**文章编号**：1003-0077(2013)01-0072-09

# BFS-CTC 汉语句义结构标注语料库

刘盈盈，罗森林，冯 扬，韩 磊，陈 功，王 倩

（北京理工大学 信息与电子学院 信息安全与对抗技术实验室，北京 100081）

**摘 要**：句义结构分析是汉语语义分析中不可逾越的重要环节，为了满足汉语句义结构分析的需要，基于现代汉语语义学理论构建了一种层次化的汉语句义结构模型，定义了标注规范和标记形式，建设了一个汉语句义结构标注语料库 BFS-CTC(Beijing Forest Studio-Chinese Tagged Corpus)。标注内容方面，基于句义结构模型的定义标注了句义结构句型层、描述层、对象层和细节层中所包含的各个要素及其组合关系，包括句义类型、谓词及其时态、语义格类型等信息，并且提供了词法和短语结构句法信息，便于词法、句法、句义的对照分析研究；语料库组织结构方面，该语料库包括四个部分，即原始句子库、词法标注库、句法标注库和句义结构标注库，可根据研究的需要，在词法、句法、句义结构标注的基础上进行深加工，在核心标注库的基础上添加更多具有针对性的扩展标注库，利用句子的唯一 ID 号进行识别和使用；语料来源和规模方面，语料全部来自新闻语料，经过人工收集、整理，合理覆盖了主谓句、非主谓句、把字句等六种主要句式类型，规模已达到 10 000 句。同其他语义标注库相比，BFS-CTC 基于现代汉语语义学，提供了多层次的句义结构标注信息，兼容进行了词法和语法标注，各类标注既可以单独使用也可综合使用进行横向分析，可用于自然语言处理多方面的研究，进一步推动汉语语义分析的研究和发展。

**关键词**：自然语言处理；语义标注；句义结构；语料库

**中图分类号**：TP391 **文献标识码**：A

## BFS-CTC: A Chinese Corpus of Sentential Semantic Structure

LIU Yingying, LUO Senlin, FENG Yang, HAN Lei, CHEN Gong, WANG Qian

(Lab of Information Security & Countermeasures Technology, School of Information & Electronics, Beijing Institute of Technology, Beijing 100081, China)

**Abstract**: Sentential semantic structure analysis is an important issue in Chinese semantic analysis. Based on the Modern Chinese Semantics, this paper establishes a hierarchical Chinese sentential semantic structure model, defines the standard and the tagset, and thus constructs a Chinese corpus of sentential senmantic structure: BFS-CTC (Beijing Forest Studio - Chinese Tagged Corpus). All sentences in this corpus are tagged on the lexical, the syntactic and the whole sentential semantic structure levels, and it is easy to analyze the relation between syntax and semantics. The core of BFS-CTC is consists of four banks: the original sentence bank (OSB), the lexical tagged bank (LTB), the syntax tagged bank (STB) and the semantic structure tagged bank (SSTB). The more than 10,000 sentences in current version come from news texts, covering six major sentence types in Chinese.

**Key words**: natural language processing; semantic analysis; sentential semantic structure; corpus

**收稿日期**：2011-03-29 **定稿日期**：2011-12-27

**基金项目**：国家 242 项目(2005C48)；北京理工大学科技创新计划重大项目培育专项计划资助(2011CX01015)

**作者简介**：刘盈盈(1987—)，女，硕士研究生，主要研究方向为中文信息处理；罗森林(1968—)，男，教授，博士生导师，主要研究方向为中文信息处理、信息安全、数据挖掘、媒体计算等；冯扬(1982—)，男，博士，高级工程师，主要研究方向为中文信息处理。

## 1　引言

语料库是为某一个或多个应用而专门收集的、有一定结构的、有代表性的、可以被计算机程序检索的、具有一定规模的语料的合集[1]。语料库按加工程度不同可分为生语料(未加标注的语料库)、词法标注语料库、句法标注语料库及语义标注语料库四类。20 世纪 80 年代以来，自然语言处理研究的重点逐渐转移到语义处理方面，而句义分析又是语义研究中的一个重要问题，因此，句义标注语料资源作为句义分析的一项不可或缺的基础资源正在受到越来越多的关注。

在汉语语料的语义标注深加工方面，主要有宾州大学汉语浅层语义标注库(Chinese Proposition Bank，CPB)[2]、山西大学汉语框架语义知识库(Chinese FrameNet，CFN)[3-4]和清华大学句法语义标注库(Syntactically and Semantically Annotated Corpus，SSAC)[5]等。其中，CPB 语料库是目前汉语句义分析研究中主要使用的语料资源。CPB 是建立在句法标注语料库 CTB(Chinese Tree Bank)基础之上，标注出部分句法成分相对于给定动词所具备的语义角色；CFN 是一个以 Fillmore 的框架语义学为理论基础、以加州大学伯克利分校的 FrameNet 为参照、以汉语真实语料为依据的汉语框架语义知识库；SSAC 以清华大学的句法树库 TCT 和句法语义链接库 SSL 数据为基础，针对目标动词，在句法依存信息基础上形成完整的句法语义信息标注句子。

