密级：　　保密期限：

北京郵電大學

硕士研究生学位论文

题目：基于结构学习的语义角色标注

学　　号：076390

姓　　名：白雪

专　　业：模式识别与智能系统

导　　师：王小捷

学　　院：计算机学院

2010年　1月　20　日

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

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。

本人签名：白雪　　日期：2010.3.12

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关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北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位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、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密在__年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名：白雪　　日期：2010.3.12

导师签名：[illegible]　　日期：2010.3.12

基于结构学习的语义角色标注

摘 要

近年来对自然语言进行浅层语义分析逐渐兴起，它已成为自然语言处理应用的重要组成之一。作为其具体实现，语义角色标注是一项定义完整，有着充实工作内容和可比较评测的任务。语义角色标注就是为句子中谓语动词的论元及附属成分标上其担任的语义角色，如施事、受事、时间和地点等等。目前英语语义角色标注已经取得了一定的成果，但大多基于要求大量标注语料的监督的机器学习算法。但汉语语义角色标注的研究才刚刚起步，可利用的语料资源非常有限。为此，本文采用半监督机器学习方法，以期在资源有限的情况下能取得比较好的标注性能。

结构学习算法是一种通过多任务学习得到“共同结构”，并利用其来提高目标任务分类器性能的一种机器学习算法。ASO 算法是最近提出的一种线性的半监督结构学习算法，能够利用大量的未标注语料，通过辅助问题抽取“共同结构”，来提高分类准确性。本文构建了一个基于 ASO 结构学习算法的中文语义角色标注系统，并在 Chinese Proposition Bank 语料上进行了实验，取得了比较好的结果。

本文构建的语义角色标注系统不是传统的基于句法树的系统，即对句法树上的节点进行语义角色识别和分类，而是以组块为基本标注单元。这一方法由于避开了句法分析这个阶段，使得语义角色标注摆脱了对句法分析的依赖，从而突破了汉语语法分析器的时间和性能限制。另外构建合适的辅助问题是 ASO 算法性能的关键，我们分析了构建辅助问题的原则和方法，并进行了一系列实验。实验结果表明，结构学习算法可以有效地利用未标注语料，提高系统的性能。

关键词：自然语言理解　语义角色标注　浅层语义分析　结构学习　半监督

SEMANTIC ROLE LABELING BASED ON STRUCTURE LEARNING

ABSTRACT

Nowadays there has been an increasing interest in shallow semantic parsing of natural language, which is becoming an important component in all kinds of natural language process applications. As a particular case, semantic role labeling (SRL) is a well-defined task with a substantial body of work and comparative evaluation. Semantic role labeling is to labels the constituents with semantic roles which have direct relation with the predicate in a sentence. The semantic roles include agent, patient, time, locations and so on. At present English SRL has achieved certain results, but is almost based on supervised methods which need a large labeled corpus. However these kinds of resources are still quite limited for Chinese. In order to deal with this problem this paper presents a semi-supervised method for SRL.

Structure learning algorithm is a multi-task machine learning

algorithm, which extracts the common structures of multiple tasks to improve accuracy of target task. ASO is a recently proposed linear semi-supervised structure learning algorithm, which extracts the common structures via the use of auxiliary problems with large number of unlabeled data. In this paper, we build a Chinese SRL system based on ASO. We carry out experiments based on Chinese Proposition Bank corpus and improve the accuracy of system.

Many SRL systems have been built on the parsing trees, in which the constituent s of the sentence structure are identified and then classified. In contrast, in this paper we use chunk as the basic labeled unit, namely the argument s of the verbs. Along with the removal of the parsing stage, the SRL task avoids the dependence on parsing, which is always the bottleneck both of speed and precision. In addition to build a suitable auxiliary problem is the key to performance of ASO algorithm. We analysis principles and methods of building auxiliary problem, and explore a number of different auxiliary problems used in a series of experiments. The results show that the structure learning algorithm can be effectively used unlabeled corpus to improve system performance.

KEY WORDS: natural language understanding shallow semantic parsing semantic role labeling structure learning semi-supervised

目录

第一章 绪论

1.1 研究背景和意义

语言是信息的重要载体，为使计算机具有理解、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。在句子层面上的语言分析一般分为三个层次：句法、语义、语用。句法分析关心的是词语如何排列形成正确的句子，并决定每个词语在句子中充当的结构角色。句法分析问题早已引起人们的广泛关注，并取得了积极的进展。所谓句子语义分析，指的是将自然语言句子转化为反映这个句子意义(即句义)的某种形式化表示。即将人类能够理解的自然语言转化为计算机能够理解的形式语言。而语用就是语言的实际应用，语用分析研究影响语言行为(如招呼、劝说)的标准和支配轮流发言的规则等，目前在自然语言处理领域的研究和应用还较少。

随着语言处理的技术发展，句子语义分析成为当前研究的一个重点。句子语义有多种定义的方式。基于格语法的句子语义是其中一种[1]，也是本文的研究重点。在格语法中，句子语义是指词语进入句子以后，词语与词语之间形成的词汇意义之外的一种关系意义，是词语在语句结构中体现出来的意义。如"老李打了小王"和"老李被小王打了"这两句中的"老李"，在前一句中是发出"打"这个动作行为的主体，我们称作"施事"；而在后一句中，则是"打"这个动作行为的承受对象，我们称作"受事"。这里的施事，受事的意义不是"老李"这个词语本身所具有的，而是在进入这两个句子后，作为动作“打”的参与者才具有的。表达动作所有参与者的名词(短语)和表达动作的动词(短语)一起构成了该句子的骨干语义结构。再如"我用这把刀切菜"这样一个句子，"我"是发出动词"切"这个动作行为的主体，"菜"则是"切"这一动作行为的承受对象，而"这把刀"则是"切"这个动作行为的凭借工具，这些部分和动作的参与者一样都是词语进入句子后才具有的，和骨干结构一起共同构成了完整的句子语义。

然而，限于目前的技术水平，经过几十年的发展，还没有太多使用学习的方法来获取详细语义理解知识的研究，深层的语义分析较难做到。研究者开始关心“浅层语义分析 (Shallow Semantic Parsing)”，一种简化了的语义分析方式。

基于格语法的句子分析就是一种浅层语义分析，它只标注与句子中谓词有关的成份的语义角色。语义角色(semantic role)又称为论旨角色或论元角色

(thematic role)，是指谓词和句子中谓词论元和附属语之间的语义关系。论元大多用名词短语表达，也可以表达成介词短语、动词短语或者从句。附属语也是语义角色标注时可能需要识别的，附属语是那些与动词联系不是很紧密的短语。它们一般是可选的，描述了动作或状态的时间、地点或方式。

主要的语义角色包括施事格、受事格、工具格、伴随格等，最初是由Gruber和Fillmore提出的。1968年Fillmore出版了著名的《The case for case》，他在传统语法范畴“格”(case)的基础上提出了一种新的“深层格”(deep case)的概念。不同于传统形态学中的“主格”、“宾格”、“与格”等，新的格标记“施事格”(agent)、“客体格”(object)等揭示了一些谓词与名词等的普遍的语义结构关系。语义角色和语法功能间通常都有一定的关系，例如施动者通常是主语， 但是也有一些例外。语义角色和语法功能之间的关系也会随着语态（主动语态或被动语态）改变而改变。主动语态对应表达一个动词论元的默认方式：施动者是主语，受动者是宾语。在被动语态中，两个论元的顺序颠倒了，受动者成为了主语，施动者降格为一个间接角色。

语义角色标注（Semantic Role Labeling，SRL）就是要为给定句子中的每个谓词（动词或名词）标注相应的语义角色，如施事、受事、工具或附加语等。例如，对于句子：

北京去年举办了奥运会。

语义角色标注的结果为：

[北京 Agent][去年 Tmp][举办 V]了[奥运会 Passive]。

其中“举办”为目标动词，“北京”、“奥运会”和“今年”分别是其施事、受事和发生的时间。

语义角色标注是自然语言理解的底层技术，最终要靠实际的应用体现其价值。在许多高层次的研究和应用上，语义角色标注都大有用武之地。语义角色标注在问答系统（Question Answering）、机器翻译（Machine Translation）、信息抽取（Information Extraction）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处理研究和应用的学者们的重视。

1. 问答系统

问答系统(Question Answering System, QA)是信息检索系统的一种高级形式。它能用准确、简洁的自然语言回答用户用自然语言提出的问题。其研究兴起的主要原因是人们对快速、准确地获取信息的需求。问答系统是目前人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向。

问答系统问答问题的类型进行区分：询问人(如:谁发现了北美洲？)、询问时间（如：人类哪年登录月球？）、询问数量（如：珠穆朗玛峰有多高？）、询问

定义（如：什么是氨基酸？）、询问地点和位置（如：芙蓉江在重庆市哪个县？）、询问原因（如：天为什么是蓝的？）。

问答系统是人们利用搜索引擎最自然，最便利的方式，用户再也不用苦恼于关键词的选择和配置。但是目前的自动问答系统多是基于文字表层的模式匹配，没有用到自然语言的深层理解技术。

Narayanan[2]等人首次将语义角色标注技术应用于自动问答系统，并且取得了不错的效果。这一结果也是显然的，因为浅层语义分析步骤能够识别出一个动词的施事、受事，以及该动作发生的时间、地点等信息，这对回答一些针对该动作的问题是非常有帮助的。

例如句子“昨天老师在办公室批评了我”的标注实例如图 1-1 所示。

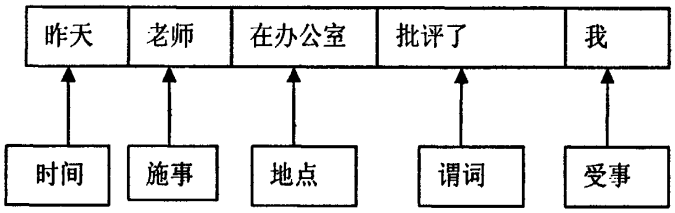

图 1-1 句子标注实例

考虑一下四个问题：

谁批评了我？

老师批评了谁？

老师什么时候批评了我？

老师在哪批评了我？

通过以谓语动词“批评”为中心标注出来的语义角色，可以很容易的回答上述问题。

2. 机器翻译

机器翻译是用计算机实现一种自然语言到另一种自然语言的转换。一般指自然语言之间句子和全文的翻译。虽然经过多年的发展，但是全自动机器翻译的性能还不能令人满意。目前，机器翻译领域的研究中统计的方法占有统治地位，越来越多的自然语言处理技术被应用于此。语义角色标注技术[3]的应用，使统计机器翻译以谓词为中心，对角色逐个对照翻译，然后按照目标语言习惯组合起来，完成整个句子的翻译。由于这是在语义的基础上完成的机器翻译，而不再仅仅局限于文字表面的匹配，可以有效利用自然语言内在的结构、语义等信息，使用较少的资源，达到更好的效果。表 1-1 是一个语义角色标注基础上的机器翻译实例[57]。

表 1-1 在语义角色标注基础上的机器翻译实例

English（SVO）	Farsi（波斯语）（SOV）
[AGENT The little boy]	[AGENT pesar koocholo]boy-little
[PRED kicked]	[THEME toop germezi]ball-red
[THEME the red ball]	[ARGM-MNR moqtam]hard-adverb
[ARGM-MNR hard]	[PRED zaad-e]hit-past

3. 信息抽取

信息抽取（Information Extraction, IE）是把文本里包含的信息进行结构化处理，变成表格一样的组织形式。输入信息抽取系统的是原始文本，输出的是固定格式的信息点，如命名实体、实体关系、事件等。这些信息可以导入关系数据库等结构化数据管理工具，然后可以方便的进行查询等操作，这就是信息抽取的主要任务。信息以统一的形式集成在一起的好处是方便检查和比较。例如比较不同的招聘和商品信息。还有一个好处是能对数据作自动化处理。例如用数据挖掘方法发现和解释数据模型。目前，信息抽取已经成为文本挖掘领域一个相当热门的研究课题。它不但能够自动的获取人们想要的信息，还能够帮助提高信息检索等技术的性能。

Surdeanu[4]等人利用浅层语义分析技术进行信息抽取，并且提高了信息抽取，特别是事件抽取系统的性能。事件抽取需要找到相应的触发动词，发生的时间、地点，参与的人物等信息，一般一个事件范围不超过一个句子。可见利用语义角色标注可以很好的完成这种事件抽取方式，我们需要找出每个事件对应的触发动词，然后标注出它的各种语义角色作为事件的参数，并根据参数的特点，进一步确认是否为需要的事件。这里需要注意的是，在实际应用中，参数往往是较长的人名、地名、机构名、时间和数字等，因此，对这些实体的识别变得异常重要，同时若实体的识别准确率提高，语义角色标注的性能也将随之提高。同时，浅层语义分析本身也可以看作是一种通用的信息抽取，只是人们事先定义的结构化信息并不局限于特定的领域。因此，浅层语义分析技术的发展，必将会促进信息抽取技术的发展。

考虑文本“Time for our daily market report from NASDAQ. London gold fell $4.70 cents to $308.45”。根据信息抽取要素的模板，以及对句子的语义角色的分析，可以得出表 1-2 的模板。

表 1-2 从以上文本总结事件信息的模板

<MARKET_CHANGE_1>:=

INSTRUMENT	London[gold]
AMOUNT_CHANGE	fell[$4.70]cents
CURRENT_VALUE	$308.45
DATE:	Daily

