

文章编号:

基于特征结构的汉语连动句语义标注研究*

陈波^{1,2}, 姬东鸿², 吕晨²

(1. 湖北文理学院 文学院, 湖北 襄阳 441053; 2. 武汉大学 计算机学院, 湖北 武汉 430072)

摘要: 对汉语特殊句型的语义分析是当前中文信息处理的难点之一。现有的传统语义分析方法存在一些问题, 不能很好的反映汉语中各个词语或成分之间的语义关联。本文以汉语连动句为例, 提出了基于特征结构模型的语义标注方法, 探讨了连动句的语义标注模型, 并在此基础上建构了一个大规模的汉语语义资源。结果表明, 特征结构模型能够对连动句中的主语与多个谓语动词、多个宾语之间的复杂语义关系进行全面准确的描述, 为面向汉语的自然语言处理提供了一种不同的语义分析方法。

关键词: 特征结构; 连动句; 语义标注; 语义资源

中图分类号:

文献标识码:

Semantic Labeling of Chinese Serial-Verbs Sentence Based on Feature Structure

Bo Chen^{1,2}, Donghong Ji², ChenLv²

(1. Department of Language and Literature, Hu Bei University of Arts & Science, Xiangyang, Hubei 441053, China;

2. School of Computer, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: Parsing Chinese Special sentence patterns is one of the major tasks in Chinese Information Processing. The Current conventional approaches in semantic parsing have some defects in that they cannot denote the semantic relatedness between Chinese words and constituents. In the paper, we choose "Serial-Verbs sentence" as the research target, propose a semantic annotation approach based on Feature Structure, and study the semantic annotation models of "Serial-Verbs sentence". Feature Structure Model provides a different semantic parse approach for Chinese Information Processing, which can represent the complicated semantic relations among the subject, the predicates and the objects of "Serial-Verbs sentence".

Keywords: Feature Structure; Chinese Serial-Verbs Sentence; semantic annotation; semantic resource

1 语言学界连动句研究

朱德熙(1985), 黄伯荣(1991), 范晓(1980), 陈昌来(2000)认为, 连动句的特点是由两个或两个以上的动词短语连用构成了句子的谓语, 这些动词短语有一个共同的主语(施事), 动词短语之间结构紧凑, 也不能有句法关系(如: 主谓关系、偏正关系、动宾关系、动补关系等), 动词短语的出现一般是按照动作发生的时间顺序或者动作的逻辑顺序依次排列的。如:

例1. 小宋取了存款去买衣服了。

例2. 小宋看论文看累了。

例3. 小宋有权利打电话投诉。

例1中, “取”和“去”、“买”是时间上先后发生的顺序; 例2, “看论文”和“看累了”是因果关系;

*收稿日期:

定稿日期:

基金项目: 国家自然科学基金资助项目(61202193,61070082); 国家哲学社会科学重大项目招标项目(11&ZD189); 湖北省重点学科建设项目学科成果; 湖北文理学院院级项目(2010YB009)。

作者简介: 陈波(1976—), 女, 博士后, 主要研究方向为自然语言处理; 姬东鸿(1967—), 男, 教授、博导, 主要研究方向为自然语言处理; 吕晨(1989—), 男, 博士研究生, 主要研究方向为自然语言处理。

例3是条件关系。

连动句可以描述为：

“主语+动词短语 1+动词短语 2+.....”

语形可以表示为：

“N1+V1+N2+V2+N3+.....”

连动句的语义关系包括两大部分：“主语与构成谓语的若干动词短语之间的语义关系”和“构成谓语的若干动词短语之间的语义关系。”在语言学领域，关于连动句的语义关系描述，成果非常丰富。李临定(1986)，邢欣(1987)，杨月蓉(1992)主要研究方法是先用“施事、受事”等语义格对主语和几个谓语动词进行描述，再用“方式、状态、时间、处所、方面、依靠、目的、讯号、动作发生时间顺序、因果关系、假设关系、同义或近义”等对谓语动词之间的关系进行描述。