根据语义学的相关理论，研究人员从不同角度构建了语义标注语料库，这些语料库在汉语的自然语言处理研究中起到了重要的作用，但是目前还缺乏深入的句义层次、以现代汉语语义学的句义结构理论为基础的句义结构标注语料库。因此，从 2009 年 10 月起，开始构建 BFS-CTC 汉语句义结构标注语料库。为汉语句义结构分析提供所需的句义结构标注信息，包括句义类型、句义成分以及各成分之间的组合关系等。

本文第 2 节讨论汉语句义结构的定义及其模型的基本形式和扩展形式；第 3 节介绍句义结构标注的内容及标记形式；第 4 节介绍语料来源、语料的具体分布情况以及语料库的组织结构；第 5 节对比分析句义标注资源 FrameNet、PropBank 与 BFS-CTC 句义标注的异同；最后在第 6 节总结全文并简介基于该语料库的汉语句义结构分析研究概况。

## 2　汉语句义结构定义及模型

句义结构标注语料库的构建，是基于现代汉语语义学的句义结构理论，该理论认为句义可以用具有一定逻辑的结构来表达，并从逻辑结构上将句义划分成话题和述题，而话题和述题又是由谓词和项组合而成[6]，如图 1 所示。

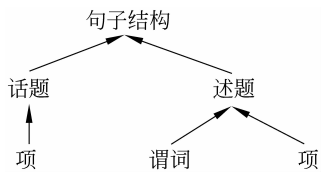


图 1　句义结构的组合关系

### 2.1　句义结构模型的基本形式

根据现代汉语语义学理论，BFS-CTC 构建了层次化的句义结构模型[7]，该模型具有可扩展、可计算的特点，图 2 所示为句义结构模型的基本形式，包含的要素有：句义的类型、句义中的话题和述题、构成句义的各个成分、谓词时态信息、成分之间的组合关系等。

句型层从句子本身出发，反映句子在句义上的类型。句义类型是对句义结构的类别描述，反映了句义结构的复杂程度、层次数目等信息。从汉语语义学的角度，汉语的句义类型分为简单句义、复杂句义、复合句义和多重句义[6]。其中简单句义只表达一个命题，结构也最简单；复杂句义、复合句义则包含一个或多个简单句充当成分句义或分句义(统称为子句义)；多重句义则由复杂句义或复合句义充当子句义，而这些子句义本身又可以找到简单句义类型的子句义。因此，定义简单句义的句义结构形式为最小完整单元，在非简单句的结构中，最小完整单元的作用与一个句义成分的作用相似。

描述层则是反映句子的描述对象(话题)以及对该对象的描述(述题)，是对句义结构的第一次划分。

对象层是针对描述层中的话题和述题的进一步划分，每一个对象对应于一个谓词或项，而充当谓词或项的词、短语、子句等用于表示这些对象。在对象层中，谓词和基本项直接与话题和述题相关联，构成句义结构的基本框架，而一般项则是用来描述和限定谓词和基本项的(如限定时间、地点、范围等，当然

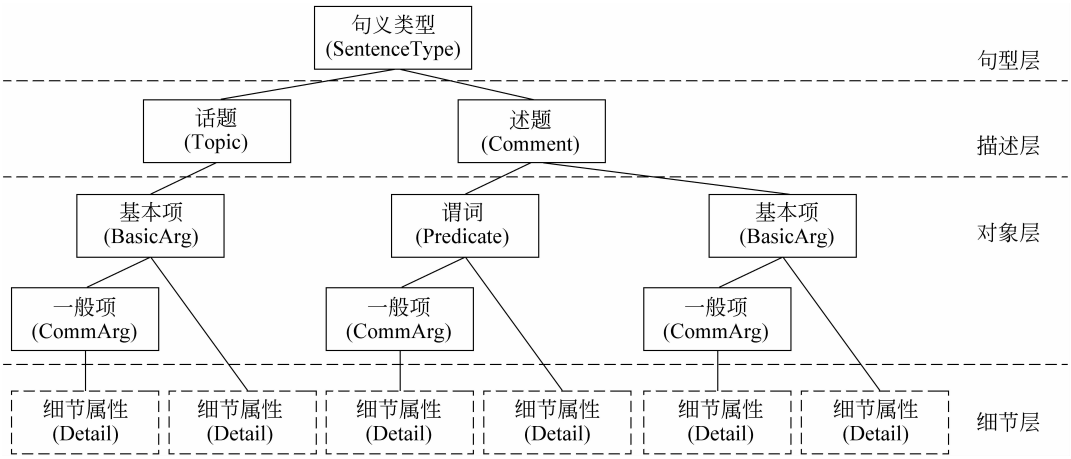


图 2 句义结构模型的基本形式

也有一般项修饰一般项的情况)。谓词是句义中说明话题的成分,是对话题指称对象的变化、运动、行为或者情感、意愿或者状态、性质等的说明,根据谓词与基本项的组合情况(即谓词能与多少个基本项组合成一个句义结构框架),谓词可以分为零目、一目、二目和多目;项在句义里与句义其他成分的组合中所体现出的不同功能和语义类型称作语义格,贾彦德先生从汉语句义结构的特点出发,提出了 21 种汉语语义格类型[8],BFS-CTC 语料库以此为基础,定义了 7 种基本格(如施事格、受事格、与格等)与 11 种一般格(如时间格、空间格、属格等)。在图 2 中用实线表示由下到上的语义组合关系,例如,话题和述题组合成句义,谓词和基本项组合成述题等。

