

BFS-CTC 汉语句义结构标注语料库构建方法

罗森林, 刘盈盈, 冯扬, 韩磊, 陈功, 王倩

(北京理工大学 信息与电子学院, 北京 100081)

摘 要: 根据现代汉语语义学, 构建了一种层次化的句义结构模型. 基于该模型构建了汉语句义结构标注语料库 (Beijing forest studio-Chinese tagged corpus, BFS-CTC). 利用自行开发的标注和管理工具, 对模型中各个句义成分及其组合关系进行快速标注, 降低培训工作和标注成本. BFS-CTC 涵盖了 6 种句式类型, 约 1 万句, 提供了符合现有规范的词法和句法标注信息与自定义规范的句义结构标注信息, 便于词法、句法和句义的对照分析研究, 以及语料的综合使用和横向分析. 此外, BFS-CTC 还具有较强的可扩展性, 可在核心标注库基础上扩展生成其它扩展库和标注资源.

关键词: 中文信息处理; 句义分析; 句义结构; 语义标注; 语料库

中图分类号: TP 391

文献标志码: A

文章编号: 1001-0645(2012)03-0311-05

Method of Building BFS-CTC: a Chinese Tagged Corpus of Sentential Semantic Structure

LUO Sen-lin, LIU Ying-ying, FENG Yang, HAN Lei, CHEN Gong, WANG Qian

(School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Based on the modern Chinese semantics, a Chinese sentential semantic mode is built, and then a Chinese tagged corpus, BFS-CTC (Beijing forest studio-Chinese tagged corpus), is built according to the Chinese sentential semantic mode. There are more than ten thousand sentences in the corpus, and the corpus contains six kinds of Chinese syntactic types. Tagging the sentence quickly and conveniently could be implemented by using the self-developed tools. BFS-CTC provides lexical, syntactic and sentential semantic structure tagging information, so that it could be used in comparative analysis of syntactic and semantic, or used for horizontal analysis. In addition, the corpus has good scalability, and it could generate more targeted extension tagged banks.

Key words: Chinese information processing; sentential semantic analysis; sentential semantic structure; semantic labeling; corpus

语料库是为某一个或多个应用而专门收集、有一定结构、有代表性、可以被计算机程序检索并具有一定规模的语料合集. 按加工程度不同可分为生语料库、词法标注语料库、句法标注语料库及语义标注语料库 4 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,

而句义分析是语义研究中的一个重要问题, 作为不可或缺的基础资源, 句义标注语料库正受到越来越多的关注.

目前, 主要的汉语语义标注语料库有宾州大学汉语浅层语义标注命题库 (Chinese proposition bank, CPB)^[1], 清华大学句法语义标注库 (syntacti-

收稿日期: 2011-04-27

基金项目: 国家“二四二”计划项目 (2005C48); 北京理工大学科技创新计划项目 (2011CX01015)

作者简介: 罗森林 (1968—), 男, 教授, 博士生导师, E-mail: luosenlin@bit.edu.cn.

通信作者: 刘盈盈 (1987—), 女, 硕士生, E-mail: yinying01@yahoo.cn.

cally and semantically annotated corpus, SSAC)^[2]和山西大学汉语框架语义知识库(Chinese frame net, CFN)^[3]等。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,它根据格语法理论,在句法标注语料库 CTB(Chinese tree bank)基础上标注出部分句法成分相对于给定目标动词所具备的语义角色,包括 81 009 个动词事件和 14 525 个名词事件;SSAC 在句法依存信息基础上,针对某个特定事件描述的目标动词生成完整的句法语义信息标注,包括 25 000 个标注句子,1 800 个存在类描述动词;CFN 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,包括框架库、句子库和词元库 3 个部分,对 1 760 个汉语词元构建了 130 个框架,标注了 8 200 条句子。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到句义层次,并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。

本文中作者构建一个能提供句义结构分析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)的语料库。具体的工作包括构建汉语的句义结构模型,建立语料库标注规范开发,标注与管理工具,收集语料及实施标注等。

