

汉语虚词用法在依存句法分析中的应用研究*

咎红英, 张静杰, 娄鑫坡

(郑州大学信息工程学院, 河南 郑州 450001)

摘要: 虚词在现代汉语中占有重要地位, 虚词与词序一起构成现代汉语的句法手段, 对句法分析有重要的影响。依存句法分析是自然语言处理领域研究的热点, 为了提高依存关系的识别效果, 本文考虑将虚词用法应用到依存关系的识别过程中。通过对虚词用法的研究, 以及对依存句法分析各种依存关系识别情况的分析, 发现并列关系与虚词中的连词关系密切。作者在并列关系识别过程中加入连词的用法信息, 从而提高了并列关系的识别效果。实验结果表明, 包含连词的并列关系的 LAS 及 UAS 分别提高了 3.43% 和 2.29%。

关键字: 虚词用法; 依存句法分析; 并列关系

Studies on the Application of Chinese Functional Words' Usages in Dependency Parsing

ZAN Hongying, ZHANG Jingjie, LOU Xinpo

(School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001)

Abstract: Functional words play an important role in modern Chinese words, and constitute the syntactic means of Chinese with the word order, so functional words have an important influence on the syntactic analysis. Dependency parsing is a hotspot of research in the field of natural language processing. In order to improve the recognition effect of dependence relation, the usages of functional words are applied to the recognition process of dependence relation in this paper. Through the study of functional words usages, as well as the analysis of dependence relation in the dependency parsing, it found that the coordination relation has close connection with conjunction. And then, the conjunction usages are considered in the recognition process of coordination relation to improve the recognition performance. The experimental results show that, through considering the conjunction usages, the LAS and the UAS of coordination relations have increased 3.43% and 2.29% respectively.

Keywords: functional words usages; dependency parsing; coordination relations

1 引言

句法分析主要是对输入的词语序列(一般为句子)判断其构成是否合乎给定的语法, 分析得出合乎语法的句子的句法结构^[1]。句法分析主要有两种思路, 短语结构句法分析和依存句法分析。其中依存句法形式简单, 只有依存弧和相应的依存关系标签, 而不包含非终结符等标记, 也不需要对比句子结构进行逐层的分析, 只需考虑两个词语之间的关系。依存句法分

基金项目: 国家高技术研究发展 863 计划 (2012AA011101); 国家自然科学基金 (60970083, 61272221); 中国科学院自动化研究所模式识别国家重点实验室开放课题基金; 以及河南省教育厅科学技术研究重点项目 (12B520055)。

作者简介: 咎红英 (1966-), 女, 河南焦作人, 博士, 教授, 主要研究方向为自然语言处理; 张静杰 (1986-), 女, 硕士, 主要研究方向为自然语言处理; 娄鑫坡 (1987-), 男, 硕士, 主要研究方向为自然语言处理。

析的这些特点更适合汉语中的意合以及句法不严格的特点，因此，汉语的依存句法分析逐渐成为研究的热点。

现代汉语中的虚词是汉语句子的的重要组成部分，同一个虚词在不同的上下文中显示不同的词性，同一个词性也可以有不同的用法，且虚词与词序一起构成中文的句法手段^[2]，因此虚词的用法，即虚词的上下文语境信息，对整个句子的理解有着重要作用。现有的依存句法分析还很少考虑到句子中虚词用法的影响，本文将虚词用法应用到汉语依存句法分析中，期望提高汉语依存句法分析的性能。

本文主要内容组织如下：第 2 部分介绍依存句法分析以及虚词用法的相关研究工作；第 3 部分主要阐述虚词用法对依存句法分析中依存关系识别的性能影响；第 4 部分为结论以及进一步的工作。

2 相关研究

依存语法理论是由法国语言学家 Lucien Tesnière 创立的^[3]，关于依存句法分析的算法主要有四类：生成式的分析方法、判别式的分析方法、决策式的分析方法和基于约束满足的分析方法。本文对汉语句子的依存句法分析，采用了哈尔滨工业大学社会计算与信息检索研究中心研制的依存树库 (HIT Chinese Dependency Treebank, HIT-IR-CDT)^[4]以及语言技术平台 (Language Technology Platform, LTP)^[5]提供的分析工具，其中的依存句法分析模块是对 McDonald 等人的最大跨度树 (Maximum Spanning Tree, MST) 模型^{[6][7][8]}的实现。