细节层是对对象层中对象的进一步描述,该层次本身不属于句义结构的框架,仅仅用于说明对象的属性、范围、性质等。细节层是一个可扩展的层次,并不是所有成分(对象)都具有细节属性,目前定义的需要描述的细节属性为谓词时态信息。谓词时态反映的是谓词描述动作、行为、变化(或状态)所处的进程状态,根据龚千言先生划分的 8 种时态类型[9],BFS-CTC 定义了将来、现在和完成 3 种时态。由于细节属性与成分之间并不构成组合关系,因此在图 2 中用虚线表示对细节属性的描述。

## 2.2 句义结构模型的扩展形式

简单句义的句义结构中只具有一个最小完整单元,故简单句义的形式化表达可直接用图 2 表示。例如简单句"委员会明天将要通过此议案。"的句义结构如图 3 所示。

除简单句义外,其他三种句义结构的模型可以通过最小完整单元扩展而来。

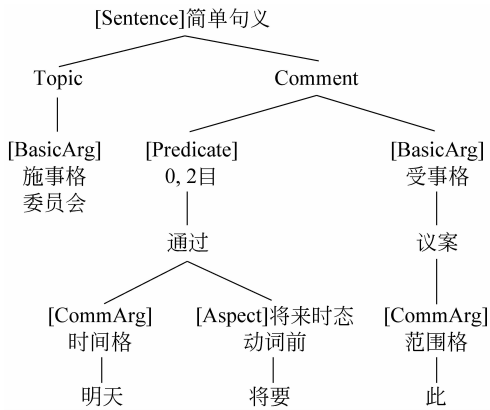


图 3 简单句义的句义结构标注示例

复杂句义中的成分句义由一个最小完整单元表示,并且在复杂句义的句义结构中作为其中的一个句义成分来使用,即简单句义充当复杂句义中的基本项或者一般项。例如复杂句义"拒绝零食完全没有必要。"的句义结构如图 4 所示。图 4 中阴影部分

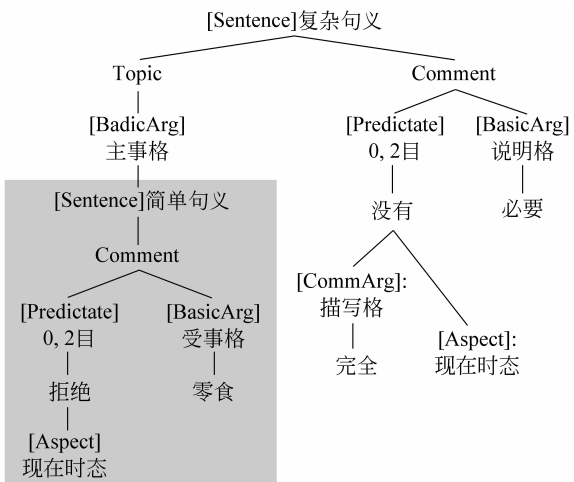


图 4 复杂句义的句义结构标注示例

为子句“拒绝零食”的句义结构，该子句在整个句义中是一个充当主事格的成分句义。

复合句义中的分句义由最小完整单元表示，在复合句义的句义结构中，分句义不是作为句义成分，而是直接组合成为复合句义。例如，复合句义“中国队输掉了比赛，球迷们很伤心。”的句义结构如图 5 所示。图 5 中阴影部分分别为子句“中国队输掉了比赛”和“球迷们很伤心”的句义结构，这两个子句在整个句义中是分句义，直接组合构成该句子的完整句义结构。

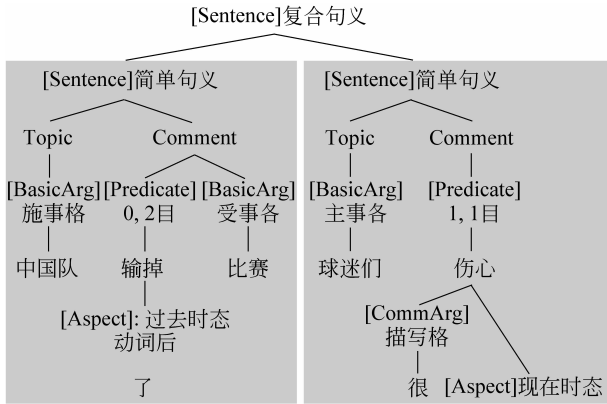


图 5　复合句义的句义结构标注示例

在多重句义中，子句义（成分句义或者分句义）本身就是一个复杂句义或者复合句义，因此多重句义的句义结构中可能会存在多层句义结构的嵌套。例如，多重句义“法塔赫把阿巴斯打造成为现在最具影响力的政治人物。”的句义结构如图 6 所示。图 6 中阴影部分为充当成分句义的复杂句义“成为现在最具影响力的政治人物”的句义结构，在该成分句义中，又由子句“现在最具影响力”充当了说明格的描写格成分。