1 汉语句义结构定义及模型

根据现代汉语语义学理论,BFS-CTC 构建了层次化、可扩展、可计算的句义结构模型(Chinese sentential semantic structure mode, CSSSM)^[4],图 1 为其基本形式,其中实线表示由下到上的语义组合关系。细节属性与成分之间并不构成组合关系,因此在图 1 中用虚线表示。

句型层反映的是句子在句义上的类型。汉语语义学将汉语句义类型划分为简单句义、复杂句义、复

合句义和多重句义 4 种^[5]。其中简单句义只表达一个命题,通过不同的组合形成其他非简单句义。因此,定义简单句义的句义结构形式为最小完整语义单元,在非简单句义的结构中,最小完整语义单元的作用与一个句义成分的作用相似。

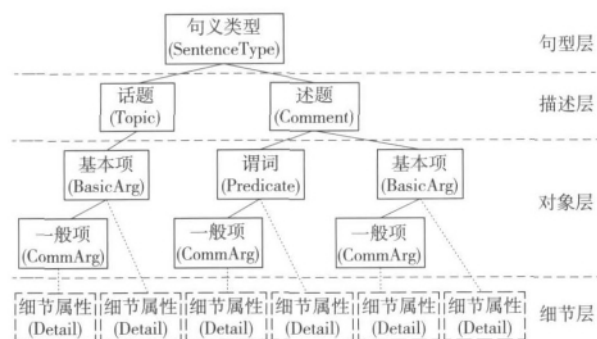


图 1 句义结构模型的基本形式

Fig. 1 Basic form of Chinese sentential semantic structure mode

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 1 次划分。

对象层是针对描述层中的进一步划分,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。谓词和基本项直接与话题和述题相关联,构成句义结构基本框架;一般项用来描述和限定谓词、基本项或一般项(如限定时间、地点、范围等)。谓词是句义中说明话题的成分,根据谓词能与多少个基本项组合成一个句义结构框架,可以被分为 0 目、1 目、2 目和多目;项在句义里与其它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的特点出发,提出了 21 种汉语语义格,BFS-CTC 语料库以此为基础,定义了 7 种基本格与 11 种一般格(如表 1 所示)。

表 1 句义结构标记中的变量说明

Tab. 1 Variables and markings in sentential semantic structure

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	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)
CommArg: case	一般格	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(其它格)
Predicate: count	谓词	0(0 目), 1(1 目), 2(2 目), multiple(多目)
Aspect: value	时态	future(将来时态), now(现在时态), past(完成时态)
Aspect: pos	时态词位置	before-predicate(谓词前), after-predicate(谓词后), sentence-top(句首), sentence-end(句尾)

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,仅用于说明对象的

属性、范围、性质等,是一个可扩展的层次。目前定义的细节属性为谓词时态信息,反映谓词所述动作、行为、变化所处的进程状态。根据龚千言先生^[6]划分的8种时态类型,BFS-CTC定义了将来、现在和完成3种时态。

其它3种非简单句义的句义结构模型可通过扩展最小完整语义单元而来^[4]。

复杂句义中的成分句义由一个最小完整语义单元担任,并且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某个基本项或一般项。例如“拒绝零食完全没有必要”中的子句“拒绝零食”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语义单元表示,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。例如“中国队输掉了比赛,球迷们很伤心”由两个分句义“中国队输掉了比赛”和“球迷们很伤心”直接组合构成。

多重句义中的子句义(成分句义或分句义)本身就是一个复杂句义或复合句义,可能存在多层句义结构的嵌套。例如“法塔赫把阿巴斯打造成为现在最具影响力的政治人物”中包含一个充当结果格成分的复杂句义“成为现在最具影响力的政治人物”,在该成分句义中,又包含一个充当描写格成分的简单句义“现在最具影响力”。

2 BFS-CTC 的构建方法和步骤

BFS-CTC 语料库构建步骤如图2所示。



图2 语料库的构建步骤

Fig.2 Steps to build BFS-CTC

2.1 标注规范制定

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^[7]以及句法标注规范^[8],最大程度上保证和目前主要的汉语加工规范的一致性,以便更好地兼容了现有汉语词法、句法的标注语料及相关标注工具。例如“委员会明天将要通过此议案。”的词法标注为“委员会/n 明天/t 将要/d 通过/v 此/r 议案/n./w”,句法标注结果为“[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]”。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,下面介绍句义结构标注的相关规范。

