文章编号：1003-0077（2011）00-0000-00

基于句式与句模对应规则的语义角色标注*

何保荣，邱立坤，孙盼盼

（鲁东大学文学院，山东 烟台 264025）

摘要：大规模语义角色标注语料库的构建可以为计算机理解自然语言的语义提供有用的训练数据。本文主要研究服务于语义角色标注语料库构建的语义角色标注规则。在人工语义角色标注的基础上，分析句式和句模的对应关系，并总结出一套基于句式的语义角色标注规则，在测试集上达到 78.73%的正确率。基于上述规则，可以在构建语义角色标注语料库时完成自动标注的工作，标注人员在此基础上进行人工校对，从而有效地减少工作量。

关键词：句模；句式；语义角色标注

中图分类号：TP391　　**文献标识码：**A

Semantic Role Labeling Based on Correspondence Rules between Syntactic Pattern and Semantic Pattern of Sentences

HE Baorong, QIU Likun, SUN Panpan

(School of Chinese Language and Literature, Ludong University, Yantai, Shandong 264025, China)

Abstract: The construction of large-scale semantic corpus can provide useful training data for computer to understand the semantics of natural language. This paper focuses on the semantic rules for the construction of semantic corpus. On the basis of artificial semantic role tagging, the corresponding relation between syntactic patterns and semantic patterns of sentences is analyzed, and a set of semantic role labeling rules based on sentence patterns is summed up, which achieves a 84% precision on the test set. Based on the above rules, we can tag the corpus automatically, and our annotators can do manual proofreading on this basis, thus effectively reducing the workload.

Key words: semantic sentence pattern； semantic role labeling； labeling rules

1 引言

语义角色标注是一种浅层语义标注，其主要内容是识别谓词的论元，并为每个论元标注一个语义角色[1]。现有研究一般将语义角色标注视为分类问题或者序列标注问题，使用最大熵模型、条件随机场模型等予以实现。在训练数据较为充足的情况下，已取得较高精度。但现有自动标注方法主要使用句法信息和词汇信息，较少考虑谓词的格框架以及语义角色与句式之间的配合关系。

在之前的研究中，我们对把字句的句式及其句模的对应关系进行了分析，总结出了把字句的语义角色标注规则[2]。本文在之前工作的基础上，进一步对现代汉语句式及其句模的对应关系进行归纳，并总结出一套语义角色标注规则。本文工作包括以下三个方面：（1）归纳现代汉语句式及句模的对应关系；（2）以基于《人民日报》新闻语料的语义角色标注语料库为依据，基于人工标注的开发集，总结出若干条语义角色标注规则；（3）对语义角色标注规则进行有效性测试。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目（61572245）

作者简介：何保荣（1990——），女，硕士研究生，主要研究方向为中文信息处理；邱立坤（1979——），通讯作者，男，博士，副教授，主要研究方向为计算语言学；孙盼盼（1991——），女，硕士研究生，主要研究方向为中文信息处理。

2 现代汉语句式及其句模

2.1 现代汉语句式

研究句式与句模的对应关系首先要分析现代汉语的句式类型。根据句子内部结构的不同，首先可以把句子分为单句和复句。单句可以表示一个或者多个命题；复句由两个或多个小句构成。单句一般只包含一种句子结构类型，而复句则可能包含多个句子结构类型。为了方便句式的划分和句模的描写，本文以单句为基本单位，将复句拆分为多个单句。

在黄伯荣、廖旭东《现代汉语》中对句式划分的基础上[3]，为了便于下文对句式及句模关系进行分析，本文在单句句式类型中增加了“共享并列句①”和“轻动词句②”两种句式。根据句子内部结构的不同，单句可以分为主谓句和非主谓句。非主谓句主要是由定中结构或状中结构的短语（好大的苹果、真冷）或者感叹词（啊呀）、拟声词（砰砰）构成。除了定中结构的非主谓句，其他类型的非主谓句一般不构成命题结构，因此本文不作讨论。本文主谓句包括四类：动词谓语句、名词谓语句、形容词谓语句和主谓谓语句。动词谓语句又包括一般主谓句、“把”字句、“被”字句、主谓句、连谓句、兼语句、双宾句、比较句、并列共享句、轻动词句等句式（图 1）。