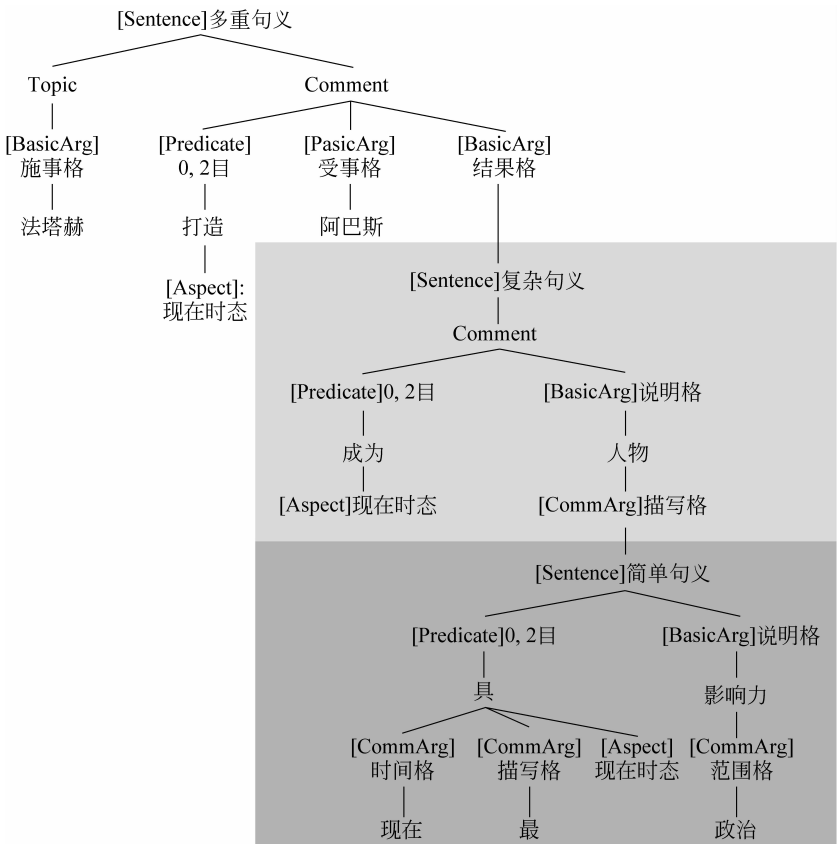


图 6　多重句义的句义结构标注示例

## 3　标注内容和标记形式说明

在 BFS-CTC 标注规范方面，对于词法和句法标注，BFS-CTC 分别采用了北京大学现代汉语语料库的词性标注规范以及句法标注规范，最大程度上保证和目前主要的汉语加工规范的一致性，以便更好地兼容现有汉语词法、句法的标注语料及相关标注工具；对于句义结构标注，BFS-CTC 遵循课题组自己建立的句义结构标注规范，并随着标注实践的

不断深入进行补充和完善。表1为BFS-CTC中采用的标注规范及相应示例。

**表1 BFS-CTC标注规范及示例**

| 类型 | 规范 | 示例 |
|---|---|---|
| 词法标注规范 | 北京大学的词性标注规范[10-11] | 委员会/n 明天/t 将要/d 通过/v 此/r 议案/n 。/w |
| 句法标注规范 | 北京大学计算语言学研究所规范[12] | [dj [np 委员会/n] [vp 明天/t [vp 将要/d [vp 通过/v [np 此/r 议案/n]]]] 。/w] |
| 句义结构标注规范 | BFS-CTC-SSTB 句义结构标注规范 | |

### 3.1 标注内容

按照句义结构的定义,句义结构的标注包括了对句子的句型层、描述层、对象层和细节层中所包含的各个要素及其组合关系的标注。

在句型层和描述层,标注句义的类型以及构成该句义的话题与述题成分。

在对象层的标注中,标注构成句义的主干部分(谓词、基本项)以及起修饰和限定作用的部分(一般项)。由于有时表主体的基本项在一定的语境中会被省略,因此,当句子中只存在表客体的基本项而不存在表与事的基本项时,谓词仍标注为2目谓词,当客体和与事都存在时,谓词即可标注为多目谓词。此外,由于有时句子中同一层次的句义结构里会出现两个或两个以上的谓词,因此需要标注出本层中谓词在时间上的先后顺序(0,1,2……),当本层句义结构中只有一个谓词时,其序号标注为0。

