildea等人[34]以及Punyakanok等人[35]的研究发现，深层句法分析结果对于语义角色标注是非常重要的信息。因此，获取高准确度的深层句法分析也是语义角色标注的一个难点。

另外，由于机器学习的技术已经日渐成熟，因此人们关注的重点从机器学习算法的改进，转移到了寻找丰富有效的特征。在Xue等人[36]的文章中，对目前的常用特征及其有效性进行了分析。在Pradhan等人[13]的文章中，则引入了更多的特征，并且对一些特征提出了几种泛化方案。然而，许多特征需要基于语言学方面的知识来构建，有些特征则是经过多次试验后的经验想出来的。语义角色标注作为一个特征空间很大的任务，特征的自动选择一直是一项难题。用CPB的语料进行实验，特征数量会达到好几百万，面对这么庞大的特征空间，通过人来进行特征筛选是耗费很大的。并且，特征与特征之间会存在相互关联，比如特征 F_A 和 F_B 单独加入，会引起系统性能下降，而当 F_A 和 F_B 同时都加入时，系统性能则提升。因此，建立一套特征的自动选择机制也是人们期望的目标。

中文语义角色标注对自然语言处理的底层模块如分词、词性标注和句法分析依赖性很强。目前分词和词性标注技术已经比较成熟，但Xue等人[37]的中文短语结构句法分析还不够成熟，性能也只能达到 80%多。目前中文语义角色标注还是仅使用短语结构句法分析[31,32]，而在英文的语义角色标注中，已经有学者使用依存句法分析[28]来做相关方面的研究。所以，基于多种句法分析的中文语义角色标注也是一个研究的热点。

1.4 本文完成的主要工作

本文完成的主要研究工作如下：

第 1 章为绪论。首先介绍课题的研究背景，以及语义角色标注在自然语言处理领域的重要意义。接着介绍中文语义角色标注的定义，标注方法包括：语料资源、机器学习方法、标注步骤、特征构造以及国际评测等。最后还介绍了该领域研究的热点和难点问题。

第 2 章介绍了基于手工标注短语结构句法和自动短语结构句法的中文语义角色标注。首先使用 CPB 的语料进行了实验，同时提出了许多新特征，通过 χ^2 显著性分析进行特征选择得到最优特征集，并对实验结果进行了深入分析。接下来使用 DBParser 自动句法分析器构建了语料资源，并使用前面的最优特征集进行了实验，得到了基于自动短语结构句法的中文语义角色标注结果，并和手工短语结构句法的结果进行了比较和分析。

第 3 章介绍了基于依存句法的中文语义角色标注，这是中文语义角色标注的一种新方法。首先用 Penn2Malt 转化 CPB 语料得到准确度较高的依存句法分析语料，同时提出丰富有效的特征并结合 χ^2 显著性分析和贪心算法进行了特征选择，构建了基于依存句法的角色标注系统并进行实验。接下来介绍了哈工大信息检索研究室的语言技术平台和依存句法，并构建了基于完全自动依存句法的语料资源，介绍了使用自动依存句法进行角色标注的方法，给出实验结果。最后把实验结果和使用短语结构句法的结果进行了比较分析。

第 4 章介绍了两种句法角色分析结果相结合的中文语义角色标注方法。首先给出这种方法的意义，能够充分利用自然语言处理的多层面信息。接下来介绍了系统的实现框架，并通过召回率上限在结合前后的变化分析了该方法的预期效果。最后介绍角色候选分类处理的结合策略并完成实验，由实验结果证实了该方法的可行性。

第2章 基于短语结构句法的中文语义角色标注

2.1 基于手工标注句法的中文语义角色标注

2.1.1 语料资源构建和标注单元选择

我们实验中使用来自Chinese Proposition Bank(CPB)的数据，文章的第二部分已经对CPB做了介绍。我们校正了CPB语料中存在一些问题，使用全部 760 个文档进行实验。其中前 100 个文档作测试语料，剩余 660 个文档作训练语料。CPB基于PCT手工标注的句法分析结果，准确率较高。它几乎对PCT中的每个动词及其语义角色进行了标注，因此覆盖范围更广，可学习性更强。表2-1和2-2列出来CPB中的核心角色和部分附加角色。

表 2-1 动词的核心语义角色

Table 2-1 The core arguments for predicate

角色标注	含义描述
ARG0	施事，动作的发出者
ARG1	受事，动作接受者
ARG2	动作的间接作用对象
ARG3	直接目的、目标等
ARG4	直接工具、方法等

表 2-2 动词的部分附加语义角色

Table 2-2 Some adjunctive arguments for predicate

角色标记	含义描述
ARGM-ADV	Adverbials（附加的，默认标记）
ARGM-BNE	Beneficiary（受益人）
ARGM-CND	Condition（条件）
ARGM-EXT	Extent（扩展）
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）

第 1 章提到了三种语义角色的标注单元。由于 CPB 的语料是基于短语的句法分析，并且为语义角色大多与句法成分对应。所以我们采用句法成分作为标注单元可获得较高的性能。每个句法成分都作为语义角色的候选，将被分到对应语义角色类别。

2.1.2 标注步骤和分类器

为了提高系统召回率，避免过滤阶段语义角色的丢失，在系统中没有使用过滤阶段。同时，由于我们采用的最大熵分类器的效率很高，并且类别数量对最大熵的分类效率影响不大，因此我们把识别和分类两阶段合二为一，对与谓语动词相关的句法成分进行预测，属于语义角色的成分被分到对应类别，不属于任何角色的成分被赋予空类别。

2.1.3 特征构造

第 1 章对中文语义角色标注中一些基本特征分三类做了介绍。为了进一步挖掘自然语言处理底层模块的信息，把它应用到语义角色标注中，这部分我们引入了的一些新特征。对新特征的构造，我们从中文分词与词性、句法结构角度来考虑。对于词与词性方面特征，通过上下文信息提取和词的搭配来构造；对于句法结构方面特征，通过句法节点、句法路径以及节点之间的关系来构造。

1. 句法成分的句法功能：CPB 手工标注的句法分析中，短语类型后缀有功能标记，比如-IO 表示间接宾语，-OBJ 表示直接宾语，-SBJ 表示主语等。这些功能标记作为特征能够有效暗示语义角色的类型。

2. 句法成分前一个词和后一个词

3. 从句层数：在Xue等人[38]有关PCT的句法标注文章中，对汉语句子提出了几种类型：带补语的子句（CP）、简单子句（IP）、不带疑问次的疑问句（IP-Q）等。我们把句法成分到谓语动词的路径上经历的子句IP、CP、IP-Q等的个数作为特征。

4. 句法成分到谓语动词的路径上出现的名词短语个数

5. 句法成分和谓语动词的相对位置：我们从三方面来考察他们的相对位置：它们是否兄弟节点关系，是否属于相同动词短语（VP）的儿子节点，是否属于相同子句 IP 或 CP 短语的儿子节点。

6. 句法成分和谓语动词的共同最近父节点

7. 谓语动词的搭配模式：CPB 语料数据中，ARG2 大多情况在含有下面 5 种结构的句子中出现：介词-动词结构、使-动词结构、把-动词结构、被-动词结构、动词-数量词结构五种搭配结构。这种搭配模式能够提高对 ARG2 的预测效果，比如对于动词“修到”，ARG2 表示修建的地点，那么在语句“把公路修到山顶上” 中“把-动词结构”就暗示句法成分“公路”属于角色 ARG2。

许多单一特征对语义角色分类已经非常有效，而把这些单一特征组合在一起时，能更有效的增强分类能力。由于最大熵分类器不能够自动地对特征进行组合，因此我们通过基础特征和扩展特征构造了一些组合特征，比如谓语动词和短语类型、谓语动词和中心词、谓语动词和位置、谓语动词类别信息和路径等。

2.1.4 后处理阶段

我们系统中会出现两个标注的句法成分互相嵌套的情况，但在 CPB 标注语料中，这是不允许的。最大熵分类器能够预测每一个类别的概率，所以当两个标注的句法成分发生嵌套时，我们保留了概率最大的那个句法成分。CPB 标注语料中允许标注为相同语义角色的多个句法成分在句子中同时出现。处理这种情况时，我们设置一个阈值，当这些句法成分预测概率都大于阈值时全部保留，否则仅保留概率最大的那个句法成分。语义角色 ARG0-PSR 和 ARG0-PSE 表示持有和被持有关系，在句子中往往成对出现。我们系统的标注结果中，可能出现一个句子只有 ARG0-PSR 或 ARG0-PSE 的情况，当对应句法成分预测概率高于某个阈值时，我们保留旧的标注，否则，把标注更新为最大熵预测的概率次高的那个语义角色类型。

2.1.5 实验结果及讨论

特征之间存在影响，也就是说特征A和B单独加入时会使性能提高，但A和B同时加入则可能降低性能，这说明加入所有特征得到的系统性能并不是最优的。首先建立一个基于基本特征的系统，称为基础系统（BaseLine）。我们在基础系统上加入全部新构造的单一特征构成系统进行了实验，结果如表2-3。从表2-3可以看出，这些特征全部加入后，系统性能提高并不明显。因此，我们通过 χ^2 显著性检验，从新构造的单一特征和组合特征选择最优特征集。

表 2-3 加入全部新单一特征前后的系统性能比较

Table 2-3 The performance comparison after adding all new single features

实验	Precision（%）	Recall（%）	F-Score（%）
基础系统	90.94	88.62	89.76
加入全部新单一特征的系统	91.02	88.75	89.87

由于特征之间影响很难预测，我们不考虑这种影响，单独把扩展特征和组合特征逐个加入基础系统中进行显著性分析的实验。表2-4列出了加入这些特征后系统性能的变化。在F-Score列中，粗体表示性能提高，前面加星号表示性能显著提高。

实验中我们采用 χ^2（自由度为 1）显著性检验，测试数据中角色总数 $n=10{,}822$， χ^2 检验的上侧 α 分位数 $\alpha=0.10$，基础系统性能为 F_b，加入一个新特征后系统性能为 F_n，则 χ^2 公式如(2-1)定义：

$$\chi^2=\frac{(nF_b-nF_n)^2}{nF_n}+\frac{(nF_b-nF_n)^2}{n(1-F_n)} \tag{2-1}$$

则仅当 $\chi^2 \geq \chi_\alpha^2(1)=2.706$ 时，性能 F_n 增加显著。

从表2-4可以看出，加入句法成分后一个词、谓语动词和短语类型的组合、谓语动词类别信息和路径的组合都显著提高了系统的性能F-Score值。其它特征或特征组合加入后，除了少数特征和特征组合的加入使得性能降低，多数都使系统性能提高。

句法成分后一个词能够显著提高系统的性能，一方面是由于汉语语法的一个重要特点就是十分重视词序，词序不同表达的意思就不同；另一方面，句法成分后一个词作为上下文特征，能够反映当前句法成分的特定语境意义。短语类型是一个非常有效的特征，这是由于不同句法类型的短语总是趋向于充当不同的语义角色，而当给定句子中谓语动词时，这种特性就更加明显，所以谓语动词和短语类型的组合显著提高了系统的性能。比如例句“今年比去年同期**增长**九十点七亿美元”，对于动词“增长”，其后的数词短语往往趋向于做ARG2。路径特征在谓语动词已知时非常有效，但是两者组合会导致数据稀疏，所以我们采用谓语动词类别信息和路径组合，显著提高了系统的性能。