2.1.1 句义结构标注内容说明

根据定义,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。

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。

对象层标注构成句义结构的主干部分(谓词、基本项)以及起修饰和限定作用的部分(一般项)。有时表主体的基本项在一定的语境中会被省略。因此当句子中存在表客体的基本项时,谓词仍标注为2目谓词;当客体与事都存在时,谓词标注为多目谓词。此外,由于有时句子中同一层次的句义结构里会出现多个谓词,因此需要标注出本层中谓词在时间上的先后顺序(0,1,2,...)。

在细节层,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^[9],包括时态副词、时态助词和时态语气词。因此除标注谓词时态外,还标注出表征该时态的时态词,以及时态词在句子中的位置。

2.1.2 句义结构标记形式说明

句义结构模型中除话题和述题外,其它元素(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,每个标记的各个变量及其取值说明如表1所示。

例如简单句义“委员会明天将要通过此议案”的句义结构如图3(a)所示,其XML标记形式如图3(b)所示。



(a) 句义结构标注形式

```

<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>

```

(b) XML表示形式

图 3 简单句义的句义结构示例

Fig. 3 A simple sentential semantic structure

2.2 语料收集和整理

BFS-CTC 中的句子均来源于新闻语料,包括人民日报 1998 年 1 月语料(共 200 多万字);搜狗新闻语料,互联网新闻报道的语料集合(共 8 万个文件);北京大学汉语语言学研究 CCL 语料库(约 2 644 万字)。

首先对原始语料进行分句,从大段的新闻语料中获得独立的句子;然后进行分拣,将句子划分到其相应的句式类别中,根据汉语句子句式的划分,BFS-CTC 收录了主谓句、非主谓句、把字句、被字句、连动句、兼语句 6 种句式类型的句子;最后进行排重。目前 BFS-CTC 中的句子以单句为主,约 1 万句,具体分布如表 2 所示。

表 2 BFS-CTC 中句子按句式类型的分布情况

Tab. 2 Distribution of the sentence type in BFS-CTC

句式	标记	数目
主谓句	SP	5 165
非主谓句	NP	1 098
把字句	BA	1 075
被字句	BEI	1 366
连动句	CA	991
兼语句	PS	979
总计		10 664

2.3 语料标注

标注工作基本由人工标注完成,标注时采用自行设计的辅助标注工具。

标注工具提供对标注工程的创建、保存、修改、合并等功能。在句子列表区,按照句式划分为 6 个标签显示句子基本信息和标注信息;在句子图形化标注区支持图形化标注以及自动插入已完成分词的句子词块信息,加快标注速度;标注结果区自动显示

标注后的标注信息,并且在切换标注标签时自动对标注的正确性与完整性进行初步检查,以保证句子文件标注信息的一致性,降低标注成本。此外还提供了标注的帮助文档,包括标注的基本概念、标注规范、标注工具的使用说明以及各种句式的标注示例等内容,降低了标注人员的培训工作量。

标注工具集成了中科院计算所的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System, ICTCLAS)的自动词法分析算法,标注人员检验分词结果及词性标注,并且校正其中的错误;校对后,对句子进行人工的句法和句义结构标注。当人员完成标注后,进行交叉验证,以保证标注质量。交叉验证通过之后,再经管理人员验收、入库。

2.4 语料存储

BFS-CTC 由 4 个核心库构成,分别为原始句子库(original sentence bank, OSB)、词法标注库(lexical tagged bank, LTB)、句法标注库(syntax tagged bank, STB)、句义结构标注库(semantic structure tagged bank, SSTB)。核心库是 BFS-CTC 的最主要内容,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 的组织结构如图 4 所示。

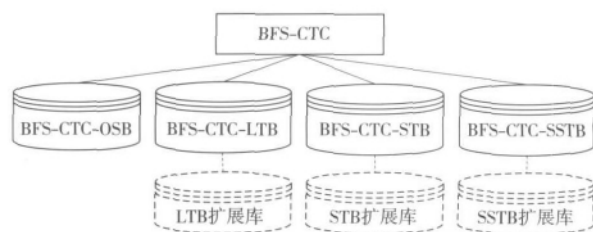


图 4 BFS-CTC 的组织结构

Fig. 4 Structure of BFS-CTC

对 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。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研究。

