北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.006

基于清华汉语树库的复句关系词识别与分类研究

李艳翠[1,2]　孙静[1]　周国栋[1,†]　冯文贺[3]

1. 苏州大学计算机科学与技术学院，苏州　215006; 2. 河南科技学院信息工程学院，新乡　453003;
3. 河南科技学院人文学院，新乡　453003; † 通信作者, E-mail: gdzhou@suda.edu.cn

摘要　根据清华汉语树库的标注方法，利用规则从中提取复句关系词并标注其类别。然后分别抽取带功能标记和不带功能标记的自动句法树的句法、词法、位置特征，进行复句关系词的识别和分类。实验结果表明，复句关系词判断准确率达 95.7%，复句关系词类别判断 *F*1 值为 77.2%。

关键词　复句关系词；清华汉语树库；关系词识别；关系词分类

中图分类号　TP391

Complex Sentence Relative Recognition and Classification Based on Tsinghua Chinese Treebank

LI Yancui[1,2], SUN Jing[1], ZHOU Guodong[1,†], FENG Wenhe[3]

1. Department of Computer Science and Technology, Soochow University, Suzhou 215006; 2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003; 3. School of humanities, Henan Institute of Science and Technology, Xinxiang 453003; † Corresponding author, E-mail: gdzhou@suda.edu.cn

Abstract　According to Tsinghua Chinese Treebank annotation methods, using the rules, the authors extracted relative words and marked their categories. Then syntax, lexical and position feature of automatic syntax tree with and without functional marke were extracted to recognize and classify relative. Experiment results show that relative recognition accuracy is 95.7%, and relative classification *F*1 is 77.2%.

Key words　complex sentence relative; tsinghua chinese treebank; relative recognition; relative classification

中文信息处理已经完成了字处理，较好地解决了词处理，目前的研究集中在句子，正在向篇章过渡。复句是从小句到篇章的过渡[1]，它连接小句和篇章并在二者之间起到很好的衔接作用，同时兼有语法、语义和语用等多方面属性。汉语的复句理解是句子语义分析研究的重点之一，在篇章分析、自动问答、信息抽取以及机器翻译等领域都有非常重要的用途。

黄伯荣等[2]对复句的定义为：复句是由两个或几个意义上紧密相关、结构上互不包含的分句构成的句子。复句表达的主要问题是它内部分句之间的语义关系。复句类别的命名大多也着眼于不同分句之间的语义关系，如并列、递进等。要理解复句就要先搞清复句内部各分句之间的语义关系。特定的复句语义关系，由特定的复句关系词语标示出来。复句关系词，在复句中联接分句，标明关系，构成特定的复句格式。邢福义[3]认为复句关系词语没有十分明确的范围，在词类系统中，不属于固定的类；在语法单位中，不处于固定的级；在造句功用上，不具有划一性。根据复句关系标志的有无，复句可以划分为有标和无标复句。

从自然语言处理的角度分析有标复句的关系，其前提是要对关系词进行正确的识别。基于复句自动分析的目的，本文主要研究有标复句关系词的识别与分类。拟从清华汉语树库中抽取复句中的关系词及关系类别，探讨使用词法、句法特征进行关系

863 计划(2012AA011102)、国家自然科学基金(61273320)和教育部人文社科青年基金(13YJC740022)资助
收稿日期: 2013–06–15; 修回日期: 2013–08–15; 网络出版时间: 2013-11-11 10:25

词的识别及分类方法，以期弄清不同复句关系标记与复句关系之间是否存在着某种制约关系，为汉语复句关系分析和篇章分析奠定基础。

1 相关工作

关于复句的分类，邢福义[3]提出了复句的三分法，将复句关系类别分类为广义因果、广义并列和广义转折三 种，并对每种类别的复句所包含的句式特点、语义特点等进行详细说明。黄伯荣等[2]将复句分为联合复句和偏正复句两种。联合复句又分为并列复句、承接复句、递进复句、解说复句和选择复句；偏正复句又分为因果复句、条件复句、假设复句、转折复句、让步复句和目的复句。根据关系标志的有无，复句可以划分为有标复句(使用关系标志、形成特定句式的复句)和无标复句(语表上不出现关系标志的复句)。对有标复句，吴锋文[4]根据复句中关系标记序列对复句层次关系标示能力的强弱，划分出充盈态有标复句和非充盈态有标复句两大类型，构建了一个基于关系标记的复句分类体系。舒江波[5]全面总结影响关系词自动标识的因素，对标记连用现象进行研究，主要研究二标记连用和三标记连用时各个标记的语法语义功能和类别，研究句式特点与关系词标识之间的关系，并对部分充盈模态和非充盈模态下关系词的标识问题进行研究。