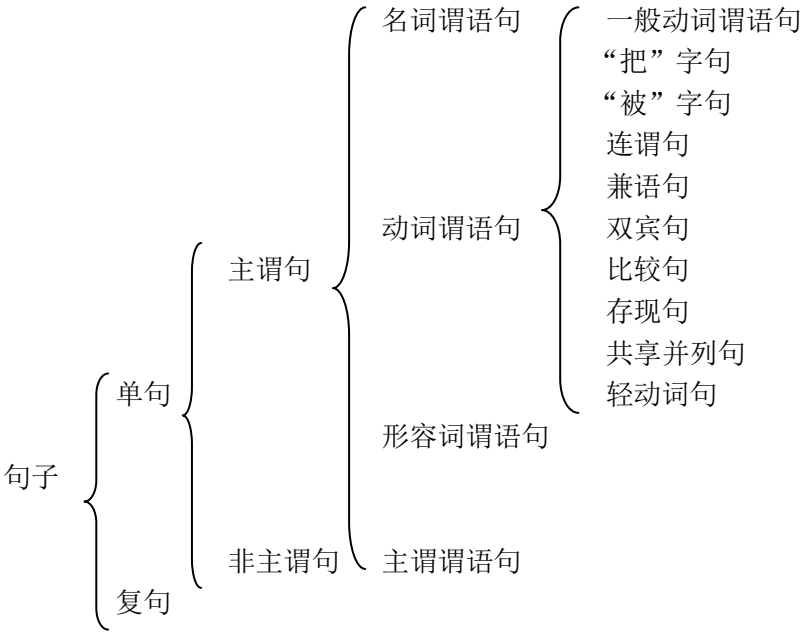

图 1 现代汉语句式

2.2 句式与句模的对应关系

本文语义角色标注是在句法角色标注的基础之上进行的[4]。句法树库中，“VV”表示连

① 共享并列句：句子的谓语核心由两个具有并列关系的动词充当，且这两个动词都可以支配句子的宾语，比如“我们要建立、健全养老机制”，“建立”和“健全”共同充当句子的谓语核心，且共享句子的宾语“养老机制”。

② 轻动词句：句子的谓语动词为轻动词，比如“我们要对学校安全设施进行检查”，“进行”为轻动词，充当句子的谓语核心，后接表示实际意义的动词宾语“检查”。

谓动词，“IC”表示小句的谓语中心语，“HED”表示整个句子的谓语中心语。相应地，在语义角色标注语料库中，“VV/IC/HED”表示命题的动词核心，主语（SBV）、宾语（VOB）、间接宾语（IOB）、状语（ADV）（副词性状语、动词性状语除外）等表示与动词核心相关的论元成分。另外，当宾语（VOB）充当主语（SBV）的父结点的时候，VOB 一般也是动词，只不过 VOB 中又包含了一个命题，比如“我们打算明天去北京”，“明天去北京”作为“打算”的“VOB”，其内部还存在一个命题“去北京”，“北京”又充当“去”的“VOB”。这种情况规律性很强，也作为本文句模研究的对象。

下面分别对动词谓语句（“把”字句、“被”字句、兼语句、连谓句、双宾句、“比”字句、并列共享句、轻动词句、一般主谓句等）、名词谓语句、形容词谓语句和主谓谓语句的句模进行分析和描写。

（1）“把”字句句模

“把”字句指的是在用介词“把”引出句子的“受事”并对“受事”加以处置的句式。表示处置义的“把”字句是最为典型的“把”字句，“把”字句的主语一般充当主事，介词“把”介引的宾语一般充当客事。此外，口语中还有表示致使义的“把”字句[5]，比如“你怎么把罪犯跑了”，但是由于这类“把”字句数量非常少，不具有典型性，所以本文暂时不作讨论。“把”字句句法结构及其句模的对应关系为：