2.5 语料管理

为有效地管理语料,在构建 BFS-CTC 的过程中开发了语料管理工具。

管理工具支持按句式类型管理和按人员管理(包括标注信息和修改信息两部分)两种模式,对句

子的入库时间、标注状态、标注进度、修改情况等进行全面管理。

3 结束语

主要讨论了基于现代汉语语义学的句义结构模型及其标注语料库,重点介绍了 BFS-CTC 的构建方法和步骤,包括标注规范、语料的收集和整理、语料的标注及标注工具、语料的存储方式和管理等。

BFS-CTC 根据现代汉语语义学构建了一种层次化的句义结构模型,涵盖了 6 种句式类型,约 1 万句,提供了词法、短语结构句法以及句义结构的标注信息,为汉语句义结构分析提供了必要的语料资源,便于词法、句法和句义的对照分析研究。BFS-CTC 的构建工程量大,难度较高,通过自行开发的标注工具和管理系统,可加快语料的标注进度,降低培训和标注成本。基于 BFS-CTC 的标注信息,已经进行了句义类型识别、谓词识别以及谓词时态识别^[10]等研究工作。语料库的标注工作还在不断进行中,同时语料库将逐步开放,供广大研究同行使用。

参考文献:

- [1] Palmer M, Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles[J]. Computational Linguistics, 2005,31(1):71-105.
- [2] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007,21(3):21-27.
Zhou Qiang. Base chunk scheme for the Chinese language [J]. Journal of Chinese Information Processing, 2007,21(3):21-27. (in Chinese)
- [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. 北京: 清华大学出版社, 2006:64-71.
Liu Kaiying, You Liping. The Chinese framenet project [C]// Proceedings of Frontiers of Chinese Information Processing, the 25th Annual Meeting of Chinese Information Processing Society of China. Beijing: Tsinghai University Press, 2006:64-71. (in Chinese)
- [4] 冯扬. 汉语句义模型构建及若干关键技术研究[D]. 北京: 北京理工大学信息与电子学院, 2010.
Feng Yang. Research on Chinese sentential semantic mode and some key problems[D]. Beijing: School of Information and Electronics, Beijing Institute of Technology, 2010. (in Chinese)
- [5] 贾彦德. 汉语语义学[M]. 北京: 北京大学出版社, 2005: 249-265.
Jia Yande. Chinese semantics[M]. Beijing: Peking University Press, 2005:249-265. (in Chinese)
- [6] 龚千言. 汉语的时相时制时态[M]. 北京: 商务印书馆, 1995.
Gong Qianyan. Phase, tense and aspect in Chinese[M]. Beijing: The Commercial Press, 1995. (in Chinese)
- [7] 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.
Yu Shiwen, Duan Huiming, Zhu Xuefeng, et al. The basic processing of contemporary Chinese corpus at Peking University Specification[J]. Journal of Chinese Information Processing, 2002,16(5):49-64. (in Chinese)
- [8] 周强. 汉语语料库的短语自动划分和标注研究[D]. 北京: 北京大学计算机科学与技术系, 2002.
Zhou Qiang. Phrase bracheting and annotating on Chinese language corpus[D]. Beijing: Department of Computer Science Technology, Peking University, 2002. (in Chinese)
- [9] 陈立民. 汉语的时态和时态成分[J]. 语言研究, 2002(3):14-31.
Chen Limin. Tense and tense component in Chinese[J]. Language Study, 2002(3):14-31. (in Chinese)
- [10] 刘莉莉. 汉语句义类型及谓词时态识别算法研究[D]. 北京: 北京理工大学信息与电子学院, 2010.
Liu Lili. Research on algorithms of chinese sentential semantic type and predicate aspect recognition [D]. Beijing: School of Information and Electronics, Beijing Institute of Technology, 2010. (in Chinese)

(责任编辑: 刘芳)