基于特征结构的汉语主谓谓语句语义标注研究*

陈 波[1,2,3]，姬东鸿[2]，孙 程[2]，吕 晨[2]

[1]武汉大学 文学院，武汉 430072

[2]武汉大学 计算机学院，武汉 430072

[3]襄樊学院 文学院，襄樊 441053

E-mail: cb9928@gmail.com; donghongji_2000@yahoo.com.cn; gensun.cc@gmail.com; lvchen1989@gmail.com

摘 要：建构大规模的汉语语义资源，是当前中文信息处理的重要任务之一。但是其中语义分析的传统方法存在一些问题，不能很好地反映汉语中各个词语或成分之间的语义关联。本文提出了基于特征结构的语义标注方法，并在此基础上建构了一个大规模的汉语语义资源。以汉语主谓谓语句为例，探讨了特征结构的标注方法。结果表明，特征结构分析解决了以往传统标注方法对汉语特殊句型无法表示的难题，包含更多的语义信息，其标注效率更高，标注精度也更高。

关键词：特征结构；主谓谓语句；语义标注；语义资源

Semantic Labeling of Chinese Subject-Predicate Predicate Sentence Based on Feature Structure

Chen Bo[1,2,3], Ji Donghong[2], Sun Cheng[3], Lv Chen[2]

[1] Center for Study of Language and Information, Wuhan University, Wuhan 430072

[2] NLP Lab, Wuhan University, Wuhan 430072

[3] Department of Language and Literature, Xiangfan University, Xiangfan 441053

E-mail: cb9928@gmail.com[1]; donghongji_2000@yahoo.com.cn[2]; gensun.cc@gmail.com[3]; lvchen1989@gmail.com[4]

Abstract: Constructing large scale Chinese semantic resources is one of the major tasks in Chinese Information Processing. However, the conventional approaches in semantic parsing have some defects in that they cannot denote the semantic relatedness between Chinese words and constituents. We propose a semantic annotation approach based on Feature Structure and accordingly construct large scale Chinese semantic resources. We chose "subject-predicate predicate sentence" as the research target in the paper, summarized seven categories of Feature Triples, and compared the results of three different analysis methods. Parsing based on Feature Structure solves the problems that there existed no appropriate ways to tackle the special Chinese sentence patterns. It contains more semantic information than conventional approaches, and simultaneously annotating efficiency and accuracy can be improved to a large extent.

Keywords: feature structure; Chinese subject-predicate predicate sentence; semantic tagging; semantic resource

1 前言

语义分析是现代语言学和计算语言学领域最具挑战性的课题之一，也是当前制约语言信息技术大规模应用的主要瓶颈。在众多语义分析的问题中，短语和句子级的语义分析是一项最基本的任务。汉语由于具有语序灵活、重视虚词等独特的特点，与英语法语相比，它的语义分析更具挑战性。在自然语言处理（Natural Language Processing，简称 NLP）中，对汉语语句的语义标注，一直是一个难点。其中，对于汉语特殊句型的语义标注，更是难中之难，例如"连动句"、"兼语句"、

* 本文承国家自然科学基金委员会重大研究计划"视听觉信息的认知计算"培育项目《汉语特征结构的资源建设和自动分析研究》（项目号 90820005）、2008-2009 年度武汉大学人文社科自主创新项目《基于语义的网络舆情智能监测平台研究》、武汉大学 985 项目《基于汉语特征结构的语义描写及其应用》（项目号 985yk004）、湖北省教育厅人文社科项目《基于依存语法的语料库标注研究》（项目号 2008q275）的资助。

“主谓谓语句”、“把字句”、“被动句”等等。这些句型，在语言学界它们本身的界定都存在很多争议，在NLP学界，处理的时候通常运用的是传统的分析方法。

在语言学界，主谓谓语句是汉语中一种具有独特特点的句型。作为汉语主谓句的下位句型,它的特点是由主谓短语做句子的谓语。语言学领域关于主谓谓语句的研究有80余年，至今什么是主谓谓语句、主谓谓语句有哪些类型尚未定论，这些争端包括：句中的成分谁是大主语谁是小主语的问题？倒装句是不是主谓谓语句的问题？