SBV+把+POB+IC/HED→ 主事+把+客事+IC/HED[③]	他把花瓶砸了。
SBV+把+POB+IC/HED（当作类）+VOB→主事+把+客事+IC/HED+结果	他把你当作好朋友。
SBV+把+POB+IC/HED+（CMP+DEI）→主事+把+客事+IC/HED；把+主事+CMP	他把花瓶砸碎了。
SBV+把+POB+IC/HED+CMP+VOB→主事+把+客事+IC/HED；把+主事+CMP+客事	他把房子改造成了仓库。

（2）“被”字句句模

本文“被”字句是指用介词“被（让、给、叫）”引出句子的施事或者单独使用介词“被”的句子。介词“被”一般出现在谓语动词的前面，“被”字句主语一般充当的是客事或对象和与事，介词“被”的介引宾语一般充当主事。“被”字句的句法结构及其句模对应关系为：

SBV+被+POB+IC/HED→客事+被+主事+IC/HED	王某被人举报了。
SBV+被+POB+IC/HED（给予类）+VOB→与事+被+主事+IC/HED+客事	他被学校授予学习标兵的称号。
SBV+被+POB+IC/HED+VOB（v）→对象+被+主事+IC/HED+客事；主事+VOB	他被检察院责令停止运营。
SBV+被+POB+IC/HED+CMP→客事+被+主事+IC/HED；主事+CMP	他被王某打伤了。
SBV+被+POB+IC/HED+（CMP+DEI）→客事+被+主事+IC/HED；主事+DEI	街道被人打扫得干干净净。
SBV+被+POB+IC/HED+CMP+VOB→客事+被+主事+IC/HED；主事+CMP+客事	他被总统任命为内阁成员。

（3）双宾句句模

双宾句指的是有两个宾语的句子，前一个宾语一般指人，称为间接宾语；后一宾语一般指物，称为直接宾语。本文中，双宾句中的直接宾语仍用“VOB”表示，间接宾语则用“IOB”来表示。双宾句的句法结构及其句模对应关系描写如下：

SBV+IC/HED +IOB+VOB→主事+IC/HED+与事+客事	他送了我两本书。
SBV+IC/HED +VOB+IOB→主事+IC/HED+客事+与事	他送了两本书给我。

（4）兼语句句模

黄伯荣、廖旭东将兼语句定义为：由兼语短语充当谓语或独立成句的句子叫做兼语句。[6]比如“经理派他去了上海”。语言学中，兼语句可以表示为：N1（主语）+V1（谓语中心语）+N2（兼语）+（V2（第二个谓词）+N3）。在本文句法标注体系中，兼语句可形式化为：SBV+IC/HED+VOB+（ACT+VOB）。根据兼语句特有的语法标签“ACT”，可将兼语句单独抽取出来。

[③]句模中的句法成分标记，在句法树库中一般由“IC/HED”“CMP”“VOB”“ATT”等（动词或者形容词）充当，这些动词或形容词在句子或者短语结构中都充当谓语核心。

兼语句中的两个动词都能构成一个命题，因此，其句模是双动核结构。一般情况下，兼语句的主语充当的是谓语动词的主事，宾语充当的是客事；同时，宾语还充当第二个动词“ACT”的主事，“ACT”和谓语动词之间的关系则一般是“目的”或“结果”；“ACT”后的宾语一般充当“ACT”的客事。兼语句的句法结构及其句模的对应关系为：

SBV+IC/HED +VOB+ACT➞主事+ IC/HED+客事；主事+ACT；IC/HED+目的/结果　经理让他立刻行动。

（5）“比”字句句模

“比”字句指的是用“比”字介词短语充当状语的句子，又称“差比句”。比较句一般包括比“比较主体、比较项目、比较对象和比较结果”四个部分。语言学中一般将“比”字句标记为：X（比较主体）+比（比较对象）+Y（谓语中心语）+W（比较结果）。比如“他比弟弟高一头”。在本文中，“比”字句句法结构及其句模对应关系表示如下：