语义角色标注与信息抽取之间的区别和联系也可以通过表 1-3 来描述。

表 1-3 信息抽取与语义角色标注之间的区别和联系

特征	信息抽取	语义角色标注
覆盖面	窄	宽
语义深度	浅层	浅层
与应用的直连程度	有时	没有
处理的单元	>句子	句子

4. 其它应用

一词多义的现象在自然语言中是普遍存在的。例如，汉语中的词“打”就有十几种意思，而英语中的“bank”既可能表示“银行”，也可能表示“河堤”。根据有关人士统计，物理学方面的文本中的多义词约占文本的 30%，其余科学文本的多义词则约为 43%。正确的区分这种一词多义现象，对于语言的深入分析和理解具有重要的作用。词义消歧（word sense Disambiguation），就是要在特定的上下文中确定多义词的意思。一般的词义消歧方法多是利用多义词的所有上下文信息，这势必会给消歧过程带来一些噪声。Dang[5]等人利用浅层语义分析的结果找出更有代表性的上下文，使得消歧效果得以提高。

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，利用浅层语义分析的结果，必然能够较好的完成复述的任务。

因此，我们说在自然语言处理领域，众多需要对句子语义有一定理解的课题中，浅层语义分析技术都将有用武之地，做好浅层语义分析必将为这些方向提供新的动力。

1.2 语义角色标注研究现状

下面的论述中我们主要从三个方面来介绍语义角色标注研究现状，其一是目前应用于语义角色标注任务的相关算法，其二是相关的语料建设，最后对汉语语义角色标注和英语语义角色标注任务进行了比较。

1.2.1 算法研究

自从 Gildea[7][8]等首先基于统计模型进行自动的语义角色标注以来，目前的大多数语义角色标注系统都采用机器学习的方法来完成。基本思想是将语义角色标注看成一个分类问题，在语义角色的识别和分类过程中，以句子中一定的连续词语为标注的基本单元，然后根据一定的语言学知识列出该单元的各种特征，并与该单元的语义角色类型（也可能不属于任何语义角色）组成学习的实例，最后使用某种学习算法对这些实例进行自动的学习，并用学习得到的模型对新的实例进行预测，得到标注结果。

目前针对英文已经进行了大量的研究，并取得了不错的进展。常用的机器学习算法，如最大熵模型（MaxEnt）和支持向量机模型（SVM）等被应用于语义角色标注任务。代表性工作包括：Gildea[7][8]等、Xue[9]等、Pradhan[10][11]等、刘挺等[12][13]。

2002 年 Gildea 和 Jurafsky 构建的语义角色标注系统，是首次基于纯概率的统计模型实现自动语义角色标注。该系统使用基于语义格的后退相对频率模型，从单一句法树中抽取各种语言学特征，在 FrameNet 近 50000 句手工标注的语料上进行了语义角色的识别。对于已标注了句法成分的测试文本，该系统的精确率可达 82%。但同时进行全自动的句法分析和语义角色标注当时只达到了 65%的正确率，并不是特别理想。该项研究的主要贡献在于：提出了一系列有助于语义角色识别的词汇化特征和句法特征，并把句法特征应用于语义角色标注研究中，为后来的研究者指明了方向。他们提出了目前 SRL 系统最常用的七个基本特征，包括：谓词、句法类型、子类框架、路径、位置、语态和中心词。后来的研究中大多都把这 7 个基本特征作为标准特征集用来形成基础系统，并扩展一些新特征以提高系统性能。他们同时考虑了有关于谓词（动词，名词，形容词）的语言知识，以及各类语义角色相结合的先验概率等信息。在此基础上，Gildea[8]等进一步在 PropBank 语料库上做了同样的试验，并提出语义角色标注需要句法分析的必要性。基于手工标注句法树，F1 值达到了 87%。

Xue 等[9]在基于单一句法树的基础上，详细验证了 Gildea 等中各个基本特征在 SRL 各阶段的贡献，并提出了新的特征：句法框架、词汇成分类型、词汇中心词、谓词与当前句法间的距离等，还有组合特征（谓词+句法类型、谓词+中心词、语态+位置等），并提出了一个有效的剪枝算法，最后使用最大熵模型进行实验。在 PropBank 语料库上的实验结果表明采用新的特征后系统性能有了显著提高，基于手工标注句法树，对 19 个角色进行分类，已知论元上的精确率为 92.92%，包括 NULL 的分类 F1 为 88.51%。该文提出了新的特征并详细分析了各个特征的作用，表明特征还有很大的开发空间，识别和分类这样不同的任务的性能提高需要不同的特征集。

Pradhan 等[10][11]使用 SVM 分类器，除了基本特征，选取了更多的特征（如动词聚类、部分路径、谓词词意信息、介词短语的中心词、当前句法成分的首词和末词及其词性、当前句法成分的父亲兄弟结点的句法类型和中心词及中心词词性、时间提示词、命名实体、中心词词性、位置次序、成分树的距离、成分相对特征、和动态类上下文等）取得了很好的性能。其识别阶段在训练语料上进行二元训练并预测，进而删除高概率为 NULL 的句法成分。保留下来的句法成分作为分类阶段的输入。在 PropBank 语料库上，基于手工句法分析的结果是 P/R/F1(%)：89/85/87，基于 Chariniak 自动句法分析的结果是 P/R/F1(%)：84/75/79。他们对语义角色标注的深入研究极大的推动了语义角色标注研究的发展。

刘挺等[12][13]选取了较多的特征（句法成分前后第一二个词，谓词词性，谓词后缀、较多的组合特征）使用最大熵分类器把识别和分类一步做训练，再做后处理的方法，在基于 PropBank 语料库的单一自动句法分析上报告了取得的最好结果。他们对语义角色标注的研究在国内是较成功的。

相比于英文的语义角色标注研究，针对中文的语义角色标注研究相对少很多，主要包括：Sun 等[14]、Xue 等[15]、刘怀军等[16]、于江德[17]等。

最早进行研究的是 Sun 等[14]，由于在当时还没有中文方面的专门语料，所以他们只是在 Pradhan 等的工作基础上，使用 SVM 分类器，选择了 10 个中文谓词和部分 Chinese Propbank 的数据进行实验，虽然不是很成系统，但是在汉语的语义角色标注研究中是一个很有意义的开端。通过对实验的分析，他们认为在中文语义角色标注中比较小的语料库就能取得较好的性能；英文语义角色标注中广泛使用的特征能较好的应用到中文任务当中；中文的语义角色标注要比英文的语义角色标注容易些。

伴随着 Chinese Propbank 的建立，中文有了规模较大的语义角色标注语料资源，出现了比较系统的中文语义角色标注的工作。Xue 等[15]比较和分析了中文和

英文语义角色标注的性能以及影响因素，在 Chinese Propbank 上进行了实验。基于手工标注句法树的标注结果，F1 值可达 91.3%；基于单一自动标注句法树的标注结果，F 值大幅降为 61.3%。该文通过实验发现，对于手工分析的句法树，实验结果基本与英文的结果相当，甚至稍微高出一点；但对于自动产生的句法树，则结果要比英文的差得多。对于这种情况，作者认为是由于句法分析假定分词和词性标注都是正确的，这导致中文分词和词性标注阶段产生的错误在句法分析阶段就很难被恢复，以致影响到整个角色标注系统的性能。而目前中文分词、词性标注和句法分析都还不成熟。

刘怀军等[16]在英文语义角色标注特征的基础上，针对中文的特点提出了一些有效的新特征和组合特征。例如，句法成分后一个词、谓语动词类别信息和路径的组合、谓语动词和短语类型的组合等，并在 Chinese PropBank 语料库上，基于手工标注句法树，使用最大熵分类器进行了实验，F1 值达到了 91.31%。

1.2.2 语义角色标注语料资源

要想进行语义角色标注，需要好的语料资源的支持。目前，英语较为知名的浅层语义分析资源有 FrameNet[18]、PropBank[19]和 NomBank[20]。

FrameNet 由 U.C.Berkeley 开发，它以框架语义为标注的理论基础，对英国国家语料库（BNC）进行部分标注。框架语义学的中心思想是：词的意义的描述必须跟语义框架相联系。框架是信仰、实践、制度、想象等概念结构或概念模式的图式表示，是言语社团中人们相互交流的基础。他们把框架网项目的任务设定为：

（1）描述给定词项所隶属的概念结构，即框架。

（2）从语料库中抽取包含某个词的句子；并按照该词的义项选择句子加以示例。

（3）对所选的句子进行框架元素标注。

（4）汇总框架元素标注结果，显示每个词项在组合上的可能性，即“配价描述”。

它试图描述每个谓词(动词、部分名词以及形容词)的语义框架，同时也试图描述这些框架之间的关系。从 2002 年 6 月发布至今，现共标注了约 49,000 句。其中每个句子都标注了目标谓词和其语义角色、该角色句法层面的短语类型(如 NP，VP 等)以及句法功能(如主语、宾语等)。FrameNet 现包含 1462 个目标谓语动词（927 个动词，339 个名词和 175 个形容词)。语义领域覆盖了：医疗保健卫生（HEALTH CARE）、机会（CHANCE）、感知（PERCEPTION）、通信

(COMMUNICATION)、交易(TRANSACTION)、时间(TIME)、空间(SPACE)、身体(BODY)、运动(MOTION)、生活阶段(LIFE STAGES)、社会语境(SOCIAL CONTEXT)、情绪(EMOTION)、认知(COGNITION)等。图 1-2 是 FrameNet 中表示身体动作的一个语义框架以及对一个句子的标注实例。

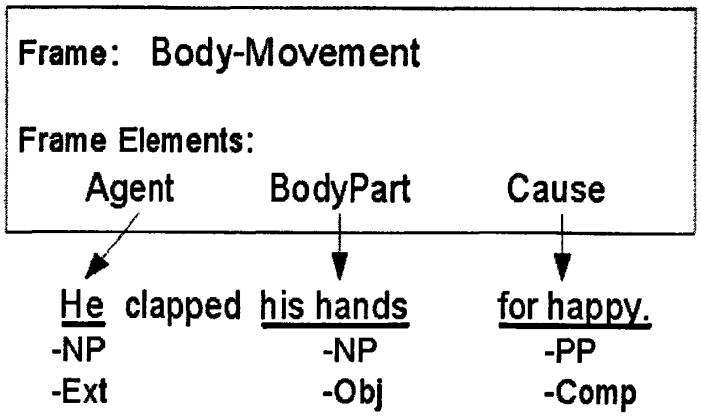

图 1-2 FrameNet 框架以及句子标注示例

PropBank 是宾州大学在其树库（Penn Treebank）句法分析的基础上，添加一层“谓词—论元”信息（或语义角色标签）。即把语义角色指派到树库的句法树的有关结点上，来实现浅层的语义表示（Shallower Level of Semantic Representation）。于是，不去涉及照应与同指（Anaphora and Coreference）、量化（Quantification）、体（Aspect）和模态（Modality）等高阶的相对深层次的语义现象。命题库（PropBank）旨在提供一个覆盖面广的用手工标注语义角色的语料库，使得开发更好的独立于领域的语言理解系统、对论元结构句法实现时发生变化的原因和方式的计量研究等成为可能。他们为每一个动词定义了一组底层的语义角色，并在宾大树库文本的每一次出现上进行角色标注。每一个动词的角色都被编了号。比如，核心论元 Arg0~Arg5。Arg0 通常表示动作的施事，Arg1 通常表示动作的影响等，Arg2～5 根据谓语动词不同会有不同的语义含义。它们的具体含义通常由 PropBank 中的 Frams(框架)文件给出，例如“buy”的一个语义框架如图 1-3 所示。与 FrameNet 不同的是，PropBank 只对动词（非系动词）进行标注，相应的被称作目标动词。与 FrameNet 相比，PropBank 基于 Penn TreeBank 手工标注的句法分析结果，因此标注的结果几乎不受句法分析错误的影响，准确率较高；而且它几乎对 Penn TreeBank 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。

RoleFrame buy.01 "purchase":
Roles:
Arg0: *buyer*
Arg1: *thing bought*
Arg2: *seller*
Arg3: *price paid*
Arg4: *benefactive*

图 1-3 "buy" 的语义框架示例

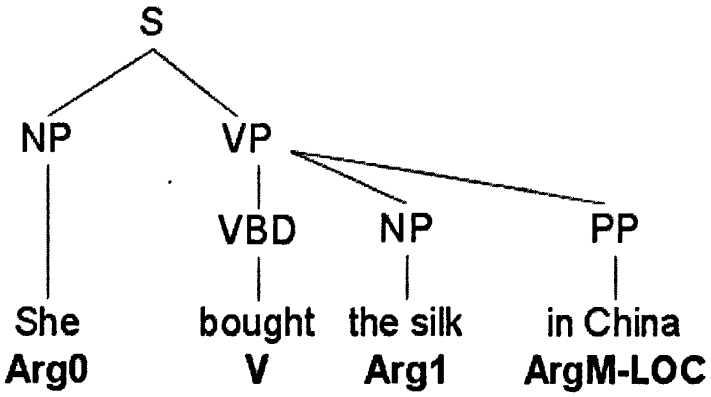

图 1-4 Propbank 中的句子标注示例

与 PropBank 标注 Penn TreeBank 中的动词做谓词不同，NomBank 标注了其中的名词作谓词的情况，参数的类别和表示同 PropBank 是一样的。例如：名词短语 "John's replacement Ben" 和 "Ben's replacement of John" 中，名词 replacement 便是谓词，Ben 是 Arg0，表示替代者，John 是 Arg1，表示被替代者。