如何寻找一种较为有效的方法，可以对这些汉语特殊句型进行更好的语义标注，对于语言学界和NLP学界，都具有重要意义。本文提出了一种新颖的“特征结构”（Feature Structure）理论的方法，进行了大规模的语义标注，建立了一个具有近两万句的汉语语义标注资源库。在此基础上，选取汉语主谓谓语句进行进一步语义分析，得到了比较好的结果。

2 汉语主谓谓语句在语言学和NLP中研究现状

2.1 语言学界汉语主谓谓语句研究现状概述

主谓谓语句的语言学本体研究成果相当丰富，但是各位专家学者的观点却不尽相同。最早对主谓词组做谓语的论述可追溯到1921年的陈承泽，“主谓谓语句”概念的正式提出源于1984年《中学教学语法系统提要》。几十年来，各家学者的探讨主要集中在对其范围的确定、结构的分析、性质及生成的探讨上。

在语言学界，主谓谓语句语形表示为：“Nx+N+V/A”。Nx指句子的主语，也称作“大主语”，N指充当句子谓语的主谓短语中的主语，也称作“小主语”，V/A指充当句子谓语的主谓短语中的谓语。充当大主语Nx的成分一般是名词、代词、动宾短语、小句等；充当小主语N的成分一般是名词、代词、动宾短语等；充当V/A的成分一般是不及物动词、及物动词、动宾短语、形容词等。本文标注时借用洪维1998的例句，主谓谓语句的句型大致上包括十三种类型：1、Nx与N具有领属关系，如：例1：他性格坚强。2、Nx前可以加上介词，如：例2：这件事我有不同看法。3、Nx与N具有施受关系，如：例3：那个人我认识。4、Nx或N的施事具有周遍性，如：例4：他一句话也不说。5、句中包含复指成分，如：例5：这样的好同志，我们喜欢他。6、 Nx与NV具有总分关系，如：例6：他写的字，有的大，有的小。7、Nx后的两个N是对举的，如：例7：咱俩谁也别忘了谁。8、Nx表处所，如：例8：北京城里树木很多。9、Nx表时间，如：例9：工作时间你严肃一点好吗？10、Nx后是组熟语，如：例10：他这个人，事事领先人人夸好。11、Nx是N的工具，如：例11：这间屋子我们堆东西。12、N是数量结构，如：例12：这种布，一尺五毛钱。13、N（动词短语）与Nx可以构成主谓关系，如：例13：你做事认真。

2.2 NLP中汉语主谓谓语句语义标注现状及问题分析

对于语言分析，有两种传统方法：短语结构分析和依存语法分析。目前的汉语标注方法主要运用的就是这两种方法。但是运用这两种方法来标注汉语的特殊句型的语句都会遇到一些问题，如图1。

例7的特点是，大主语与充当谓语的主谓句中的主语和宾语之间是任指的关系，“咱俩”任指“谁1”、“谁2”。例12的特点是，由复杂名词短语组合而成的主谓谓语句，没有谓语动词。传统的依存语法的标注方法，在标注例7、例12的时候对于一些词语之间语义关系，无法处理，丢失了很多词语与词语之间的语义信息。可以看出，传统的方法存在以下问题：

（1）现存的句法分类仅仅适用于典型的句法结构，对于汉语中一些特殊的句型，如复杂名词

短语，动补结构，很难判定词语之间的句法特征。在复杂名词短语中，每个词都是一个名词，很难确定不同的句法特征。

（2）依存结构只能表示出两个词语之间的依存关系，一个句子中非中心词(non-head)总是依存于中心词(head)。依存语法通常被认为是单一依存，即一个节点只能依存于另一个节点。依存语法假设词语的地位是不平等的。但是，对于汉语的某些特殊结构，如复杂名词短语，有一些例外情况。比如：有时每个名词的地位是平等的，因此没有中心词。有时存在多重依存关系，即一个节点是可以依存于两个以上的节点。因此，传统句法结构分析和依存语法分析都不适合汉语语义分析。