SBV+比+POB+IC/HED➞比较主体+比+比较对象+IC/HED　他比我高。

SBV+比+POB+IC/HED +VOB➞比较主体+比+比较对象+IC/HED+比较结果　他比我高十公分。

TPC+SBV+比+POB+IC/HED+VOB➞比较主体+比较项目+比+比较对象+IC/HED　他身高比我高十公分。

（6）连谓句句模

连谓句指的是由连谓短语来充当句子的谓语或者由连谓短语单独成句的句子，连谓句中的谓词大都能体现出时间上的先后。

通过语料标注，我们发现绝大多数连谓句的谓语都是双动词结构，即一般是由两个时间上具有先后顺序的动词构成的。在本文句法树库中，连谓可以用“VV”这一标签来表示，比如“小王拿了电脑就走了”，“拿”和“走”句法上标为“VV”。连谓句可以构成两个（或多个）命题，故其句模也是双（多）动核结构。连谓句句法结构及其句模的对应关系为：

SBV+VV+IC/HED+VOB➞主事+VV；主事+IC/HED+客事　他上街买菜。

（7）共享并列句

有种句式的谓语结构是由两个具有并列关系的动词构成的，这两个动词共享后边的宾语，本文称之为“共享并列句”。这种句式有两个谓语动词，可构成两个命题，句模为双动核结构。比如“有关部门要制定并落实每项政策”。“制定”和“落实”共享宾语“每项政策”。共享并列句的句法结构及其句模的对应关系为：

SBV+COS+IC/HED+VOB➞主事+COS+客事；主事+IC/HED+客事　我们要建立、健全养老机制。

（8）轻动词句

轻动词是一种比较特殊的动词，其意义较虚，且后面一般要与表示实在意义的动词组合构成动宾结构，动宾结构的宾语是由表示实在意义的动词充当的。较为常用的轻动词有“进行、予以、作”等。比如“媒体对该事件进行了大肆宣传”。其中，“进行”只表示主体实施了某项行为或动作，但具体行为或动作则由动词宾语“宣传”来承担。该句所表达的意思其实是“媒体大肆宣传了该事件”。轻动词句的句法结构及其句模的对应关系为：

SBV+ADV+IC/HED（轻动词）+VOB➞主事+对象+IC/HED+客事；主事+客事+VOB　我们对此事进行了调查。

（9）名词谓语句

名词谓语句指的是以名词或名词性短语充当谓语的句子，比如“明天端午节”。名词谓语句实际上是动词谓语句的一种变体或者说一种省略形式。在进行语义角色标注时，本文仍按照省是句来标注。名词谓语句的句法结构和句模描写情况如下：

SBV+IC/HED（n）➞当事+IC/HED　经济总量1亿元。

（10）形容词谓语句

形容词谓语句是指由形容词或者形容词性短语充当谓语的句子，形容词或形容词性短语用来表示主语的性质或者状态。形容词谓语句中，主语充当谓语的当事，比如“这支花美丽极了”。形容词谓语句的句法结构及其句模的对应关系为：

SBV+IC/HED（a）➞当事+IC/HED　景色很美丽。

（11）主谓谓语句

由主谓短语充当句子谓语的句式称为主谓谓语句。本文将主谓谓语句的第一个“主语”（大主语）标注为话题（TPC），主谓短语中的主语标为主语。主谓谓语句的句法结构和句模的对应关系描写如下：

TPC+SBV+IC/HED+VOB➞接事+主事+IC/HED+客事　　她双手举着火炬。

（12）一般主谓句

除了上述几种特殊句式外，动词谓语句还有大量句法形式上无标记的句式，我们称之为一般主谓句，即简单的主谓宾句。主谓宾句的主语一般充当主事，宾语一般充当客事，比如“他得到了一笔巨款”。一般主谓宾句的句法结构及其句模的对应情况描写如下：