包含复句标记的语料库目前主要有汉语复句语料库和清华汉语树库。汉语复句语料库(the Corpus of Chinese Compound Sentences)①由华中师范大学语言与语言教育研究中心开发，是一个面向汉语复句研究的专用语料库。该语料库是在邢福义教授的指导和支持下设计和开发的，已收有标复句 658447 句，约 44395000 字。语料来源以《人民日报》和《长江日报》为主，收入各种句式的现代汉语有标复句。清华汉语树库(Tsinghua Chinese Treebank)[6]中标出了复句内各分句之间的关系信息，复句分类采用比较常用的并列关系、连贯关系、递进关系、选择关系、因果关系、目的关系、假设关系、条件关系、转折关系分类方法。但清华汉语树库中没有标注特定复句关系所对应的复句关系词。

目前，很多研究者对汉语复句语料库进行了研究。胡金柱等[7]讨论利用关系词库中的信息来判断关系词的搭配关系、连用形式以及单用形式的方法，作者从自建的 5000 条三句式复句语料集中抽取 1000 条进行关系词标注试验，正确标注的分句有 2073 个，分句的正确标注率达 69.1%。胡金柱等[8]结合词性标记和关系词搭配理论，提出正向选择算法提取关系词。通过测试可知，关系词提取的正确率达到 89.8%。并非复句中出现的关系标记都是关系词，胡金柱等[9]利用汉语复句语料库和关系词库，提出一种基于规则的连用关系标记的自动标识算法，从中识别出真正的关系词，该算法结合关系词库和关系词提取技术，分析其连用特征，对连用关系标记标识准确率达 72.9%。周文翠等[10]尝试一种基于机器学习的自动判别并列复句的方法，对缺乏明显标记的复句，利用其主干句子成分的语义关系来判定句子间的语义关系。从《人民日报》语料库中抽取 4 万多个句子作为训练集和测试集，选取其主语、谓语等相关特征并根据《知网》将特征量化，然后使用支持向量机进行训练，在开放测试中获得 84%的准确率。洪鹿平[11]在清华汉语树库上做汉语复句关系自动判断研究，作者穷尽式地收集关系词语，并把关系词语标注上联合和偏正两种类型，然后抽取特征利用 CRF 模型进行分类。王东波等[12]基于条件随机场进行有标记联合结构的自动识别，使用北京大学《人民日报》语料和清华汉语树库，分别用基于复杂特征的特征模板和增加语言学特征的特征模板在含有嵌套的联合结构、无嵌套联合结构和最长联合结构语料上进行实验，得到 $F1$ 值分别为 88. 21%、87. 85%和 84. 42%。

综上可知，目前关于复句的研究还比较初步，大部分研究集中在对复句关系词的识别上。在复句分类方面，目前所做的复句类别较粗，需要进一步细分。

2 清华树库介绍

清华汉语树库[6]是从大规模的经过基本信息标注(分词和词性标注)的汉语平衡语料库中，提取出 100 万汉字规模的语料文本，经过自动断句、自动句法分析和人工校对，形成的高质量汉语句法树库语料。经统计，目前语料中不同文体语料所占比例为文学 37.25%、学术 4.64%、新闻 25.83%、应用 32.28%，具体情况如表 1 所示。

① http://ling.ccnu.edu.cn:8089/jiansuo/TestFuju.jsp

表 1　清华汉语树库统计数据

Table 1　Statistics of Tsinghua Chinese Treebank

文体	文件数	文件所占比例/%	句子数(单句数/复句数)	复句所占比例/%	平均句子长度(单句/复句)(词/句)
文学	225	37.25	24799(10614 / 14185)	57.2	21.4(12.0 / 28.4)
新闻	156	25.83	6773(2822 / 3951)	58.3	25.0(14.3 / 32.7)
学术	28	4.64	9395(4387 / 5008)	53.3	28.1(18.0 / 36.9)
应用	195	32.28	3169(1869 / 1300)	41.0	21.0(12.9 / 32.6)
合计	604	100	44136(19692 / 24444)	55.4	23.3(13.7 / 31.1)