在细节层的标注中,由于现代汉语中的时态是通过围绕动词的时态成分来体现的[13],时态成分包括时态副词(曾经、已经、正在、将、要……)、时态助词(着、了、过……)和时态语气词(了、来着……),因此,除标注谓词时态之外,还标注出了表征该时态的时态词,以及时态词在句子中的位置,包括谓词前、谓词后、句首和句尾四种,当不存在时态词时,则不进行标注。

### 3.2 标记形式

句义结构模型中除了话题和述题之外,其他元素(句义类型、谓词、语义格、谓词时态等)都具有自身的类型和范围。这些元素的定义与现代汉语语义学中的定义一致,在语料库中对其范围和类型进行了一些规约和修改,并定义了每一类元素的标记,如表2所示,每个标记中的各个变量及其取值说明如表3所示。

**表2 句义结构中的标记说明**

| 标记 | 含义 | 说明 |
|---|---|---|
| SemanticTag | 语义标注标记 | 作为整个文件的标识符 |
| Sentence | 句子标记 | 包含句子标号(id)、句义类型(type) |
| Topic | 话题标记 | |
| Comment | 述题标记 | |
| Predicate | 谓词标记 | 包含谓词目数、顺序,数据为句子中承担谓词职责的词 |
| BasicArg | 基本格标记 | 包含基本格类型(case),数据为句子中承担基本格职责的词 |
| CommArg | 一般格标记 | 包含一般格类型(case),数据为句子中承担一般格职责的词 |
| Aspect | 时态标记 | 包含时态属性(value),时态词位置(pos),数据为句子中反映时态的词 |

**表3 句义结构标记中的变量说明**

| 变量 | 含义 | 取值(说明) |
|---|---|---|
| Sentence::type | 句义类型 | SIMPLE(简单句义),COMPLEX(复杂句义),COMPOUND(复合句义),MULTIPLE(多重句义) |
| BasicArg::case | 基本格类型 | AGENTIVE(施事格),RENCONTRE(遭遇格),OBJECTIVE(受事格),RESULT(结果格),SUBJECTIVE(主事格),EXPLAIN(说明格),DATIVE(与格) |

续表

| 变量 | 含义 | 取值(说明) |
|---|---|---|
| CommArg::case | 一般格类型 | RANGE(范围格),TIME(时间格),SPACE(空间格),TOOL(工具格),MODE(方式格),CAUSE(根由格),ATTACH(属格),DESCRIPTION(描写格),PARITY(同位格),STANDARD(基准格),OTHER(其他格) |
| Predicate::count | 谓词类型 | 0(0 目谓词),1(1 目谓词),2(2 目谓词),MULTIPLE(多目谓词) |
| Aspect::value | 时态类型 | FUTURE(将来时态),NOW(现在时态),PAST(完成时态) |
| Aspect::pos | 时态特征词位置 | BEFORE-PREDICATE(谓词前),AFTER- PREDICATE(谓词后),SENTENCE-TOP(句首),SENTENCE-END(句尾) |

在句义标注文件内部,以 XML 形式标记句义结构的标注结果,图 3 中例句"委员会明天将要通过此议案"即可表示为以下形式。

```
<Sentence id="SP00000001" type="SINGLE">
    <Topic>
        <BasicRole case="AGENTIVE">
            委员会
        </BasicRole>
    </Topic>
    <Comment>
        <Predicate count="2" sequence="0">
            通过
            <CommRole case="TIME">明天
            </CommRole>
            <Aspect value="FUTURE" pos=
            "BEFORE-VERB">将要</Aspect>
        </Predicate>
        <BasicRole case="OBJECTIVE">
            议案
            <CommRole case="RANGE">此
            </CommRole>
        </BasicRole>
    </Comment>
</Sentence>
```

## 4　语料来源、规模和语料库组织结构

BFS-CTC 中的句子均来源于新闻语料中的句子,包括《人民日报》1998 年 1 月份的语料,(共 200 多万字,约 27 000 句);搜狗新闻语料(互联网新闻报道语料的集合,共 80 000 个文件);北京大学汉语语言学研究中心的 CCL 语料库(共 838 803 906 字节,汉字总字数:264 444 436)。

认知语言学派认为,语法与语义是密不可分的,两者之间存在内在的联系。句式是具有某种结构特征的句子格式,其中有以某种结构类型为特征的,有以特定的标志词为特征的,是语法的一项重要内容[14]。不同句式的句子,其句义结构也不尽相同。例如,在非主谓句的句义结构中不存在话题,在被字句的句义结构中通常主体格出现在述题中,在连动句的句义结构中通常是由两个或两个以上的简单句构成主句的述题等。因此,BFS-CTC 对句子按照其句式类别进行了划分,共分为主谓句、非主谓句、把字句、被字句、连动句、兼语句这六种典型的、句义结构特点较为显著的句式。通过对这些语料进行人工筛选,BFS-CTC 语料的标注规模为 10 000 句,目前 BFS-CTC 中的句子以单句为主,句子长度的分布情况及相应实例如表 4 所示。