除英语外，许多其它语言也建立了各自的语义角色标注库，例如：SALSA[21]是基于 FrameNet 标注体系的德语语料库；Prague Dependency Treebank[22]项目进行了大量的句法和语义标注(捷克语)，甚至包括指代消解的标注等。

下面介绍几个重要的中文语义角色标注库。

Chinese Proposition Bank (CPB)[23]是 Upenn 基于 Penn Chinese Treebank 标注的汉语浅层语义标注资源，在 Penn Chinese Treebank 句法分析树的对应句法成分中加入了语义信息。与英语的 Propbank 不同的是在语义标注时保留了宾州中文树库的句法标记。

CPB 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。它共有 10364 个句子，4854 个不同的谓词，词语数量达到 250K。用于标注的语义角色数目有 27 个，可以分为两大类，核心论元和非核心论元。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。由于 PropBank 中的论元划分依据的 Dowty 的原型理论，所以施事、受事等角色包括的范围都是很广的。相同的语义角色对于不同目标动词有不同的语义含义。其余的语义角色为非核心论元，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。表 1-4 为 CPB 的 18 种附加语义角色标记。

表 1-4 CPB 的附加标记列表

语义附加成分的 11 个附加标记	
角色名称	说明
ArgM-ADV	adverbial,default tag(附加的，默认标记)
ArgM-BNF	beneficiary(受益人)
ArgM-CND	condition(条件)
ArgM-DIR	direction(方向)
ArgM-DGR	degree(程度)
ArgM-FRQ	frequency(频率)
ArgM-LOC	location(地点)
ArgM-MNR	manner(方式)
ArgM-PRP	purpose or reason(目的或原因)
ArgM-TMP	temporal(时间)
ArgM-TPC	topic(主题)
谓语动词作为参数的 1 个附加标记	
PRD	predicate
习惯动作的 6 个附加标记	
AS	为，是，作，做
AT	在，于
INTO	成，入，进
ONTO	上
TO	到，至
TOWARDS	向，往

CPB 基于 Penn Chinese Treebank 手工标注的句法分析结果，准确率较高。

它几乎对 Penn Chinese Treebank 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。图 1-5 是 CPB 中的一个例子(chtb_433.fid 第 1 句)。在这个例子中，核心动词是“提供”。“提供”只有一个子语类框架，这个子语类框架包含三个论元成分：“提供者”，“被提供物”，分别对应原型施事和原型受事，在 CPB 中标记为 Arg0 和 Arg1。此外还有一个与事成分，在 CPB 中标记为 Arg2。在图 1-4 中，“保险公司”是“提供者”，“保险服务”是“被提供物”，“三峡工程”则是与事成分，是服务的接受者。“截至目前”表示“提供”的时间信息，标记为“ArgM-TMP”，同类的论元成分在该句中还有“已”，它被标记为“ArgM-ADV”，表示了一个与时间有关的成分。

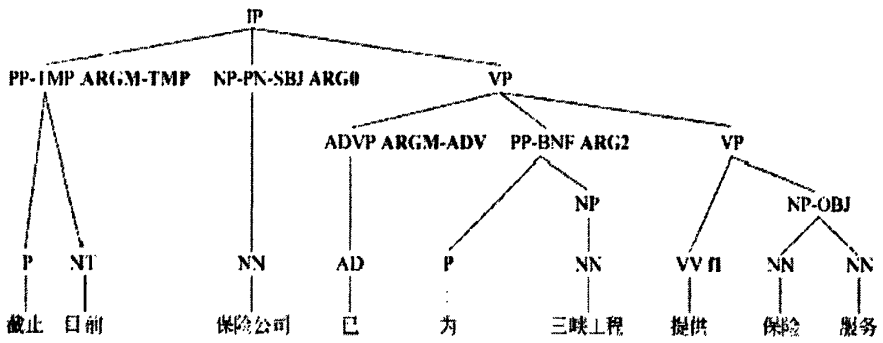

图 1-5 CPB 标注实例

Chinese Nombank[24]把传统 English Proposition Bank 和 English Nombank 的标注框架，扩展到对中文名词性谓词的标注。Chinese Nombank 在 PCT 数据上加入了语义角色层的标注信息，与 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 还标注了名词性谓词的框架，不过规模只是 CPB 中对应动词性谓词标注框架集的一小部分。Chinese NomBank 中的角色位置有两类情况。第一类，角色在以名词性谓词为核心词(Head word)的名词短语中。第二类，当以名词性谓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 1-6 是 Chinese NomBank 的一个标注实例。

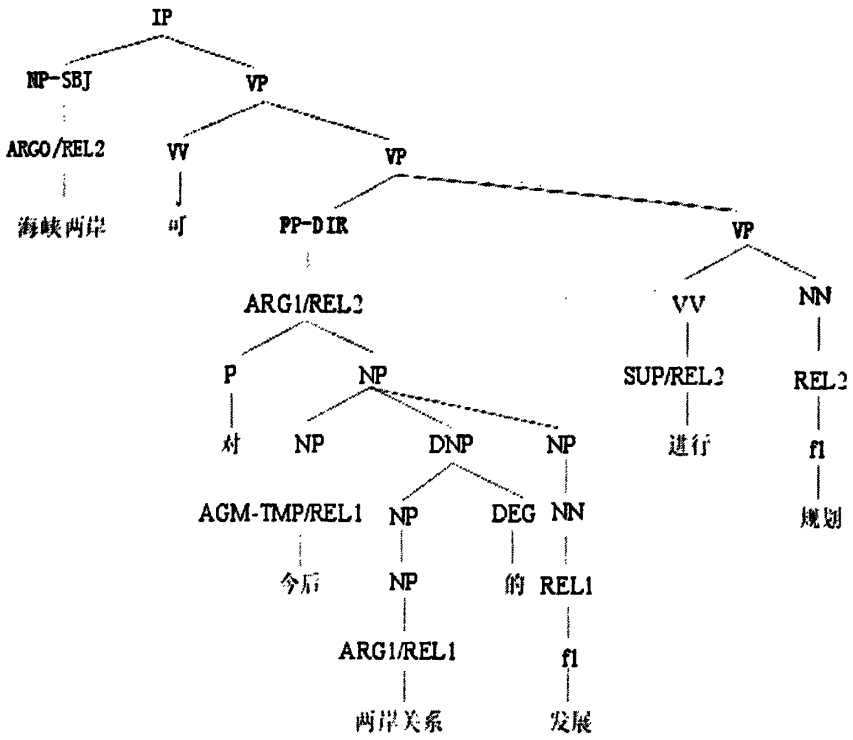

图 1-6 Chinese NomBank 中的一个标注实例

台湾中研院的陈凤仪[25]建立的 Sinica Treebank(中文句结构树资料库)，是一个语义标记和句法标记混合的语料库。它是以讯息为本的格位语法(Information-based Case Grammar)的表达模式为基本框架的，主要是对小句进行标注。目前已标注了 61087 个句子，包含了 361834 个词语。语义角色标记共有 50 多个，基本沿袭了格语法的标记体系，如：感受格(experieneer)、受益格(benefaetor)。图 1-7 是它的一个树结构图。

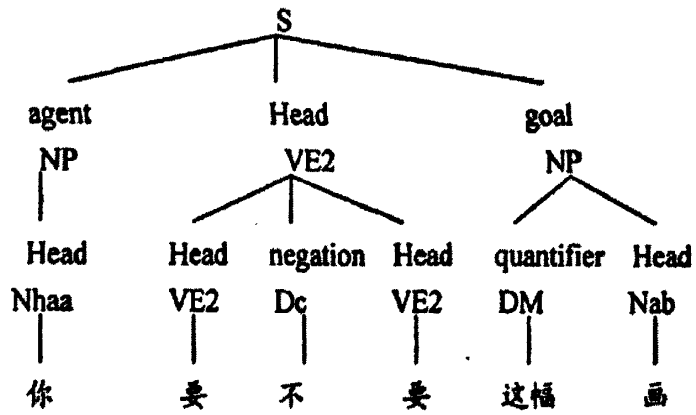

图 1-7 Sinica Treebank 中的句子标注

山西大学构建的汉语框架网络（CFN,Chinese FrameNet）工程[26] 是以框架语义学为理论基础、以 FrameNet 为参照、以汉语真实语料为依据的供计算机使用的汉语词汇语义知识库。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。

汉语框架网络知识库包括框架库、句子库和词元库三部分组成。目前，CFN 已经对 2610 个词元构建了 230 个框架，标注了 15000 条句子，包含认知领域用词、科普文章常用谓词以及部分中国法律用词。框架库以框架为单位，对词语进行分类描述，明确给出框架的定义和这些词语共有的框架元素，并进而描述框架和其他框架之间的概念关系；句子库记录带有框架语义标注信息的句子，即按照框架库所提供的框架和框架元素类型，标注句子的框架语义信息和句法信息，它可以作为训练数据供计算机处理语言使用；词元库记录的词元的语义搭配模式和框架元素的句法实现方式。表 1-5 为 CFN 框架库实例。

表 1-5 CFN 框架库实例

框架名	查看（Scrutiny）	
定义	该框架指的是认知者(人或其他智能生命体)对某事物，即现象，给予密切关注，比起一般的感知觉活动，查看隐含了人的分析、思考过程。	
核心框架元素	认知者（Cog）	认知者关注某实体。 例：所领导接过小包查看，内装有数十万元人民币。
	现象（Phen）	现象是查看活动所针对的对象。 例：我们的航标船要迎风向外走，去查看航标
非核心框架元素	背景（Ground）	认知者对背景进行关注，以找到目标事物(现象)。
	媒介（Med）	认知者在一个文本片段或者一部著作里关注一个实体，以获得某些发现，这个文本片段或者著作就是媒介。
	程度（Degr）	指认知者对背景的关注程度。
	方向（Dir）	方向描述的是感知者的注意力在感知活动中朝向哪里。

	修饰（Manr）	该框架元素是一个杂类，表示对动作行为本身的一般性描述，是除了程度、手段、目的等的各种描述，如难易、快慢等。
	方法（Mns）	该框架元素表示认知者采取什么技术或行动以查看某背景。
	目的（Purp）	该框架的一些词语伴随着一个句子成分，表示查看产生的预期结果。 例：你应该检查一下你的邮件以确定一下转帐结果。
框架关系	父框架	注意
	子框架	核实
	总框架	
	分框架	自主感知
	总域	自主感知
	分域	
	后续过程	
	结果状态	
	参照	
词元	检查 v，打量 v，视察 v，端详 v，查看 v，察看 v，探查 v，校阅 v，侦查 v，巡视 v，勘探 v，查 v，审视 v，审看 v，审查 v 等	

1.2.3 中文语义角色标注特点

英语的语义角色标注的研究已发展得较为成熟，但针对汉语的标注才刚刚起步，无论是在语料库还是在标注方法上，都需要进一步进行探索。现代汉语在词汇和语法上的特点，都使得在标注汉语的语料时面临了一些不同于英语的困难[27]：比如汉语中缺乏形态标志和形态变化，不利于区分一些谓词的论元角色。英语中有些用介词或形容词表达的内容，在汉语中常常用动词。英语中有形式主语 it 和 there be 句型等，但在汉语中这样的主语常常省略，使得确定谓词与主语的语义关系时缺乏了一个标记。汉语中存在大量的分句，使得句子的长度较长，而且主语和宾语等常常承前或蒙后省略，这也增加了标注的困难。句子中的一个成分可能同时是几个谓词的论元角色等等。

但尽管汉语有着自身的这些特点，反映在中文语料库中具有汉语特色的工作还不多，这类工作大多还停留在词典层面（如汉语语法信息词典、知网）。语料库建设的大部分工作还在模仿西方的做法。以语言学理论为指导的大规模语料库建设仍然比较少见，影响也不大。国外的工作如 PropBank、FrameNet 等，都有很强的西方语言学背景，未必适合于汉语。比如 CPB, 虽然是汉语语义角色标注研究中最常用的语料，但它把许多汉语独具特色的描述信息硬纳入英语的描述框架, 总给以汉语为母语的人许多生硬别扭的感觉。国内这方面工作已经开始起步，如清华周强的工作、宋柔的工作, 但语言资源建设与算法模型研究仍然处于割裂状态，语言学家和计算机科学家的交流太难、太少，中文语料库的规模较小。因此目前中文语料库资源的缺乏严重影响了中文语义角色标注的发展。

1.3 本文研究内容和组织

如上所述，语义角色资源的缺乏使得充分利用未标注语料来进行语义角色标注成为一个重要的研究方向。本文在这个方向开展研究，探索一种多任务结构学习算法——交互结构最优化（Alternating Structure Optimal，ASO）算法在中文语义角色标注中的应用。ASO 算法中，构建辅助问题是关键，本文讨论了如何为语义角色标注任务设计辅助问题，并实验了一些比较适合的辅助问题，提高了系统的性能。系统采用 CPB 语料库，以组块为标注单元进行实验。实验表明，相比于基线的线性分类器而言，本文的方法取得了很大的提高。