表1 连动句语义关系研究分类表

	主语与谓语之间的语义关系		谓语动词之间的语义关系	
范晓	2种	主语为主事；主语为客事。	5种	先后；动作方式；目的；因果；互补。
陈昌来	3种	施事-动作、受事-动作；主语是一个V的施事，同时又是另一个v的受事。	6种	先后、方式、目的、因果、互补、“有”字型
吴启主	4种	主语为施事、受事、施受同体，非施非受	9种	先后、目的、方式、原因、工具、手段、处所、时间、情况的说明等等
杨月蓉	4种	施事-动作；受事-动作；主语是V1的施事v2的受事；主语是V2的施事v1的受事		略
黄伯荣			8种	同陈六种.还包括两种：V1=V2，V1P 动宾，V2P 动补，表示动作的对象和结果；V1V2 正反两面描述主语
李临定	2种	施事-动作；受事-动作	12种	方式、状态、时间、处所、方面、依靠、目的、讯号、动作发生时间顺序、因果关系、假设关系、同义或近义

从表1可以看出，语言学上的这种语义关系的描述还是很细致的，对主语(N1)、动词(V1)、名词(N2)、动词(V2)、名词(N3)之间的语义关系从多个角度进行了分析。尽管，这种分类在语言学领域很有意义，但是面向自然语言处理时，这种语义分类的成果却难以形式化或制定严格的分类标准，描述时颗粒度过细会造成研究的复杂和繁琐；描述的颗粒度过粗又会丧失研究意义。

2 NLP 中连动句语义标注难点所在和问题分析

语言处理的分析方法以短语结构分析法和依存分析法为代表。

2.1 连动句的短语结构标注

短语结构语法主要是从表层的句法结构层面进行分析。如：

例4. 我开车去车站接他

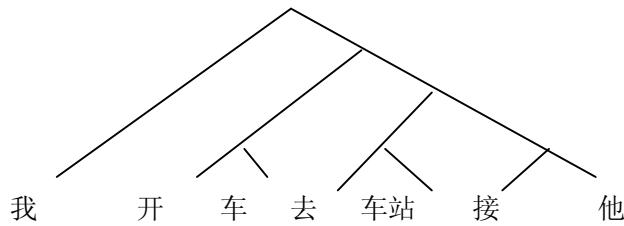


图 1 例 4 短语结构分析图

短语结构分析法能够很好地把连动句表层的句法结构描述出来，却没有涉及语义信息。

2.2 连动句的依存分析标注

传统依存公理认为，一个句子中有且只能有一个中心动词，其他词语都依存于该中心词。这个公理在处理汉语连动句时带来很多难题。汉语连动句至少包含两个动词 V1、V2，有时还有更多的动词 V3、V4 等。如果要运用传统的依存语法来分析汉语连动句，就必须强制地规定出一个中心动词来，然后其他的动词都依附于它。那么如何来寻找这个中心动词呢？

陈波(2007)研究表明，目前国内建设标注语料库时，大多运用的是传统依存语法。传统依存语法从法语和英语的标注研究中发展而来，没有遇到过汉语连动句这种有两个或两个以上的动词短语连用构成句子的情况。国内研究机构在研究汉语连动句时发现，连动句中 V1、V2 之间的语义关系很复杂，很难确定出这些连续的动词的语义地位孰轻孰重，很难确定哪一个是中心词。但是如果运用传统依存分析方法来描述连动句，就必须找一个动词中心词来。

在当前的中文信息处理中，针对连动句，目前采取的方法是：人为地规定，连续出现的若干个动词 V1、V2 等，第一个动词 V1 被默认为中心词，后面的动词 V2 等都依存于它，如哈工大(1996)。传统依存语法分析连动句方案如下图：

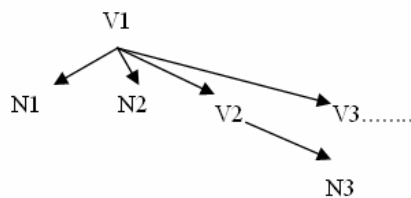


图 2 连动句传统依存分析方案

因此运用传统依存分析法，例 4 的分析结果为：

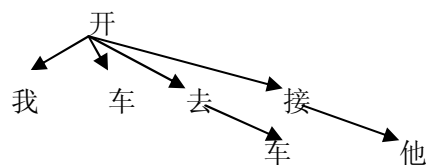


图 3 例 4 的传统依存分析树

例 4 中，主语 N1 是“我”；第一个动词短语 V1 是“开”，N2 是“车”；第二个动词短语 V2 是“去”，N3 是“车站”；第三个动词短语 V3 是“接”，N4 是“他”。