LTP 提供了一系列的自底向上的汉语语言处理模块，其中包括分词、词性标注、命名实体识别、依存句法分析、词义消歧、语义角色标注等六项处理技术，为汉语的相关研究建立了平台^{[9][10]}。其中依存句法分析依赖分词和词性标注的结果，而本文着重研究虚词的用法对依存句法分析的影响，所以采用分词和词性标注正确的结果进行依存句法分析，减少因分词和词性标注错误对依存句法分析的影响，进而方便对依存句法分析结果的分析和研究。

在虚词用法的研究方面，俞士汶等^[11]提出了“三位一体”构建现代汉语广义虚词知识库的想法和思路，咎红英等^[12]对现代汉语虚词的用法进行了大量的分析和研究，并完成了虚词知识库的构建工作。在这些研究的基础上，周丽娟^[13]等对虚词用法进行了自动识别研究，并取得了较好的效果。

3 虚词用法在依存句法分析中的应用

3.1 LTP 中依存关系识别情况

HIT-IR-CDT 中的依存关系共有 24 种，作者对该树库中最后 1000 个句子的测试实例进行依存句法分析，依存关系的识别情况如表 1 所示，其中“-R-”表示依存关系的父节点识别正确，但是依存关系识别错误的个数，“-PR-”表示父节点识别错误的个数，LAS 表示依存关系的识别准确率，UAS 表示依存弧的识别准确率。

表 1 24 种依存关系的识别情况

关系类型	符号	结点数	-R-	-PR-	UAS(%)	LAS(%)
定中关系	ATT	4755	11	174	96.34	96.11
状中结构	ADV	2704	43	64	97.63	89.39
动宾关系	VOB	2358	62	66	97.20	94.57
主谓关系	SBV	1613	42	48	97.02	94.42
“的”字结构	DE	1081	1	41	96.21	96.11
核心	HED	1000	0	19	98.10	98.10
介宾关系	POB	817	4	21	97.43	96.94
连动结构	VV	730	141	26	96.44	77.12

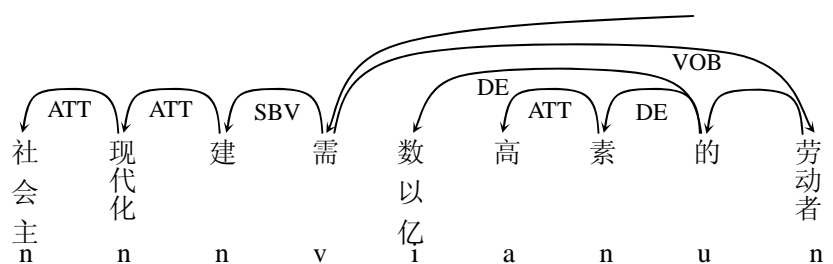
并列关系	COO	701	54	58	91.72	84.02
数量关系	QUN	618	5	4	99.35	98.54
独立分句	IC	334	110	28	91.62	58.68
语态结构	MT	308	0	14	95.45	95.45
前附加关系	LAD	262	2	8	96.95	96.18
动补结构	CMP	233	16	5	97.85	90.99
关联结构	CNJ	226	6	3	98.67	91.59
后附加关系	RAD	111	0	3	97.30	97.30
独立结构	IS	89	14	7	92.13	76.40
同位关系	APP	77	14	7	90.91	72.73
“地”字结构	DI	33	0	1	96.97	96.97
依存分句	DC	20	15	1	95.00	20.00
“得”字结构	DEI	12	0	0	100.00	100.00
比拟关系	SIM	4	0	1	75.00	75.00
“把”字结构	BA	0	0	0	——	——
“被”字结构	BEI	0	0	0	——	——

在 24 种依存关系中，定中关系（ATT）、后附加关系（RAD）、前附加关系（LAD）等依存关系的识别效果较好，并列关系（COO）、依存分局（DC）等的识别情况较差。