表 2-4 新特征对性能的影响

Table 2-4 The effect of new features on the performance

特征	Precision (%)	Recall (%)	F-Score (%)
基础系统	90.94	88.62	89.76
逐个加入扩展特征			
+ 句法成分的功能	90.99	88.77	**89.87**
+ 句法成分前一个词	91.15	88.70	**89.91**
+ 句法成分后一个词	91.58	89.14	***90.34**
+ 句法成分前一个词的词性	90.97	88.71	**89.82**
+ 句法成分后一个词的词性	91.05	88.86	**89.94**
+ 从句层数	91.01	88.91	**89.95**
+ 谓词到句法成分的路径上名词短语个数	90.93	88.69	**89.80**
+ 句法成分和谓语动词的相对位置	90.97	88.81	**89.88**
+ 句法成分和谓语动词的共同最近父节点	90.87	88.68	89.76
+ 谓语动词的搭配模式	91.08	88.81	**89.93**
逐个加入组合特征（冒号前后为组合对应的单一特征）			
+ 谓语动词 ： 短语类型	91.64	89.19	***90.40**
+ 谓语动词 ： 中心词	91.48	88.75	**90.09**
+ 谓语动词 ： 中心词 ： 中心词词性	91.46	88.71	**90.06**
+ 谓语动词 ： 位置	91.27	88.79	**90.01**
+ 谓语动词 ： 路径	91.44	88.98	**90.19**
+ 中心词 ： 位置	90.98	88.44	89.69
+ 谓语动词类别信息 ： 路径	91.58	89.06	***90.30**
+ 谓语动词类别信息 ： 句法成分的句法框架	91.17	88.83	**89.98**
+ 短语类型 ： 左兄弟成分的类型	90.84	88.63	89.72
+ 短语类型 ： 右兄弟成分的类型	90.90	88.78	**89.83**
+ 谓语动词的搭配模式 ： 位置	91.01	88.45	89.71
+ 句法成分的句法框架 ： 短语类型	90.81	88.62	89.70
+ 谓语动词 ： 句法成分的句法框架	91.22	88.87	**90.03**
+ 中心词 ： 中心词词性 ： 路径	91.15	88.59	**89.85**
+ 短语类型 ： 路径	91.09	88.75	**89.90**

我们在基础系统上加入了使性能提高的扩展特征和组合特征，构成了新系统。表2-5列出了基础系统和新系统的性能。从表2-5可以看出：尽管每个特征单独加入后，系统性能的增加幅度不是很大，但这些特征全部加入后，系统的性能就有了明显的改进，增加了 1.55 个百分点。

表 2-5 加入扩展特征和组合特征前后的系统性能比较

Table 2-5 The performance comparison after adding the new features

实验	Precision（%）	Recall（%）	F-Score（%）
基础系统	90.94	88.62	89.76
新系统	92.68	89.97	91.31
Xue等人的系统	91.40	91.10	91.30

从我们实验的结果可以看出，在汉语手工标注句法语料上性能能够达到91.31%，和 Xue 等人的性能一样。主要在于如下几个因素：

首先，动词类别信息能够有效提高系统性能。因为汉语中动词一词多义的现象比较少，在所有语料数据集中，共有 4,858 个动词。其中，只有 62 个动词有 3 个或 3 个以上的框架集（Framesets），如表2-6所示。这样对于大量只有 1 个框架集的动词，其语义角色就有相对固定的句法形式，在手工标注的准确句法分析情况下，角色预测就比较容易。同时，汉语中形容词作谓语的情况很多，这样谓语动词的角色相对单一，句法实现也简单。在CPB语料中，谓语动词的词性有VA，VV，VC，VE四种，其中形容词性谓词VA占有很大比例，这样对应谓语动词的语义角色就很容易预测。

表 2-6 不同框架集数的动词分布

Table 2-6 The verb distribution by different framesets

框架集个数	动词个数
1	4,511
2	285
3	41
4	13
5	5
6	2
7	1
≥8	0

对表2-6进一步进行比较，如图2-1所示。可以看出框架集数多的动词只是占了很少一部分比例。这部分动词对应的角色也很少，这样大部分语义角色是属于简单框架集的动词。因而系统性能较高，这也是一个重要影响因素。

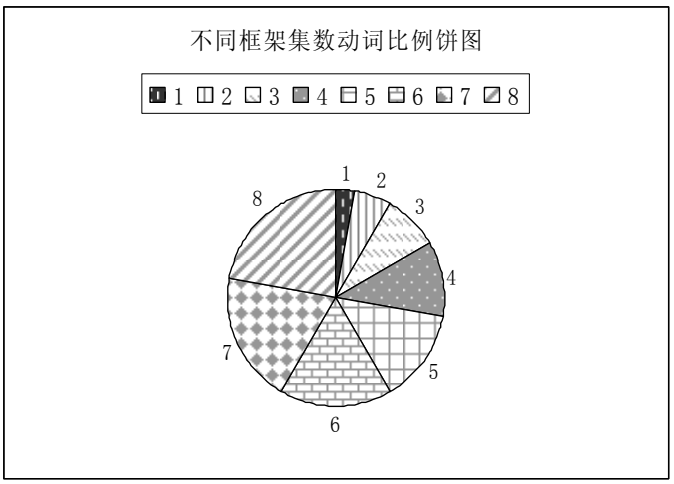

图 2-1不同框架集数的动词分布

Figure 2-1 The verb distribution of different framesets

其次，Penn Chinese Treebank使用了更加层次化的结构方式，在完全句法分析树中，使用了许多空标记（-NONE-）来表示深层的含义。如图2-2句法分析中：（-NONE- *-1）和（-NONE- *-2）表示指代关系的索引，指代节点（NR 张三）。

```
(IP (NP-PN-SBJ-1 (NR 张三))
    (VP (VV 好像)
        (IP (NP-SBJ-2 (-NONE- *-1))
            (VP (SB 被)
                (VP (VV 打)
                    (AS 了)
                    (NP (-NONE- *-2)))))))
```

图 2-2 CPB 中的空标记

Figure 2-2 Null elements in CPB

并且，在中文语义角色标注的训练语料数据中，有更多附加角色ARGM-X，相对少的而又难分辨的核心角色ARG3，ARG4，如表2-7所示。这样使得角色的识别和分类更加容易。

通过对分类错误的语义角色进行分析，这些错误主要有如下几方面引起：首先，一般动词的主语（Subject）被标为ARG0，宾语（Object）被标为ARG1。但也有一些动词例外，比如“出现”。例如：这支新队伍以新面孔**出现**在世人面前。其中“这支新队伍”做主语却被标注为ARG1。其次，占比例较高的角色ARG2 的召回率较低。在汉语里中，ARG2，ARG3，

ARG4 这几类角色非常灵活，对于不同的动词表示不同的含义，这种灵活性增加了分析的难度。下面是角色ARG2 的例子：

1．他们都**给**我肯定的答复。

2．外商独资企业**增加**了百分之四点一二，达八千四百八十四个。

3．中国对外贸易合作部**派驻**澳门的直属企业。

在上面的三个句子中，例句 1 中 ARG2 表示“给”的接受者；例句 2 中 ARG2 表示“增加”的数额；例句 3 中 ARG2 表示“派驻”的地点。

表 2-7 训练集中主要角色的分布情况

Table 2-7 List of main roles in training data

角色类型	数量	所占比例(%)
ARG0	23,239	34.36
ARG1	21,018	31.08
ARG2	2,428	3.59
ARG3	188	0.28
ARG4	26	0.04
ARGM-ADV	9,003	13.31
ARGM-MNR	1,264	1.87
ARGM-LOC	1,717	2.54
ARGM-TMP	4,334	6.41

2.2 基于自动句法DBParser的中文语义角色标注

2.2.1 DBParser介绍

DBParser (Daniel Bikel's Parser)[39] 是Daniel M. Bikel设计实现的一个多语短语结构句法分析器。它提供多种已实现的统计分析模型，比如对Mike Collins的模拟，它也能很方便的扩展到多个领域和多种语言。目前的DBParser句法分析器通过Java实现，提供了英语、汉语和阿拉伯语的设置文件很相关资源，并且分析性能较高。

DBParser支持用户提供训练数据来生成需要的训练模型。训练数据一般是Penn TreeBank的句法数据格式，如图2-3所示。在进行句法分析时，输入的数据支持两种类型：

1. 只有分词结果。格式是 (word1 word2 … wordN)

2. 提供分词结果和词性标注结果。格式是 ((word1 (pos1)) (word2 (pos2)) …(wordN (posN)))

```
( (IP-HLN (NP-PN-SBJ (NR 中国))
          (VP (NP-TMP (NT 去年))
              (VP (VV 发现)
                  (NP-OBJ (QP (CD 十)
                              (CLP (M 个)))
                          (DNP (NP (NP (QP (CD 亿)
                                           (CLP (M 吨)))
                                       (NP (NN 级)))
                                   (NP (NN 储量)
                                       (NN 规模)))
                               (DEG 的))
                          (NP (NN 油气区)))))) )
( (IP-HLN (NP-TMP (NT 一九九七年))
          (NP-SBJ (NP (NP (NN 内地))
                      (CC 与)
                      (NP-PN (NR 香港)))
                  (NP (NN 经贸)
                      (NN 交流)))
          (VP (VA 活跃))) )
```

图 2-3 DBParser 训练数据格式

Figure 2-3 Train data format in DBParser

使用CPB语料的句法数据作训练时，其中有许多短语类型后缀有功能标记，比如图2-3中NP-OBJ。由于DBParser的局限，在自动句法的数据里只有短语类型，而没有功能标记。

2.2.2 语料资源构建

前面 2.1 节我们使用的语料资源是基于手工标注短语结构句法的CPB(Chinese Proposition Bank)语料。本节需要由前一节 CPB 语料来构建基于自动短语结构句法 DBParser 的语料资源，我们把它称为 DB-CPB (Chinese Proposition Bank Based DBParser)。CPB 语料共有 760 个文档，我们把前 100 个作为测试数据，剩余 660 个作为训练数据。

首先构建自动句法分析资源。用训练数据部分 660 个文档的句法分析结果(句法标注文件 chtb_101.fid 到 chtb_760.fid)，通过 DBParser 来生成训练模型，然后再对测试数据部分的 100 个文档进行自动的句法分析。在训练模型文件时，需要删除 CPB 句法结果中空节点， 这是由于空节点是为了层次结构的完整而手工加入的，而自动分析时无法得到。对空节点的处理规则：

1. 如果空节点有非空兄弟节点，则直接删除空节点

2. 如果空节点没有非空兄弟节点，则删除空节点并删除其父节点。同时进行自动句法分析的测试数据中也不再含有空节点信息。

其次构建语义角色标注的资源。把训练数据中所有句子中的空节点删除，并处理对应的角色标记经过处理后的语料做训练语料。遵守如下三个规

则：

1. 如果空节点是一个完整的角色类型或者不是角色类型，删除对应角色标记。

2. 如果空节点是角色的开始，则角色开始标记移到后一个词。

3. 如果空节点是角色的结尾，则角色结尾标记移到前一个词。

根据空标记处理规则，处理测试数据中角色和分词结果的对齐，并加入自动句法分析得到句法结果得到测试数据。同时，我们保持DB-CPB中谓语动词和CPB中谓语动词一致。图2-4是其中一个句子的标注结果。