根据语义关联，我们按照语义认知来分析，例 4 中具有语义关系的词语对(word pair)至少有：(我，开)；(我，去)；(我，接)；(开，车)；(去，车站)；(接，他)。

除了上述信息外，该句还包含其他语义信息。比如，我们还可以提问：

在哪里接? 答案: 车站。因此, 可描述为:

[接, (在), 车站]

怎么接? 答案: 用车接。 可描述为:

[接, (用), 车]

怎么去? 答案: 开车。 可描述为:

[去, (方式), 开车]

图 3 是传统依存语法对例 4 的分析结果, 可以看到, 运用依存分析法的分析结果, V1“开”被默认为中心词, V2“去”、 V3“接”都被默认地依存于“开”。主语“我”仅跟 V1“开”发生语义联系, 而“我”与 V2“去”、 V3“接”的语义关系却没有表示出来。而“接”和“车”、“接”和“车站”的语义关系也没有表示出来。

因此, 从例 4 可以看出, 现有的传统依存分析法在分析汉语连动句时丢失了很多语义信息, 为进一步的语言处理工作如信息抽取等带来一定的困难。

3 连动句的特征结构标注

3.1 特征结构模型

冯志伟(2004)认为, 找出句子中单词之间的语义关系才是机器翻译中语义分析的最终目的。这种语义关系与单词本身的语义特征不同, 而且与单词在句子和文中的上下文密切相关。陈波(2011,2012)提出, 对句子或短语内部概念关联和关联的种类的描写是特征结构资源的基本任务。

例5. 从广州飞 飞到武汉

例 5 描述为(飞, 从, 广州), (飞, 到, 武汉), “飞”和“广州”、“飞”和“武汉”都存在概念关联, 如果在这两个关联对中添加上关联种类的话, 就是“飞-从(起点)-广州”, “飞-到(终点)-武汉”, 直观上, 这个三元组正好可以理解为“实体、特征、特征值”关系。一般来说, 一个短语或句子可表示成一个由实体(Entity)、特征(Feature)和特征值(Value)组成的三元组(Triple)的集合, 这个集合我们称之为该短语或句子的特征结构(Feature Structure), 如下:

特征结构三元组: [实体, 特征, 特征值] Feature Triple: [Entity, Feature, Value]

一个特征三元组反映一对语义关联, 例 5 可以用特征三元组表示为:

[飞, 从, 广州] [飞, 到, 武汉]

姬东鸿(2000)提出, 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。形式表达上, 如图 4 所示, 特征三元组可以用一个无向图来表示, 并允许递归。

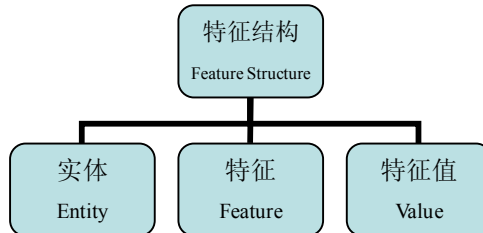


图 4 特征结构三元素组成图

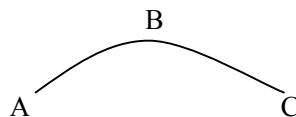


图 5 一个特征三元组形式化表示

如图 5 所示, A 表示实体, B 表示特征, C 表示特征值。B 反映了 A 和 C 之间的语义关联。

上图语义是，实体 A 的特征 B 的值是 C。陈波(2008, 2011)提出，特征结构图是一种可递归的、允许多重关联的、无向的、图结构，如例 6 所示：

例6. “小王说自己是大学教师”

其特征结构三元组如下，特征结构图如图六：

- [说, , 小王]
- [说, , 自己是大学教师]
- [自己小王, , 小王]
- [是, , 自己]
- [是, , 教师]
- [教师, , 大学]