论文组织如下：

第一章 序言

阐述了论文的研究背景及其意义，介绍了语义角色标注的研究现状，并概述了论文的主要工作。

第二章 结构学习算法-ASO

主要介绍了半监督学习的基本知识，然后重点介绍本文实验采用的结构学习的基础理论知识和 ASO 算法的基本步骤，然后对 ASO 算法中最关键的辅助问题的选择进行了分析。

第三章 中文语义角色标注

首先介绍了我们基于结构学习的语义角色标注系统。

然后介绍了语料上进行的预处理。

第三介绍了语义角色标注的基本过程、标注单元和评价指标等相关技术。

第四，介绍了我们所使用的特征。

第五，介绍了我们构建辅助问题的原则和最后使用的辅助问题。

最后，介绍了本文实验和其在测试语料上的结果并进行分析。

第四章 工作总结及展望

总结全文，并展望下一步继续要研究的工作。

第二章　结构学习算法-ASO

2.1 半监督学习

自然语言处理中的标注问题很多时候可以归结为分类的问题。比如词性标注是将每个词分到一个词性类中，组块标注是将每个组块分到一个组块类中。最初人们使用基于规则的方法来解决分类问题，即先依据某种语言理论建立语言模型，再从语言模型构造规则系统。然而模型永远是已有现实的抽象，在面对已有事实时显然具有适应性。但在面对不断进化的事实时（如人类语言表述），模型会被不断加入各种用于妥协的规则，最终旧的模型被修改的面目全非而被新的模型所替代。而新的模型则又仅仅是已观察事实的抽象。此方法需要专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动，代价高昂。同时，随着知识库的增加，矛盾和冲突的规则也随之产生。

为了克服知识库方法的缺点，人们后来使用机器学习的方法来解决此问题。该方法的优点是不需要有专业技能的专家书写知识库，只需要有一定专业知识的人对任意一种语言现象做出适当的分类即可。然后以此为训练数据，再使用各种学习方法构造性能卓越的分类器。按照传统的机器学习理论框架，机器学习可以分为有监督学习(Supervised Learning)和无监督学习(Unsupervised Learning)两类。在有监督学习中，分类器利用已标注示例进行学习，从而建立模型用于预测未见示例的标注。这里的“标注”（label）就是示例所对应的输出，可以是分类问题中示例的类别，也可以是回归问题中示例所对应的实值输出。近些年来，随着数据采集和存储技术的发展，我们可以很容易的获取大量未标注示例。但获取已标注示例则还相对比较困难，因为需要耗费一定的人力和物力。例如，在基于内容的图像检索中，已标注图像的数目比较少，但图像库中却存在着大量的未标注图像。

显然，仅使用少量“昂贵的”已标注示例而不利用大量“廉价的”未标注示例，是对数据资源的极大的浪费；另外，如果只使用少量的已标注示例，那么利用它们所训练出的学习系统往往很难具有强泛化能力。因此，在已标注示例较少时，如何利用大量的未标注示例来改善学习性能已成为当前机器学习研究中最受关注的问题之一。

从没有经过任何加工的原始语料中，机器就可以学到很多书面语言的知识，

例如汉字频度、常用的汉字串（组块）及其频度、汉字串与汉字串的搭配以及搭配强度等，甚至通过聚类方法也可以区分（或者说“辨析”，也是某种意义上的“学习”）词语的义项乃至文本的内容，所以人们试图使用未标注的语料库直接进行学习，这种方法被称作无监督学习。但无监督学习只使用大量未标注示例而会忽略已标注示例的价值，这显然也是一种不合理的浪费。因此，研究如何综合利用少量已标注示例和大量的未标注示例来提高学习性能的半监督学习(Semi-supervised Learning)已成为当前机器学习研究中最受关注的问题之一。

一般认为，半监督学习的研究始于 B.Shahshahani 和 D.Landgrebe 的工作[36]，但早在上世纪 80 年代末一些研究者就已经意识到了未标记示例的价值[37]。随着统计学习技术的不断发展，以及利用未标注示例这一需求的日渐强烈，半监督学习在近年来逐渐成为一个研究热点。

半监督学习的基本设置是给定一个来自某未知分布的有标注示例集 $L=\{(x_1, y_1), (x_2, y_2), \cdots, (x_{|L|}, y_{|L|})\}$ 以及一个未标注示例集 $U=\{x'_1, x'_2, \cdots, x'_{|U|}\}$，期望学得函数 f: X→Y 可以准确地对示例 x 预测其标注 y。这里 $x_i, x'_j \in X$ 均为 d 维向量，$y_i \in Y$ 为示例 x_i 的标记，|L|和|U|分别为 L 和 U 的大小，即它们所包含的示例数。

很多研究者从理论上探讨了为什么可以利用未标注示例来改善学习性能。实际上，只要能够合理建立未标注示例分布和学习目标之间的联系，就可以利用未标注示例来辅助提高学习性能。例如，D. J. Miller 和 H.S.Uyar [38] 从数据分布估计的角度给出了一个直观的分析。他们假设所有数据服从于某个由 L 个高斯分布混合而成的分布，即

$$f(x/\theta)=\sum_{l=1}^{L}\alpha_l f(x/\theta_l) \qquad \text{式（2-1）}$$

其中 $\sum_{l=1}^{L}\alpha_l=1$ 为混合系数，$\theta=\{\theta_l\}$ 为参数。这样，标注就可视为一个由选定的混合成分 m_i 和特征向量 x_i 以概率 $P(c_i/x_i,m_i)$ 决定的随机变量。于是，根据最大后验概率假设，最优分类由式（2-2）给出：

$$h(x)=\arg\max_k \sum_j P(c_i=k|m_i=j,x_i)P(m_i=j|x_i) \qquad \text{式（2-2）}$$

其中，

$$P(m_i=j|x_i)=\frac{\alpha_j f(x_i|\theta_j)}{\sum_{l=1}^{L}\alpha_l f(x_{il}|\theta_l)}$$

这样，学习目标就变成了利用训练示例来估计 $P(c_i=k|m_i=j,x_i)$ 和

$P(m_i = j|x)$。这两项中的第一项与类别标注有关，但第二项并不依赖于示例的标注。因此，如果有大量的未标注示例可以利用，就意味着能用于估计第二项的示例数显著增多，使第二项的估计变得更加准确，进而导致式（2-2）更加准确，也就是说，分类器的泛化能力得以提高。此后，T.Zhang 和 F.J.Oles[39]进一步分析了未标注示例在半监督学习中的价值，并指出如果一个参数化模型如果能够分解成 $P(x,y|\theta) = P(y|x,\theta)P(x|\theta)$ 的形式，那么未标注示例的价值就体现在它们能够帮助更好地估计模型参数从而导致模型性能的提高。

在更一般的情况下，建立未标注示例和目标之间的联系需要在某些假设的基础上。目前，在半监督学习中有两个常用的基本假设，即聚类假设（Cluster Assumption）和流形假设（Manifold Assumption）[63]。

聚类假设是指处在相同聚类（cluster）中的示例有较大的可能拥有相同的标注。根据该假设，决策边界就应该尽量通过数据较为稀疏的地方，从而避免把稠密的聚类中的数据点分到决策边界两侧。在这一假设下，大量未标注示例的作用就是帮助探明示例空间中数据分布的稠密和稀疏区域，从而指导学习算法对利用有标注示例学习到的决策边界进行调整，使其尽量通过数据分布的稀疏区域。

流形假设是指处于一个很小的局部邻域内的示例具有相似的性质，因此，其标注也应该相似。这一假设反映了决策函数的局部平滑性。相对聚类假设着眼整体特性，流形假设主要考虑模型的局部特性。在该假设下，大量未标注示例的作用就是让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性，使得决策函数能够更好地进行数据拟合。

值得注意的是，一般情形下，流形假设和聚类假设是一致的。由于聚类通常比较稠密，满足流形假设的模型能够在数据稠密的聚类中得出相似的输出。然而，由于流形假设强调的是相似示例具有相似的输出而不是完全相同的标注，因此流行假设比聚类假设更为一般，这使其在聚类假设难以成立的半监督回归中仍然有效[40][41]。

根据半监督学习算法的工作方式，可以大致将现有的很多半监督学习算法分为三大类。第一类算法以生成式模型为分类器，将未标注示例属于每个类别的概率视为一组缺失参数，然后采用 EM 算法来进行标注估计和模型参数估计。此类算法可以看成是在少量已标注示例周围进行聚类，是早期直接采用聚类假设的做法。第二类算法是基于图正则化框架的半监督学习算法。此类算法直接或间接地利用了流形假设，它们通常先根据训练例及某种相似度度量建立一个图，图中结点对应了（已标注或未标注）示例，边为示例间的相似度，然后，定义所需优化的目标函数并使用决策函数在图上的光滑性作为正则化项来求取最优模型参数。第三类算法是协同训练（co-training）算法。此类算法隐含地利用了聚类假设或

流形假设，它们使用两个或多个分类器，在学习过程中，这些分类器挑选若干个置信度高的未标注示例进行相互标注，从而使得模型得以更新。

2.2 结构学习算法

假设一个好的分类器是可以用一些潜在的函数结构所表示的，结构学习算法的[42]基本思想是通过同时考虑多个预测问题来学习到这个结构，又可以叫做一种多任务学习，但是结构学习这个名字更能准确反映它的机制。从直觉上来说，当我们对不同的问题得到多个分类器的时候，我们就有了一个好的潜在预测空间的示例，可以从中分析出这些分类器所共享的“共同结构”。当预测空间中发现一个重要的预测结构时，我们就可以用这个信息来提高每一个预测问题。

我们首先描述一个简单的结构学习模型。考虑 m 个学习问题（l=1, …, m），每个问题有 n_l 个例子，分别独立的符合分布 D。对于每个问题 L，假设我们有一个候选的预测空间 H，共同的结构参数是所有问题共享的 H_θ。

现在，对于第 l 个问题，我们感兴趣的是找到一个 H 中的分类器 F:X->Y，使 D 上的预期损失最小化。为了使符号简单，我们假设所有问题有着相同的损失函数（尽管这在我们的分析中并不是必需的）。给一个固定的结构参数，每个问题的分类器可以用经验风险最小化在预测空间 H 上得出。

$$\hat{f} = \arg\min_{f \in H} \sum_{i=1}^{n_l} L(f(X_i^l), Y_i^l), \qquad (l=1,...m) \qquad \text{式（2-3）}$$

结构学习的目标就是找出一个最优的结构参数可以使 m 个问题上分类器的平均预测风险最小。

如果我们用交叉验证对结构参数进行选择，我们可以立即注意到通过多个学习任务一起学习可以得到一个更稳定的预测。实际上，如果对每个问题 L，我们对 N 个例子有一个确定的关系集（X,Y），然后对于结构学习，关系集的总数是 $\sum_{l=1}^{m} n_l$。因此我们对于选择最优的共享假设空间结构的这个目的就有了更充足的数据。这表示即使对于每个个别的问题示例很少，只要 m 很大，我们都可以找到正确的最优 θ。

一般而言，预测空间决定学习到的分类器的函数结构。θ 参数可以作为一个连续参数来反映我们对好的分类器应该的样子的假设。如果我们有一个很大的参数空间，那我们可以构建很多可能的函数结构，这种观点暗示了当问题的个数 m

很大时发现最优的共享结构是可能的。

特别要指出的是，结构学习并不必须是半监督学习，我们可以在未标注语料上也可以在标注语料上通过同时预测多个问题来寻找这个“共同结构”。但对于本文来说，我们主要专注于结构学习应用在未标注语料的预测问题上，即半监督学习。

下面提供一个直觉上的讨论，为什么原则上存在着被多个任务共享的好的函数结构（好的预测空间）。概念上，我们可以考虑简单的例子$H_{1,\theta}= H_{\theta}$，即不同的问题共享完全相同的预测空间。

给一个任意的不包含任何已知结构的输入空间 X，我们认为可能从多个预测问题中学习到一个“好”的分类器。关键的原因在于实际上，并不是所有的分类器都同样的“好”。在现实世界的应用中，人们通常观察到分类器反应了输入空间中某一潜在的距离。一般来说，如果两个点在这种潜在距离中是相近的，那一个好的分类器在这些点上的结果也应该是相近的。而完全随机的分类器是“坏”的预测器，在实际的应用中很少用到。

然而，实际上经常不是很清楚应该如何测量这种潜在距离。例如，在自然语言处理中，空间 X 由离散的点（如词）组合而成，并没有一种可以很清晰定义的距离。即使对于连续向量值的输入点，也很难判断欧氏距离是不是就比其他的距离好。更进一步说，即使在一个好的距离函数被选择以后，我们也很难能对于距离定义适当的预测条件。如果我们观察多个任务，那么重要的共同结构可以从数据中学习到的多个分类器中分析出来。如果这些任务是与实际我们感兴趣的学习任务非常相似的，那我们就可以从发现的结构中显著的受益。即使这些任务不是直接相关，找到的结构也仍然是有用的，这是因为一般说来，分类器对于潜在输入空间中固有的特定的距离趋向共享相似的预测条件。

用图 2-1 来描述我们的主要论点，我们考虑一个有 6 个点的离散输入空间 X={A,B,C,D,E,F}。假设我们从三个不同的预测问题中得到三个函数估计，函数的值与输入的值如图所示。在这个例子中，我们能注意到 A,C 和 D 上的函数值是相似的，同时 E 和 F 上的函数值是相似的。因此用观察到的估计函数，我们可以得出在 X 上，A,C 和 D 是互相接近的，而 E 和 F 是互相接近的。一个 X 上好的分类函数应该能够反应这种固有的距离。