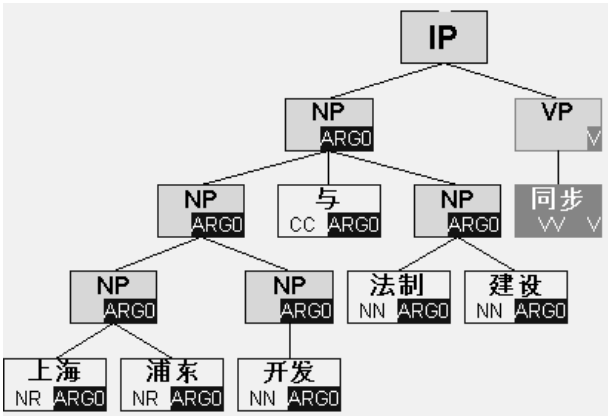

图 2-4 DB-CPB 中的一个标注实例

Figure 2-4 A instance illustrating in DB-CPB

2.2.3 语义角色标注系统

由于DB-CPB语料由DBParser自动分析得到，标注单元也以句法成分为单位，但自动句法分析的句法成分并不能完全和CPB语料的句法成分对应。所以在DB-CPB语料中，存在一部分语义角色找不到句法成分与之对应，召回率受到限制，召回率上限计算公式如(2-2)。

$$Recall_{max} = \frac{ArgNum_{findCand}}{ArgNum_{all}} \tag{2-2}$$

其中，$Recall_{max}$为召回率上限，$ArgNum_{findCand}$为语料中能找到对应标注单元的语义角色数，$ArgNum_{all}$为语料中全部语义角色数。

我们对系统进行了召回率的上限进行计算，如表2-8所示。

表 2-8 系统召回率上限

Table 2-8 Maximal recall of the system

角色	$ArgNum_{all}$	$ArgNum_{findCand}$	$Recall_{max}$ (%)
总体	11,586	10,292	88.83
ARG0	2,646	2,343	88.55
ARG1	3,639	3,153	86.64
ARG2	545	456	83.67
ARGM-ADV	2,063	1,991	96.51
ARGM-LOC	349	318	91.12
ARGM-MNR	325	290	89.23
ARGM-TMP	1,007	938	93.15

从表2-8可以看出，系统的召回率上限能达到 88.83%还是比较高的。这主要由于DBParser是使用了CPB的手工标注句法做训练语料，自动分析得到的短语结构句法其句法成分就能和CPB中句法成分很好的对应。并且，DBParser的句法分析是使用的CPB的分词、词性标注结果，并不是完全自动的，这就减少了由于分词、词性标注错误而导致的召回率上限的降低。

我们选择最大熵分类器，并且参数设置同前一节，系统的实现步骤也同前面一样，采用角色的识别和分类一步实现，并加入对应后处理。前一节对短语结构句法的中文语义角色标注进行了特征的选择，得到了最优特征集。我们直接使用此特征集来实现基于自动句法的中文语义角色标注。对DB-CPB语料进行实验，结果如表2-9所示。

DB-CPB 的语料是使用了 CPB 的分词、词性标注，假定了分词、词性标注完全正确下进行的实验。DBParser 的自动句法分析结果和 CPB 句法结果有很大不同，前面我们分析了系统的召回率上限是 88.83%。这样实验的结果中，只能标出 DBParser 句法分析里有句法成分与之对应的角色。

表 2-9 DB-CPB 和 CPB 语料的最优性能比较

Table 2-9 The best performance comparison between DB-CPB and CPB

系统	Precision（%）	Recall（%）	F-Score（%）
DB-CPB 语料的系统	75.70	67.03	71.10
CPB语料的系统	92.68	89.97	91.31

可以看出，基于手工句法分析的角色标注结果精确率和召回率相差不是

很大，仅有 3%；而基于自动句法分析的角色标注结果，精确率和召回率则相差 7.70%。因此，召回率一直是基于自动句法的语义角色标注的瓶颈。DB-CPB 语料的系统和 CPB 语料系统相比，性能差近 20%。主要影响因素有：

1. CPB 句法分析中，句法节点类型有丰富的功能后缀，比如 NP-SBJ、NP-OBJ、PP-MNR、LCP-TMP 等，这些功能后缀给角色标注提供了丰富的信息，非常有利于角色标注。而在 DBParser 自动句法里，这些后缀信息是无法得到的。

2. CPB 句法为了句子层次结构完整而加入的空标记，而自动句法的结果里这种层次上的完整化信息也是无法得到的。

3. DBParser 句法分析的准确性还无法和 CPB 句法相比，会引入许多错误信息。

因此，语义角色标注是对自动句法分析的一个考验和挑战，同时高质量的自动句法分析结果也是提高角色标注性能的一个努力方向。

2.3 本章小结

本章主要是使用手工标注短语结构句法和自动短语结构句法进行中文语义角色标注。首先使用 CPB 的语料进行了实验，同时提出了许多新特征，通过 χ^2 显著性分析进行特征选择得到最优特征集，并对实验结果进行了深入分析。接下来使用 DBParser 自动句法分析器构建了语料资源，并使用前面的最优特征集进行了实验，得到了基于自动短语结构句法的中文语义角色标注结果，并和手工短语结构句法的结果进行了比较和分析。

第3章 基于依存句法的中文语义角色标注

3.1 基于Penn2Malt依存句法的中文语义角色标注

3.1.1 Penn2Malt介绍

Penn2Malt[3]是一个Växjö大学的计算语言学教授Joakim Nivre[4]实现的一个java工具包，能够把Penn TreeBank短语结构句法，通过规则的方法转化成依存句法，并且转化结果非常好。目前能够对中文和英文的短语结构句法进行转化。软件包Penn2Malt.jar(需要jdk 1.5 及以上)，Penn TreeBank的规则文件和待转化的Penn TreeBank短语结构句法文件。结果输出分 3 个文件，转化的依存文件后缀tab，词性列表文件后缀pos，依存类型文件后缀dep。

Penn2Malt 转化过程中，保留了 CPB 的分词、词性标注结果不变，在词语之间加入了依存弧以及依存关系类型。Penn2Malt 转化 Penn TreeBank 得到的依存句法有 12 类依存关系，分别是 PMOD，OBJ，VMOD，SUB，P，PRD，NMOD，AMOD，ROOT，DEP，VC，SBAR，其中 ROOT 是根节点的依存类型。

3.1.2 语料资源构建

第 2 章我们研究了基于短语结构句法的中文语义角色标注，分别使用了手工标注的 CPB 语料和基于 DBParser 自动句法分析的 DB-CPB 语料。本章我们需要构建基于依存句法的语义角色标注语料资源，通过 Penn2Matl 工具把 CPB 短语结构句法转化成依存句法来实现，我们把它称为 PM-CDB(Chinese Dependency Bank based Penn2Malt)。

和 CPB-DB 一样，我们把 CPB 语料的 760 个文档前 100 个作为测试数据，剩余 660 个作为训练数据来构建 PM-CDB。首先用 Penn2Malt 把 CPB760 个文档的全部短语结构句法转化成依存句法，转化过程中删去由于层次完整而手工加入的空节点。然后把得到的数据和 CPB 中角色对齐，并

[3] http://w3.msi.vxu.se./~nivre/research/Penn2Malt.html
[4] http://w3.msi.vxu.se/users/nivre/

且保持谓语动词和 CPB 的一致。

3.1.3 系统的召回率上限分析

第 2 章我们是以句法成分为标注单元，提取可能为语义角色的句法成分作为角色候选。本章我们以依存句法来实现语义角色标注，是把依存关系作为标注单位，提取依存句法的依存关系对应节点作为角色候选。

由于CPB语料在标注语义角色时，是以句法成分为单元的，每个语义角色必然对应一个句法成分。而短语结构句法转化成依存句法后，并不是每个句法成分都能找到对应的依存关系，也就必然有一部分语义角色无法召回了。我们对PM-CDB的语料进行了分析，通过公式(2-2)得到了系统召回率的上限。表3-1列出来总体召回率上限以及约占全部角色 90%的几类主要角色的召回率上限。

表 3-1 系统召回率上限

Table 3-1 Maximal recall of the system

角色	$ArgNum_{all}$	$ArgNum_{findCand}$	$Recall_{max}$ (%)
总体:	11,586	10,805	93.26
ARG0	2,646	2,272	85.90
ARG1	3,639	3,407	93.63
ARG2	545	536	98.39
ARGM-ADV	2,063	2,055	99.66
ARGM-LOC	349	342	98.25
ARGM-MNR	325	305	94.07
ARGM-TMP	1,007	990	98.34

由表3-1可以看出，总体召回率上限约 93%，而和CPB语料召回率上限100%相比还存在一定距离。同时，占总体 34%的ARG0 召回率上限才约86%，对系统的性能会有很大影响。不同类型的召回率上限不同，跟短语结构句法到依存句法转化的规则密切相关，说明依存分析还有待提高。

3.1.4 语义角色标注系统

第 2 章 2.1 节详细介绍了基于短语结构句法的中文语义角色标注系统的构建，包括标注单元的选择、标注步骤和分类器、特征选择和后处理。本章

构建的基于依存句法的中文语义角色标注系统的构建与之相似，我们以依存关系为标注单元，采用相同标注步骤，选择最大熵分类器并且参数设置相同。在特征选择上，跟短语结构句法的语义角色标注有了很大不同，同时后处理规则进一步细化完善。

3.1.4.1 特征构造

基于依存句法的语义角色标注和基于短语结构句法的语义角色标注相比，特征差别很大。我们先简单介绍在Hacioglu[28]文章中提到的几个特征。图3-1是PM-CDB中的一个标注实例，以依存关系SUB为角色候选实例。

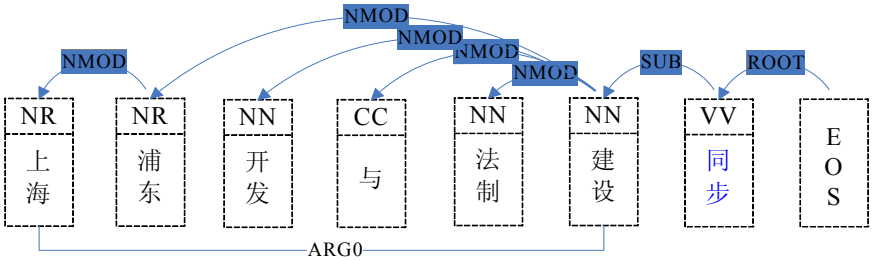

图 3-1 PM-CDB 中的一个标注实例

Figure 3-1 A instance illustrating in the PM-CDB

- 依存关系结构特征

1. 依存关系类型。角色候选实例的依存关系类型为“SUB”。
2. 核心词：依存关系中被修饰词。角色候选实例的核心词是“同步”。
3. 依存词：依存关系中修饰词。角色候选实例的依存词是“建设”。
4. 核心词词性和依存词词性

- 谓语动词结构特征

1. 谓词子节点的词性模式：谓词所有直接子节点的词性，按照从左到右的顺序构成的模式。角色候选实例的谓词子节点词性模式是“NN”。

2. 谓词子节点的依存关系模式：谓词所有直接子节点和谓词的依存关系，按照从左到右的顺序构成的模式。角色候选实例的谓词子节点依存关系模式是“SUB”。

3. 谓词兄弟节点的词性模式：谓词所有直接兄弟节点的词性，按照从左到右的顺序构成的模式。角色候选实例的谓词由于没有直接兄弟节点，该特征用标记“NPPS”。

4. 谓词兄弟节点的依存关系模式：谓词所有直接兄弟节点和其父节点

的依存关系类型，按照从左到右的顺序构成的模式。角色候选实例的谓词由于没有直接兄弟节点，该特征用标记“NTPS”。

- 谓语动词-依存关系结构特征

1. 家族关系：依存节点和谓词的家族关系，包括父子关系、兄弟关系、祖先和后继的关系、其他关系。角色候选实例的家族关系特征是“父子关系”。

2. 位置：该特征和基于短语结构句法的语义角色标注中相同。

3. 路径：依存分析树中从依存关系到谓词的依存关系路径。角色候选的路径特征是“SUB↑ROOT”。

语义角色标注中的特征绝大部分是句法层次的特征。在依存分析中，这些特征主要从两个角度来提供角色判别的信息：依存弧方向和依存关系类型。依存弧方向可以得到依存关系对应的词序列，直接影响系统的召回率。多样化的依存关系类型了包含了角色类别之间的差异性信息，直接影响系统的精确率，因此依存句法分析的性能直接影响语义角色的性能。同时也暗示了特征构造时应该从这两个角度考虑。上述特征中，依存关系结构特征充分考虑了这两个因素，能够有效挖掘依存句法中用于角色标注的信息。