由表 1 可知, 清华汉语树库中文学句子数最多(24799 句, 占 56.2%), 学术所占比例最低(3169 句, 占 7.2%)。语料中共有 4 万多个句子, 其中复句占 55.4%, 单句和短语占 44.6%, 表明在真实文本的汉语句子中, 复杂句子占大多数。不同文体中, 复句占比例最高的是新闻(58.3%), 占比例最低的为应用(41.0%)。清华汉语树库中复句的平均长度是 31.1 个词语, 其中学术长度最长, 达 36.9 个词语, 文学长度最短, 为 28.4 个词语。

综合利用现有的汉语语法学研究成果, 清华汉语树库目前的标注体系中, 为复句设计了两套标记。一方面, 用标记“fj”标注所有的复句, 以此作为复句的外部语法功能的载体。另一方面, 通过一组结构关系标记描述复句内部各个分句间的复杂逻辑语义关系, 这是复句标注研究的重点所在。关联词语是复句的分句间逻辑语义关系的重要载体, 它包括连词、副词以及一些常用的固定短语。目前, 清华汉语树库根据各自的特征关联词语将复句的内部结构分为并列、连贯(顺承)、递进、选择、因果、目的、假设、条件、转折等 9 类。而对于子句间没有关联词语的复句, 则遵循以下处理原则: 1) 对于通过词汇手段(指代、反复)和结构手段(对偶、排比)等连接起来的复句, 深入分析其中隐含的子句逻辑语义关系, 把它们归入上面的 9 种结构关系子类中; 2) 对于两类特殊的复句(解注复句和流水复句), 用 JZ 和 LS 予以标注; 3) 对于其他不能归入以上 11 类的复句, 标注缺省结构标记“XX”。清华汉语树库的结构关系标记集如表 2 所示。

3　复句关系词的抽取与分类

复句关系词, 是复句中用来联结分句标明关系的词语。它们是根据联结分句、标明相互关系、形成复句格式的共同特点组合起来的一些词语, 没有十分明确的标准, 因而也没有十分明确的范围。大体上有以下几种: 句间连词, 它们通常连接分句、不充当句子成分, 如“因为、所以、虽然、但是”等; 关联副词, 一般起关联作用, 又在句子中充当状语, 如“就、又、也、还”等; 超词形式, 如“如果说、总而言之”等。

清华汉语树库中的复句关系词主要有连词(c)、副词(d)和连接语(l)。对于复句关系词, 清华汉语树库中可以推断其所属复句的具体关系类别, 具体标注实例如例 1 和 2 所示。标注为 c, d 和 l 的词很多, 有些是复句关系词, 有些不是, 同一个词在不同情况下有时可以作为复句关系词, 有时则不是。如清华汉语树库的标注实例 1“民不益赋而天下用饶”中的“而”表示句子内部的转折, 在此情况下“而”是关系词但不是复句关系词。

例 1　47　[zj-XX [fj-MD [fj-DJ [dj-ZW 他/rN [vp-ZZ [pp-JB 根据/p [np-DZ 先秦/nR 的/u [np-DZ 有关/b [np-LH 理论/n 和/c 原则/n]]]]] , /, [vp-PO [vp-AD [vp-LH 制订/v 和/c 推行/v] 了/u] [np-LH [np-DZ [np-LH 盐/n 铁/n] 官营/vN] 、/、 [np-DZ 酒类/n 专卖/vN]]]]] , /, [vp-XX 并/c

表 2　汉语复句结构关系标记集

Table 2　Chinese complex sentence relation tag set

序号	标记符号	关系类型	序号	标记符号	关系类型	序号	标记符号	关系类型
1	BL	并列关系	5	YG	因果关系	9	ZE	转折关系
2	LG	连贯关系	6	MD	目的关系	10	JZ	解注复句
3	DJ	递进关系	7	JS	假设关系	11	LS	流水复句
4	XZ	选择关系	8	TJ	条件关系			

[vp-PO 首创/v [np-LH 均输/n 和/c 平准/n]]]] , /, [vp-XX 以/c [vp-PO 实现/v [np-DZ [yj-XX [yj-BH “/“ [fj-ZE [dj-ZW 民/n [vp-ZZ 不/dN [vp-PO 益/v 赋/n]]] [dj-XX 而/c [dj-ZW 天下/n [dj-ZW 用/n 饶/v]]]] ”/”] [dlc-BC(/([np-BH 《/《 [np-DZ 史记/nR·/·平准书/nR] 》/》])/)]] 的/u 目的/n]]]] 。/。] (BAIKE001.pct)