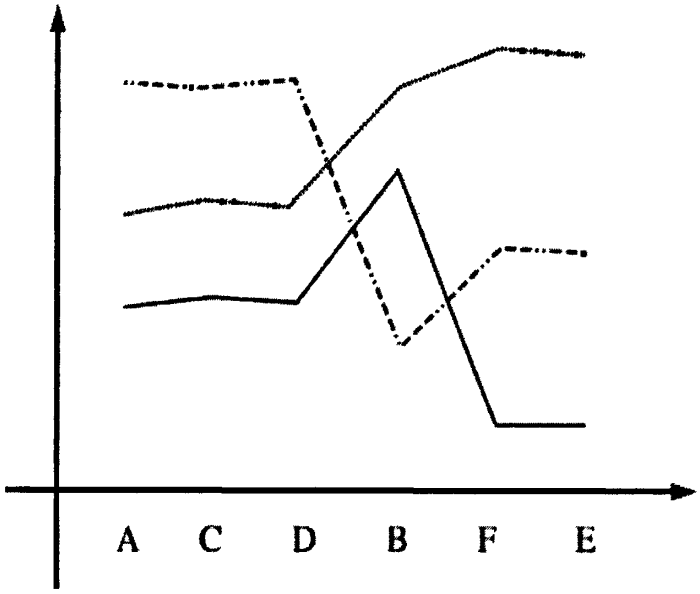

图 2-1 多个预测任务中的结构函数

2.3 ASO 算法

ASO 算法[42][43]即“交互结构最优化”(Alternating Structure Optimal)，是一种结构学习算法，它通过在未标注数据上同时建立多个和目标任务相关的辅助问题并同时学习这些分类器来找出其中具有很强预测能力的共享结构，从而对提高目标问题的分类性能。ASO 的性能主要依赖于构造的辅助问题和目标问题的相关性，相关性越大，它的性能也就越好。ASO 目前在英文自然语言处理领域已经有一些应用，比如：词性标注、命名实体识别、语法组块分析、语义角色标注[44]、文本分类、词义消歧[48]等等，但中文方面还很少见其应用。

下面给出了 ASO 算法的简单介绍。

给定包含 n 个特征向量的例子的训练集$\{X_i,Y_i\}$和它们对应的二元标记，其中$i\in\{1,...,n\}$，每个X_i是一个p维向量，一个二元线性分类器近似的将这种未知的关系表示为$Y=u^TX$。如果u^TX是正数则输出表示为+1，反之则表示为-1。其中 u 称为权值向量。一种常用的来寻找最优预测模型u的方法是加入了正则化因子的经验风险最小化（empirical risk minimization），如式（2-4）。其中，按照经验值，我们将正则化因子固定为$\lambda=10^{-4}$。$\|u\|^2$定义为$\sum_{i=1}^{p}u_p^2$。我们选择 L-BFGS

作为训练算法，L-BFGS 是一种充分利用以前的梯度和修改值来近似曲率值的二阶方法，只要求提供似然函数的一阶导数，它对于正则化的凸的 ERM 的优化问题有很好的效果。式右边第二项为正则化因子。

$$\hat{u} = \arg\min_{u} \frac{1}{n}\sum_{i=1}^{n} L(u^T X_i, Y_i) + \lambda \|u\|^2 \qquad \text{式(2-4)}$$

$L(p,y)$ 叫做损失函数，就是衡量预测的标记和真实的标记之间的差异造成的损失的函数。我们使用的是 Huber 损失函数的修改版本，是一种常用的损失函数，如式（2-5）所示。

$$L(p,y) = \begin{cases} -4py & if\ py < -1 \\ (1-py)^2 & if\ -1 \le py < 1 \\ 0 & if\ py \ge 1 \end{cases} \qquad \text{式(2-5)}$$

ASO 算法可以看成是一个改进的多任务线性分类器。其中最关键的部分就是用来得到结构参数 θ 的二元分类问题不必须是那些我们要最终要解决的问题。事实上，可以为得到一个更好的 θ 而专门设计一些新的问题。因此，在 ASO 中我们将这两种问题分开，为得到 θ 而提出的问题，我们称之为辅助问题。而为了解决我们最终要求的问题，我们称之为目标问题。在 ASO 中，为求解目标问题，需要引入一些辅助问题。这需要有一个基本的假设就是，辅助问题和目标问题存在一个共同的低维预测结构。

其基本过程是，首先对每个辅助问题由 $Y = u^T X$ 求解权值 u。对 m 个辅助问题，用一个 h*p 维的矩阵 θ 来表示 m 个权重向量 u_l 的共同结构，其中 $l \in \{1,...,m\}$（$h \le m$）。基于结构化因子 θ，定义目标问题的权值为式(2-6)，并根据式(2-7)，通过联合经验风险最小化（Joint empirical risk minimization）迭代求出最优解，优化算法可用 L-BFGS 算法。。

$$u_l = w_l + \theta^T v_l \qquad \text{式(2-6)}$$

$$[\{\hat{w}_l, \hat{v}_l\}, \hat{\theta}] = \arg\min_{\{w_l, v_l\}, \Theta} \sum_{l=1}^{m} \left(\frac{1}{n}\sum_{i=1}^{n} L((w_l + \theta^T v_l)^T X_i^l, Y_i^l) + \lambda \|w_l\|^2\right) \qquad \text{式(2-7)}$$

$$s.t. \quad \theta\theta^T = I_{h\times h}$$

假设用 k 个目标问题和 m 个辅助问题，简化的 ASO 算法的求解步骤如下：

1 对于 m 个辅助问题中的每一个问题，根据式（2-4）学习到权重 u_l

2 用 u_l 组成矩阵 $U = [u_1, u_2, ..., u_m]$，一个 $p \times m$ 的矩阵。这是对 Ando 的 ASO

算法的一个简化版本，这样可以使所用的辅助问题使用相同的λ。

3 在 U 上用 SVD 分解(Singular value decomposition，奇异值分解)：$U=V_1SV_2^T$，其中V_1是$p\times m$维的矩阵。取出V_1^T的前h行作为结构化因子θ，用它来刻画大规模的未标注语料中的“共同结构”。

4 对于给定的θ，我们用最小化经验风险的方法求得 k 个目标问题中每个问题的 w 和 v

$$\frac{1}{n}\sum_{i=1}^{n}L((w+\Theta^Tv)^TX_i,Y_i)+\lambda\|w\|^2 \qquad \text{式(2-8)}$$

5 目标问题的权重可表示为

$$u=w+\theta^Tv \qquad \text{式 (2-9)}$$

通过选择一个凸的损失函数，如式（2-5）中所示的，步骤 1 和 4 可以作为一个凸函数的优化问题而得到解决。

步骤 3 抽取出辅助问题的预测器所分享的最关键的部分，希望它对目标问题的预测同样有效。

需要说明的是，ASO 方法并不等同于以前的主成分分析(PCA)方法，PCA 可认为在数据空间里降维，而 ASO 是在预测器空间进行降维，可以认为是在很多预测器上找到“主成分”，即找到一个具有最高预测能力的低维结构。

图 2-2 为 ASO 算法的基本流程图。

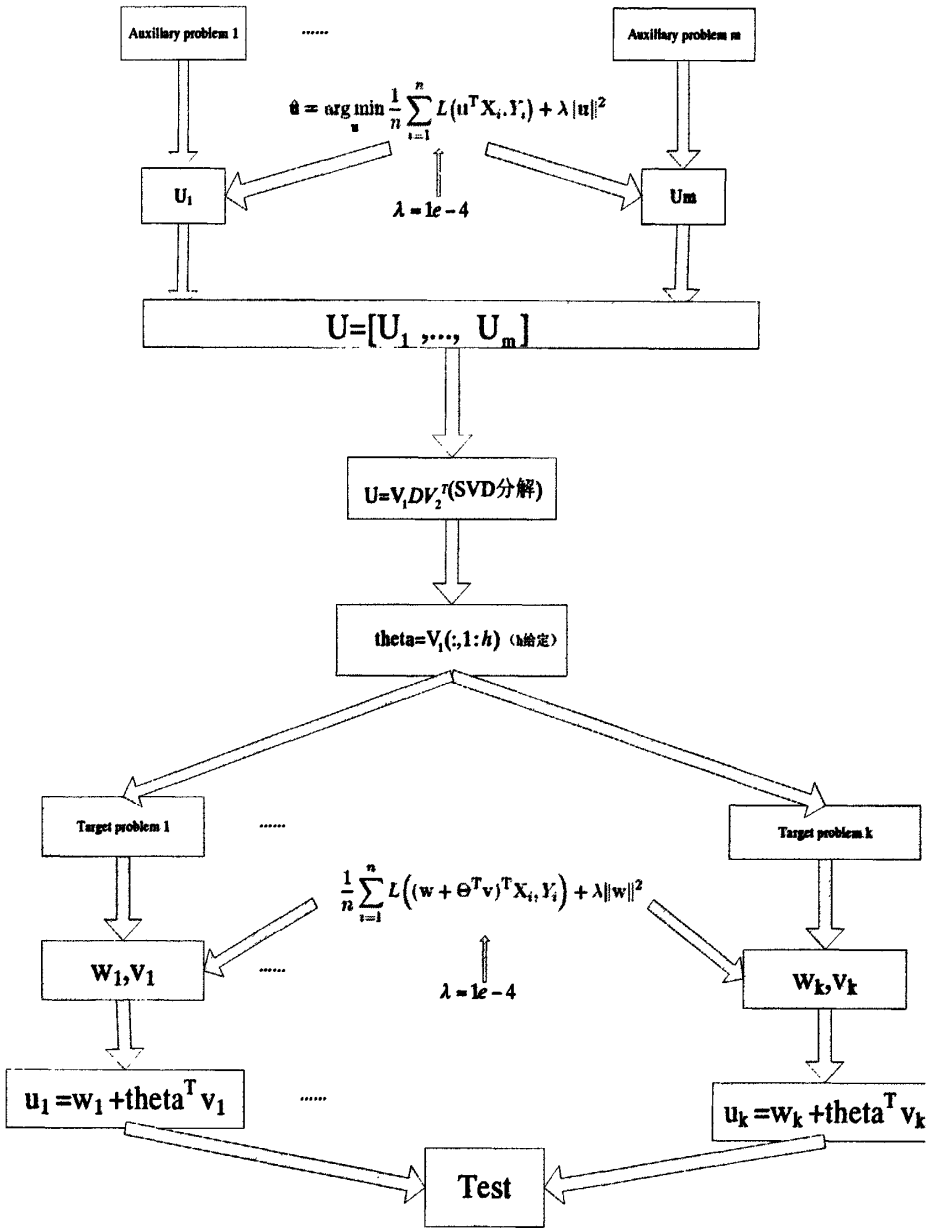

图 2-2 ASO 算法的基本流程图

2.4 辅助问题

ASO 算法重要思想就是——提取 θ 的辅助问题不必是我们所要求解的问题

本身，而可以是与目标问题具有相关性的问题。所以对 ASO 算法来说如何构建“好”的辅助问题是关键。辅助问题的好坏主要看它与目标问题的相关性，显而易见的是，如果辅助问题和目标问题的相关性越大，那么辅助问题对目标问题的就越有帮助。但是如何确定辅助问题与目标问题的相关性，目前并没有好的方法，一般还是基于经验来进行筛选。

2.4.1 辅助问题分类

辅助问题可以是无监督的，半监督的，也可以是全监督的[44]，下面我们分别进行介绍：

无监督辅助问题

无监督的辅助问题就是问题的结果是可以直接从原始语料中得到的，而不用提供附加的额外标注。比如用组块的其他特征来预测组块的第一个词是不是“中国”这样的问题，答案是可以直接从原始语料中得到的。

创建这种辅助问题的难度相当于通常的为监督问题选择特征，所以我们可以根据对这个目标问题的了解来创建辅助问题。而且就算我们选择的辅助问题对目标问题没什么帮助，也只不过是引入了一些无效特征，并不会严重的损害到 ERM 学习器。另一方面来说，一旦选择的辅助问题是合适的，那得到的收益将相当显著。

另外要注意到的是，由于未标注语料本身的丰富性，相较标准的监督集上的特征工程，我们可以有更广泛的选择。比如，在监督集上会造成严重的数据稀疏问题的高维特征却可以用于辅助问题，因为大量的未标注语料可以提供更可靠的统计。这样可以从高维特征中得到其低维预测结构用于监督任务，而不会造成数据稀疏问题。

半监督辅助问题

半监督的辅助问题是问题的结果虽然不是直接可以从原始语料中得到的，但是需要的是低级的标注。比如用组块的其他特征来预测组块的第一个词的词性是不是“名词”这样的问题，虽然不能直接从无标注语料中得到，但是需要的只是词性标注而不是语义角色标注。一般来说，有低级的标注的语料数量远远高于有高级的标注的语料，使用这种辅助问题也能得到很好的效果。

全监督辅助问题

由于结构学习本身并不局限于半监督学习算法，所以辅助问题完全可以建立在标注语料上，成为全监督的辅助问题。Chang Liu[44]在语义角色标注任务上就应用了监督的辅助问题。比如在语义角色标注问题中，目标问题为预测当前单元