不同谓语动词的角色框架不同，所以挖掘谓语动词的深层信息也很重要。谓语动词结构特征恰好挖掘了这部分信息，其中子节点的模式特征体现了谓词的语法框架，兄弟节点的模式特征体现了其上下文信息。谓语动词-依存关系结构特征则综合考虑了上述因素，能够有效预测角色类别。我们根据这些信息，并借鉴基于短语结构句法的语义角色标注中使用的特征，提出了一些更有效丰富的特征，同时也引入了一些组合特征，下面将详细介绍。

- 依存关系结构特征

1. 前一个词、后一个词、第一个词、最后一个词其词本身、词性及对应依存关系类型。

2. 左右兄弟节点的依存关系，左右兄弟节点对应的词及词性。

3. 依存关系词序列的词性模式和依存关系模式。这两类的特征的构造方法是：取依存关系第一个、最后一个词的词性或依存关系作模式边界，中间取其余词序列不重复出现的词性或依存关系。由于角色类型 ARG2 对应的词序列一般是以 “介词-名词序列”模式出现，所以此特征能够有效预测该类角色。

- 谓语动词结构特征

1. 谓语动词类别信息，谓语动词

2. 谓语动词子节点、兄弟节点其词性、依存关系模式类特征的简化模式。这几类特征是前述对应特征的泛化，由于前面的这几类模式特征数据稀疏，所以对其进行泛化能够提高其对系统性能的贡献。简化方法是：遍历对应的序列模式，对于相邻多次出现的同一词性或依存类型，只保留一个。

- 谓语动词-依存关系结构特征

1. 最近公共父节点的词性，最近公共父节点的依存关系。

2. 部分路径和词性路径。该特征是对路径特征的泛化，前面的路径特征是以依存关系进行提取的，为了加入词性信息，我们也加入了词性的路径特征。同时，为了进一步解决数据稀疏问题，也对路径特征进行了简化，简化方法同谓语动词序列模式特征的简化。

3. 泛化后的距离。由于距离特征稀疏，效果并不明显，所以采用分段方式进行泛化，即考察距离以 5 为差，分 0-5，6-10，11-15 等。

4. 距离相对节点长度。如果角色候选对应的节点词序列很长，对应的距离值就大，所以为了消除候选节点长度对距离的影响，采用距离值除候选节点长度作为特征。表3-2给出了部分特征名称及其标记。

前面介绍了许多有效的单一特征，有些单一特征和其他特征组合在一起时，对角色预测更加有效，因此就引入了许多组合特征。特征组合可以选择两个特征组合，也可以选择多个特征进行组合，我们主要使用了两个特征的组合。组合方法是：依存关系结构特征、谓语动词结构特征、谓语动词-依存关系结构特征 3 种结构的特征任意两种(包括同种结构)进行组合，得到 6 种组合：依存关系结构特征-依存关系结构特征组合、依存关系结构特征-谓语动词结构特征组合等。下面简单列举几个组合特征。

- 组合特征

1. 谓语动词-依存关系类型
2. 谓语动词类别-路径
3. 谓语动词-路径
4. 谓语动词-核心词
5. 前一个词性-路径词性模式的简化
6. 谓语动词-谓语动词子节点的词性模式
7. 依存关系类型-节点的词序列模式
8. 其他组合特征

表 3-2 部分特征名称及标记

Table 3-2 The name and label of some features

依存关系结构特征			
特征名称	标记	特征名称	标记
依存类型	Type	核心词	Headword
前一个词	PreWord	最后一个词	LastWord
前一个词词性	PosOfPW	核心词词性	PosOfHW
依存词	Depword	依存词词性	PosOfDW
候选词序列的词性模式	BegEndPosPattern		
谓语动词结构特征			
特征名称	标记	特征名称	标记
谓词类别	PdClass	谓词	Predicate
子节点依存类型模式简化	RelSimpPatOfPdChildren		
谓语动词-依存关系结构特征			
特征名称	标记	特征名称	标记
家族关系	FamilyShip	位置	Position
部分路径的依存类型模式简化	SimpRelPathPattern	部分路径的词性模式简化	SimpPosPathPattern
路径	Path		

3.1.4.2 特征选择

前面介绍了许多单一特征和组合特征，但并不是每个特征都是有效的，需要选择合适的特征集。特征之间的相互影响是非常复杂的，一个非常有效的特征，在某个特征出现时可能会降低系统性能，反之亦然。因此自然语言处理领域特征选择一直是个很难的问题，第 2 章我们通过 χ^2（自由度为 1）显著性检验来选择特征，本章我们采用 χ^2 显著性检验和贪心算法[40]相结合的方法来选择特征。

选择前面介绍的 45 类单一特征，并且构建了 52 类组合特征作为特征候选集按如下算法进行选择：

1. 记 97 类单一和组合特征候选集为 FC。选择 10 类单一特征建立 BaseLine 特征集记为 FB，FB = { Type，Headword，Depword，PosOfHW，PosOfDW，PdClass，Predicate，FamilyShip，Position，Path }；

2. 从候选集中去掉 BaseLine 特征集，则 $FC = FC - FB$。

3. 从 FC 中逐次选一个特征加入 BaseLine 中，对该特征加入后系统性能变化进行 x^2 显著性检验(上侧 α 分位数 $\alpha = 0.05$)，得到一组性能显著提高的

特征集记为 FI 。

4. 把 FI 中使性能提高最高的那个特征 f_max 加入 FB ，即 $FB = FB \cup \{f\_max\}$，候选集 $FC = FI - \{f\_max\}$。

5. 重复步骤 3，4 直到候选集 FC 为空。

3.1.4.3 后处理规则

在后处理上，保留了第 2 章对互相嵌套标注单元的处理和相同角色类型多次出现的处理方法。对 X-PSR 和 X-PSE 类型的处理进行了改进和扩展，也引入了对 X-QTY 类型的处理，其中 X 分别为 ARG0，ARG1，ARG2。

1. 对 X-PSR 和 X-PSE 类型的处理。如果在句子标注结果里 X-PSR 和 X-PSE 没有同时出现，比如只出现 X-PSR，这时做如下处理：如果所有出现的 X-PSR 第一预测概率都小于某个阀值，则把 X-PSR 类型更新为对应第二概率预测的非空类型；否则，从第一概率预测为空角色但第二概率预测为非空角色的候选中，找 X-PSE 类型插入标注结果中。对 X-PSE 的也采用相同的处理方法。

2. 对 X-QTY 类型的处理的处理。如果预测为 X-QTY 的候选，不满足：其对应词序列的词性是 AD(副词)、CD(基数词)、M(量词)其中一种，则处理如下：如果第一预测概率超过某个阀值，则把 X-QTY 更新为 X 类型；否则把 X-QTY 更新为第二概率预测为非空的类型，第二概率预测为空角色时删除该 X-QTY 类型。

3.1.5 实验结果及分析

通过前面特征选择算法，对 97 类单一和组合特征进行选择，得到了最优特征集 F_{best} ， $F_{best} = FB \cup \{$ BegEndPosPattern, RelSimpPatOfPdChildren, Predicate+Type, PdClass+PreWord, PdClass+Type, Position+Path, PdClass +LastWord, SimpPosPathPattern+PosOfPW, PdClass+SimpPosPathPattern, Predicate+SimpRelPathPattern$\}$。 使用BaseLine特征的性能和最优特征集 F_{best} 的性能如表3-3。

表 3-3 特征选择前后系统性能比较

Table 3-3 The performance comparison after feature selecting

系统	Precision（%）	Recall（%）	F-Score（%）
BaseLine 系统	85.28	74.82	79.70
F_{best} 系统	90.01	80.91	85.22

由表3-3可以看出，BaseLine系统只选用了 10 个单一特征，F-Score就能达到 79.70%，主要由于这 10 个单一特征分别选自 3 类结构特征，从不同角色挖掘了词、词性、依存句法的深层信息，这些信息对于角色预测是非常有效的。通过特征选择，加入选择后的最优特征后，系统性能提高了近6%，可见构造有效的特征对于语义角色的预测是至关重要的。特征选择的结果只是从 97 个特征或特征组合中选择了 10 个，这是由于特征之间的相互作用非常复杂，有些特征的加入会带给分类器错误信息导致系统性能降低，而有些特征的加入则会带给分类器允余信息，这时系统性能变化不大，但是增加了特征数据规模导致系统的效率降低。

第 2 章用CPB的语料实现了基于短语结构句法的中文语义角色标注，在通过特征选择后，使用最优特征集得到了系统的最高性能 91.31%。这部分我们用构建的PM-CDB语料实现基于依存句法的中文语义角色标注，也作了特征选择，在最优特征集上系统性能是 85.22%。两者性能比较如表3-4。

表 3-4 PM-CDB 和 CPB 系统性能比较

Table 3-4 The performance comparision between PM-CDB and CPB

系统	Precision (%)	Recall (%)	F-Score (%)
PM-CDB 语料的系统	90.01	80.91	85.22
CPB语料的系统	92.68	89.97	91.31

从表3-4中可以看出，PM-CDB语料的系统精确率可以和CPB语料的系统精确率相比，但是召回率差了约 9%。这主要由于语料的质量不同，CPB语料是手工标注的高质量语料，而PM-CDB中依存句法是有CPB短语结构句法用规则的方法转化得到。规则的方法一个局限性就是规则有限，无法覆盖所有可能的情况，并且规则正确性直接影响结果。前面我们得到了系统的召回率上限是 93.26%，这样在由短语结构句法到依存句法的转化过程中就有近 7%的角色由于没有标注单元与之对应而无法召回。

3.2 基于HIT-IR自动依存句法的中文语义角色标注

3.2.1 HIT-IR语言技术平台

语言技术平台LTP是一套面向Web 基于XML 的中文语言处理平台[41]。语言理解是一个复杂的分层互动式系统，从以句子为处理单元的词法、词义、句法、句义、语用分析，到以篇章为处理单元的指代消解、自动文摘、文本分类，再到以篇章集合为处理单元的多文档文摘、文本检索等，构成了一个复杂的认知体系。LTP 包含5 项主要内容：语言技术置标语言LTML、基于DOM Tree 的一套DLL 模块、一套可视化工具、基于LTML 的语料库资源、以及基于Web Service 的网络应用。目前实现了包括断句、分词、词性标注、命名实体识别、词义消歧、依存句法分析、语义角色标注、指代消解、自动文摘和文本分类等10 项中文处理技术。LTP 解决了语言处理的多层次（句、篇，文本集合）机内表示问题，语言处理结果的可视化问题，以及各处理模块避免重复调用的问题，方便了学者们致力于自然语言处理领域的研究。图3-2是LTP的在IE上的部分显示结果。