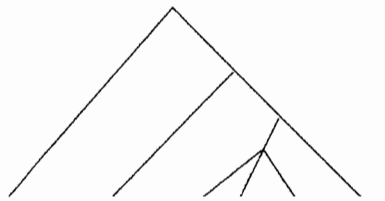
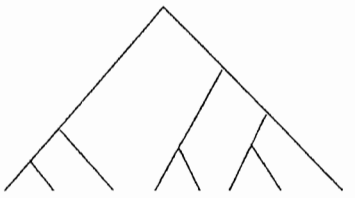
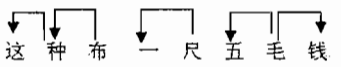
例句 / 方法	例 7：咱俩谁也别忘了谁	例 12：这种布一尺五毛钱
短语结构树		
依存结构树		
存在问题	依存语法的方法无法表示出句中 “咱俩” 与 “忘” 的语义关联、“咱俩” 与 “谁$_1$”、“谁$_2$” 的语义关联。	依存语法的方法无法表示出句中 “一尺” 与 “钱” 的关系，“一尺”、“钱” 与 “布” 之间的语义关联也无法表示。

图 1　主谓谓语句句法分析图

3　特征结构理论

特征结构 (Feature Structure) 在现代语言学和计算语言学领域并不是一个新术语。语音学很早就采用类似特征结构的机制描述音节，后来形式句法理论如 GPSG 和 LFG 又采用复杂特征集描述句法结构，复杂特征集也类似于特征结构。这两种情况都是定义一组特征用以区分音节和句法结构，分别在生成语音学和生成语法领域产生了很大影响。可是至今为止，还未见到利用特征结构进行大规模的语义描述及语义分析的尝试。

针对语义分析的两种传统方法分析汉语时遇到的难题，我们提出了“特征结构”方法来解决。

通常，一个短语或句子可以用一个特征三元组集合来表示：[实体，特征，特征值]，我们称之为这个短语结构或句子结构的 “特征结构” 集合。正如语言中有很多词语描述实体概念一样，语言中也有很多词语描述实体的特征。这些词通常称为特征词。英语 WordNet 和汉语的同义词词林都有一部分专门列出这些特征词。这里说的 “特征” 并不仅限于严格意义上的特征词，也包括那些抽象名词和虚词等，只要它们用来反映概念关联，在特征结构中就作为特征。

例 14：

1) 红颜色汽车　　2) 红汽车

在 1)中，“汽车” 是实体(entity)，“颜色” 是 “汽车” 的特征(feature)，“红” 是特征 “颜色” 的值(value)。“颜色” 一边联系 “汽车”，一边联系 “红”，因此它可作为“汽车”和“红”概念关联种类的标记。这样，1)表示成一个三元组如 1’)：

1') [汽车，颜色，红]

在 2)中，“汽车”是实体，“红”是特征“颜色”的值，值得注意的是，这里“汽车”的特征词“颜色”并没有出现。这种情况下，我们约定其特征结构中的特征为空。这个约定的好处在于不必去设计一个一般性的特征词表，而是根据具体应用的需求而制定相应的特征词表。特征词表牵涉到泛语言的范畴(包括语义格等)，如果脱离具体应用而试图设计一个一般性的特征词表，就如设计格系统一样会有很多争议。另一方面，在具体应用中只需标注少许例子，这些空的特征就可以从这些标注例子中被激活出来。根据此约定 2)表示成 2')。

2') [汽车， ，红]

例 15：他说他是大学教师。

该句的特征三元组表示为：

[说， ，他]； [说， ，他是大学教师]； [是， ,教师]； [教师， ，大学]； [是， ，他]

从例 15 我们可以发现，特征和特征值都可以作为实体出现在特征结构中。这从它们都可带一定修饰语判断出来。

“他”是“说”的特征值。“他是大学教师”是“说”的另一个特征值。这里“他是大学教师”是作为一个整体，和“说”产生语义关联。并且，特征值“他是大学教师”本身也是一个特征结构。其中，“是”是实体，“大学教师”是特征值，“他”是“是”的另一个特征值。另外，特征值的节点“大学教师”本身也是一个特征结构，“教师”是实体，“大学”是它的特征值。

形式上，一个三元组可看作两个“点”(node)和连接它们的“边”(edge)，其中的“节点”表示实体或特征值，“边”表示特征。特征一定是某个节点的特征，这个节点就作为特征拥有者，另一个节点就作为特征值。于是一个特征结构可看作一个图，而且是无向图 (undirected graph)。考虑到特征值也可是另外一个特征结构，因此特征结构可看作一个递归图，意即节点本身又可是一个图。

简言之，同句法结构相比，特征结构和依存结构类似，都主要描述词汇之间的关系，因此不用定义句法范畴。即便在递归性的特征结构中，也不用定义特征结构的类别。和依存结构相比，特征结构一方面允许嵌套，另一方面允许多重关联；另外特征结构既注重描述概念是否关联，也同时注重关联的种类。