例 2 87 [zj-XX [fj-LG [vp-ZZ [pp-JB 从/p [np-DZ 这样/rV 的/u 论点/n]] 出发/v] , /, [fj-LS [dj-ZW 他们/rN 主张/v] [fj-BL [dj-ZW [np-DZ 政府/n 的/u 任务/n] [vp-XX 不是/c [vp-LW 去/v [vp-PO 刺激/v 需求/n]]]] , /, [vp-XX 而是/c [vp-LW [vp-ZZ 应/vM [vp-PO 注意/v [np-DZ 政策/n [np-DZ [pp-JB 对/p [np-DZ [np-DZ 生产/vN 经营/n] 活动/n]] 的/u 作用/n]]]] [vp-XX 以/c [vp-PO 刺激/v 供给/n]]]]]]] 。/。] (BAIKE001.pct)

例 1 中标注了目的(fj-MD)、递进(fj-DJ)和转折(fj-ZE)3 种关系, 表示目的的关系词是“为”, 表示递进的关系词是“并”, 表示转折的关系词是“而”。其中“为”和“并”是复句关系词, 但“而”表句子内部转折, 属于句内关系词。例 2 中标注了并列关系(fj-BL), 所对应的关系词是“不是”和“而是”。

清华树库中虽然可以推断出特定词语是否为复句关系词, 但并没有专门的标记记录某个词是否为复句关系词和其具体所代表的复句关系类别。为了实现复句中的关系词的抽取与分类, 本文对清华汉语树库的标注进行分析, 根据规则, 抽取出树库中标注的复句关系词及关系类别。

3.1 复句关系词及类别抽取

根据文献[11]搜集的关系词进行统计, 发现词类标注为 c, d 和 l 的词在复句关系词中所占比例达 89%, 故本文只对这 3 类词进行处理。如果某个词是复句关系词, 则在清华树库中一定可以找到其具体复句类别的标注, 反过来, 对于有标复句, 如果标注了复句类别, 就一定存在一个复句关系词。本文在制定复句关系词抽取方法时人工考察了大量标注实例, 发现其部分句法树一般含有一个标点、一个关系词标记和一个复句类别标记。如果关系词上层节点有多个复句类别标记, 采用最邻近原则确定复句类别。图 1 是例 1 中的两个复句关系词“以”和“并”的部分句法树。从图 1 可知, “以”位于逗号的右边, 其上层节点标注为 fj-MD, 表示目的关系; “并”邻近的上层复句功能标记节点为 fj-DJ, 表示递进关系。图 2 是例 2 中表并列关系(fj-BL)的部分句法树, 表并列关系的词是“不是”和“而是”。

分析语料中的复句关系词标注情况, 可以得到如图 3 更通用的句法结构树。图 3 的通用句法树覆盖大多数复句。有些关系词是单个出现的, 如例 1 中的“以”“并”, 有些则是成对出现的, 如例 2 中“不是…..而是”。本文主要研究关系词的识别及分类, 对于搭配出现的情况没有特别处理, 在生成实例和判断其是否为关系词时, 本文对每个关系词分别进行处理, 没有将其合起来处理为一个实例。

根据以上介绍, 可以抽取关系词所在部分的句法及功能标注信息, 进而推断其是否为复句关系词, 并抽取“fj”后的标记作为其具体关系类别。本文抽取复句关系词所连接的两个子句之间一定有标点信息。由于位于句子开头的关系词很多情况下表示