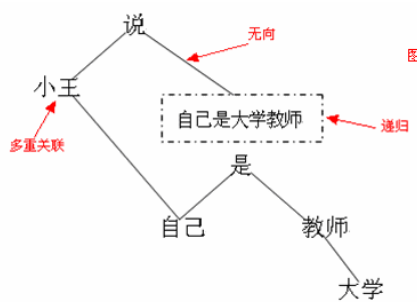


图 6 例 6 的特征结构图

3.2 连动句特征结构模型分析

我们以典型意义的连动句为语义描述对象，《现代汉语句型》（李临定，1986）对连动句分类较详细，我们选取了其中的例句，加上目前语言学界争议较多的语例，共 16 句作为考察对象，运用特征结构模型细致地分析了这 16 个例句的特征结构三元组和特征结构图。因篇幅关系，选取三个例句分析如下：

例7. 我推开门走出去。

这是语言学家公认的一个典型的连动句。动词短语 1 是“推开门”，动词短语 2 是“走出去”，两者是按时间顺序发生的连续动作。它的特征结构三元组和特征结构图如下：

- [推, , 我]
- [推, , 开]
- [推, , 门]
- [走, , 我]
- [走, , 出去]

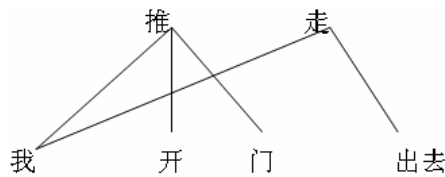


图 7 例 7 特征结构图

例8. 我买了碗面吃。

例 8 的动词短语 1“买了碗面”，“面”是 V1 的宾语；动词短语 2 没有宾语，只有一个及物动词“吃”。例 8 的特点是 V1 的宾语“面”是 V2 吃的受事，两者存在着语义关联。它的特征结构三元组和特征结构图如下¹：

¹ 本例的另一特点是“一碗面”中的数词“一”省略了，这也是汉语的一个特点，在图中我们用“\$”表示。

[买, , 我]
 [买, , 了]
 [买, , 面]
 [面, 碗, \$]
 [吃, , 我]
 [吃, , 面]
 [买, , 吃]

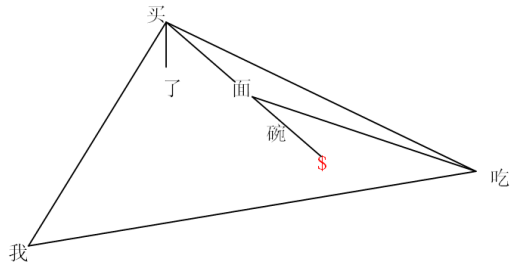


图8 例8 特征结构图

例9. 他穿着一件大衣走进来。

例9的动词短语1“穿着一件大衣”，动词短语2“走进来”，动词短语1是对主语状态的描述，与动词短语2是无语义关联的。它的特征结构三元组和特征结构图如下：

[穿, , 他]
 [穿, , 着]
 [穿, , 大衣]
 [大衣, 件, 一]
 [走, , 进来]
 [走, , 他]

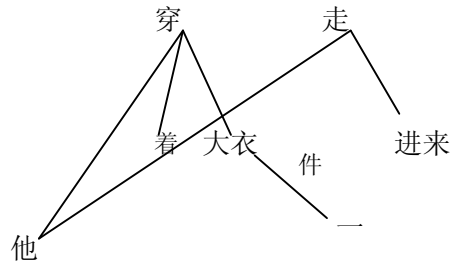


图9 例9 特征结构图

4 不同理论的连动句标注分析比较

4.1 标注结果对比

我们以例8为例，进行特征结构分析与传统依存分析，将分析结果进行对比，来评测对连动句的语义分析效果。

表2 两种分析结果对比表

传统依存分析树	特征结构图

分析结果对比图：

词语与词语之间的语义关系		特征结构分析结果	传统依存方法分析结果
有	无		
我, 买		+	+
买, 了		+	+
买, 面		+	+
(一) 碗, 面		+	+
吃, 面		+	-
我, 吃		+	-
	买, 吃	-	+

图 10 两种分析结果对比表

如图, 传统依存分析结果没有把“我”和“吃”、“吃”和“面”的语义关系表示出来, 但是却把没有语义关系的“买”和“吃”之间的语义关系标注出来。因此它漏掉了两个语义关系对, 还标注了一个没有语义关系的语义关系对。