4 基于特征结构理论的汉语主谓谓语句标注

基于特征结构理论，我们运用标注软件对这十三类主谓谓语句的语料进行了语义关系的标注，共概括出了七类标注图，如图 2。

运用特征结构的方法对例 7 和例 12 进行标注，得到图 3。

图 3 中，特征结构图可以把主谓谓语句中的每一个词和与其有语义关联的词都能用弧表示出来，如：“一尺”与“布”之间的语义关联。图中用虚线框表示的是主谓谓语句中的大主语、小主语和谓语，即 Nx+N+V/A 中的“Nx”“N”和“V/A”。因此，特征结构图包含了多重信息：第一，语句表层的结构信息；第二，传统的依存结构信息；第三，以往的方法会漏标、错标、无法标注的语义信息。我们可以发现，第一，特征结构的方法可以避免汉语语序灵活造成的句子形式多样的问题，不管句中词语的位置在哪里，都可以表示出词语之间的概念关联。这可以解决汉语中大量的倒装句的标注问题。第二，与句法结构标注和传统的依存语法相比，特征结构的方法可以表示出更加丰富的语义信息。以往标注中无法表示出的语义关系、漏标的语义关系以及错误标注的语义关系，运用特征结构的方法都可以表示出来。第三，特征结构的方法避免了很多在语言学界的纯理论的争端，如：主谓谓语句中到底哪个词语是大主语，哪个词语是小主语？

类型	例句	语义分析
[N1,N2,A]	这个人个子很高 [这个人,个子,很高]	N1 是实体，N2 是 N1 的特征，A 是 N1 的特征值。这里的 N2 是一个外在的特征，例如：数字，颜色，高度，年龄，等等；或者是一个抽象的特征，如：精神，脑子，干劲等。
[A, ,N2] [N2, ,N1]	屋里空气很沉闷 [沉闷, ,空气] [空气, ,屋里]	A 是一个实体，N2 是 A 的特征值，在下一个递归图中，N2 是实体，N1 是它的特征值。这里的 N1 表示的是 N2 的空间。
[N1, ,N2] [A, ,N3] [N3, , N1N2]	所长太太手里人还多着呢 [所长太太, ,手里] [多, ,人] [人, ,所长太太手里]	A 是实体，N3 是 A 的特征值。在下一个递归图中，N3 是实体，N1 和 N2 同时是 N3 的特征值。在第三个递归图中，N1 是实体，N2 是它的特征值。这里 N1 和 N2 合在一起，表示 N1 的空间。
[V, ,N1] [V, ,N2] [V, ,N3] [N1, ,N2]	农民们谁也没租过房子住 [租, ,农民们]; [租, ,谁] [租, ,房子]; [农民们, ,谁]	V 是实体，N1/ N2/ N3 是它的特征值。在第二个递归图中，N1 是实体，N2 是它的特征值。N2 和 N1 的关系是任指或者虚指。
[V, ,N1] [V, ,N2] [V, ,N3] [N1, ,N2] [N1, ,N3]	我们谁 $_1$ 也离不开谁 $_2$! [离不开, ,谁 $_1$] [离不开, ,谁 $_2$] [离不开, ,我们] [我们, ,谁 $_1$][我们, ,谁 $_2$]	V 是实体，N1/ N2/ N3 分别是它的特征值。在第二个递归图中，N1 是实体，N2 和 N3 是它的特征值。N2N3 和 N1 之间的关系是任指或者虚指。
[V, ,N2] [N1, ,N2]	这石象眼睛直盯着远方。 [盯, ,眼睛] [石象, ,眼睛]	V 是实体，N1/ N2/ N3 分别是它的特征值。在第二个递归图中，N1 是实体，N2 是它的特征值。N2 是 N1 的一部分，这里 N1 大多是人类或者其他生物，N2 属于 N1 身体的一部分。N1 和 N2 是领属的关系。
[N1,N2,N3] [V, ,N1] [V, ,N3]	这个宣传委员名叫周天桂。 [宣传委员, 名,周天桂] [叫, ,宣传委员]; [叫, ,周天桂]	V 是实体，N1N3 分别是它的特征值，在第二个递归图中，N1 是实体，N2 是它的特征，N3 是它的特征值。