依存句法是由法国语言学家L.Tesniere 在其著作《结构句法基础》（1959 年）中提出，对语言学的发展产生了深远的影响，特别是在计算语言学界备受推崇。依存语法通过分析语言单位内成分之间的依存关系揭示其句法结构，主张句子中动词是支配其他成分的中心成分，而它本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[42]。二十世纪七十年代，Robinson提出依存语法中关于依存关系的五条公理, 在处理中文信息的研究中，中国学者提出了依存关系的第五条公理[43]，它们是：

1. 一个句子中只有一个成分是独立的；
2. 其它成分直接依存于某一成分；
3. 任何一个成分都不能依存于两个或两个以上的成分；
4. 如果 A 成分直接依存于 B 成分，而 C 成分在句中位于 A 和 B 之间，那么 C 或者直接依存于 B，或者直接依存处于 A 和 B 之间的某一成分。
5. 中心成分左右两边的其它成分相互不发生关系。

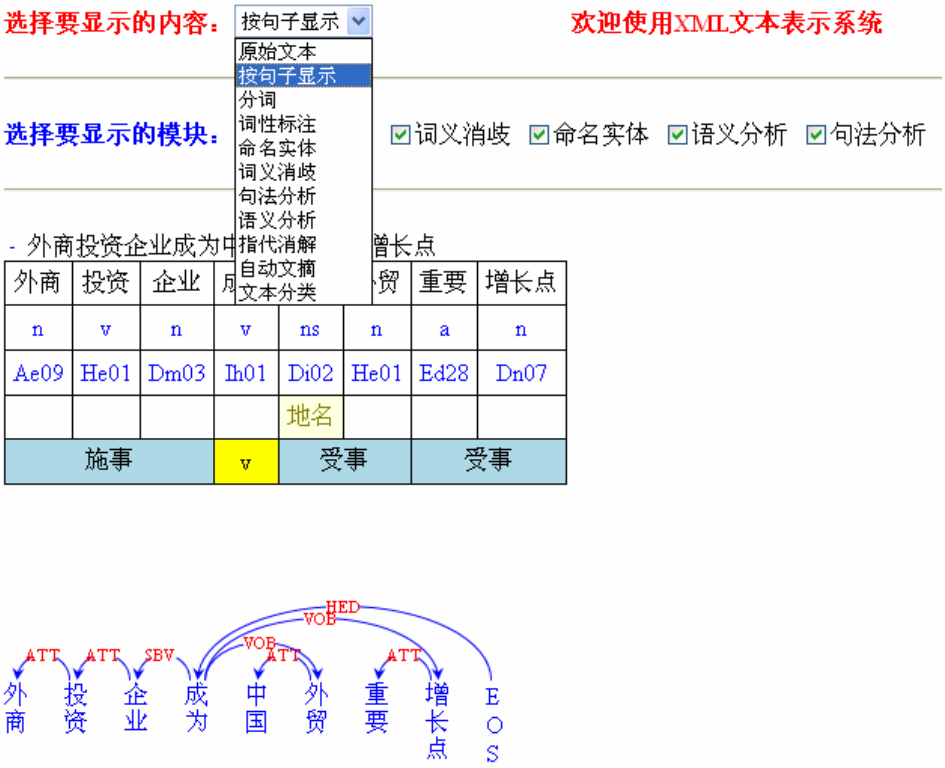

图 3-2 HIT-IR LTP 处理结果在 IE 上的部分显示

Figure 3-2 Some processed result of HIT-IR LTP displayed on IE

哈尔滨工业大学信息检索研究室的中文依存句法分析系统用于对汉语进行句法分析，将句子由一个线性序列转化为一棵结构化的依存分析树[44]，通过依存弧反映句子中词汇之间的依存关系。示例如图3-2。

由图3-2可以看出利用LTP对该例句的分析结果。依存分析的结果，直观的表示可以用依存文法弧线图。图中的弧线从某一受支配成分指向其支配成分。弧上的标记表示依存关系的类型。例如，“企业”和“成为”之间存在依存关系SBV（主谓关系），其中，“成为”是这个关系的核心成分，“企业”依存于“取消”。在此系统中，依存关系类型共有 24 种。该图中共有 7 个弧，句子的末尾增加一个句尾标志“<EOS>”，由其支配全句的核心词。

3.2.2 语料资源构建

前一节我们通过基于规则方法实现的软件包Penn2Malt，把CPB基于短语结构句法的语料转化成了基于依存句法的语料，但是Penn2Malt结果是保

留了CPB的分词、词性标注不变基础上得到了依存句法。我们需要实现一个基于完全自动依存句法分析的中文语义角色标注系统，也即分词、词性标注、依存分析以及其他的自然语言技术都自动获得。对此，我们使用信息检索研究室的语言技术平台LTP（Language Techonology Platform），建立一个基于依存句法分析的中文语义角色标注的语料资源，称为HIT-IR CDB（Chinese Dependency Bank based HIT-IR LTP）。图3-3是HIT-IR CDB中一个标注实例。

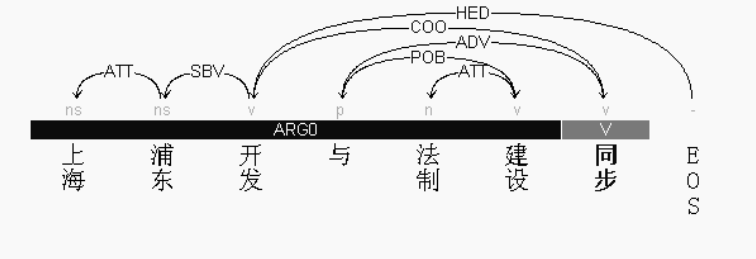

图 3-3 HIT-IR CDB 中一个标注实例

Figure 3-3 A instance illustrating in HIT-IR CDB

构建语料资源的思路基本是：

1. 获取 CPB 语料 760 篇文档的不包含空标记的原始句子。

2. 对这些句子通过 LTP 得到自动分词、词性标注、命名实体识别、依存句法分析的结果。

3. 把 CPB 语料中的语义角色信息对应添加到到第 2 步自动分析的结果重。

在构建过程中，需要考虑和处理下面四个问题：

1. 动词识别：由于 LTP 的分词和 CPB 分词结果不完全相同，LTP 的词性标注体系 CPB 词性标注体系的也不同，需要在 LTP 自动分析后的语料中识别语义角色标注谓语动词。目前谓语动词的自动识别很难准确实现，LTP 词性标注存在一些错误，不能直接把 LTP 标为动词的词作为谓语动词。我们处理方案是：CPB 中谓语动词能够在 LTP 自动分词结果中找到的词，在 HIT-IR CDB 作为谓语动词

2. 空标记处理：CPB语料中为了对句子进行深层次分析，加入许多空标记，比如：*T*, *PRO*, *pro*, *OP*, *RNR*等。而LTP自动分析结果中这些空标记是无法得到的，如图3-4 所示。对此，需要对CPB中空标记对应的语义角色进行处理，方案是：

(1) 如果空标记是一个完整语义角色，则删除对应角色

(2) 如果空标记是语义角色的开始，则对应角色起始位置移到下一个词

(3) 如果空标记是语义角色的结尾，则对应角色结束位置移到前一个词

3. 语义角色对齐：由于 LTP 分词和 CPB 分词不同，需要把 CPB 标注中的语义角色对齐到自动分析的语料上。处理方案是：对于 CPB 中所有角色，从 LTP 分词的词序列中查找。如果找到则进行标注，找不到则角色丢失。

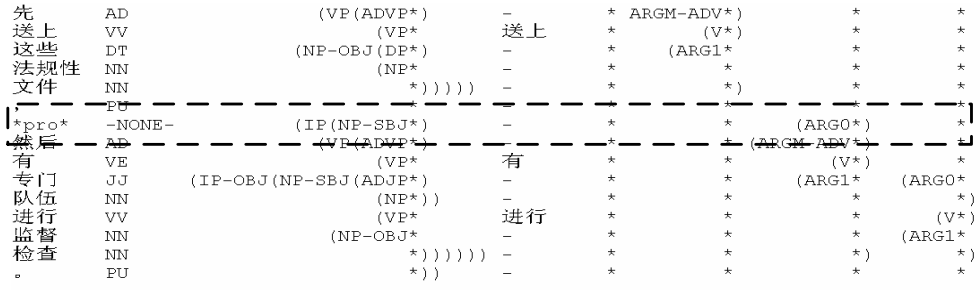

图 3-4 CPB 中语料数据格式

Figure 3-4 Data format in CPB

4. 一些非汉字符号的处理：CPB 中有许多非汉字符号，比如“，”，“…”，“『”，“＊”，“＜”等，LTP 的词性标注为标点符号(/wp)，而 LTP 的依存分析规定：词性为 wp 的词不依存于任意词，任意词也不依存于它。对此，需要对含有这些符号的角色进行处理。处理方案同空标记对应角色的处理。

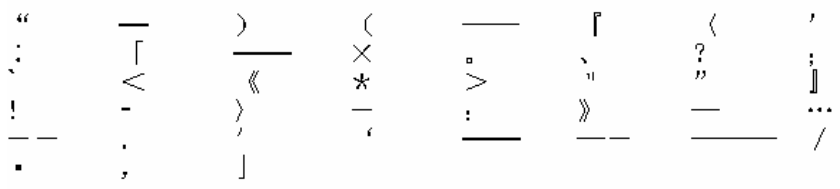

图 3-5 CPB 中一些非汉字符号

Figure 3-5 Some Non-Chinese characters in CPB

对 760 篇文档的全部句子，从CPB到CDPB转化后，我们对转化数据分别进行了语义角色和动词统计和分析，如表3-5和3-6。

从表3-5可以看出，转化后角色损失比较大，达 25.86%。引起语义角色损失的因素主要如下几方面：

1. 角色损失主要由空标记删除和谓语动词损失引起。空标记删除引起的损失是不可避免的，谓语动词的损失主要是由于分词错误引起，导致对应

的谓词找不到，此谓词对应的全部语义角色也随之丢失。

2. 由于分词不对齐引起的角色缺失很少，说明 HIT-IRLAS 的分词性能很高，基本和 CPB 的分词接近。

3. 非汉字符号处理引起的损失。这部分角色损失很少，但统计显示以非汉字符号为边界的角色占全部 2.23%，这部分如果不加处理在 HIT-IR CDB 中是无法找到的，因此处理后可以提高系统召回率。

表 3-5 从 CPB 到 HIT-IR CDB 转化语义角色统计

Table 3-5 Statistics of arguments converting from CPB to HIT-IR CDB

角色	数量	比例 (%)	累计 (%)
HIT-IR CDB 全部角色	77,327	74.14	—
删除空标记丢失角色	20,652	19.80	25.86
动词丢失引起丢失角色	5,736	5.50	
由于分词不对齐丢失角色	554	0.53	
处理非汉字符号引起丢失角色	32	0.00	
以非汉字符号为边界的角色	2325	2.23%	—
CPB 全部角色	104,304		

在转化过程中，由于分词不同有部分谓词无法找到。由于词性标注的不同，而一部分谓词没有被标为动词。如表3-6：

表 3-6 从 CPB 到 HIT-IR CDB 转化中谓语动词的统计

Table 3-6 Statistics of predicate converting from CPB to HIT-IR CDB

动词	数量	比例 (%)
HIT-IR CDB 分词不同找不到的谓词	132	0.36
HIT-IR CDB 词性标注没标为动词“v”的谓词	2,422	6.68
CPB 的谓语动词	36,252	