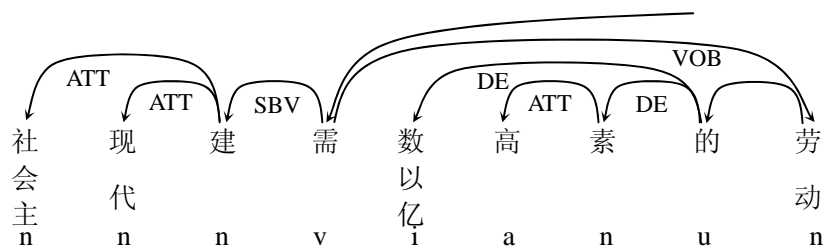
如果 LAS 和 UAS 两个指标的差别较大，则该依存关系的识别问题主要出在依存关系的确定上，如关联结构（CNJ）；倘若两个指标的值基本相同，则问题主要在父节点的识别错误上，如定中关系（ATT），对定中关系 ATT 的错误情况进行分析，发现当中心词前出现多个修饰语时，修饰语之间以及修饰语与中心语之间的依存关系容易识别错误，如例句（1）：

（1）**社会主义现代化建设**需要数以亿计高素质的劳动者和数以千万计的专门人才，除了思想和专业方面的要求外，还应当使他们具有较高的语言文字能力。

该例句在 LTP 中依存句法分析的部分结果如图 1 a 所示，名词“建设”在句子中作主语，“现代化”是“建设”的定语，“社会主义”又是“现代化”的定语。分析发现，“社会主义建设”和“现代化建设”更符合语言表达，即“社会主义”和“现代化”同时作为定语修饰“建设”，标准的句法分析结果为图 1 b。



(a)



(b)

图 1 例句（1）的句法分析结果

定中关系主要涉及名词、代词等实词，与虚词及其用法没有明显的联系。而并列关系 COO 一般与连词一起出现，与虚词的关系比较密切，且并列关系的识别效果较差，因此主要分析连词及其用法对并列关系识别的影响。在此基础上，简单探讨其他虚词对依存关系识别的影响。

3.2 连词用法在依存句法分析中的应用

并列关系的句子中的并列成分与连词结合共同构成连词结构短语，如果可以先识别出连词结构短语，然后确定其中的并列成分，最终就可以根据并列关系的左核心原则识别并列成分之间的并列关系。

3.2.1 并列关系的识别情况

对并列关系识别的错误类型进行总结（表 2），其中“COO-R-”表示并列关系的父节点识别正确，但依存关系识别错误；“COO-PR-”表示父节点识别错误，例如“COO-PR-ATT”表示某一结点因父节点识别错误而将依存关系误标为定中关系 ATT。

表 2 并列关系识别错误分类

错误类型	错误依存关系	个数	总计
COO-R-	VV	39	54
	IC	6	
	APP	2	
	VOB	6	
	CMP	1	
COO-PR-	VOB	9	58
	ATT	18	
	COO	9	
	VV	7	
	DE	6	
	SBV	5	
	ADV	2	
	POB	1	
	HED	1	

由表 2 可以看出，并列关系识别错误的主要原因有以下两个：

a) 当两个或两个以上的动词并列连用时，易识别为连动结构，如例句（2）：

（2）经过电话及时指导，便可立即**排除故障**或**进入**正常操作，减少处理过程和维修的盲目性，提高了医疗设备的使用率。

该句在 LTP 中依存句法分析的部分结果如图 2a 所示，句子中“排除”和“进入”虽都为动词，但由“或”连接，两者属于选择结构，并非两个动作的承接，因此，两者是并列关系，标准的依存句法分析结果如图 2b 所示。

b) 当复合短语中的修饰部分包含并列关系时，易识别为并列成分对核心词的修饰关系，如例句（3）：

（3）这个举动，体现了该站抓好下岗职工基本生活**保障**和再**就业**工作的决心。

该句在 LTP 中依存句法分析的部分结果如图 3a 所示，句子中的“下岗职工基本生活保障”和“再就业”是构成并列关系的两个并列成分，虽然识别出了并列关系，但是两个并列成分识别错误，标准的依存句法分析结果如图 3b 所示。