是否为 arg0 时，辅助问题可以是预测当前单元是否为 arg1, arg2, ...。但由于本文我们主要关注的是半监督学习算法，所以对这种辅助问题没有进行深入探讨。

2.4.2 构建辅助问题的方法

Ando[42]提出了几种构建辅助问题的方法。

1 预测一个用到的特征：将某一用到的特征“盖住”，用其余所有特征预测现在“不可见”的特征。

比如目标问题用到的特征有 10 项，辅助问题我们可以用 10 项特征中的 9 项预测剩下的那个特征的结果。在实际实验中，我们简单的将要被预测的（盖住的）那个特征置为 0 即可。

这种辅助问题与目标问题的相关性是显而易见的，因为它预测的就是能够预测目标问题的特征。但我们还是希望能够尽量选择到具有最强预测能力的特征来作为辅助问题的答案。

2 预测没用到的特征：比如有些特征可能在语义角色标注这个任务上并不是很有效，所以在目标问题的分类器中并没有用到。但是它还是与目标问题相关的，所以我们在辅助问题中预测它。

3 预测分类器的结果。

这种方法有些类似于协同训练（co-training）。我们用两个（或更多）不同的特征映射集：Φ_1：$X->F$ 和 Φ_2：$X->F$。首先，我们对目标问题训练一个分类器，用特征映射集 Φ_1 和标注数据，辅助任务则是用 Φ_2 来预测这个分类器的行为（如预测出的标注，指派的置信值等）。要注意的是不同于协同训练的是，我们用这个分类器只是作为一种创建辅助问题的手段，以此来达到相关性的要求，而不是用它来 bootstrap 标记。

第三章　中文语义角色标注

语义角色标注可以看作是分类问题。也就是说，人们可以逐一判断一个标注单元是否是某一动词的语义角色，然后再更进一步的预测其属于何种具体的语义角色。

一般的语义角色标注系统一般分为 4 个步骤：剪枝(Pruning)、识别(Identification)、分类(Classification)和后处理(Post-processing)。其中，剪枝是指根据启发式规则，删除大部分不可能成为语义角色的标注单元，这样可以大幅减少待识别实例的个数，提高系统的效率。识别过程一般是对一个标注单元是否是语义角色加以判别，并保留识别成语义角色的标注单元，待下一步进一步分类究竟属于哪个语义角色类，这样也可以减少进入分类判别的实例的个数，加快处理速度。最后根据语义角色之间的一些固有约束进行后处理。这些约束通常包括，一个谓语动词不能有重复的核心语义角色并且语义角色不存在相互重叠或嵌套等等。

然而，并非所有的系统都必须包括以上 4 个步骤，特别是前两个步骤，其主要目的是提高处理效率，但随之带来的是召回率的下降，即损失了一些本应是语义角色的标注单元。而且如后面我们会具体介绍的，我们的标注单元是组块而不是传统的句法成分，非语义角色的情况比例没有那么大。因此，我们的系统中，去除了剪枝步骤，合并了识别和分类步骤，直接对语义角色进行分类。也就是将非语义角色的标注单元也看成是一类，将非角色和所有角色一起判断，使用多类分类器，将组块划分为具体的语义角色或 NULL。在多分类器中，应用了 ASO 算法，从而有效地利用了未标注语料。

在本章内，将详细的介绍我们的基于 ASO 算法的结构学习语义角色标注系统的构建过程，包括我们所使用的语料，评测函数，标注单元，特征，辅助问题等。最后对实验结果进行了分析，探讨结构学习算法在语义角色标注任务上起到的作用。

3.1 ASO 语义角色标注系统

既然我们将语义角色标注问题看成一个多分类问题，很自然可以用 ASO 算法来解决。对于我们的语义角色标注任务来说，目标问题就是将当前标注单元分给一个合适的语义角色类别，即（Arg0, Arg1, ..., ArgM-LOC, ..., NULL）其中之一。而辅助问题我们在后面的章节中给出具体的构建。

我们的实验步骤如下：

1 对 CPB 语料进行处理，将树库的形式转化为以组块为单位的扁平形式，并提取我们需要的特征。

2 将特征构建为稀疏矩阵格式。

3 在未标注语料上对辅助问题用线性分类器进行分类，每个辅助问题都是一个二分类问题，最后从这些问题的 u 中提出共同结构参数θ。

3 在标注语料的训练语料上对目标问题进行训练，将上一步得到的θ代入线性分类器得到 ASO 分类器。

4 在测试语料上分别用 ASO 分类器和线性分类器进行测试，进行对比。这是一个多分类任务，结果有 19 种可能，即每个组块分类为（Arg0, Arg1, ..., ArgM-LOC, ..., NULL）其中之一。

经过分类步骤产生的语义角色标注结果中，会有个别不满足语义角色标注约束的情况。在语义角色标注中，一般对于一个谓词，不能够存在两个或两个以上相同的语义角色(重复)。我们在后处理阶段使用 winner take all 的方法，只保留分类阶段输出的概率最大的角色。

3.2 语料处理和评测函数

我们的实验使用了 CPB 的数据并对原始语料进行了预处理。

既然是语义角色分类，那么首先要解决的是有多少个类的问题。我们并没有直接使用 PropBank 中的论元标记，而是对其中的论元标记进行了化简合并。最后如表 3-1 所示共 18 个论元。这样分类的依据是：对于 Arg0-Arg4 这一类的核心论元，虽然有些也包含二级的功能标记，但是毕竟是少数；并且功能标记不同的同类核心论元之间区别不大，所以二级功能标记对核心论元起的作用有限，所以我们将其二级功能标记去除。

表 3-1 CPB 论元列表

标签	描述

Arg0	动作的施事或起因
Arg1	主题
Arg2	范围
Arg3	动作开始点
Arg4	结束点
ArgM-ADV	状语
ArgM-BNF	受益人
ArgM-CND	条件
ArgM-DIR	方向
ArgM-EXT	程度
ArgM-FRQ	频率
ArgM-LOC	位置
ArgM-MNR	方式
ArgM-PRP	目的
ArgM-TMP	时间
ArgM-TPC	主题
ArgM-CRD	并列论元
ArgM-PRD	次谓词
ArgM-PSR	所有人
ArgM-PSE	被占有者

CPB 语料的格式如图 3-1 所示。

Predicate 摆放

```
Frameset: [1

   ARG0: agent
   ARG1: theme
   ARG2: location

   :

( (IP (NP-SBJ (NN 贺礼))
      (VP (ADVP (AD 日後))
          (VP (VV 摆放)
              (NP-OBJ (NN 博物馆))))
      (PU 。)))

ARG1:  贺礼
ARG2:  博物馆
REL:  摆放
```

图 3-1 CPB 语料示例

经过我们处理后的语料文件格式如表 3-2 所示。

表 3-2 处理后的语料格式

第 1 列	组块所在的文件号
第 2 列	组块所在的语句号
第 3 列	当前词的序号
第 4 列	当前词的句法成分
第 5 列	当前词的词性标注
第 6 列	当前词的内容
第 7 列	当前词所在的组块类型
第 8 列	当前词所在组块中心词，若不存在用???表示
第 9 列	中心词的序号，若不存在用???表示

表 3-3 是一句处理后的语料的例子。

表 3-3 处理后的语料的示例

0007	9	64	NP	NN	祖国	SBJ	祖国	64
0007	9	65	VP	VV	尊重	PRD	尊重	65
0007	9	66	NP	NN	香港人	OBJ	香港人	66
0007	9	67	O	PU	、	NULL	???	???
0007	9	68	VP	VV	相信	PRD	相信	68
0007	9	69	NP	NN	香港人	OBJ	香港人	69
0007	9	70	O	PU	、	NULL	???	???
0007	9	71	VP	VV	爱护	PRD	爱护	71
0007	9	72	NP	NN	香港人	OBJ	香港人	69

需要特殊说明的是，CPB 中有一类特殊的空语类标记，在我们的系统中没有去除这些空语类标记而是也将其作为一个特征使用。

下面选择 CPB 中的一些例句对于空语类进行说明：

“富于爱心的社会”。“富于”前面有一个空语类的标记“*T*”，并且“*T*”在树库中带有主语标记，谓词“富于”的 Agr0 被标注为“*T*”。

“产品的国内市场占有率和品种及盖率分别达到百分之七十和百分之九十，并返翎日本，出口韩国和印度等国家”。在“并”的前面有空语类标记“*Pro*”，并且判断谓词“返销”的 Argl 为“*Pro*”。

“芬兰的零售商们为方便公众购物，…”。介词短语中的谓词“方便”前面有一个“*Pro*”，并带有主语标记，所以只需要将“方便”的 Arg0 标为“*Pro*”即可。但介词短语中的谓词是否与前面的主语有语义关系是不确定的，使用“*Pro*”标记实际上是忽略了这种区分。

另外，虽然 CPB 中承前省略需要找出省略的成分，但蒙后省略都有“*Pro*”作为标记，不需要判断出省略的成分。

我们根据常用的评测标准：准确率 precision、召回率 recall 和综合评价 $F_{\beta=1}$ 来进行性能评测。

$$准确率=\frac{标注正确的语义角色的数目}{标注的语义角色总数目}$$

$$召回率=\frac{标注正确的语义角色的数目}{应当标注的语义角色的总数目}$$

$$F_{\beta=1}=\frac{2*召回率*准确率}{召回率+准确率}$$

3.3 标注基本单元

自动标注的基本单元一般可以是句法成分（Constituent）、短语（Phrase）、词（Word）或者依存关系（Dependency Relation）等等。

在图 3-2 的短语结构句法分析树中，每个非终结节点，如 S，NP，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。也就是说一个语义角色必然对应一个句法成分，反之未必。如在图 3-3 的例子中，Arg0 对应一个 NP，ArgM-LOC 对应一个 PP 等等。然而，我们很难自动的获得这种深层句法分析的结果，尤其是除英语外的其它语言，而且现有的句法分析系统，在通用领域表现也不尽如人意。

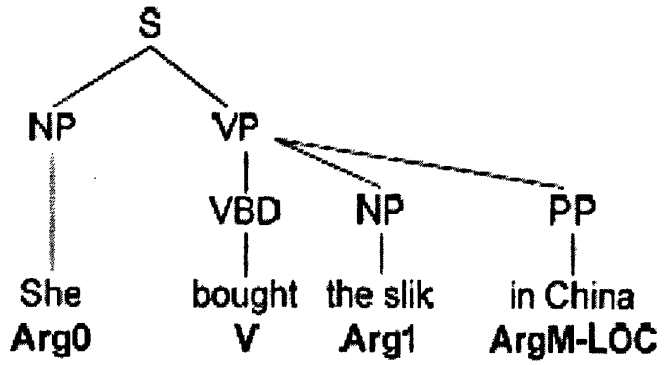

图 3-2 PropBank 中的句子标注

为此有人试图将浅层语义分析建立在浅层句法分析的基础之上的，毕竟浅层句法分析的鲁棒性要好于深层句法分析。虽然通过浅层句法分析只能获得非嵌套短语的信息，而不能获得全部的句法分析结果，也就是不能获得句法成分的分析结果，但是我们一般认为一个非嵌套的短语属于同一语义角色，因此产生了使用短语作为浅层语义分析的基本元的系统。词是比短语更细的语言单元，有些系统也使用其作为标注的基本单元，然而效果并不理想。

还有人基于依存句法分析结果进行浅层语义分析，也取得了可以与基于短语结构句法分析相似的效果。此时的语义标注单元为依存关系。我们可以直接使用依存句法分析器获得依存句法分析的结果，也可以转化短语结构句法分析的结果为依存句法分析结果。与基于短语结构句法分析的方法相比，基于依存句法分析不但可以利用短语之间的依存特征，而且只需要学习和预测与目标动词有依存关系的短语为某种语义角色即可，因此极大地加快了计算的速度。

我们的系统就是将语义角色标注的工作建立在浅层语法分析之上，不再对树上节点进行分类，而是利用组块进行语义角色标注，希望利用相对更准确些的组块分析结果提升语义角色标注准确率，绕过分析准确率相对较差的完全句法分析。

组块分析（chunk）又称为浅层句法分析或部分句法分析，是自然语言处理的重要任务之一。句法分析是自然语言理解的基础，但由于目前对大规模真实文本进行完整的句法分析遇到很多困难，许多研究人员尝试着把一个完整的句法分析问题分解为几个易于处理的子问题，以降低完整句法分析的难度。组块分析就是这样一个子问题，它的句法描述能力介于词性标记序列和完整句法树表示之间，可作为最终形成句子的句法层次树的一种预处理手段，为后续的句法分析提供了基础、依据和辅助信息。组块分析在机器翻译、信息检索、信息抽取、文本分类及语音识别等领域具有重要的应用价值。

Abney[61]最早就英语提出了一个完整的组块描述体系，对组块有着权威性的定义。他把组块定义为子句内的一个非递归的核心成分。这种成分包含核心成分的前置修饰成分，而不包含后置附属结构。组块不一定覆盖整个句子，例如常有一些介词、连词等不属于任何一个组块。

周强等[62]对整理和加工中文组块库做了大量工作，建立了一个完整的组块划分体系，设计了 8 个标记的组块标记集（包括主语块、述语块、宾语块、兼语块、状语块、补语块、独立块、语气块）。本文中对组块的定义与此类似。