转化过程中，未被标为动词的谓词有四种词性：VA，VV，VC，VE，其分布比例如表3-7。结合表3-7对动词损失因素进行了分析：

1. 词性标注不同引起。CPB 中标为动词而在自动分析中被标注为习语（i），如“有机可乘/i”，“纸上谈兵/i”等。CPB 中的形容词性动词 VA 在 LTP 词性标注中许多被标为形容词（a），如“安宁/a”，“便利/a”等。

2. 分词不同引起。如：CPB 中动词“评出”，LTP 分析时分词“评/vg 出/vq”，“预估”自动分析为“预/d 估/vg”。

分词不同导致的谓词无法找到，但词性没标为动词的谓词我们在系统中也把他们作为了谓词来处理。对于 760 篇文档转化得到的 HIT-IR CDB 语

料，我们仍然把前 100 篇作为测试数据，后 660 篇作训练数据来进行实验。

表 3-7 从 CPB 到 CDPB 转化中未标为动词“v”的谓词分类统计

Table 3-7 Statistical categories to lost Verbs converting from CPB to HIT-IR CDB

损失动词	2,422
VA	768
VV	1,643
VC	5
VE	6

3.2.3 语义角色标注系统

由于标注单元由CPB的句法成分转为HIT-IR CDB的依存关系，我们也用前一节PM-CDB语料的系统上限分析方法，对HIT-IR CDB语料进行了系统上限分析，如表3-8所示。

表 3-8 系统的召回率上限分析

Table 3-8 Maximal recall of the system

角色	$ArgNum_{all}$	$ArgNum_{findCand}$	$Recall_{max}$ (%)
总体:	11,002	7,379	67.07
ARG0	2,516	1,630	64.79
ARG1	3,465	2,023	58.38
ARG2	507	317	62.52
ARGM-ADV	1,925	1,685	87.53
ARGM-LOC	335	237	70.75
ARGM-MNR	313	177	56.55
ARGM-TMP	981	798	81.35

召回率上限与Penn2Malt语料系统的召回率上限作了比较，如表3-9。可以看出，召回率上限只有 67.07%，和基于PM-CDB的系统召回率上限93.26%相比差很多。充分说明依存句法分析的的质量对系统性能的影响。

表 3-9 召回率上限比较

Table 3-9 Maximal recall comparision

系统	召回率上限 (%)
PM-CDB 的系统:	93.26
HIT-IR CDB 的系统	67.07

由表3-8可以看出主要角色的召回率上限差距很大，最高的ARGM-ADV能达到近 90%，而最低的ARGM-MNR才近 57%。系统召回率主要受这些低召回率角色的制约。系统召回率上限很低，主要是由于LTP的依存分析结果不是很准确，词语之间的依存关系弧的错误引起了标注单元词序列的错误。目前完全自动的中文依存句法分析还不是很成熟，有待进一步的改进。

我们仍然以依存关系作为角色标注单元，选用最大熵分类器，参数配置同前一节基于 PM-CDB 的系统，标注步骤也与之类似，但加入了过滤规则。规则如下：

1. 如果当前节点的父节点编号为负数，则过滤掉。这个规则主要过滤一些非汉字的标点符号节点，句子中依存关系为“HED”的核心动词节点。

2. 如果当前节点为对应谓语动词，则过滤掉

3. 如果当前节点是谓词节点的父节点以及祖先节点，则过滤掉。语义角色标注中，包含谓词的词序列是不能作为语义角色的。

前面也对基于依存句法的中文语义角色标注构造了许多有效的特征，并且进行了特征选择，找到了最优特征集 F_{best}。这里我们把 LTP 中命名实体的结果作为特征加入，NE(Name Entity)特征定义为：如果标注单元对应词序列是一个命名实体，则把命名实体类型作为特征，否则把特征赋予非命名实体标记，这样 $F_{best}=F_{best}\cup\{\text{NE}\}$。

3.2.4 实验结果及分析

我们也使用前一节特征选择的BaseLine特征集和加入命名实体后得到的最优特征集分别进行了实验，结果如表3-10所示。并对已有系统的性能进行了比较，如表3-11所示。

表 3-10 HIT-IR CDB 系统特征选择前后系统性能比较

Table 3-10 The performance comparison of HIT-IR CDB after feature selecting

系统	Precision (%)	Recall (%)	F-Score (%)
BaseLine 系统	68.45	30.82	42.50
F_{best} 系统	77.22	37.92	50.86

表 3-11 各个系统性能的比较

Table 3-11 The performance comparison of each system

系统	Precision (%)	Recall (%)	F-Score (%)
HIT-IR CDB 系统	77.22	37.92	50.86
DB-CPB 系统	75.70	67.03	71.10

从表3-11可以看出，最优特征集的系统性能和BaseLine系统相比，F-Score提高了 8%，相比于PM-CDB语料系统提高的约 6%还多些。对于基于完全自动依存句法的角色标注系统来说，特征的选择更加重要，因为完全自动依存句法分析不是很准确，只有选择丰富有效的特征，才能减少由于句法分析而引起的分类器预测错误。

实验结果显示系统性能很低，尤其是召回率才 37.92%。前面对系统的召回率上限做过分析，只能达到 67.07%。召回率一直是基于自动依存的中文语义角色标注的瓶颈。LTP 的依存句法，主要分析词语之间的依存关系及依存类型，依存句法分析的两个主要方面是词语之间依存弧和依存关系的类型。其中，依存弧是否正确，直接影响标注单元对应的词序列，是影响系统的召回率的一个重要因素；依存关系类型在特征构造中多次使用，是构成许多特征的重要成分，对于角色类别的预测至关重要。因此 LTP 依存分析的准确性是造成系统性能很低的主要因素。

系统的精确率 77%还相对比较高，一方面由于选用了最优的特征，从不同角度挖掘了对应语义角色类型涵盖的信息。再一方面也由于 LTP 的依存句法分析有许多丰富的依存关系类型，这些依存关系类型共有 24 种，包括定中关系 ATT，状中结构 ADV，数量关系 QUN，动宾关系 VOB，主谓关系 SBV，“的”字结构 DE， “把”字结构 BA，“被”字结构 BEI 等。这些依存关系对于与之相关的角色类型如 ARG0，ARG1，ARG2，ARGX-QTY(X 为 0,1,2)，ARGM-TMP，ARGM-MNR 等，能够给出充足的预测信息。

实验结果显示，系统整体性能还是较低，通过对错误进行分析，主要在于这几个方面：

1. 词性标注的错误。由于某些语义角色，偏向于一定词性的词来充当。比如 ARG0，ARG1 一般都是名词性词或有修饰成分的名词序列来充当。因此，词性标注错误会导致语义角色预测错误。例如 LTP 标注句子“上海/ns 浦东/ns 开发/v 与/p 法制/n 建设/v 同步/v”，而在 CPB 语料中“开发”和“建设”都是标为名词“NN”。

2. 依存句法分析性能不是很高。依存节点不能和语义角色很好对应，

所以导致召回率很低。实验中使用自动的依存句法分析，而句法分析性能有限，对于短距离的依存关系和关系类型分析较准，但句子如果很长时，长距离的依存分析就不准了。并且，由于词性标记错误，会级联导致依存弧的错误。如图3-6所示。

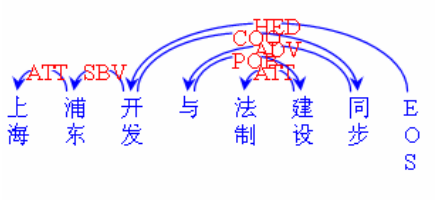

图 3-6 依存弧错误例图

Figure 3-6 A instance of dependency arc error

3. 还有一些角色本身难以和类似的角色区分。比如 ARGX-PSR，ARGX-PSE 等这些类角色和对应 ARGX(X 为 0,1,2)对于人来说都难以区分，机器自动分析时更为不易。这些难以区分的角色比较多，整体分析准确度低。

3.3 本章小结

本章介绍了使用依存句法进行中文语义角色标注的这种新方法。首先用 Penn2Malt 转化 CPB 语料得到准确度较高的依存句法分析结果，同时提出丰富有效的特征并结合 χ^2 显著性分析和贪心算法进行了特征选择，构建了基于依存句法的角色标注系统并进行实验。接下来介绍了哈工大信息检索研究室的语言技术平台和依存句法，并构建了基于完全自动依存句法的语料资源，介绍了使用自动依存句法进行角色标注的方法，给出实验结果。最后把实验结果和使用短语结构句法的结果进行了比较分析。

第4章 短语结构句法和依存句法相结合的中文语义角色标注

4.1 多句法结合的意义

我们语义角色标注使用了最大熵分类器，也有学者们使用了多分类器融合，对多个分类器输出的结果进行融合，这种方法往往获得比其中任何一个分类器好的效果，因此经常被应用于各种评测之中。Ngai等[45]、以及等Tsai[46]就利用多分类器融合的方法参加了SENSEVAL-3 和CoNLL2005 语义角色标注的比赛。多分类器融合是依赖不同分类器的预测结果来综合判断一个角色候选是否某类型角色，它既不能增加角色候选的数量，也不能丰富角色候选的特征信息。因此这种融合只是使用了不同的机器学习方法，而没有使用更多的自然语言处理技术来指导语义角色标注。由于机器学习方法日渐成熟，并且不同机器学习方法适用于具体的问题，因此这种方法对性能的提高是有限的。

为了使用更多自然语言处理技术来提高语义角色标注的性能，我们采用多句法分析结果相结合的方法，目前使用短语结构句法和依存句法结果相结合。在英文语义角色标注中，Koomen等人[15]、M`arquez等人[47]以及Pradhan等人[48,49]就采用多句法结果融合的方法，而且也取得了非常好的效果。短语结构句法和依存句法从不同角度对句子的语法关系进行了分析，短语结构句法以短语块为单位，能很好描述语义角色的边界；而依存句法有丰富的依存关系类型，对于语义角色类别的预测能提供有用的信息。首先，短语结构句法和依存句法分别以句法成分和依存关系为标注单元，能弥补单一句法角色候选有限的缺陷。其次，两种句法分析从不同角度对角色候选提供预测信息，使得分类结果更加准确可信。当然，多句法结合的语义角色标注需要不同句法分析结果都准确度较高，才能取得不错的效果。

4.2 系统构建

4.2.1 系统实现框架

CPB语料中语义角色标注是以句子为单位的，因此我们的系统也以句子为单位进行角色标注，图4-1是处理一个句子的实现框架。

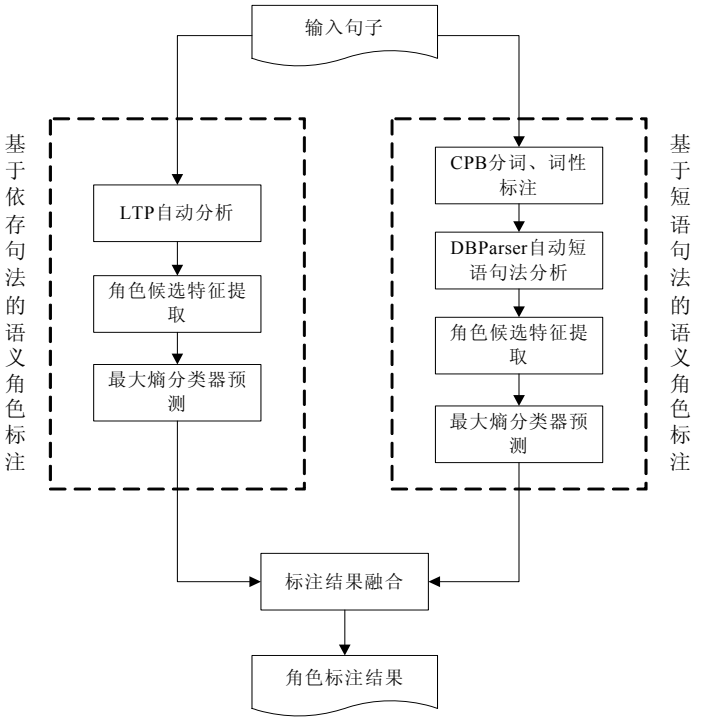

图 4-1 系统的实现框架

Figure 4-1 System framework

由于 LTP 的分词结果和 CPB 的分词结果不同，而角色候选的位置信息是以角色在分词结果中起始位置记录的，因此在标注结果融合之前需要做一个预处理。处理方法如下：