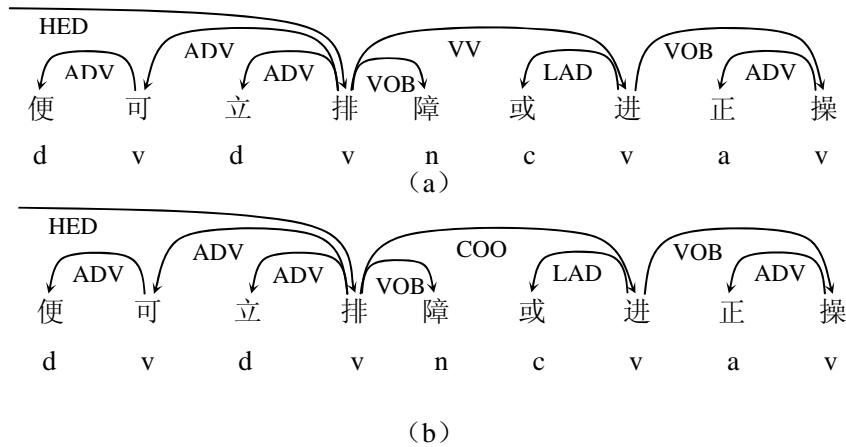


图2 例句 (2) 的句法分析结果

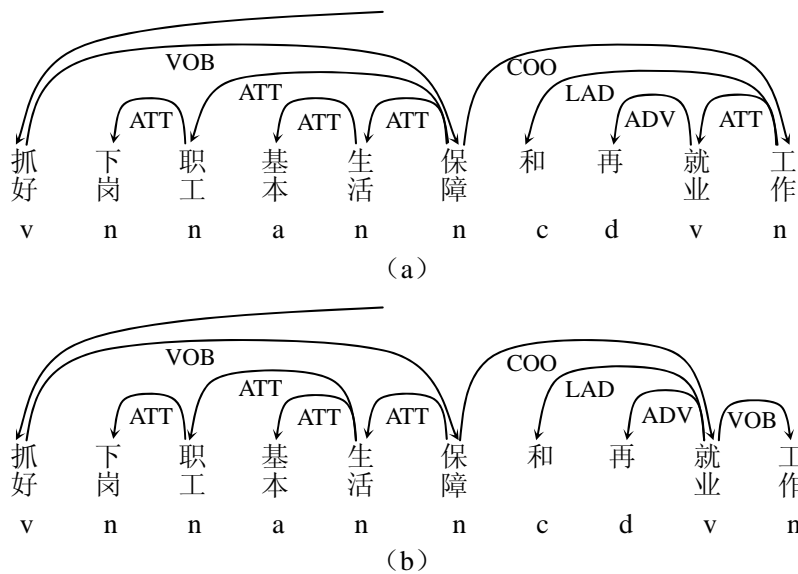


图3 例句 (3) 的句法分析结果

经过对包含并列关系句子的分析,发现这些句子中除了有连词标记,或者顿号等标点符号标记,并列成分之间的依存关系还满足左核心原则,即由多个并列成分时,后面并列成分均依存于第一个并列成分,且并列标记(连词)依存于其右侧的并列成分。如例句(4):

(4) 徐先生还具体帮助他确定了把画雄鹰、鳊鱼、斑鸠、松鼠、麻雀和竹、梅、松、柏作为主攻目标。

其句法分析部分结果图4所示。

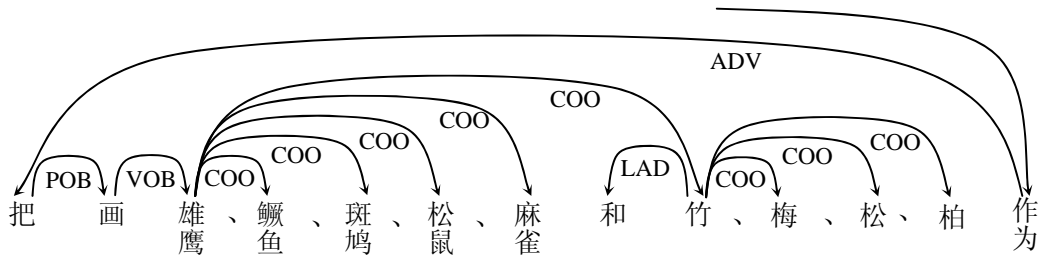


图4 例句 (4) 的依存句法分析结果

当然,在实际的文本中,并非所有的并列关系都包含上述并列标记,也并非都满足上述原则,另外还有上下文语境的信息,这些也可能对并列关系的识别产生影响。

3.2.2 连词用法在并列关系识别中的应用

并列关系相关的连词与并列成分一起构成并列结构短语或者选择结构短语,关于并列关系以及并列结构的研究也有很多^{[14][15]}。本文主要采用周丽娟等^{[13][16]}对连词的用法及连词结构短语识别的相关研究,对依存句法分析的结果进行优化,提高并列关系的识别效果。具体的流程如图 5 所示。