4.2 特征结构模型的优点

一、特征结构模型能够描述更多的语义关系对, 因此包含更加丰富的语义信息。

(一) 传统依存语法无法表示连动句中主语和除第一个谓语动词之外的其他谓语动词之间的语义关系。特征结构模型可以完整地表示连动句中主语和所有谓语动词之间的语义关系。

如: “我买了碗面吃”中“我”和“吃”的语义关系, 传统依存分析无法表示但是特征结构模型可以表示出来。

(二) 传统依存语法无法表示连动句中某个谓语动词的宾语与其他谓语动词之间的语义关系, 特征结构模型可以完整地表示它们之间的语义关系。

如: “我买了碗面吃”中“面”和“吃”的语义关系, 传统依存分析无法表示但是特征结构模型可以表示出来。

(三) 连动句中, 两个或多个谓语动词之间, 有时存在语义关系, 有时不存在语义关系。传统依存语法无法辨析这两种情况, 不管实际的语义关系是否存在, 为了标注的方便, 传统依存语法强制规定第一个动词为全句中心词, 其他动词都直接依存于它。这种处理方法, 显然违背了语言处理的目标“语义理解”。特征结构模型可以按照语言实际情况, 如实地反映这些谓语动词之间的语义关系。

如: 例句“他站着吃饭”中“站着”和“吃”之间恰巧有语义关系, 传统依存分析方法能把它表示出来; 例句“我买了碗面吃”中“买”和“吃”没有直接的语义关系, 而传统依存分析方法仍然把它表示了。特征结构模型的处理结果是前者有语义关系而后者没有语义关系。

二、特征结构模型能够对传统依存分析法不能解释的语言现象做出解释。

(一) 有助于对连动句句式的判定

在语言学界, 争议较大的连动句类型, 一种是“N1+V1+着+V2+N2”, 这类句型的争论焦点是: 到底是连动句还是“V1+着”充当状语修饰 V2 的一般主谓句? 另一种是有关联词的连动句, 到底是连动句还是紧缩复句?

我们先来看看连动句典型的特征结构图。每一种句式都有着自己的独特特点, 即使偶尔有变形的情况, 也是基于独特特点的情况下进行的变形。连动句的典型特征结构图是:

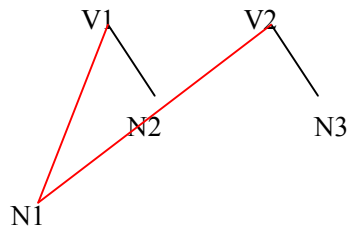


图 11 连动句的典型特征结构图

如图 11 所示，连动句的特点是要求主语 N1 与动词 V1、动词 V2 之间必须存在语义关联。

先来看第一种“N1+V1+着+V2+N2”句式，这类句式的例句有很多：

例10. 他站着吃饭。

例11. 他躺着看小说。

例12. 他穿着一件大衣走进来。

这三句是不是连动句呢？我们先不管动词 V1 和动词 V2 之间的语义关系，关键来看主语“他”和动词短语 1 中的“动词”是否存在语义关联。这三个句子的特征结构图是：