**表 4　BFS-CTC 中句子句式类型及其长度的分布情况**

| 句式 | 数目 | 最短字数 | 最长字数 | 平均字数 | 实例 |
|---|---|---|---|---|---|
| 主谓句 | 5 100 | 3 | 51.5 | 16.5 | 史迪威来华标志着中美合作进入了新的阶段。 |
| 非主谓句 | 1 100 | 3 | 37 | 11.5 | 坚决禁止挪用财政性经费举办文化娱乐等活动。 |
| 把字句 | 1 073 | 6 | 51 | 20.4 | 他把自己的英文专著送给随行的王光美。 |
| 被字句 | 1 361 | 8 | 49.5 | 17.4 | 该车在北美市场的销售前景被一致看好。 |
| 连动句 | 1 058 | 4 | 53 | 20.8 | 1961 年 5 月 16 日,朴正熙少将率领韩国军人发动政变。 |
| 兼语句 | 975 | 6 | 47 | 20.8 | 普京亲自任命伊万诺夫为该委员会主席。 |

BFS-CTC作为句子在句义层次上的深加工语料库，不仅提供了完整的句义结构标注信息，同时也提供了句义结构分析所需的词法标注以及基于短语结构语法的句法标注信息。BFS-CTC由4个核心库构成，其组织结构如图7所示。

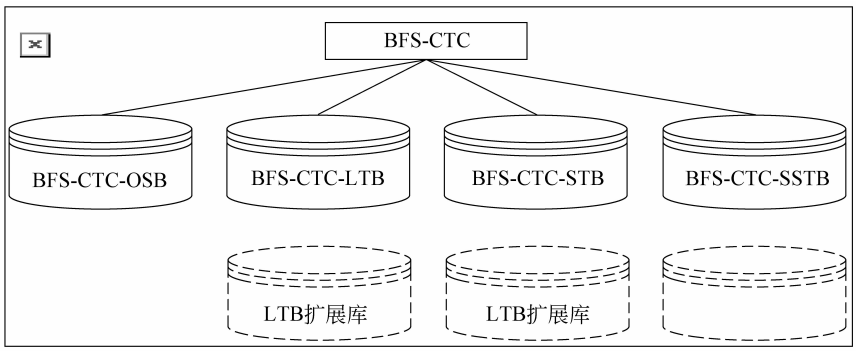


图7 BFS-CTC的组织结构

原始句子库(OSB, Original Sentence Bank)、词法标注库(LTB, Lexical Tagged Bank)、句法标注库(STB, Syntax Tagged Bank)、句义结构标注库(SSTB, Semantic Structure Tagged Bank)，这4个核心库是BFS-CTC的最主要内容，对于BFS-CTC中收录的句子，采用深度优先的方式进行标注。每个句子都会进行词法、句法、句义结构的三层标注，包括原始句子在内的所有标注信息将会被分别存储到4个核心库中，每个句子都对应一个唯一的ID号。在此基础上，可以根据研究的需要，在词法、句法、句义结构标注的基础上进行深加工，得到更多的针对性更强的扩展标注库。BFS-CTC中的各类标注既可以单独使用，也可综合使用进行横向分析，可用于自然语言处理中多方面的研究。

## 5 BFS-CTC语料库特色及对比分析

语料库作为自然语言处理研究的基础，近年来受到了广泛的关注。语料库按加工程度不同可分为生语料(未加标注的语料库)、词法标注语料库、句法标注语料库及语义标注语料库四类。

在汉语句法方面，目前公开的比较著名的有宾州大学中文树库(CTB)、清华大学周强建立的汉语树库(TCT)、北京大学的现代汉语树库以及中国台湾“中央”研究院的Sinica树库，另外还有哈尔滨工业大学的汉语依存树库、中国传媒大学的依存树库、中国科学院计算技术研究所的汉语树库等。

在汉语语义方面，目前使用的语料主要有宾州大学构建的中文PropBank(CPB)，山西大学参照加州大学克伯利分校的FrameNet而构建的汉语框架网络知识库(CFN)。由于CPB与FrameNet是目前两大主要的句义标注体系，因此本文将从以下三个方面同BFS-CTC语料库进行比较分析。

(1) 理论基础

FrameNet[15]以框架语义学为标注的理论基础，描述每个谓词的语义框架，以及这些框架之间的关系，通过框架对句子进行标注(如图8所示)；CPB[2]根据格语法理论，在中文TreeBank上对目标动词添加了一个语义角色标注层(如图9所示)；BFS-CTC根据现代汉语语义学理论[6]，标注汉语句义结构中所包含的各个句义成分及其之间的组合关系(如图3所示)。FrameNet和CPB主要针对的是语义角色的标注研究，而BFS-CTC更侧重的是完整的汉语句义结构的分析研究，语义角色标注只是