SBV+IC/HED+VOB➞主事+IC/HED+客事　　我们完成了这项任务。

2.3 关系结构

本文句模研究除了各种句式之外，还涵盖包含关系从句的关系结构。“关系从句”是一种语言中普遍存在的、特殊的并带有一定标记的重要结构。确切地说，所谓的关系从句其实是一种短语结构，而非真正意义上的句子，为了便于理解，本文引用陈宗利的“关系结构”这一说法：“关系结构”是指包含关系从句的名词性成分，由关系从句和中心语两部分构成，关系从句和中心语也可以带数量词和限定词等修饰成分。[7]

不同于其他 SO 语序的语言，汉语关系结构的语序比较特殊，关系从句处于核心名词之前，比如“他去过的城市”，“他去过”在“城市”之前。汉语最普遍的关系结构标记类型就是定语标记“的”[8]，比如“他写过的书”和“他的书”共用一个“的”，且二者都是定中短语结构，但两者定语部分的性质不同。前者是关系从句作定语，后者是的字短语作定语，在句法结构上二者比较容易区分：关系从句是“NP+V+的+NP”，后者是“NP+的+NP”。

关系结构中有两个“NP”，我们用“NP_1”和“NP_2”来表示，即“NP_1+V+的+NP_2”。一般情况下，NP_1 充当 V 的主事，NP_2 充当 V 的客事、与事或者外围语义角色，或者 NP_2 与 V 不存在语义关系。“NP_1+V+的+NP_2”对应到句法树库中的形式化表示为：SBV（NP_1）+ATT（V）+的+ DE（NP_2）。下面分是关系结构的句模类型：

（1）主事+V+的+客事

施事+V+的+受事：他吃的苹果

施事+V+的+系事：他买的书

施事+V+的+内容：他说的话

施事+V+的+对象：他批评的人

（2）主事+V+的+外围语义角色

施事+V+的+路径：他走过的路

施事+V+的+材料：他画画的颜料

（3）主事+V

施事+V：他出门的时间/他说谎的目的/飞机降落的地点

上述四种关系结构有着同样的句法结构，但句模却不同。这主要是由于关系结构中动词的价不同。（1）中的动词都是二价动词，比如“吃”“买”“说”“批评”等；（2）（3）则是一价动词，比如“走”“画画”“出门”“说谎”“降落”等。在语义角色自动标注过程中，根据动词给定的格框架，采取动词左侧句法成分优先标注的原则将格框架中的语义角色赋予关系结构中的 NP_1 和 NP_2。综合上述三种句模，关系结构的句模可归纳为：

主事+ATT+的+客事/外围语义角色

3 语义角色标注规则

根据上述句式和句模的对应关系，本文总结出一套语义角色标注规则。该规则旨在对大多数句子进行语义角色自动标注，以降低人工标注的工作量、提高语义角色自动标注的准确率。

“在实际的语义角色标注过程中，规则的使用具有先后顺序。局部规则优于全局规则。”[9]现代汉语句子中，除了一般主谓宾句没有特殊的标记之外，其他句子都是有标记的句子。比如，“把”字句、“被”字句、“比”字句可以通过介词“把”“被”和“比”来辨别；双宾句可以通过间接宾语的标记“IOB”辨别；兼语句通过“ACT”辨别；连谓句通过“VV”辨别；共享并列句通过“COS”辨别；主谓谓语句通过“TPC”区分；轻动词句可以通过直接限定有限的轻动词与其他句式区分出来；名词谓语句和形容词谓语句的谓语中心语的词性分别为名词和形容词，也可以区分出来，其他句式即主谓宾句式。因此，本文语义角色标注规则的运行顺序是：首先处理“把”字句、“被”字句、连谓句等特殊句式；然后处理一般主谓宾句。至于关系结构，它有可能出现在所有句式的句子当中，所以，在每个句子中都要检索是否存在关系结构。根据局部规则优于全局规则的原则以及关系结构的特点，本文语义角色标注规则归纳如下：