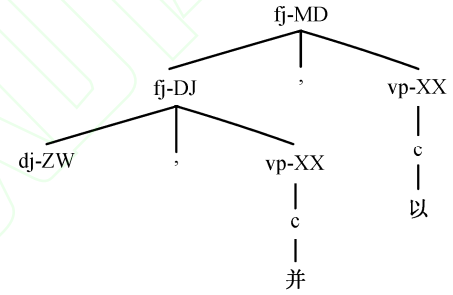

图 1 例 1 中的目的关系和递进关系

Fig. 1 Purpose and progressive relations in Example 1

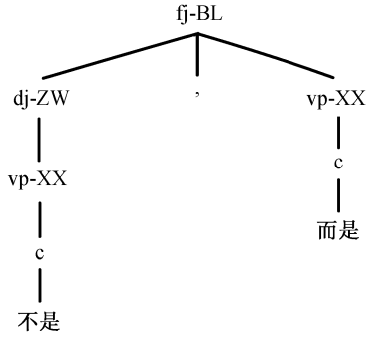

图 2 例 2 中的并列关系

Fig. 2 Coordinate relation in Example 2

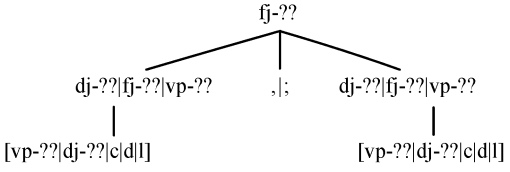

图 3 清华汉语树库中的复句关系标注

Fig. 3 Annotation of complex sentence relation in Tsinghua Chinese Treebank

与前面句子的关系, 清华树库中并没有标注, 故本文无法抽取其类别。在具体抽取时, 本文采用以下抽取算法, 输入是句法树, 输出是复句关系词及类别。

输入: 清华汉语树库的句法树;

输出: 复句关系词及类别;

步骤 1 初始化。 对句子中标注为 c、d、l 的每一个节点 cNode, cp = cNode.nodeParent, 序号 iIndexChild = cp.iIndexChild, 句法模式 pattern = "";

步骤 2 沿句法树向上。If cp 不包含"fj|np" 或 cp 包含"dj": cp=cp.nodeParent, iIndexChild = cp.iIndexChild

步骤 3 得到句法模式。If cp 为空: pattern = cNode.nodeParent.nodeValue

If iIndexChild >= 2: pattern = cp.nodeValue +"_"+ cp.childList[iIndexChild-2]. nodeValue +"_"+cp.childList [iIndexChild-1].nodeValue

Elif iIndexChild == 1: pattern = cp.nodeValue +"_"+cp. childList.nodeValue

步骤 4 得到复句关系词及类别。if pattern.startswith ("fj") and indexLeaf > 0 and pattern.find(", |; ") !=−1 and pattern.startswith ("fj-LS|fj- JZ|fj-XX") == False

return 复句关系词 cNode, 关系词类别 pattern[3:5]

3.2 复句关系词统计

采用 3.1节抽取 算法, 本文共抽取出 72662 个词, 其中复句关系词 19119 个, 非复句关系词 53543 个。由于清华树库没有标注句子之间的关系, 而出现在句子开头的连接词通常表示句间关系, 抽取不出其类别。这种情况的词共有 4973 个, 本文复句关系词分类实验时并没有将其进行分类。清华汉语树库中复句关系词词性及关系类别分布情况如表 3 所示, 其中“其他”包含复句关系词在开头及关系类别为解注和流水的情况。出现频次最多的不同词性关系词及出现次数如表 4 所示。

4 实验及分析

4.1 所用的特征

根据以上分析, 本文抽取简单的词法、句法和位置特征进行复句关系词识别和分类。所用特征如下, 括号中内容以图 1 中的“并”为例进行说明。

词法特征: 1) 关系词及词性(并 c); 2) 关系词前后的 2 个词及词性(专卖 vN , , 首创 v 均衡 n); 3) 关系词所在的句子位置: 开头、句中、句尾(句中)。

句法特征: 1) 关系词节点句法信息(vp-XX); 2) 关系词节点父亲节点信息(fj-DJ); 3) 关系词节点左兄弟信息(,); 4) 关系词节点右兄弟信息(NULL)。

位置特征: 1) 关系词是否位于句首(否); 2) 关系词前是否有标点(是)。

4.2 实验结果

利用 3.1 节中的复句关系词及类别抽取方法, 本文抽取所有标注为 c, d 和 l 的词是否为复句关系词, 对于复句关系词, 抽取出其具体的类别的实例。共抽取出 72662 个实例, 其中是复句关系词的实例 19119 个, 非复句关系词的实例 53543 个。数据准备完毕后, 采用机器学习的方法分别判断关系词是否为复句关系词(二元分类)和复句关系词的类别(多元分类)。

利用 4.1 节的特征, 采用 Mallet 工具包[13]进行实验。实验采用 10 倍交叉验证, 句法特征对复句关系词判断最重要, 本文采用清华汉语树库, 利用伯克利句法分析器分别重新训练了带功能标记和不带功能标记的句法模型(如 vp-ZZ 带功能标记, vp 不带), 然后给定分好词的清华树库中的句子, 利用训练好的模型分别自动标注带功能标记和不带功能标记的句法树。然后抽取 4.1 节所述特征进行实验, 对复句关系词识别的结果如表 5 所示。