图 2　主谓谓语句特征三元组

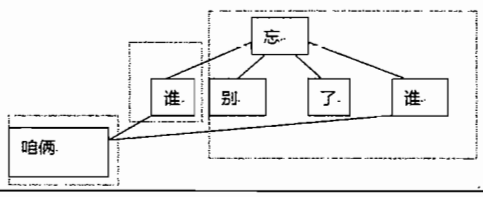
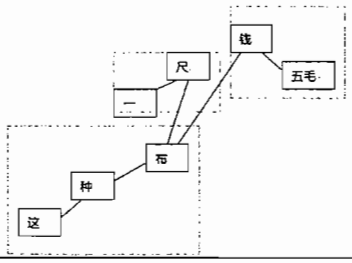
例句	特征结构图
咱俩谁也别忘了谁。	
这种布，一尺五毛钱。	

图 3　主谓谓语句特征结构图

5 结论及展望

我们可以看到，运用特征结构对句子进行标注可反映出哪些成分充当实体，哪些充当特征，哪些充当特征值，这些词语之间的语义关系也很清晰地反映出来。今后运用特征结构标注的资源，通过训练，就有可能抽取出句子中隐含的语义关系。

特征结构分析有如下优点：

（1）标注的是语义关联，而非句法关联。我们标注的是句中词语与词语之间的语义关联，跟句子表层的句法结构无关，因此跳过了句法层面的分析。

（2）标注的是“关联”而非“依存”。我们表示的是语义上的关联，而不是传统的依存关系。因此我们的标注图用“无向图”表示，也弱化了中心词的概念。

（3）标注效率更高。特征结构的方法不牵涉词性争议、结构歧义等问题，也无需判断中心词，因此标注效率比句法标注和依存标注要高。

（4）标注的结果一致性高。我们的判断标准是基于关联，经过人工标注，最后得到的标注结果分歧较少。

特征结构的理论是我们的一个新尝试，现在我们已经建立了特征结构的基本概念和描述框架，建构了一个大规模的汉语语义资源，并且应用到了食谱分析、国家安全信息收集和分析、汽车市场情报分析等领域，取得了比较好的效果。

但是在标注过程中，仍然存在一些不可避免的难题，如：不断发展变化的语言永远无法穷尽列举，真实语料中会出现很多语言的临时用法和特例，针对这类极少部分的语例，我们该如何制定规则确定特征结构？这是我们下一步工作要解决的问题。

参 考 文 献

[1] A. Spencer. Phonology[M]. MA: Blackwell Publishing Ltd, 1996.

[2] J. Bresnan. Lexical Functional Syntax[M]. MA: Blackwell Publishers, 2001.

[3] G. Gerald, E. H. Klein, G. K. Pullum, I. A. Sag. Generalized Phrase Structure Grammar[M]. MA: Harvard University Press, 1985.

[4] M. Dalrymple. Lexical Functional Grammar[C]. No. 42 in Syntax and Semantics Series. New York: Academic Press, 2001.

[5] J. F Allen. Natural Language Understanding[M]. 1987, 2nd ed. Benjamin Cummings, 1994.

[6] N. Chomsky. The minimalist program[M]. MIT Press, Cambridge, MA, 1995.

[7] N. Chomsky. Syntactic Structures[M]. Berlin: Mouton de Gruyter, The Hague, 1957.

[8] N. Chomsky. Some Concepts and Consequences of the Theory of Government and Binding[M]. MA: MIT Press,1982.

[9] J. M. Eisner. Bilexical grammars and their cubic-time parsing algorithms[C]. In Harry Bunt and Anton Nijholt, editors, Advances in Probabilistic and Other Parsing Technologies, Dordrecht: kluwer Academic Publishers, pp. 29-62, 2000.

[10] C. Fellbaum . WordNet: An Electronic Lexical Database[M]. MIT Press. May 1998.

[11] I. Mel'˘cuk. Dependency Syntax: Theory and Practice[M]. State University of New York Press. 1988.

[12] 李临定. 现代汉语句型[M]. 北京: 商务印书馆, 1986.

[13] 陆俭明. 新中国语言学 50 年[J]. 当代语言学, 1999, (4).

[14] 朱德熙. 语法答问[M]. 北京: 商务印书馆, 1985.

[15] 范晓. 关于汉语的语序问题(一) [J]. 汉语学习, 2001, (5).