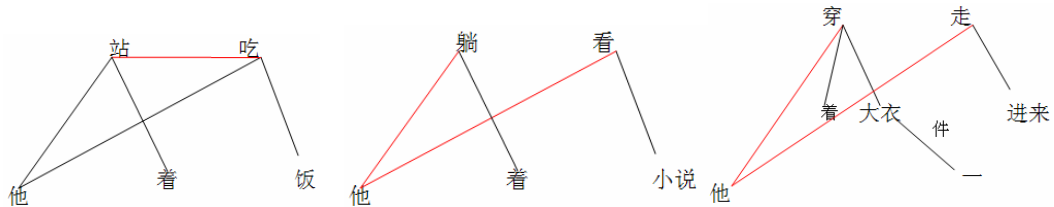


图 12 例 10-12 的特征结构图

显然，这三个例句中主语和第一个动词之间存在着语义关联，因此，这三个例句应该都属于连动句。

我们再来看第二种有关联词语的句式的辨析，下面有四个例句：

例13. 张三一来李四就走了。

例14. 张三一来就哭了。

例15. 李四拿了笔记本就走了。

例16. 李四拿了笔记本走了。

按照有的学者的观点，例 13-15 三句中含有关联词语“一……就”和“就”，所以是紧缩复句；例 16 句没有关联词语，所以是连动句。这种观点是否有道理呢？

如果运用传统依存分析法，这四句的依存分析如下：

由于传统依存分析法无法处理成对的关联词语“一……就”，例 13、14 两句的分析结果为：

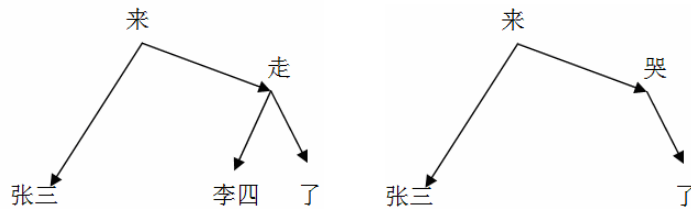


图 13 例 13-14 的依存树

例 15、16 两句的依存分析树为：

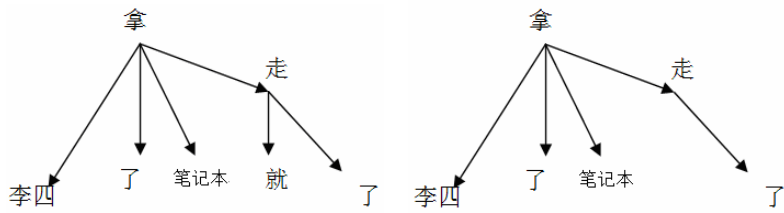


图 14 例 15-16 的依存树

通过图 13、图 14 的四个依存分析树我们可以发现，传统依存分析法无法分析有着成对关联词语的句子，因而无法分析这些分句之间的语义关系。在分析有着单个关联词语的句子时，也无法表示出主语和两个谓语动词之间的语义关系；而且不管这两个动词之间是否具有语义关系，都把二者标注在一起。

我们运用特征结构模型来分析例 13-16，结果如图 15：

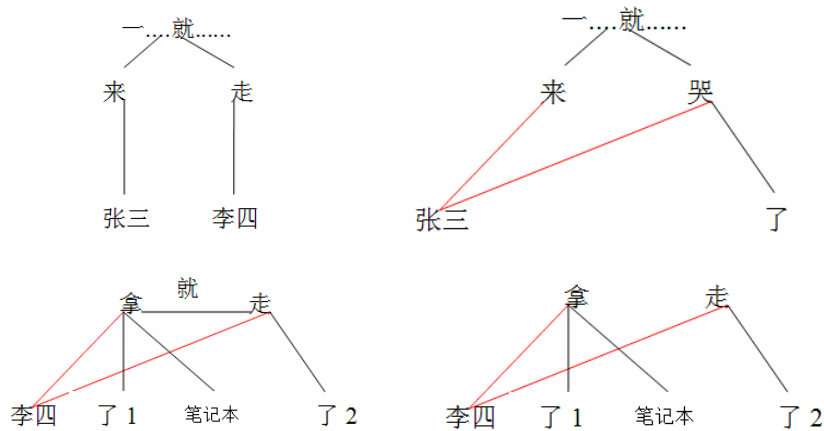


图 15 例 13-16 的特征结构图

根据分析图我们发现，很明显例 13 不符合典型的连动句特征结构图，应该属于紧缩复句，而例 14-16 三句虽然互不相同但是都具有连动句的特点，因此应该属于连动句的类型。