从表 5 可以看出, 使用决策树效果最好, 说明复句关系词识别问题并不太复杂, 可以抽取出一定的规则。单纯的词汇特征对复句关系词的识别也有一定作用, 不带功能标记识别准确率为 71.7%, 带功能标记准确率最高为 82.2%。由实验可知, 利用词汇、句法和位置特征的组合进行复句关系词的识别效果最好, 带功能标记和不带功能标记时复句关系词判断的准确率分别为 95.7%和 92.1%。带功能

表 3 关系词及类别分布情况

Table 3 Distribution of relative and relative categories information

词性	是/否为复句关系词	复句关系词类别									
		并列	连贯	递进	选择	因果	目的	假设	条件	转折	其他
c	8727 / 15421	561	627	1031	87	767	157	206	67	1711	3515
d	9985 / 37935	1222	2913	439	42	244	20	330	204	189	4382
i	407 / 187	12	22	0	1	5	0	4	2	3	358

标记由于所带信息较多，对复句关系词识别的总体性能较好。

利用 4.1 节中给出的所有特征，使用决策树对复句关系词类别进行分类，结果如表 6 所示。从表 6 可以发现，使用带功能标记的自动句法树结果明显优于不带标记的自动句法树。在所有复句关系中，目的、因果、转折关系识别效果较好，主要原因是这几种关系通常有比较明显且无歧义的关系词。条件、选择、并列关系识别效果较差，主要原因是这几种关系的复句关系词比较多，很多词可以表示多种关系。使用不带功能标记的自动句法树，复句关系类别判断平均 *F*1 值为 62.2%，使用带功能标记的自动句法树 *F*1 值为 77.2%，结果相差接近 15%，在复句关系词识别时相差仅为 3.6%，说明功能标记对复句类别划分作用非常明显。

4.3 错误分析

分析实验结果发现，对关系词分类错误的原因主要有抽取复句关系词错误及复句关系类别判断错误，下面分别举例说明。

1) 抽取复句关系词错误

如图 4 所示，图中的“则”按照本文的抽取方

表 4 出现次数最多的复句关系词

Table 4 Most frequency relatives and it's number

d	次数	c	次数	l	次数
也	896	但	1176	例如	43
还	551	而	724	总之	28
又	548	并	564	据说	25
却	191	而且	380	看来	25
同时	190	但是	348	比如	18
才	170	因为	275	看见	10

表 5 是否为复句关系词识别准确率

Table 5 Complex sentence relative recognition accuracy

特征	自动句法树(不带功能标记)			自动句法树(带功能标记)		
	最大熵	决策树	贝叶斯	最大熵	决策树	贝叶斯
词汇	69.5	71.7	65.8	82.2	81.6	78.5
句法	90.8	90.7	90.6	94.8	95.0	93.5
词汇+句法	91.1	92.0	88.3	95.5	95.6	93.3
词汇+句法+位置	91.2	92.1	88.1	95.5	95.7	93.6

表 6 复句关系词类别识别结果

Table 6 Relative word category recognition results

		并列	连贯	递进	选择	因果	目的	假设	条件	转折	平均
自动句法树(不带功能标记)	召回率	42.7	49.7	80.1	52.2	74.4	66.6	48.7	53.1	85.8	61.5
	准确率	40.8	58.8	73.5	64.4	71.2	70.9	67.3	59.8	67.9	63.8
	*F*1 值	41.8	53.9	76.7	57.6	72.8	68.7	56.5	56.2	75.8	62.2
自动句法树(带功能标记)	召回率	65.8	76.5	85.1	62.2	78.1	85.8	68.9	67.9	86.5	75.2
	准确率	70.0	80.8	74.9	74.4	83.9	86.9	83.3	85.9	80.0	80.0
	*F*1 值	67.9	78.6	80.0	67.8	80.9	86.4	75.5	75.9	82.6	77.2