1. 由角色候选的位置信息，在 LTP 的分词结果中找到对应词序列。
2. 根据该词序列，在 CPB 的分词结果中找该词序列的起始位置。
3. 如果能找到，则由此起始位置更新角色候选位置信息；否则，删除该角色候选。

比如句子“目前，王翔又开始了九龙街的建设。”，LTP 的分词结果是

“目前 ， 王翔 又 开始 了 九龙 街 的 建设 。”，而 CPB 的分词是“目前 ， 王翔 又 开始 了 九龙街 的 建设 。”这样，对于前者结果中的角色候选“九龙”在 CPB 中是无法找到对应词序列的。这部分候选肯定不是语义角色，因此删除这些候选能提高系统精确率。

4.2.2 召回率上限分析

第 2 章实现了的基于自动短语结构句法DBParser的角色系统，和第 3 章实现了的基于自动依存句法的角色系统，并且都对应的进行了系统的召回率上限分析。由于前者是以依存关系为标注单元，而后者以句法成分为标注单元，这样它们的会有一部分角色候选对应不同的词序列，两者结合后有望找到更多的候选参与角色预测。因此，对结合后系统的召回率上限作了分析，如表4-1所示。

表 4-1 结合后系统召回率上限

Table 4-1 Maximal recall after combination

角色	$Recall_{max}$ (%)
总体	90.89
ARG0	90.60
ARG1	88.47
ARG2	85.90
ARGM-ADV	97.87
ARGM-LOC	94.27
ARGM-MNR	91.16
ARGM-TMP	94.26

在此，我们对两者组合后的系统的召回率上限进行了比较，如图4-2。从图4-2可以看出，不论是总体的召回率还是主要角色类型的召回率都有一定程度的提高，总体提高约 2%。可见基于单一句法的标注结果，其角色候选相互补充，能够提高系统的召回率。由于短语结构句法的系统以句法成分为标注单元，而依存句法的系统以依存关系为标注单元。两者分析的角色候选中有相当一部分对应了不同的词序列，这样在语义角色标注时就能找到更多的角色候选来参与预测。因而召回率有了提高。

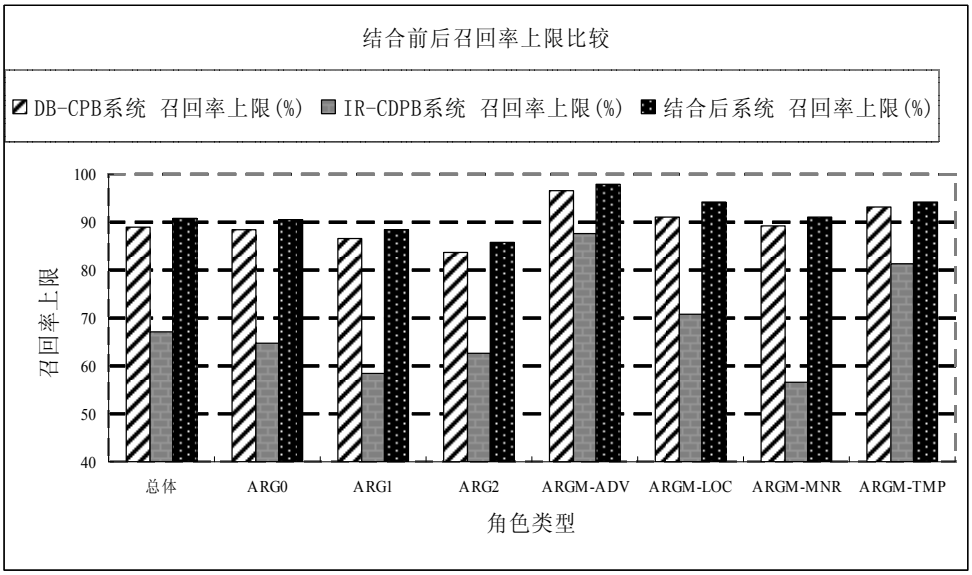

图 4-2 结合前后系统召回率上限比较

Figure 4-2 Maximal recall comparision after combination

4.2.3 结合策略

基于自动短语结构句法的角色标注系统 DB-CPB 系统，由于它和 CPB 语料一样，是以句法成分为语义角色的标注单元，因而具有召回率较高的优点，但精确率不是很高。基于自动依存句法的角色标注系统 HIT-IR CDB 系统，由于依存关系类型丰富，并且许多依存关系类型很大程度上能找到对应语义角色类型，因而其优点是有较高的精确率。但是其中许多语义角色找不到标注单元与之对应，因而召回率却很低。这样，我们结合两者的优缺点，给出如下的启发式策略。

首先，对各个系统中的角色候选分成两类。第一类公共候选，即在 DB-CPB 系统中和 HIT-IR CDB 系统中都对应相同词序列的角色候选。第二类差异候选，两个系统中代表不同词序列的候选。这样可以通过公共候选来增强角色类型的预测，通过差异候选召回单个系统无法找到的角色。

其次，我们先处理公共候选。由于公共候选只能有唯一一种角色类型标注，非语义角色或者某种类型语义角色。而 HIT-IR CDB 系统和 DB-CPB 系统可能会有不同的预测结果。我们处理如下：

1. 如果两个系统都预测为同一类角色，但预测概率可能不同，该候选

标注对应类型，并且预测概率设置为最大的那个。

2. 如果两个系统预测为不同类型的角色，考虑两个单一系统的性能特点来综合处理。HIT-IR CDB 的精确率比较高，而 DB-CPB 的召回率比较高，这样处理方法是：如果 HIT-IR CDB 以较大概率预测为非空角色类型，则候选标为该类型并设置为对应预测概率；如果 HIT-IR CDB 预测以较小概率预测为空类型，但是 DB-CP 以较大概率预测为非空角色类型，则候选标为 DB-CPB 预测的类型并设置预测概率。其他情况，根据两者预测概率差值是否超过某个阈值来选择对应的预测结果。

最后，我们把两个系统的差异候选和公共候选得到的标注结果，结合前面使用的后处理规则，进行来综合考虑，得到了最终的语义角色标注结果。

4.3 实验结果及分析

对两个系统结合后的标注结果进行了评测，实验结果如表4-2：

表 4-2 结合后系统性能

Table 4-2 The performance after combination

系统	Precision (%)	Recall (%)	F-Score (%)
相结合后系统	77.37	69.05	72.97

我们对基于单一句法的系统和结合后的系统性能进行了比较，如图4-3。从图4-3可以看出，结合后的系统精确率提高不明显，但召回率提高了约 2%，F-Score提高了 1.87%，性能的提高还是比较明显的。

由召回率上限分析可以看出，由于两种句法提供了不同的角色候选，能够使得召回率提升空间增加，并且在结合策略中找出了两个系统的差异候选，这些差异候选能够提高系统的召回率。

系统的精确率提高不明显，一方面由于两个系统的精确率比较接近，说明它们对公共候选的预测类型比较相近。在一方面在结合策略中，处理公共候选时更多的考虑了角色的召回。同时也由于两类自动句法分析都不是很准确，在预测时会引入错误，导致精确率的提高有限。

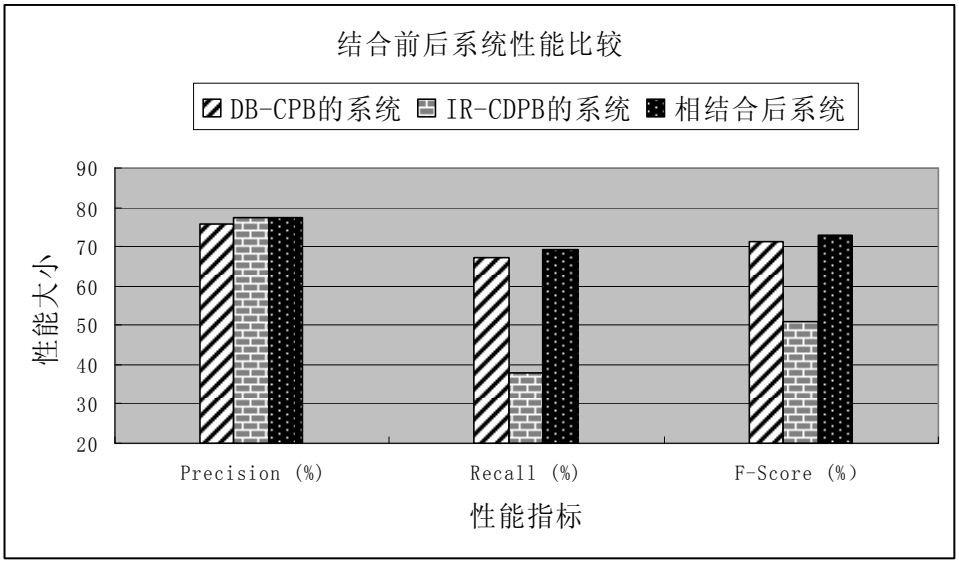

图 4-3 结合前后系统性能比较

Figure 4-3 The performance comparison after combination

系统最后F-Score提高虽比较明显，但幅度不大。这主要是句法分析准确度有限。LTP的依存分析依存弧方向有很多错误，比如图3-6所示。其中“建设”应该依存于“同步”，而不是“与”。这种依存弧的错误导致了角色找不到对应的标注单元，因此召回率很低。如果句法分析准确度很高，这种多句法标注结果相结合的方法能够获得很好的效果。

4.4 本章小结

本章介绍了两种句法角色分析结果相结合的中文语义角色标注方法。首先给出这种方法的意义，能够充分利用自然语言处理的多层面信息。接下来介绍了系统的实现框架，并通过召回率上限在结合前后的变化分析了该方法的预期效果。最后介绍角色候选分类处理的结合策略并完成实验，由实验结果证实了方法的可行性。

结论

中文语义角色标注是近年来中文信息处理中的一个热点和难点。句法分析对语义角色标注的影响很大，目前学者们主要通过使用不同的句法分析方法进行语义分析的研究。因此本文主要针对基于单一句法的中文语义角色标注中特征构造和选择、使用依存句法进行中文语义角色标注以及多句法分析结果相结合进行了深入研究。

对于基于单一句法的中文语义角色标注，本文构造了丰富有效的单一特征和组合特征，并采用 χ^2 显著性检验、贪心算法的特征选择方法，引入了有效的后处理方法，在高质量句法分析的语料资源上进行实验并取得了较高的性能。实验表明，通过新特征的加入以及特征的选择，能够有效地提高中文语义角色标注系统的性能。本文还实现了基于依存句法分析的中文语义角色标注，分别在高质量句法分析和完全自动句法分析上进行实验，并对实验结果进行了比较，证实了句法分析质量对语义角色标注性能的至关重要。同时也对两种句法的标注结果进行了比较分析，比较结果表明，短语结构句法的标注结果能够给出较高的召回率，而依存句法的标注结果能够给出较高的精确率。本文最后采用自动短语结构句法和自动依存句法相结合的中语义角色标注方法，综合利用单一句法标注的优点，并给出有效的结合策略，有效提高了系统的性能。证实了这种多句法相结合的中文语义角色标注方法是有效的。