(二) 有助于对杂糅句式的分析

汉语与其他语言相比，特别注重意合，是一种语义关系非常复杂的语言。有时，一个句子可能杂糅着若干个句型，句中各个词语之间的语义关系错综复杂。如例 17：

例 17. 首长要他们派车送二排长入院。

例 17 由四个动词短语构成，每个动词短语各自带有宾语，这些词语之间是什么语义关系呢？这个句子包含哪些句型呢？

如果运用传统依存分析法，它的依存树如下：

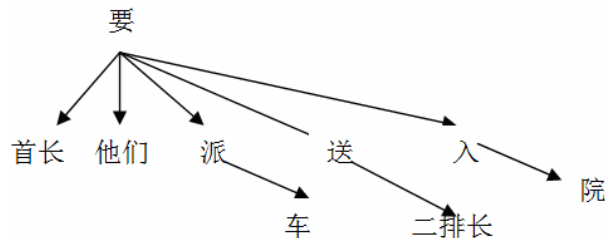


图 16 例 17 的依存树

如图 16，传统依存分析法无法表示出句子中“他们”和“派”、“车”和“送”、“他们”和“送”、“二

排长”和“入”之间大量的语义关系。

运用特征结构模型，它的特征结构图如下：

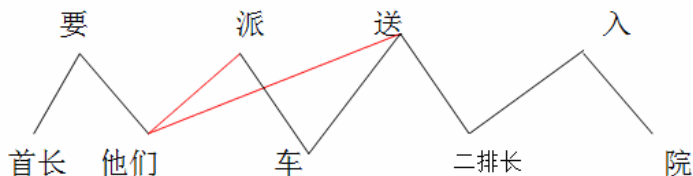


图 17 例 17 的特征结构图

通过特征结构分析，我们可以发现该句中的“他们派车送二排长”这部分属于连动句式。整个句子分为三个部分，第一部分“首长要他们派……”是一个兼语句式，中间“他们派车送二排长”是一个连动句式，第三部分“送二排长入院”又是一个兼语句式，因此通过特征结构分析，我们可以辨析出：该句由两个兼语句式和一个连动句式杂糅而成。

5 结论

本文回顾了语言学界对连动句的研究成果和争论焦点，总结了汉语连动句的语法特点，然后分析了面向自然语言处理时连动句的标注难点。运用特征结构模型对语言学界讨论较多的 16 个连动句分别进行了细致地语义描述和分析，总结出了四类语义模型。将现有的传统依存分析方法和特征结构分析方法对连动句的分析结果进行了对比，结果表明，传统依存语法无法表示连动句中主语和除第一个谓语动词之外的其他谓语动词之间的语义关系，无法表示连动句中某个谓语动词的宾语与其他谓语动词之间的语义关系，也无法准确表示两个或多个谓语动词之间的语义关系。与传统依存分析法相比，特征结构模型能够描述更多的语义关系对，因此包含更加丰富的语义信息。另外，特征结构模型能够对传统依存分析法不能解释的语言现象做出解释，比如对连动句句式的判定、对连动句和紧缩复句的区分、对复杂的杂糅句式的语义分析等。特征结构模型为面向汉语的自然语言处理提供了一种不同的语义分析方法。

参考文献

- [1] 朱德熙. 语法问答[M].北京:商务印书馆,1985: 95-110.
- [2] 黄伯荣. 现代汉语[M].北京:高等教育出版社,1991:122-129.
- [3] 范晓. 汉语的句子类型[M].太原:书海出版社,1980:68-78.
- [4] 陈昌来. 现代汉语句子[M].上海:华东师范大学出版社,2000:167-171.
- [5] 李临定. 现代汉语句型[M].北京: 商务印书馆,1986: 302-307.
- [6] 邢欣. 简述连动式的结构特点及分析[J].新疆大学学报,1987,(1):116~122.
- [7] 杨月蓉. 连动句和兼语句中的语义关系——兼论连动式与兼语式的区别[J].西南师范大学学报,1992,(4):96~100.
- [8] 陈波. 基于语料库的依存语法标注研究[C].中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集,北京: 电子工业出版社,2007:286~289.
- [9] 冯志伟. 机器翻译研究[M].北京:中国对外翻译出版公司,2004:467-469.
- [10] Ji Donghong. Semantic annotation of Chinese phrases using recursive-graph[C]. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong: Association for Computational Linguistics, 2000:101~108.
- [11] Chen Bo. Labeling corpus with dependency relations[C]. Recent Advance of Chinese Computing Technologies——Proceeding of ICC 2007.Singapore: Chinese and Oriental Language Information Processing

Society, 2008:205~208.

- [12] Chen Bo, Ji Donghong, Zhang Mingyao. Chinese Semantic Resource Construction Based on Feature Structure[C]. The 11th Chinese Lexical Semantics Workshop (CLSW2011).TaiBei.2011: 324~331.
- [13] Chen Bo, Ji Donghong. Chinese Semantic Parsing Based on Dependency Graph and Feature. [C] In Proceedings of the First International Conference on Electronic & Mechanical Engineering and Information Technology (EMEIT 2011), Haerbin, China. 2011: 1730-1734.
- [14] 陈波. 特征结构及其汉语语义资源建设[D].武汉大学,博士论文,2011.
- [15] 陈波. 基于特征结构的汉语主谓谓语句标注[J].中文信息学报,2012,(3):22~25.