组块的研究还没有得到充分的重视，目前还没有可用的具有相当规模的组块库。CPB 本身并不是一个组块库，但我们可以通过程序自动实现从中文树库到组块的转换， 每次从一颗句法树的底层开始抽取，也就是得到最小的短语类型，并适当的转换，生成相应类型的组块。

在下面的实验中，我们以组块为基本标准单元，用结构学习算法进行实验，探讨语义角色标注的特点。

我们认为，语法分析应该为语义分析提供帮助，因此希望对于组块的研究能更好地揭示组块与论元结构之间的内在联系，从而以组块描述体系作为出发点，建立汉语的句法、语义、语用分析的紧密结合体。据此，我们定义的组块类别主要关注词语在句子中的功能成分，表 3-4 是我们实验中用到的组块类型和解释。

表 3-4 汉语组块的类别

序号	类别	解释
1	SBJ	基本主语组块
2	OBJ	基本宾语组块基本
3	PRD	谓语组块
4	ADV	基本状语组块
5	PPOBJ	介词宾语组块
6	PPSBJ	介词主语组块
7	Null	非组块

3.4 特征选择

特征一直是决定统计自然语言处理系统性能的重要因素，对于语义角色标注，特征同样起着举足轻重的作用。相比特征空间较小的较底层的自然语言处理任务，如分词、词性标注和命名实体识别，语义角色标注任务的一个显著特性就

是特征空间很大。英文语义角色标注基础系统中常用的有七个特征（谓词、位置、路径、中心词、子类框架、短语类型、语态），一般的标注系统都用它们作为基本特征。但与传统的方法不同，基于组块的语义角色标注系统不需要句法分析。这在提高了系统的效率同时，也给特征的选择带来了一定困难。传统方法中句法相关的特征，比如路径，动词子类框架等都无法使用。这使得我们必须多利用词一级的特征。我们实验中参考了英文语义角色标注基础系统中常用的特征并引入了词和词性特征。另外由于中文谓词不同于英文的谓语动词，没有主动被动的语态之分，因此我们加入了“把”和“被”作为简单的语态特征。以上这些特征，都从不同的侧面反映了待标注单元的语义角色信息。

最后我们确定使用了 14 个特征，具体介绍如下：

1 组块类型：组块类型对语义角色往往有较为明显的指示作用，比如组块类型是 SBJ 语义角色经常为 Arg0。

2 核心词：标注单元的核心词。带特定核心词的组块很可能是特定类型的论元。

3 核心词词性：此特征是对核心词的泛化，能够缓解一定的数据稀疏问题。

4 组块中第一个词。

5 组块中第一个词词性。

6 组块中最后一个词。

7 组块中最后一个词词性。

8 谓词：即待标注的目标动词。直接决定了可能的论元。

9 语态：CPB 标注集中词性为 BA 表示“把”，词性为 LB 或 SB 表示“被”。一般“把”是主动语态，“被”是被动语态，会影响到 Arg0、Arg1 等在谓词前后的位置。

10 位置：待标注的组块与目标动词之间的相对位置，这是二元特征，分别为“前”和“后”。对于“覆盖”的情况，我们根据启发式规则直接将其忽略，因为这种情况下，组块不可能成为谓词的角色。位置特征对于语义角色也具有较强的指示作用，例如施事往往在“前”。

11 组块前一个词：若组块前面不存在任何词，则为 NULL。

12 组块前一个词词性。

13 组块后一个词：若组块后面不存在任何词，则为 NULL。

14 组块后一个词词性。

3.5 辅助问题构建

在ASO算法中，辅助问题的好坏直接决定了最后标注结果的好坏。合适的辅助问题能从未标注语料中得到大量有用的信息，使我们的目标分类器性能显著提高。而不好的辅助问题会引入“噪声”，甚至可能对最后的目标分类器性能造成一定损害。所以如何构建对于语义角色标注最适当的辅助问题是我们研究的关键。

对于我们的语义角色标注任务，由于我们主要是研究半监督学习算法，所以并没有采用全监督的构造辅助问题的方法。我们的实验遵循以下原则构造辅助问题：

1、相关性：辅助问题要和目标问题相关。

比如说，预测辅助问题和目标问题时用到的特征要一致，它们要共享相同的预测空间。这是构建所有辅助问题最重要的原则，只有满足了相关性的要求，我们在辅助问题上所得到的“共同结构”才会对目标问题的求解产生积极的影响。

2、自动标注：结果可以从原始语料直接观察到，或者是低级已标注的结果。

这是我们半监督学习算法的性质所决定的，满足这一条件，才能尽量充分的利用大量的未标注语料，弥补中文语义角色语料不足的问题。

3、简单性：线性分类器是简单分类器，所以我们的辅助问题也要尽量简单。

因为我们的分类器是简单的线性分类器，如果辅助问题非常复杂的话，线性分类的效果可能不会很好，而且也会大大增加求解问题需占用的时间和机器资源。相对于目标问题而言，如果选择的辅助问题足够简单，就能够利用大规模的未标注语料来方便地构建该类问题的分类器。

辅助问题理论上可以创建几千个，但是考虑我们的语料规模比较小，另外太多的辅助问题要求的机器资源也很惊人，我们实验中使用了Ando构造的辅助问题的第一种方法，即用其余所有特征预测某一特征。为了得到更多正例，我们预测出现比较频繁的特征。辅助问题如下：

1 用特征1（组块类型）以外的特征预测特征1，因为特征1可能的值比较少，所以我们全部预测，共7个问题。

2 用特征4（组块中的第一个词）以外的特征预测特征4，因为特征4的特征空间很大，所以我们只预测最频繁的前100个词，共100个问题。

3 用特征5（组块中的第一个词的词性）以外的特征预测特征5，我们只预测最频繁的前30个词性，共30个问题。

4 用特征6（组块中的最后一个词）以外的特征预测特征6，因为特征6的

特征空间很大，所以我们只预测最频繁的前 100 个词，共 100 个问题。

5 用特征 7（组块中的最后一个词的词性）以外的特征预测特征 7，我们只预测最频繁的前 30 个词性，共 30 个问题。

综上所述，我们一共创建了 267 个辅助问题。可以看出这些辅助问题可以归结为 3 类，预测组块类型，预测词，预测词性，其中预测词是无监督的辅助问题，预测词性和组块类型需要一定的标注，属于半监督的辅助问题。

3.6 实验结果及分析

我们的所有实验均将语料分成 5 份，4 份做训练语料，1 份做测试语料，进行了 5 折交叉验证。正则化因子 $\lambda = 10^{-4}$。参数估计均用 L-BFGS。

我们首先用很少的标注语料，大概占 CPB 的 2%，用所有的 CPB 语料做未标注语料进行了实验，参数 h=5，θ 由所有辅助问题得到的 u 混在一起得到，进行实验。结果如下：

表 3-5 标注语料为 CPB 的 2%时的结果

	准确率	召回率	$F_{\beta=1}$
baseline	0.6544	0.5836	0.6170
ASO	0.6800	0.7444	0.7107

然后我们将 CPB 所有的语料做未标注语料，标注语料为语料总量的 20%。对于我们使用的 5 种辅助问题，分别进行了实验。然后将所有辅助问题得到的 u 混在一起得到一个 θ，进行实验。参数 h=5。表 3-6 是 baseline 和总的 ASO 结果的对比，表 3-7 是每种辅助问题单独的结果：

表 3-6 标注语料为 CPB 的 20%时的结果

	准确率	召回率	$F_{\beta=1}$
baseline	0.6662	0.5601	0.6086
ASO	0.6838	0.5716	0.6229

表 3-7 标注语料为 CPB 的 20%时的 ASO（其余全部特征）（h=5）

辅助问题种类	准确率	召回率	$F_{\beta=1}$
第一类	0.6795	0.5703	0.6201
第二类	0.6810	0.5700	0.6206
第三类	0.6773	0.5863	0.6285
第四类	0.6800	0.5684	0.6192
第五类	0.6786	0.5748	0.6224

改变 h 的设置，将其设置为 1 的结果如表 3-8：

表 3-8 标注语料为 CPB 的 20%时的 ASO（其余全部特征）(h=1)

辅助问题种类	准确率	召回率	$F_{\beta=1}$
第一类	0.6811	0.5712	0.6213
第二类	0.6810	0.5708	0.6210
第三类	0.6791	0.5724	0.6212
第四类	0.6811	0.5712	0.6213
第五类	0.6809	0.5695	0.6202

根据之前对单个特征实验的研究，ASO 并不是对所有特征都起作用，低维特征甚至可能损害了最后的性能。所以我们进一步在辅助问题的训练上进行了特征选择，去除低维特征，结果如表 3-9 所示：

表 3-9 标注语料为 CPB 的 20%时的 ASO（高维特征）

辅助问题种类	准确率	召回率	$F_{\beta=1}$
第一类	0.6816	0.5724	0.6222
第二类	0.6801	0.5716	0.6211
第三类	0.6772	0.5936	0.6327
第四类	0.6823	0.5720	0.6223
第五类	0.6803	0.5720	0.6215

进而，我们继续将标注语料的规模扩大，达到 CPB 的 40%，得到以下的实验结果：

表 3-10 标注语料为 CPB 的 40%时的结果

	准确率	召回率	$F_{\beta=1}$
baseline	0.6511	0.5079	0.5707

ASO	0.6625	0.5243	0.5854

表 3-11 标注语料为 CPB 的 40%时的 ASO（其余全部特征）

辅助问题种类	准确率	召回率	$F_{\beta=1}$
第一类	0.6620	0.5196	0.5822
第二类	0.6626	0.5201	0.5828
第三类	0.6651	0.5154	0.5808
第四类	0.6636	0.5178	0.5817
第五类	0.6616	0.5145	0.5789

表 3-12 标注语料为 CPB 的 40%时的 ASO（高维特征）

辅助问题种类	准确率	召回率	$F_{\beta=1}$
第一类	0.6636	0.5196	0.5828
第二类	0.6642	0.5189	0.5826
第三类	0.6595	0.5218	0.5826
第四类	0.6607	0.5228	0.5837
第五类	0.6634	0.5204	0.5833

通过对实验结果的对比分析，我们可以得到以下结果：

1、ASO 算法可以提高语义角色标注分类器的效果，对比 baseline 和 ASO 结果，我们可以看到 ASO 普遍的提高了分类器的效果。这充分说明 ASO 算法的确利用了未标注语料的信息，对目标问题产生了好的影响。在各个种类的辅助问题上单独进行实验的效果并不显著，但综合在一起时显著的提高了分类器的性能，f 值最高提高了 10%。这说明结构学习算法主要依赖于辅助问题的规模，辅助问题越多，效果越好。

2、ASO 算法随着标注语料规模的扩大效果逐渐下降，但仍然有效。在标注语料非常小而未标注语料相对大的时候，ASO 使分类器性能提高了 10%，但随着标注语料比重提高，ASO 的提高幅度越来越小。这很符合我们直觉上的推断，因为标注语料丰富时只用标注语料就能得到不错的结果，如果未标注语料规模和标注语料处于一个数量级，并不能发挥充分利用未标注语料的作用。

3、对于辅助问题来说，低维特征的确不如高维特征有效。实验中去除了低维特征的性能提高了，说明可能低维特征损害了 ASO 的性能。但并不是所有结果都是这样，需要进一步研究。

4、在我们的实验中，参数 h 变化对结果影响不大，这与 Ando[42]中提到的 h

对结果影响不大可以设为固定值相符。但这也可能是由于我们实验用到的辅助问题比较少的原因，需要进一步进行研究。

第四章　工作总结及展望

4.1 课题总结

在当前自然语言处理研究中，人们发现语义理解对问题的解决能起到关键的作用。浅层语义分析作为语义研究的第一阶段，其重要性也凸显出来。语义角色标注是浅层语义分析的一种具体实现方式，以词性标注、句法分析等其它自然语言处理技术为基础，使用统计学习的技术进行分析。但由于语义角色标注是一个新兴的研究课题，亟待解决的问题还有很多，如：标注步骤的制定、标注单元的选取以及标注模型的使用等各个方面。本文的系统基于 Chinese PropBank 语料库，对语义角色标注的几个方面做了一些有益的研究，重点解决的问题如下：

1. 在语义角色标注任务上应用一种半监督的结构学习算法——ASO 算法，实验证明该方法可以有效的利用未标注语料中与标注问题相关的共同结构，提高系统的性能。

2. 对算法中关键的各种辅助问题进行了详尽的分析，并构建了比较有效的辅助问题。另外通过对不同的辅助问题进行分析，找出对辅助问题有效的特征。

3. 和以前对语法树上的节点进行分类不同，我们以组块为基本标注单元直接对语义角色进行分类。这样做简化了系统，并获得了较高的准确率。

4. 进行多组实验，分别从辅助问题选择，标注语料与未标注语料比例，辅助问题特征选择，参数设置等多个角度进行对比，深入研究结构学习算法的实现机制。

4.2 未来工作

对语义角色标注的研究还在继续进行中，我们的实验仍有很多方面需要调整完善。虽然本文的研究内容较好将结构学习算法应用于语义角色标注系统，但是通过对语义角色标注问题和结构学习较长时间的深入研究，我们认为对以下几个问题，还需要做进一步的研究：