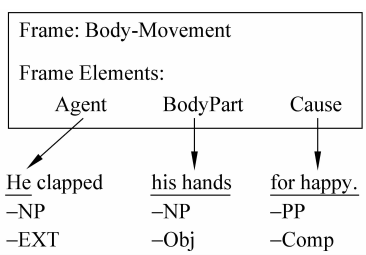


图8 FrameNet框架以及句子标注示例

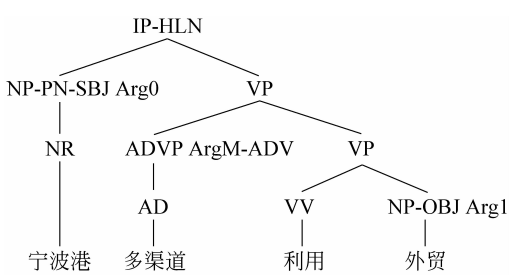


图9 中文PropBank的一个句子标注示例

其中的一部分。

2. 标注形式

在语义方面，FrameNet 利用构建的语义框架为句子中的词语进行语义角色标注；CPB 利用句子的句法标注结果，为句法成分添加其相对给定目标动词的语义角色类别；BFS-CTC 根据汉语句义结构模型，为句子构建层次化的句义结构，标注句义类型、话题述题、谓词与时态、语义格类型、句义成分组合关系等多项信息。

在语法方面，FrameNet 和 CPB 均提供了各语义角色在句法层面的短语类型（如 np、vp 等）和句法功能（如主语、宾语等）；BFS-CTC 不仅提供了句子的短语结构句法标注信息，句法成分的短语类型，还提供了句子的句式类型信息，包括主谓句、非主谓句、把字句、被字句、连动句和兼语句这六种在句义结构标注中具有典型特点的句式。

BFS-CTC 较 FrameNet 和 CPB 提供了更多、更丰富的标注信息，层次化的句义结构标注方式，加深了标注的深度，对于句式的划分能够为语法与语义两者之间内在联系研究提供语料资源。

3. 语义角色标记集

从语义角色标记集来看，FrameNet 通过基于场景的语义框架来定义语义角色，框架中的框架元素便是 FrameNet 的角色，各个框架中语义角色的含义和数量是不同的；CPB 包括 20 多种语义角色，其中核心的语义角色为 Arg0-5 六种，相同的语义角色对于不同目标动词有不同的语义含义，它们的具体含义由框架文件给出；BFS-CTC 采用的语义格概念属于中观层级的语义角色[16]，根据现代汉语语义学中定义的语义格类型，结合汉语的特点，又将其具体划分为七种基本格与 11 种一般格（如表 3 所示），每一种语义格在不同句子中所体现的语义含义是相同的（例如，“施事格”是发出或发生谓词所表行为、运动、变化的主体，施事、行动者[6]）。例如，下面两个句子

Chuck bought a car from Jerry.（Chuck 向 Jerry 买了一辆汽车）

Jerry sold a car to Chuck.（Jerry 卖了一辆汽车给 Chuck）

在 FrameNet 中，这两句属于 Commerce（商务活动）框架，两个句子中的 Chuck 均被标注为框架元素，Buyer，Jerry 均被标注为框架元素 Seller，car 均被标注为 Goods；在 PropBank 中，第一句的 Chuck 被标注为 Arg0，car 被标注为 Arg1，Jerry 被标注为 Arg2，而第二句的 Jerry 被标注为 Arg0，car 被标注为 Arg1，Chuck 被标注为 Arg2；在 BFS-CTC 中，两个句子中的汽车（Car）均被标注为受事格，Chuck 分别被标注为施事格和与格，Jerry 分别被标注为与格和施事格。

可以看出，BFS-CTC 采用现代汉语语义学中对语义格的定义方式，其语义格标注形式与 PropBank 类似，不受语义框架的约束，较 FrameNet 更便于比较不同句子的语义角色特点。然而，FrameNet 较 PropBank 与 BFS-CTC 具有更好的对于语言现象的描述能力和解释能力。

通过以上三个方面的对比分析，以现代汉语语义学为基础构建的 BFS-CTC 语料库，较 FrameNet 和 CPB 更适合汉语的句义结构分析研究，其层次化的标注方式、语义格类型的标注方法以及对于六种典型句式的划分方式，能够为汉语句义分析提供更全面的标注信息。

## 6　结束语

随着统计自然语言处理的不断深入，语料库资源的缺乏逐渐成为制约语义分析技术发展的“瓶颈”问题之一。为了满足汉语句义结构研究的需要，根据汉语语义学的句义理论，提出了层次化、可扩展、可计算的汉语句义结构模型，构建了 BFS-CTC 汉语标注语料库，涵盖了主谓句、非主谓句、把字句等六种句式，提供了词法、句法及句义结构的标注信息，目前规模约为 10 000 句。由于基于手工标注，语料标注的准确率高，其覆盖范围较广，可学习性强，能够满足汉语句义结构分析的需求。以该语料库提供的句义结构标注为基础，目前已经进行了句子的句义类型识别、谓词识别及谓词时态识别[17]、汉语语义格类型识别和汉语句义结构框架提取方法等研究工作。语料库的标注工作还在不断进行中，同时语料库将面向社会开放，供广大研究同行研究使用。

## 参考文献

[1] 何婷婷. 语料库研究[D]. 华中师范大学，2003.