为了清楚的分析连词及其用法对并列关系识别的影响,以例句(5)为例进行详细的分析。

(5) 认为李自成占领北京后,中国面临的是统一还是分裂问题。

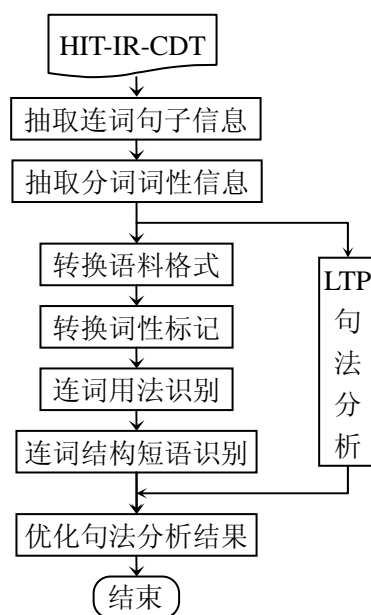


图 5 并列关系识别优化流程图

步骤一:从 HIT-IR-CDT 中抽出包含并列关系信息连词的句子,包含句子的分词、词性标注、句法分析结果等信息,将其作为实验的对比语料。例句(5)对应的依存句法分析结果如图 5 所示,包含的信息如下:

<sent id="9437" cont="认为李自成占领北京后,中国面临的是统一还是分裂问题。">

```

<word id="0" cont="认为" pos="v" parent="-1" relate="HED" />
<word id="1" cont="李自成" pos="nh" parent="2" relate="SBV" />
<word id="2" cont="占领" pos="v" parent="4" relate="ATT" />
<word id="3" cont="北京" pos="ns" parent="2" relate="VOB" />
<word id="4" cont="后" pos="nd" parent="7" relate="ADV" />
<word id="5" cont="," pos="wp" parent="-2" relate="PUN" />
<word id="6" cont="中国" pos="ns" parent="7" relate="SBV" />
<word id="7" cont="面临" pos="v" parent="8" relate="DE" />
<word id="8" cont="的" pos="u" parent="9" relate="SBV" />
<word id="9" cont="是" pos="v" parent="0" relate="VOB" />
<word id="10" cont="统一" pos="n" parent="13" relate="ATT" />
<word id="11" cont="还是" pos="c" parent="12" relate="LAD" />
<word id="12" cont="分裂" pos="n" parent="10" relate="COO" />
<word id="13" cont="问题" pos="n" parent="9" relate="VOB" />
<word id="14" cont="。" pos="wp" parent="-2" relate="PUN" />
  