规则 1：判断当前句子中有无标记词介词“把”，如果有，则进入规则 2；如果没有检索到介词“把”，则进入规则 7；

规则 2：若句式为“SBV+把+POB+IC/HED+（CMP+DEI）”，则句模为“主事+把+客事+IC/HED；把+主事+CMP”，并进行规则 31；如果不是，则进入规则 3；

规则 3：若句式为“SBV+把+POB+IC/HED+CMP+VOB”，则句模为“主事+把+客事+IC/HED；把+主事+CMP+客事”，并进行规则 31；如果不是，则进入规则 4；

规则 4：若句式为“SBV+把+POB+IC/HED+CMP”，则句模为“主事+把+客事+IC/HED；把+主事+CMP”，并进行规则 31；如果不是，则进入规则 5；

规则 5：若句式为“SBV+把+POB+IC/HED+VOB”，且 IC/HED 为“当作、称作、称为、叫做”等三价动词，则句模为“主事+把+客事+ IC/HED+结果”，并进行规则 31；如果不是，则进入规则 6；

规则 6：若“把”字句句式为“SBV+把+POB+IC/HED”，则句模为“主事+把+客事+IC/HED”，并进行规则 31；如果不是，则进入规则 7；

规则 7：判断当前句子中有无标记词介词“被”，如果有，则进入规则 8；如果没有检索到介词“被”，则进入规则 14；

规则 8：若“被”字句句式为“SBV+被+POB+IC/HED”，其句模为“客事+被+主事+IC/HED”，并进行规则 31；如果不是，则进入规则 9；

规则 9：若句式为“SBV+被+POB+IC/HED+VOB”，且 IC/HED 为“给予、授予、赋予、赠予”等动词，其句模为“与事+被+主事+ IC/HED+客事”，并进行规则 31；如果不是，则进入规则 10；

规则 10：句式为“SBV+被+POB+IC/HED+VOB”，且 VOB 的词性为“v”（动词），则句模为“对象+被+主事+ IC/HED+客事；主事+VOB+客事”，并进行规则 31；如果不是，则进入规则 11；

规则 11：若句式是“SBV+被+POB+IC/HED+CMP”，句模为“客事+被+主事+ IC/HED；主事+CMP”，并进行规则 31；如果不是，则进入规则 12；

规则 12：若句式为“SBV+被+POB+IC/HED+（CMP+DEI）”，句模为“客事+被+主事+IC/HED；主事+DEI”，并进行规则 31；如果不是，则进入规则 13；

规则 13：若句式为“SBV+被+POB+IC/HED+CMP+VOB”，则句模为“客事+被+主事+

IC/HED；主事+CMP+客事”，并进行规则 31；如果不是，则进入规则 14；

规则:14：判断当前句子中有无介词“比”，若有，则进入规则 15；如果没有，则进入规则 18；

规则 15：如果“比”字句句式为“SBV+比+POB+IC/HED”，其句模为“比较主体+比+比较对象+ IC/HED”，并进行规则 31；如果不是，则进入规则 16；

规则 16：如果句式为“SBV+比+POB+IC/HED +VOB”，其句模为“比较主体+比+比较对象+ IC/HED+比较结果”，并进行规则 31；如果不是，则进入规则 17；

规则 17：如果句式为“TPC+SBV+比+POB+IC/HED+VOB”，其句模为“比较主体+比较项目+比+比较对象+ IC/HED”，并进行规则 31；如果不是，则进入规则 18；

规则 18：判断当前句子的句法成分中有无“IOB”，如果有，则进入规则 19；如果没有则进入规则 21；

规则 19：若双宾句的句式为“SBV+IC/HED +IOB+VOB”，其句模为“主事+ IC/HED+与事+客事”，并进行规则 31；如果不是，则进入规则 20；

规则 20：若句式为“SBV+IC/HED +VOB+IOB”，其句模为“主事+ IC/HED+客事+与事”，并进行规则 31；如果不是，则进入规则 21；

规则 21：判断当前句子的句法成分中有无“ACT”，若有，则判断当前句的句式是否为“SBV+IC/HED +VOB+ACT+（VOB）”，如果是，则其句模为“主事+ IC/HED+客事；主事+ACT+（客事）；IC/HED+目的/结果”，并进行规则 31；若不是，则进入规则 22；