本文的独创性工作体现在以下几点：

首先，本文针对短语结构句法和依存句法的中文语义角色标注，提出了丰富有效的新特征和组合特征，并通过 χ^2 显著性检验、贪心算法的特征选择方法得出了最优特征集，有效提高了系统性能。

其次，本文采用依存句法进行了中文语义角色标注，并通过哈工大信息检索研究室的语言技术平台，实现了基于完全自动依存句法的角色标注系统，有效利用依存句法的丰富依存关系类型特点提高了系统的精确率。

最后，综合利用单一句法标注的优点，本文给出了两种自动句法相结合的中文语义角色标注方法，有效提高了系统性能。

参考文献

1 E. Charniak and Y. Wilks. Computational Semantics. Amsterdam: North-Holland, 1976

2 R. C. Schank. Conceptual Information Processing. Elsevier Science Inc., 1975

3 C. D. Manning and H. Schütze. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press, 1999

4 M. Surdeanu, S. Harabagiu, J. Williams, et al. Using Predicate-Argument Structures for Information Extraction. In Proceedings of ACL 2003, 2003

5 J. Hajic, M. Cmejrek, B. Dorr, et al. Natural Language Generation in the Context of Machine Translation. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2002

6 D. Beaugrande, R. Alain and W. Dressler. Introduction to Text Linguistics. London; New York: Longman, 1981

7 车万翔, 刘挺, 李生. 浅层语义分析. 全国第八届计算语言学联合学术会议, 南京, 2005, 154~160

8 N. Xue and M. Palmer. Annotating the Propositions in the Penn Chinese Treebank. Q. Ma, F. Xia, (Editors) Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, 2003, 47~54

9 N. Xue. Annotating the Predicate-Argument Structure of Chinese Nominalizations. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 2006

10 L. You and K. Liu. Building Chinese FrameNet Database. Natural Language Processing and Knowledge Engineering, 2005, 301~306

11 M. Palmer, D. Gildea and P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. 2005, 31(1):71~106

12 A. Meyers, R. Reeves, C. Macleod, et al. The NomBank Project: An Interim Report. In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation, Boston, Massachusetts. 2004

13 S. Pradhan, K. Hacioglu, V. Krugler, et al. Support Vector Learning for Semantic Argument Classification. Machine Learning Journal, 2005,

60(1~3):11~39

14 N. Kwon, M. Fleischman and E. Hovy. Senseval Automatic Labeling of Semantic Roles Using Maximum Entropy Models. R. Mihalcea, P. Edmonds, (Editors) Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain: Association for Computational Linguistics, 2004, 129~132

15 P. Koomen, V. Punyakanok, D. Roth, et al. Generalized Inference with Multiple Semantic Role Labeling Systems. In Proceedings of CoNLL-2005, 2005, 181~184

16 V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag,Berlin, 1995

17 T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. C. N´edellec, C. Rouveirol, (Editors) Proceedings of ECML-98, 10th European Conference on Machine Learning, 1398, Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998, 137~142

18 C. Cortes and V. Vapnik. Support Vector Networks. Machine Learning, 1995, 20:273~295

19 A. L. Berger, S. A. Della Pietra and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics. 1996, 22(1):39~71

20 X. Carreras and L. M`arquez. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005, 152~164

21 M. Fleischman, N. Kwon and E. Hovy. Maximum Entropy Models for FrameNet Classification. M. Collins, M. Steedman, (Editors) Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, 49~56

22 A. J. Carlson, C. M. Cumby, N. D. Rizzolo, et al. SNoW User Manual. In Proceedings of CoNLL-04, 2004

23 R. E. Schapire and Y. Singer. Improved Boosting Algorithms Using Confidencerated Predictions. Mach. Learn. 1999, 37(3):297~336

24 M. Surdeanu and J. Turmo. Semantic Role Labeling Using Complete

Syntactic Analysis. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005, 221~224

25 K. Hacioglu and W. Ward. Target Word Detection and Semantic Role Chunking Using Support Vector Machines. In Proc. of HLT/NAACL-03. 2003, 25~27

26 K.Hacioglu. A Lightweight Semantic Chunking Model Based on Tagging. In Proceedings of HLT/NAACL-04. 2004

27 K. Hacioglu, S. Pradhan, W. Ward, et al. Semantic Role Labeling by Tagging Syntactic Chunks. H. T. Ng, E. Riloff, (Editors) HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004), Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, 110~113

28 K. Hacioglu. Semantic Role Labeling Using Dependency Trees. Proceedings of CoNLL-2004，2004, 1273~1276

29 D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. Comput. Linguist. 2002, 28(3):245~288

30 S. Pradhan, H. Sun, W. Ward, et al. Parsing Arguments of Nominalizations in English and Chinese. Proceedings of the HLT/NAACL 2004, 2004

31 N. Xue and M. Palmer. Automatic Semantic Role Labeling for Chinese Verbs. In Proceedings of IJCAI2005，2005，1160~1165

32 H. Sun and D. Jurafsky. Shallow Semantic Parsing of Chinese. In Proceedings of NAACL 2004, Boston, USA, 2004，192~199

33 X. Carreras and L. M`arquez. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. H. T. Ng, E. Riloff, (Editors) HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004), Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, 89~97

34 D. Gildea and M. Palmer. The Necessity of Parsing for Predicate Argument Recognition. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, 239~246

35 V. Punyakanok, D. Roth and W. Yih. The Necessity of Syntactic Parsing for Semantic Role Labeling. IJCAI-2005, 2005, 1117~1123

36 N. Xue and M. Palmer. Calibrating Features for Semantic Role Labeling. In Proceedings of the EMNLP-2004, 2004, 88~94
37 N.Xue. Semantic Role Labeling of Nominalized Predicates in Chinese, in Proceedings of HTL-NAACL 2006. New York City. 2006
38 N. Xue and F. Xia. The Bracketing Guidelines for the Penn Chinese Treebank, IRCS Report 00-08 University of Pennsylvania, 2000
39 D. M. Bikel. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In HLT 2002, 2002, 24~27
40 Z. Jiang and H. Ng. Semantic Role Labeling of NomBank: A Maximum Entropy Approach. In Proceedings of EMNLP 2006. Sydney, Australia. 2006, 138~145
41 郎君, 刘挺, 张会鹏. LTP:语言技术平台. 第三届学生计算语言学研讨会, 2006, 64~68
42 刘海涛. 依存语法和机器翻译. 语言文字应用. 1997, 3: 89~93
43 郭艳华, 周昌乐. 一种汉语语句依存关系网协动生成方法研究. 杭州电子工业学院学报. 2000, 20(4): 24~32
44 马金山, 张宇, 刘挺. 利用三元模型及依存分析查找中文文本错误. 情报学报, 2004, 723~728
45 G. Ngai, D. Wu, M. Carpuat, et al. Semantic Role Labeling with Boosting, SVMs, Maximum Entropy, SNoW, and Decision Lists. R. Mihalcea, P. Edmonds, (Editors) Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain: Association for Computational Linguistics, 2004, 183~186
46 T. Tsai, C. Wu, Y. Lin et al. Exploiting Full Parsing Information to Label Semantic Roles Using an Ensemble of ME and SVM via Integer Linear Programming. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005, 233~236
47 L. M`arquez, P. Comas, J. Gim´enez, et al. Semantic Role Labeling as Sequential Tagging. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005, 193~196
48 S. Pradhan, W. Ward, K. Hacioglu, et al. Semantic Role Labeling Using

Different Syntactic Views. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), University of Michigan, 2005, 581~588

49 S. Pradhan, K. Hacioglu, W. Ward, et al. Semantic Role Chunking Combining Complementary Syntactic Views. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan: Association for Computational Linguistics, 2005, 217~220

攻读学位期间发表的学术论文

1 Ting Liu, Wanxiang Che, Sheng Li, Yuxuan Hu and Huaijun Liu, Semantic Role Lableing System using Maximum Entropy Classifier, CoNLL-2005, 2005, 189-192

2 刘怀军，车万翔，刘挺. 中文语义角色标注的特征工程. 第三届学生计算语言学研讨会，2006 (获优秀论文)

3 刘怀军，车万翔，刘挺. 中文语义角色标注的特征工程. 中文信息学报，第 21 卷第 1 期，2007

哈尔滨工业大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《中文语义角色标注的方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读硕士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。

作者签字：　　日期：　　年　　月　　日

哈尔滨工业大学硕士学位论文使用授权书

《中文语义角色标注的方法研究》系本人在哈尔滨工业大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归哈尔滨工业大学所有，本论文的研究内容不得以其它单位的名义发表。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，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅，同意学校将论文加入《中国优秀博硕士学位论文全文数据库》和编入《中国知识资源总库》。本人授权哈尔滨工业大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

哈尔滨工业大学硕士学位涉密论文管理

根据《哈尔滨工业大学关于国家秘密载体保密管理的规定》，毕业论文答辩必须由导师进行保密初审，外寄论文由科研处复审。涉密毕业论文，由学生按学校规定的统一程序在导师指导下填报密级和保密期限。

本学位论文属于　　保密□，在　　年解密后适用本授权书

　　不保密□

（请在以上相应方框内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

致谢

值此论文完成之际，谨向给与我无私帮助的老师、同学、朋友以及我的亲人致以诚挚的谢意！

衷心感谢我的导师刘挺老师，感谢老师车万翔。感谢他们在生活上无微不至的关怀和学习上悉心的指导。感谢他们在我的毕业设计过程中，给我做详细的指导，指引我毕业设计的目标和努力的方向。他们一丝不苟的科研精神、严谨的治学态度，正是我学习的榜样。在整个毕业设计过程中，我不仅学到了知识，更学到了方法。

感谢信息检索研究室，感谢信息检索研究室张宇、秦兵老师，他们在学习上生活上给了我莫大的帮助。在这个新的环境里，我有了家一样的感觉。在信息检索，我学习生活。在这里我学到了许多基本理论、编程技巧和方法、科研思路和方法，也学到了如何做人，在我的人生道路上它使我获益匪浅。有了“友爱精诚乐学日新”的 HIT-IR 精神，我们每个 HIT-IR 人都茁壮成长。

感谢百忙中为我审稿、答辩的各位老师。他们在炎热的夏天，为我们的顺利毕业忙碌，默默的作出奉献。

感谢我的父母。虽然他们已经年近花甲，但是为了我的学业，他们不啻辛劳的耕作在土地上；为了我的前途，他们吃尽了苦，耗尽了心血，却没有半句怨言。感谢他们对儿子的支持和鼓励。感谢我的亲姐和亲哥，他们给我很大的帮助，也在家里照顾着年迈的父母，让远在他乡的我心里有了一丝安慰。游子的心，天上的云，家里的平安才能让我得以安稳。

感谢马金山、李维刚师兄，他们为人热情，生活乐观，给我很大的帮助。感谢跟我一起进入信息检索研究室的赵妍妍、龚诚、林建国、祝慧佳、常宇、潘树燊,高立琦、李洋等同学，他们和我一起度过了硕士生活。感谢曲晓航、孙军同学，在英语学习上给我很大帮助。感谢舍友陈玉平、李海洋等同学，使我在一个温馨的宿舍里生活成长。

感谢所有帮助过我的人，感谢所有将来帮助我的人，感谢每一位好心人。

中文语义角色标注的方法研究

作者：刘怀军
学位授予单位：哈尔滨工业大学

引用本文格式：刘怀军 中文语义角色标注的方法研究[学位论文]硕士 2007