```

</sent>

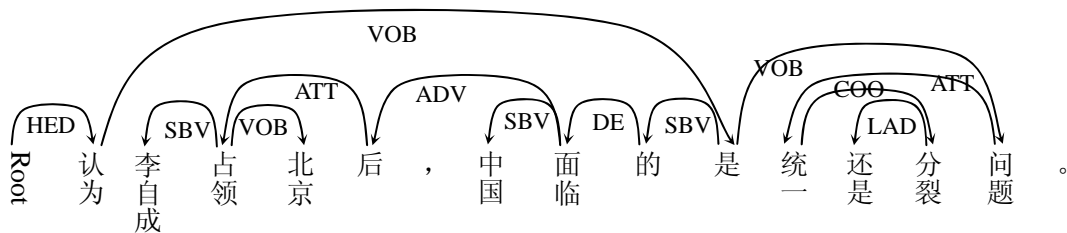


图 5 例句 (5) 标准依存分析图

步骤二：从步骤一生成的语料中抽取句子及其分词和词性标注信息作为新的语料，以克服因分词和词性标注错误对依存句法分析的影响。

步骤三：以步骤二的结果作为输入语料，对其进行依存句法分析，得到 LTP 的句法分析结果，其对应的依存分析图如图 6 所示。

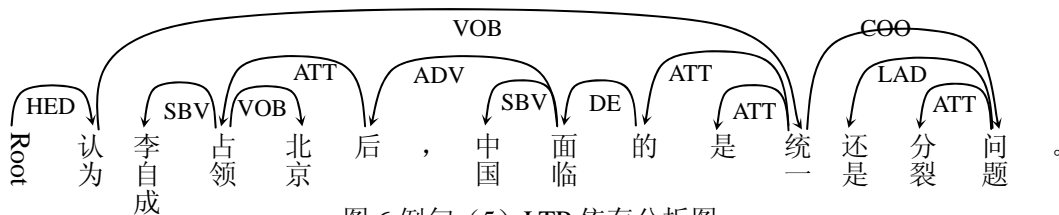


图 6 例句 (5) LTP 依存分析图

步骤四：为了便于对句子中的连词进行用法标注，对步骤二生成的语料进行格式转化，将其转化为北京大学《人民日报》中分词和词性标注形式，即：

认为/v 李自成/nh 占领/v 北京/ns 后/nd ， /wp 中国/ns 面临/v 的/u 是/v 统一/n 还是/c 分裂/n 问题/n 。 /wp

步骤五：因为依存树库的词性标注与北京大学的词性标注集有一些不同^[17]，而上下文的词性信息对虚词用法标注有影响，因此，需要将步骤四产生的语料进行词性转换，即：

认为/v 李自成/nr 占领/v 北京/ns 后/f ， /wd 中国/ns 面临/v 的/ud 是/v 统一/n 还是/c 分裂/n 问题/n 。 /wj

步骤六：对步骤五产生语料中的连词进行用法标注，其标注结果如下：

认为/v 李自成/nr 占领/v 北京/ns 后/f ， /wd 中国/ns 面临/v 的/ud 是/v 统一/n 还是/c<c_hai2shi4_1a> 分裂/n 问题/n 。 /wj

步骤七：对步骤六的语料进行连词结构短语识别^[16]，标注句子中的连词结构短语，即：

认为/v 李自成/nr 占领/v 北京/ns 后/f ， /wd 中国/ns 面临/v 的/ud 是/v <CP_bl> 统一/n 还是/c<c_hai2shi4_1a> 分裂/n </CP_bl> 问题/n 。 /wj

其中<CP_bl>和</CP_bl>分别表示并列结构短语的起始和结束位置，即并列结构短语为“统一还是分裂”。

步骤八：根据识别的连词结构短语对步骤三生成的依存句法分析的结果进行优化。因为只考查依存关系中的并列关系，所以只优化连词结构短语包含的词语之间的依存关系，具体方法的优化方法如下：

1) 找到连词结构短语中连词的前一个词语，该词即为并列关系中的一个并列成分，如“统一”；

2) 继续向前查找连词结构短语中的并列标记——顿号，如果有顿号，则顿号前的词语也属于并列成分，继续查找直至连词结构短语的第一个词语，最后查找到的那个词语便为第一个并列成分，即其他并列成分的父节点；否则，执行 3)；

3) 查找连词结构短语中的最后一个词语为最后一个并列成分，如“分裂”；

4) 将除第一个并列成分外的所有并列成分的“parent”重置为第一个并列成分的“id”，“relate”重置为“COO”，如“分裂_10_COO”；

5) 将连词的“parent”重置为最后一个并列成分的“id”，“relate”重置为“LAD”，如“还是_12_LAD”。

例句(5)的优化后的结果如图7所示：

词语“统一”和“分裂”是并列关系，且为了满足左核心原则，将“统一”作为“分裂”的父节点，并列标记“还是”依存于其后的并列成分“分裂”。

对比图5、图6、图7，发现通过连词用法标注及连词结构短语的标注，不仅正确识别了并列关系COO，同时也可以影响前附加关系LAD的识别效果。

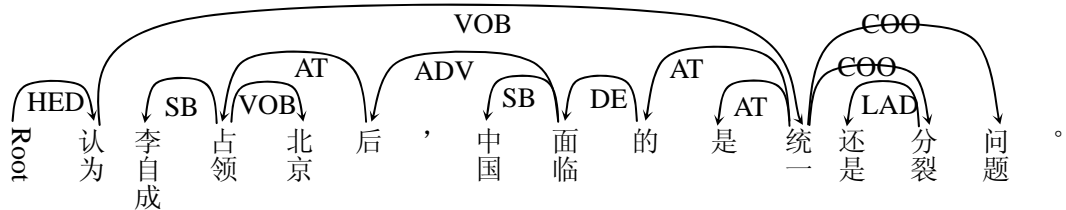


图7 例句(5)优化后的依存分析图

以连词“和”、“及”、“与”、“还是”为例采用上述步骤对抽取的语料进行实验，实验结果如表3所示，“LTP + usage”表示加入虚词用法信息后的实验结果。当加入连词用法后，这些句子中包含的并列关系的LAS及UAS分别提高了3.43%和2.29%，说明连词的用法可以应用到依存句法分析中，提高了并列关系的识别效果。