[2] M Palmer，D Gildea，P Kingsbury. The Proposition Bank：An Annotated Corpus of Semantic Roles[J]. Computational Linguistics，2005，31(1)：71-105.

[3] 刘开瑛，由丽萍. 汉语框架语义知识库构建工程[C]. 中文信息处理前沿进展——中国中文信息学会二十五

周年学术会议，2006.
[4] 由丽萍，杨翠. 汉语框架语义知识库概述[J]. 电脑开发与应用，2007，20(6)：2-7.
[5] http://cslt.riit.tsinghua.edu.cn/～qzhou/chs/Resources.htm.
[6] 贾彦德. 汉语语义学[M]. 北京：北京大学出版社，2005：249-265.
[7] 冯扬. 汉语句义模型构建及若干关键技术研究[D]. 北京理工大学，2010.
[8] 贾彦德. 对现代汉语语义格的认识与划分[J]. 语文研究，1997，(3)：23-29.
[9] 龚千言. 汉语的时相时制时态[M]. 北京：商务印书馆，1995.
[10] 俞士汶，段慧明，朱学锋，等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报，2002，16(5)：49-64.
[11] 俞士汶，段慧明，朱学锋，等. 北京大学现代汉语语料库基本加工规范(续)[J]. 中文信息学报，2002，16(6)：58-64.
[12] 周强. 汉语语料库的短语自动划分和标注研究[D]. 北京大学，2002.
[13] 陈立民. 汉语的时态和时态成分[J]. 语言研究，2002，(3)：14-31.
[14] 周一民. 现代汉语(修订版)[M]. 北京：北京师范大学出版社，2006.
[15] C Baker，C Fillmore，J Lowe. The Berkeley FrameNet Project[C]//Proceedings of COLING/ACL[C]. Montreal，Canada，1998，86-90.
[16] 袁毓林. 语义角色的精细等级及其在信息处理中的应用[J]. 中文信息学报，2007，21(4)：10-20.
[17] 刘莉莉. 汉语句义类型及谓词时态识别算法研究[D]. 北京理工大学，2010.
[18] 郝晓燕，李伟，李茹等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报，2007，21(5)：96-100.
[19] 张惠春，由丽萍. 基于中文框架网络的浅层语义分析模型[J]. 电脑开发与应用，2009，22(8)：4-6.

(上接第 6 页)

[13] Chang.P，Tsengb，H，Jurafskya，D and Manning，C. Discriminative Reordering with Chinese Grammatical Relations Features[C]//Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation at NAACL HLT 2009.51-59.

## 附录 A

### 词语相似度示例(前 10 个相似词)

哀思：情思-0.088 乡愁-0.084 乡思-0.084 忧思-0.075 爱意-0.069 英灵-0.066 情愫-0.065 痛楚-0.064 敬意-0.063 深情厚意-0.063

轮船：游轮-0.117 货轮-0.114 客轮-0.108 海轮-0.102 渡轮-0.097 海船-0.096 货船-0.096 航船-0.095 船只-0.089 江轮-0.087

泥土：浮土-0.128 灰土-0.104 沙土-0.096 沙子-0.094 淤泥-0.091 沙砾-0.091 表土-0.090 土块-0.088 泥巴-0.087 尘土-0.086

床单：被单-0.151 床罩-0.136 被褥-0.135 枕巾-0.134 枕套-0.127 蚊帐-0.121 浴巾-0.119 棉被-0.112 毛巾被-0.107 褥子-0.103

老练：老到-0.122 老辣-0.122 练达-0.109 干练-0.096 沉稳-0.092 老成-0.083 娴熟-0.083 狡诈-0.0821 镇定-0.080 精明-0.078

犀利：尖刻-0.122 锐利-0.117 凌厉-0.098 老辣-0.093 精辟-0.087 敏锐-0.087 泼辣-0.082 简练-0.081 机智-0.081 机敏-0.080

优雅：高雅-0.135 典雅-0.130 幽雅-0.129 柔美-0.121 淡雅-0.118 雅致-0.115 素雅-0.113 婉约-0.109 文雅-0.103 温婉-0.102

解救：营救-0.076 援救-0.068 挽救-0.067 搭救-0.064 拯救-0.063 抢救-0.063 脱身-0.057 安抚-0.056 救护-0.053 救活-0.053

纠正：改正-0.103 制止-0.087 订正-0.083 改掉-0.077 克服-0.075 更正-0.069 补救-0.067 矫正-0.066 督促-0.065 批评-0.064

善待：关爱-0.109 关心-0.102 爱护-0.096 对待-0.092 爱惜-0.092 体谅-0.090 珍惜-0.089 感化-0.085 体恤-0.083 尊重-0.083