1、辅助问题的构建。结构学习的基本思想就是利用非目标任务来定义一个功能结构，如果所定义的结构可以很好地概括潜在的分类器，那么它就可以提高目标任务的性能。但是如何衡量两个任务的相关程度，选取和目标任务在各个方面相关的更加有效的辅助问题，目前还没有比较好的方法，通常都是按照经验做法逐一实验，本实验中只用到了 Ando 构造辅助问题的第一种方法，即用其余所有特征预测某一特征。而其它方法构建的辅助问题的效果还不知道，需要今后做进一步的实验。而 Ando 构造辅助问题的方法也只是一家之言，如何量化目标问题和辅助问题的相关度，以便寻找更多更好的辅助问题，是今后工作的重点。

2、领域自适应。为了更好的应用语义角色标注，必须解决领域自适应问题，也就是说解决测试语料和训练语料属于不同的领域，其性能下降较多的问题。但是，我们不可能针对不同的领域，都标注大量的训练语料，所以领域的自适应成为一种必需。

对这一问题半监督的结构学习算法具有天然的优势，因为它本身就大量利用了未标注语料，这对提高分类器的泛化能力显然具有积极的作用。但如何具体将其应用于领域自适应并验证其效果，是我们今后工作的发展方向之一。

另一种思路是对于全监督的辅助问题，可以在一种语料上训练目标问题，而在另一种语料上训练辅助问题，比如在 PropBank 语料上训练目标问题，而在 FrameNet 或 NomBank 语料上训练辅助问题。这可能会提高系统的自适应性能。但目前还只是想法，并没有进行具体实验。

3、机器学习参数微调。实验表明，合适的参数配置在一定程度上能明显提高性能。ASO 算法中的 h 参数的作用，我们虽然进行了初步的验证，但实验进行的还不够全面，还不能很确定的证明其作用，也没有进行详细的理论分析。在今后的工作中可以进行更全面的实验以找出最优的参数设置。

4、进行特征选择。由于本文主要为了建立一个基于结构学习的语义角色标注系统，所以特征选择并不是我们研究的重点。我们主要根据经验，只进行了初步的特征选择。而特征选择对机器学习的影响至关重要，再进一步的研究中，这也将成为我们的一个重点。我们的语义角色标注系统的基本单元是组块，而不是目前常见的句法成分。但是与报告过的基于句法分析的方法的系统相比，我们的性能并没有表现出它的优越性。这主要可能就是由于我们使用的特征并不是非常适合基于组块的方法。今后的发展方向是提取更多的更适合基于组块的特征以提高系统的性能，另外可以利用其他的自然语言处理任务比如命名实体识别，语法组块分析以及更多的语言资源，可能会为我们的任务提供很大的帮助。

如本文开始所述，人们对于自然语言理解的最终目标是真正的深层次的语义分析，以期进行自动的知识获取，推理等等。因此在浅层语义分析的基础上，进

行深层的语义分析将成为未来研究的重点。为达此目标，我们必须建立性能更加卓越的浅层语义分析系统。

参考文献

[1] 张潮生. 格语法与自然语言处理[J]. 中文信息学报, 1988, (04).

[2] S.Narayanan, S.Harbabagiu. Question Answering Based on Semantic Structures. Proceedings of Coling 2004, 2004.

[3] J.Hajic, M.Cmejrek, B.Dorr, et al. Natural Language Generation in the Context of Machine Translation. Technical Report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2002.

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

[5] H.T.Dang, M.Palmer. The Role of Sematic Roles in Disambiguating Verb Senses. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics(ACL'05), University of Michigan, 2005,42-49.

[6] D.Beaugrande, Robert.Alain, W.Dressler. Introduction to text linguistics. London; New York: Longman, 1981.

[7] D.Gildea, D.Jurafsky. Automatic labeling of semantic roles[J]. Computational Linguistics,2002a, 28(3):245-288.

[8] D.Gildea, M.Palmer. The necessity of syntactic parsing for predicate argument recognition[C]. In Proc.of ACL-2002,2002b,239-246

[9] N.Xue, M.Palmer. Calibrating features for semantic role labeling[C]. In Proc.of EMNLP-2004, 2004.

[10] S.Pradhan, W.Ward, K.Hacioglu, et al. Shallow semantic parsing using Support Vector Machines[C]. In Proc.of HLT/NAACL-2004, 2004b.

[11] S.Pradhan, K.Hacioglu, V.Krugler, et al. Support vector learning for semantic argument classification[J]. Machine Learning Journal, 2005a, 60(3):11-39

[12] T.Liu, W.Che, S.Li, et al. Semantic role labeling system using maximum entropy classifier[C]. In Proc.of CoNLL-2005, 2005, 189-192.

[13] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

[14] H.Sun, D.Jurafsky. Shallow semantic parsing of Chinese[C]. In Proc.of NAACL-2004, 2004.

[15] N.Xue, M.Palmer. Automatic semantic role labeling for Chinese verbs[C]. In

Proc.of IJCAI-2005, 2005.

[16] 刘怀军，车万翔，刘挺．中文语义角色标注的特征工程[J]．中文信息学报，2007，21(1)：79-84.

[17] 于江德, 樊孝忠, 庞文博, 余正涛．基于条件随机场的语义角色标注[J]．东南大学学报, 2007 , 23 (3) ： 3612364.

[18] C.F.Baker, C.J.Fillmore, J.B.Lowe. The Berkeley FrameNet project[C]. In Proc.of COLING-ACL-1998, 1998, 86-90.

[19] M.Palmer, D.Gildea, P.Kingsbury. The Proposition Bank: An annotated corpus of semantic roles[J]. Computational Linguistics, 2005, 31(1): 71-106.

[20] A.Meyers, R.Reeves, C.Macleod, et al. Annotating noun argument structure for NomBank[C]. In Proc.of LREC-2004, 2004.

[21] K.Erk, A.Kowalski, S.Pado, et al. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. Proceedings of at ACL 2003.Sapporo, 2003.

[22] E.Hajicova. Prague Dependency Treebank:From Analytic to Tectogrammatical Annotation. Proceedings of the First Workshop on Text,Speech, Dialogue. 1998:45–50

[23] N.Xue, M.Palmer. Annotating the Propositions in the Penn Chinese Treebank. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan, 2003.

[24] N.Xue. Annotating the Predicate-argument Structure of Chinese Nominalizations. Proceedings of the LREC 2006.Genoa, Italy, 2006: 1382–1387

[25] 陈凤仪，蔡碧芳，陈克健，黄居仁．1999．中文句结构树资料库的构建

[26] L.You, K.Liu. Building Chinese Framenet Database. Conference on Natural Language Processing and Knowledge Engineering(IEEE NLP-KE). 2005:301–306.

[27] 陈丽江．汉语真实文本的语义角色标注[D]．南京师范大学，2007

[28] Z.-H.Zhou. Learning with unlabeled data and its application to image retrieval. Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), Guilin, China, LNAI 4099, 2006, 5-10.

[29] O.Chapelle, B.Schölkopf, A.Zien. Semi-Supervised Learning, Cambridge, MIT Press, 2006.

[30] X.Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Apr.2006.

[31] V.N.Vapnik. Statistical Learning Theory, New York: Wiley, 1998.

[32] T.Joachims. Transductive inference for text classification using support vector machines. Proceedings of the 16th International Conference on Machine Learning (ICML'99), Bled, Slovenia, 1999, 200-209.

[33] H.Seung, M.Opper, H.Sompolinsky. Query by committee. Proceedings of the 5th ACM Workshop on Computational Learning Theory (COLT'92), Pittsburgh, PA, 1992, 287-294.

[34] D.Lewis, W.Gale. A sequential algorithm for training text classifiers. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), Dublin, Ireland, 1994, 3-12.

[35] N.Abe, H.Mamitsuka. Query learning strategies using boosting and bagging. Proceedings of the 15th International Conference on Machine Learning (ICML'98), Madison, WI, 1998, 1-9.

[36] B.Shahshahani, D.Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[37] R.P.Lippmann. Pattern classification using neural networks. IEEE Communications, 1989, 27(11): 47-64

[38] D.J.Miller, H.S.Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. Advances in Neural Information Processing Systems, M.C. Mozer, M.I. Jordan, T. Petsche, eds., vol. 9, pp. 571-577, 1997.

[39] T.Zhang, F.J.Oles. A probability analysis on the value of unlabeled data for classification problems. Proceedings of the 17th International Conference on Machine Learning (ICML'00), San Francisco, CA, 2000, 1191-1198.

[40] Z.-H.Zhou, M.Li. Semi-supervised learning with co-training. Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, 2005, 908-913.

[41] Z.-H.Zhou, M.Li. Semi-supervised learning with co-training style algorithm. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(11).

[42] Rie Kubota Ando, Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6(Nov):1817–1853. 2005.

[43] Rie Kubota Ando, Tong Zhang. High performance semi-supervised learning for text chunking. In Proceedings of ACL-2005. 2005.

[44] Chang Liu, Hwee Tou Ng. Learning Predictive Structure Role Labeling of Nombank. In Proceedings of ACL-2007.

[45] Hacioglu K.Semantic role labeling using dependency trees[C].In Proc.of CoNLL-2004,2004.

[46] 于江德, 樊孝忠, 庞文博. 事件信息抽取中语义角色标注研究. 计算机科学. 2008, 35(3): 155 - 157.

[47] C.Liu, W.Ng. Learning predictive structures for semantic role labeling of NomBank[C], In Proc.of ACL-2007,2007.

[48] Rie Kubota Ando. Applying Alternating Structure Optimization to Word Sense Disambiguation. Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X).

[49] A.Meyers, R.Reeves, C.Macleod, et al. Annotating noun argument structure for NomBank[C]. In Proc.of LREC-2004, 2004.

[50] M.Surdeanu, J.Turmo. Semantic role labeling using complete syntactic analysis[C]. In Proc.of CoNLL-2005,2005,221-224.

[51] Saike HE , Taozheng ZHANG, Xiaojie WANG, Xue BAI and Yuan DONG, Incorporating Multi-task Learning in Conditional Random Fields for Chunking in Semantic Role Labeling, NLPKE-09.

[52] N.Xue, F.Chiou, M.Palmer. Building a large-scale annotated Chinese corpus[C]. In Proc.of COLING-2002, 2002.

[53] Dan.Bikel, On the Parameter Space of Generative Lexicalized Statistical Parsing Models[D]. Ph.D.thesis, the University of Pennsylvania.2004.

[54] 车万翔，刘挺，李生. 自动浅层语义分析[C]. 中国中文信息学会二十五周年学术会议，2006.

[55] 丁金涛. 基于特征向量和全局模型的语义角色标注[D]. 硕士论文，2008. 苏州大学。

[56] 丁金涛，王红玲，周国栋. 语义角色标注中有效的识别论元算法研究[J]. 计算机工程与应用，2008.

[57] S.Yih, K.Toutanova, Automatic Semantic Role Labeling, Human Language Technology Conference-North American chapter of the Association for Computational Linguistic annual meeting,June,2006,New York City.

[58] Rie Kubota Ando, Tong Zhang. Two-view Feature Generation Model for Semi-supervised Learning. Proceedings of the 24th International Conference on Machine Learning. 2007.

[59] 车万翔. 基于核方法的语义角色标注研究. 博士论文，2008. 哈尔滨工业大学

[60] 蔡洁. 中文子句语义角色标注系统实现研究. 硕士论文，2008. 北京邮电大学

[61] Abney Steven. Parsing by Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.). Principle-Based Parsing. Kluwer Academic Publishers. 1991, pp.257-278.

[62] 周强，詹卫东，任海波. 构建大规模的汉语语块库. 清华大学出版社：自然语言理解与机器翻译. 2001，pp102-107.

[63] 周志华. 半监督学习中的协同训练算法. 清华大学出版社：机器学习及其应用. 2007, 259-275.

致谢

岁月荏苒，时光如梭，毕业论文即将完成之时，也意味着充实而愉快的硕士研究生生活已将结束，而我也要离开这深爱的校园。在此，谨向给予我关心、指导与帮助的人们表示由衷的感谢。

首先感谢我的导师王小捷教授的悉心指导和积极鼓励。王老师踏实严谨的治学态度和平易近人的为人方式对我这三年的研究生涯以及未来的人生道路起了异常重要的影响。王老师不但指导我拓宽专业知识面，增强科研能力，而且还教会我们做人。他渊博的专业知识，严谨的治学态度和勤奋的工作作风，使学生受益终身。

同时，我必须感谢和我的同学们以及师兄师姐们，他们给了我很多的帮助，无论在研究工作中还是在学习生活中:于立平、段恋、王序文、刘丽娜、张春宇、杨宇、徐文智、张超辉等同学，张韬政师姐、袁彩霞师姐、张碧川师兄等。

感谢智能科学中心的钟义信老师、李蕾老师、周延泉老师、谭咏梅老师、李睿凡等老师认真细致的教授我们专业知识，拓展了我们的视野。

感谢我的母校北京邮电大学，在这里我渡过了近七年的宝贵时光。校园的一草一木都让人那么的留恋和不舍。我永远是北邮人，祝愿北邮的明天更加美好！

最后要感谢我的父母，谢谢你们一直以来对我的支持和疼爱，我一定会继续努力，好好报答你们的养育之恩，也祝愿你们身体健康，一切平安。

攻读硕士学位期间发表论文

[1] 白雪，王小捷，张韬政等. 基于ASO的汉语组块分析. 北京邮电大学出版社：中国人工智能进展（2009）. 2009. 609-615.

基于结构学习的语义角色标注

作者：白雪
学位授予单位：北京邮电大学