规则 22：判断当前句子的句法成分中有无“VV”，若有，则判断连谓句的句式是否为为“SBV+VV+IC/HED +VOB”，如果是，则其句模为“主事+VV；主事+IC/HED+客事”，并进行规则 31；若不是，则进入规则 23；

规则 23：判断当前句子的句法成分中有无“COS”，若有，则判断并列共享句的句式是否为“SBV+COS+IC/HED +VOB”，如果是，其句模为“主事+COS+客事；主事+ IC/HED+客事”，并进行规则 31；如果不是，则进入规则 24；

规则 24：判断当前句子的句法成分中有无“TPC”，若有，则判断主谓谓语句的句式是否为“TPC+SBV+IC/HED+VOB”，若是，则其句模为“接事+主事+IC/HED+客体”，并进行规则 31；若不是，则进入规则 25；

规则 25：判断当前句子的“IC/HED”（谓语核心动词）是否为“进行、给予、作”等动词，且宾语词性为“v”（动词），如果是，则判断轻动词句的句式是否为“SBV+ADV（p+POB）+IC/HED+VOB”，其句模为“主事+对象+ IC/HED+客事；主事+客事+VOB”，并进行规则 31；如果不是，则进入规则 26；

规则 26：判断当前句子的“IC/HED”（谓语核心动词）的词性是否为“n”（名词），如果是，则判断名词谓语句的句式是否为“SBV+ IC/HED”，其句模为“当事+ IC/HED”，并进行规则 31；若不是，则进入规则 27；

规则 27：判断当前句子的“IC/HED”（谓语核心动词）的词性是否为“a”（形容词），如果是，则进入规则 28；若不是，则进入规则 30；

规则 28：如果形容词谓语句的句式为“SBV+ IC/HED”，其句模为“当事+ IC/HED”，并进行规则 31；如果不是，则进入规则 29；

规则 29：如果形容词谓语句的句式为“SBV+ ADV（p+POB）+IC/HED”，其句模为“当事+对象/客事+ IC/HED”，并进行规则 31；如果不是，则进入规则 30；

规则 30：判断当前句子的句式是否为“SBV+IC/HED+VOB”，若是，则当前句子的句模为“主事+IC/HED+客事”并进行规则 31；

规则 31：判断当前句子中是否存在句法结构为“SBV+ATT+的+DE”，且“ATT”的词性为“v”（动词）的结构，如果存在，则该结构的句模为“主事+ATT+的+客事/外围语义角

色”；如果不存在该结构，则不标注。

4 标注规则的有效性测试

4.1 实验设置

本文所使用的语料库为 2000 年 1 月份《人民日报》，该语料库的句法标注体系及构建过程可参见邱立坤等（2010）[10]。在此基础上，我们对该语料库的前 30000 句进行了语义角色标注。在使用基于规则的方法进行自动标注时，我们使用前 20000 句作为开发集，后 10000 句作为测试集。

为了与基于统计的标注方法进行比较，我们还使用 Mate-tools 的语义角色标注模块[11]进行了对比实验，同样选择前 2000 句作为训练集，后 10000 句作为测试集。该实验全部使用默认参数，不需要调试参数，因此未设置开发集。

在上述实验中，我们用带标签正确率（labeled precision，LP）、带标签召回率（labeled recall，LR）和不带标签正确率（unlabeled precision，UP）、不带标签召回率（unlabeled recall，UR）来评价标注质量。UP 和 UR 仅考虑弧的正确与否，即两个词之间是否存在语义依存关系；LP 和 LR 则在考虑弧的基础上，还要考虑语义角色标签的正确与否。

4.2 实验结果及分析

表 1 自动标注结果

	UP(%)	UR(%)	LP(%)	LR(%)
基于规则的方法(本文)	93.10	52.48	78.73	44.38
基于统计的方法（mate-tools）	87.02	79.69	74.55	68.27