表3 加入连词用法前后COO的识别结果对比

词语	LTP		LTP + usage	
	LAS (%)	UAS (%)	LAS (%)	UAS (%)
和	96.21	97.15	99.05	99.05
及	85.00	90.00	90.00	90.00
与	96.30	96.30	100.00	100.00
还是	75.00	75.00	100.00	100.00
总计	95.04	96.18	98.47	98.47

依存句法分析中的每个词语只能有一个父节点，为了满足这个条件，在对其中的并列关系进行优化时，也可能对其他的依存弧进行修改。也就是说，加入虚词用法不仅会影响并列关系的识别情况，也会对句子中其他依存关系的识别情况产生影响，如例句(5)中的前附加关系LAD。

3.3 其他虚词用法在依存句法分析中的应用

在上节的研究中，主要考察了连词用法对并列关系识别结果的优化，实际上其他虚词的用法也可能影响依存句法分析中依存关系识别情况，如副词等。下面分析例句(6)和(7)：

(6) 最矮的那个人是小明。

(7) 个子最矮的那个人是小明。

这两个例句的依存句法分析结果分别如图8和图9a所示，两个句子中的词语“最”都为副词，并且都与形容词“矮”构成定中结构ADV，然后与助词“的”构成“的”字结构DE，作为主语“人”的修饰成分（与“人”构成定中关系ATT）。

例句(7)中副词“最”与形容词“矮”一起构成“最矮”作为补语修饰主语“个子”，然后与“的”构成新的修饰成分“个子最矮的”修饰主语“人”，这与图9a的结果不同，即例句(7)的LTP依存句法分析结果存在错误，标准的句法分析结果如图9b所示。

在进行依存句法分析时，主要分析的是两个词语之间的关系，如果将例句(7)按照图9a进行依存句法分析，就会导致名词“个子”与“的”组成“的”字结构DE作为定语修饰中心语“人”，短语“个子的人”显然不合中文的表达习惯。因此一定要将“矮”与其依存

关系比较紧密的词语“个子”先构成主谓关系，然后在与其依存关系相对较远的词语“人”构成定中关系 ATT。

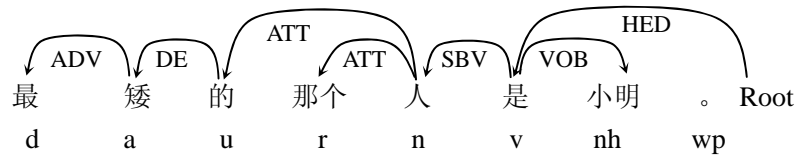
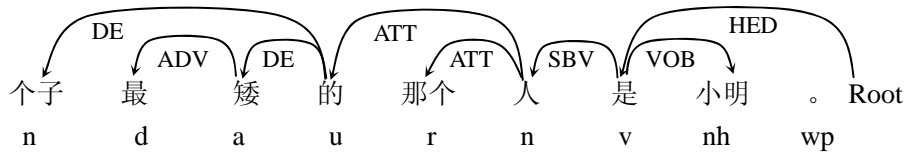
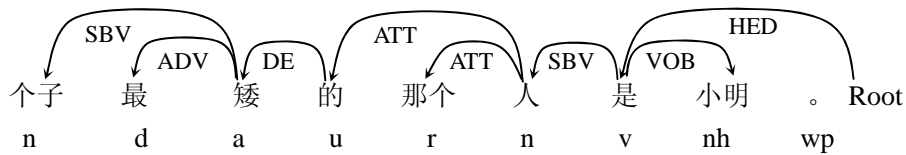


图 8 例句 (6) 的句法分析结果



(a)



(b)