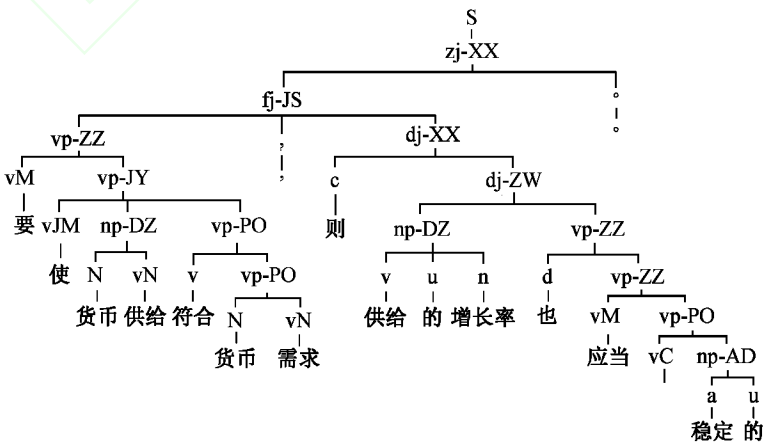

图 4 复句关系词抽取错误

Fig. 4 Relative extracts error

法，抽取结果为是复句关系词，复句关系词类别为假设(fj-JS)。但在本例中，“则”不能体现假设关系，故不是本假设关系的复句关系词，“则”在真实文本中多表示并列、转折关系，所以本实验中的“则”也不能自动分类为假设关系。

2) 复句关系词类别判断错误。

这类错误产生的主要原因是同一关系词可以表示多种关系，在判断具体关系时往往趋向于标注次数较多的类别。如“并”可以表示递进、并列关系。以下例 3 中的“并”表示并列关系，例 4 中“并”表示递进关系。在本实验中将例 4 错误判断为递进关系。

例 3 但财政的存在和发展又必须以劳动者能够为国家提供剩余产品为条件，**并**随着生产力的发展和占主导地位的生产关系的变化而发展。(递进关系 BAIKE001.fid)

例 4 民族资产阶级作为剥削阶级，同帝国主义、封建主义、官僚资本主义有着千丝万缕的联系，**并**同中国革命的领导阶级——产阶级有尖锐的矛盾。(并列关系 BAIKE028.fid)

5 结论

本文主要进行汉语复句关系词的识别与分类。实验采用清华汉语树库，首先抽取出所有标注为连词、副词和连接词的词，然后根据清华汉语树库中复句的标注结果，判断抽取出的词是否为复句关系词，若是复句关系词，根据标注结果为其标上类别，形成实验数据。数据抽取完毕后，使用清华汉语树库分别训练了带功能标记和不带功能标记的句法模型，然后利用训练好的句法模型对树库中分好词的句子进行句法分析，使用自动句法树中的句法、词性、位置特征进行关系词的识别与分类。实验表明，带功能标记的自动句法树效果明显较好。使用带功能标记的自动句法树，关系词的识别准确率达 95.7%，关系词类别分类平均 *F*1 值为 77.2%。下一步准备进一步完善工作，以便更准确地抽取复句关系词及其类别。

参考文献

[1] 徐阳春. 现代汉语复句句式研究. 北京: 中国社会科学出版社, 2002: 8–11

[2] 黄伯荣, 彦序东. 现代汉语(下册). 北京: 高等教育出版社, 2002

[3] 邢福义. 汉语复句研究. 北京: 商务印书馆, 2001: 1–37

[4] 吴锋文. 基于关系标记的汉语复句分类研究. 汉语学报, 2011(3): 63–73

[5] 舒江波. 面向中文信息处理的复句关系词自动标识研究[D]. 武汉: 华中师范大学, 2011

[6] 周强. 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1–8

[7] 胡金柱, 吴锋文, 李琼, 等. 汉语复句关系词库的建设及其利用. 语言科学, 2010, 9(2):133–142

[8] 胡金柱, 舒江波, 姚双云, 等. 面向中文信息处理的复句关系词提取算法研究. 计算机工程与科学, 2009, 31(10): 90–93

[9] 胡金柱, 陈江曼, 杨进才, 等. 基于规则的连用关系标记的自动标识研究. 计算机科学, 2012, 39(7): 190–194

[10] 周文翠, 袁春风. 并列复句的自动识别初探. 计算机应用研究, 2008, 25(3): 764–766

[11] 洪鹿平. 汉语复句关系自动判断研究[D]. 南京: 南京师范大学, 2008

[12] 王东波, 陈小荷, 年洪东. 基于条件随机场的有标记联合结构自动识别. 中文信息学报, 2008, 22(6): 3–8

[13] McCallum A K. Mallet: a machine learning for language toolkit [CP/OL]. (2002)[2012–02–28]. http://mallet.cs.umass.edu