实验结果如表 1 所示。整个测试集中，人工标注的弧和标签个数为 78917 个，基于规则自动标注的弧个数是 44484 个，正确个数为 41415，召回率为 52.48%，正确率为 93%；基于规则标注的标签正确个数为 35024，正确率为 78.73%，召回率为 44.38%。假定弧正确的情况下，标签的正确率为 84.57%。

与之相比，基于统计的方法召回率较高，但正确率较低。虽然基于规则的自动标注方法召回率不是很高，但其正确率却达到了较高的水平，运用该规则可以降低大约一半的标注工作量，因此本文语义角色标注规则在人工构建语料库时是可行的。

根据我们的初步分析，基于统计的方法的标注结果差异较大，人工校对时需要修改的地方较多；基于规则的方法的结果一致性比较高，人工校对时需要修改的地方较少，而且比较一致，但是需要添加的弧更多一些。

错误分析表明，目前的规则还有待进一步细化，比如可根据动词的类是一价动词、二价动词或三价动词总结出更细致的规则。

5 结语

本文在总结句式与句模对应关系的基础上，归纳出一套语义角色标注规则，并对该规则的有效性进行了测验。本文总结的句式主要是主谓句，包括动词谓语句、名词谓语句和形容词谓语句。其中，动词谓语句包括一般动词谓语句、“把”字句、”被”字句、兼语句、连谓句、双宾句、“比”字句等句式。根据句法、词类、词性上有无标记，将不同句式的句模进行归纳，总结出一套语义角色标注规则（共 31 条）。最后，在测试集上进行了验证。测试结果证明该规则具有较好的正确率，可以在构建语料库的人工标注过程中发挥该方法正确率高的优点，降低人工标注的工作量。

与统计方法相比，本文基于规则的方法优点在于从整体上考虑句子的结构，但由于规则考虑的因素还不够细致，整体精度与统计方法相比并无优势。在今后的工作中，我们计划进一步探讨规则方法和统计方法的融合。

参考文献

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

[2]何保荣，邱立坤，徐德宽. 基于规则的把字句语义角色标注[J]. 中文信息学报，2017，31(1)：84-93.

[3]黄伯荣，廖旭东. 现代汉语（增订四版）[M]. 北京：高等教育出版社，2007：102.

[4]Likun Q, Yue Z, Meishan Z. Dependency Tree Representations of Predicate-Argument Structures[C]//In Proceedings of AAAI-16, 2016: 2645-2651.

[5]范晓. 三个平面的语法观[M]. 北京：北京语言学院出版社，1996，201-209.

[6]黄伯荣，廖旭东. 现代汉语（增订四版）[M]. 北京：高等教育出版社，2007：90.

[7]陈宗利.汉语关系从句的位置与关系结构的特点[J].语言科学, 2009(2)：155-164.

[8]刘丹青.汉语关系从句标记类型初探[J].中国语文，2005(1)：3-15.

[9]詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[J]. 北京：清华大学出版社，2000：37.

[10]邱立坤，史林林，王厚峰. 多领域中文依存树库构建与影响统计句法分析因素之分析[J]. 中文信息学报，2015，29(5)：71-77。

[11]Björkelund A, Hafdell L, and Nugues P. Multilingual semantic role labeling[C]// In CONLL 2009, 2009: 43–48.

何保荣（1990——）女，硕士研究生，主要研究方向为中文信息处理，地址：山东省烟台市芝罘区红旗中路 186 号鲁东大学文学院，264025，17853519358，hebaorong123@sina.com

邱立坤（1979——）通讯作者，男，博士，副教授，主要研究方向为计算语言学，地址：山东省烟台市芝罘区红旗中路 186 号鲁东大学文学院，264025，13583566312，

qiulikun@pku.edu.cn

孙盼盼（1991——），女，硕士研究生，主要研究方向为中文信息处理，地址：山东省烟台市芝罘区红旗中路 186 号鲁东大学文学院，264025，18253560281，1030158547@qq.com