图 9 例句 (7) 的句法分析结果

上述内容主要是依据中文的表达习惯和依存句法分析的原则来分析的，考虑到两个例句中都存在副词“最”，可以从副词的用法角度进行分析。

两个例句中副词“最”的用法标注情况如下：

(6) 最/d<d_zui4_1aa> 矮/a 的/ud 那个/rz 人/n 是/v 小明/nr 。/wj

(7) 个子/n 最/d<d_zui4_1ab> 矮/a 的/ud 那个/rz 人/n 是/v 小明/nr 。/wj

两个例句中的副词“最”用法标注不同，其中用法<d_zui4_1aa>后跟形容词，主要修饰名词，一般与“的”搭配，用法<d_zui4_1ab>后也跟形容词，可以做谓语或者补语。也就是说，例句(7)中副词“最”的用法标注“最/d<d_zui4_1ab>”可以将“个子”和“最矮”结合在一起，作为一个整体再参与其他词语的句法分析。这样可以同时影响主谓关系 SBV 和“的”字结构 DE 两种依存关系的识别效果，由此，虚词的用法信息可以作为依存句法分析的影响因素。

4 结语

本文主要分析了 LTP 中依存句法分析的 24 种依存关系的识别情况，并讨论汉语虚词用法信息对依存句法分析的影响。包含并列关系的句子中多数含有连词，并且并列关系是依存句法分析的难点，因此着重考虑了连词用法在并列关系识别中的应用和影响。实验表明，考虑到连词用法的因素后，并列关系的识别情况有明显提高，同时也对与并列关系相关的依存关系有一定的影响，肯定了虚词用法对并列关系的影响。在此基础上，简单探讨了其他虚词用法信息对依存句法分析其他依存关系的影响。

在下一步的研究工作中，将具体分析虚词用法信息对其他依存关系的影响，并在更多的语料上进行实验研究。同时，进一步研究虚词用法在自然语言理解其他领域的应用。

参考文献

- [1] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008
- [2] 董振东, 董强, 郝长伶. 下一站在哪里[J]. 中文信息学报, 2011, 25 (6): 3~11
- [3] Tesnière L. Elements de syntaxe structurale. Editions Klincksieck. 1959.

- [4] Ting Liu, Jinshan Ma, Sheng Li. Building a Dependency Treebank for Improving Chinese Parser[C]. Journal of Chinese Language and Computing, 16 (4) :207~224
- [5] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[C]. In Proceedings of the Coling 2010.13~16
- [6] Ryan McDonald, Koby Crammer, Fernando Pereira. Online Large-Margin Training of Dependency Parsers[C]. Association for Computational Linguistics (ACL). 2005. 91~98
- [7] Ryan McDonald, Fernando Pereira. Non-projective Dependency Parsing using Spanning Tree Algorithms[C]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005.523~530
- [8] Ryan McDonald, Fernando Pereira. Online Learning of Approximate Dependency Parsing Algorithms[C]. European Association for Computational Linguistics (EACL). 2006.81~88
- [9] 李正华, 车万翔, 刘挺. 基于柱状搜索的高阶依存句法分析[C]. 中国计算机语言学研究前沿进展(2007-2009). 烟台: 鲁东大学, 2009, 117~122
- [10] 孟雷, 丁效, 秦兵, 刘挺. 基于依存句法和短语结构句法结合的金融领域事件元素抽取[C]. 中国计算机语言学研究前沿进展(2009-2011). 洛阳: 中国人民解放军外国语学院, 2011, 432~437
- [11] 俞士汶, 朱学锋, 刘云. 现代汉语广义虚词知识库的建设[J]. 汉语语言与计算学报, 2003, 2 (1): 89~98
- [12] 咎红英, 朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建[J]. 当代语言学, 2009, 11 (2): 124~135
- [13] 周丽娟. 现代汉语连词用法的自动识别及应用研究[D]. [硕士学位论文]. 郑州: 郑州大学, 2012
- [14] 赵怿怡, 高松, 刘海涛. 基于依存语法的汉语并列结构自动分析研究[C]. 中国计算机语言学研究前沿进展(2007-2009). 烟台: 鲁东大学, 2009, 148~153
- [15] 李文杰, 穗志方. 基于并列结构的概念实例和属性的同步提取方法[C]. 中国计算机语言学研究前沿进展(2009-2011). 洛阳: 中国人民解放军外国语学院, 2011, 382~387
- [16] 咎红英, 周丽娟, 张坤丽. 基于用法的现代汉语连词结构短语识别研究[J]. 中文信息学报, 2012, 23 (6): 72~78
- [17] 马金山. 基于统计方法的汉语依存句法分析研究[D]. [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2007