基于清华汉语树库的复句关系词识别与分类研究

李艳翠 孙静 周国栋 冯文贺

引言

意义

- 汉语的复句理解在篇章分析、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。
- 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。

现状

- 复句分类
  二分(邢福义) 三分(黄伯荣)
  有标分为充盈态非充盈态(吴锋文)
  舒江波对标记连用进行研究
- 复句语料及相关研究
  汉语复句语料库(胡金柱等)
  清华树库（洪鹿平等）

不足

- 复句研究多集中在对复句关系词的识别上
- 复句分类类别较粗，需要进一步细分

清华树库介绍

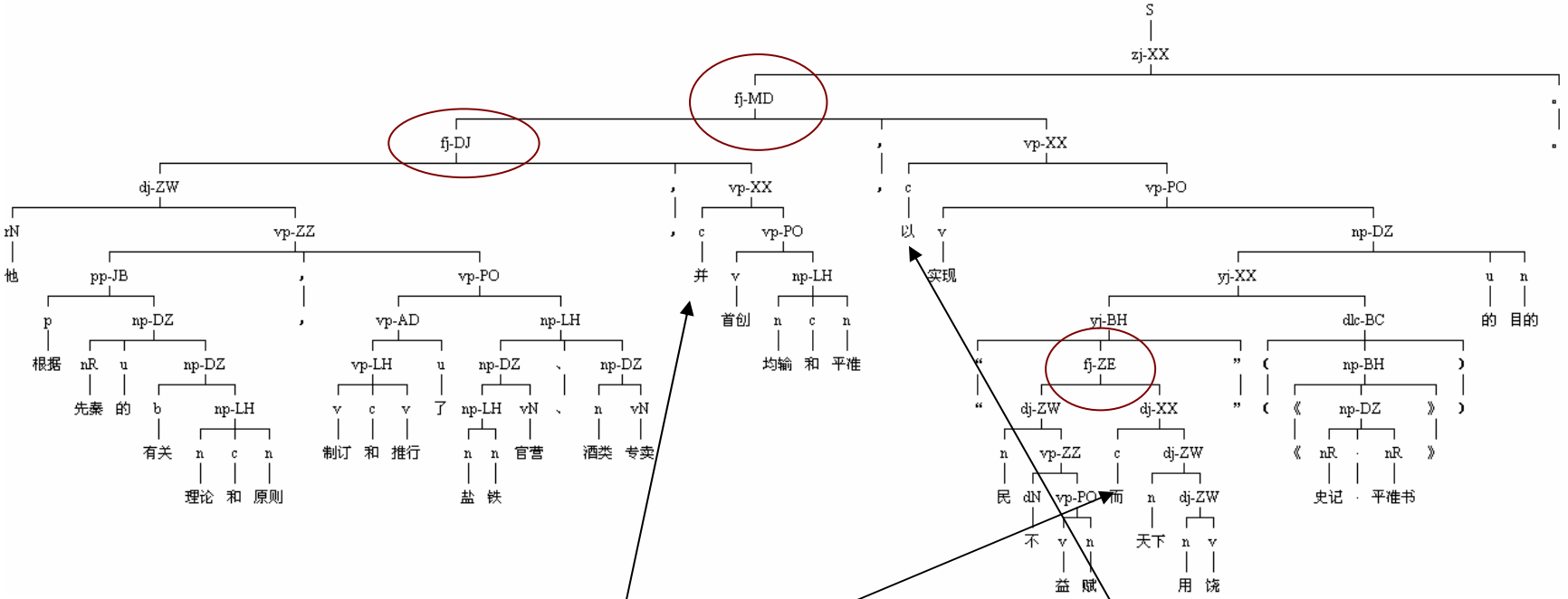
表1 清华汉语树库统计数据

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	24 799（10 614 / 14 185）	57.2	21.4（12.0 / 28.4）
新闻	156	25.83	6 773（2 822 / 3 951）	58.3	25.0（14.3 / 32.7）
学术	28	4.64	9 395（4 387 / 5 008）	53.3	28.1（18.0 / 36.9）
应用	195	32.28	3 169（1 869 / 1 300）	41.0	21.0（12.9 / 32.6）
合计	604	100	44 136（19 692 / 24 444）	55.4	23.3（13.7 / 31.1）

表2汉语复句结构关系标记集

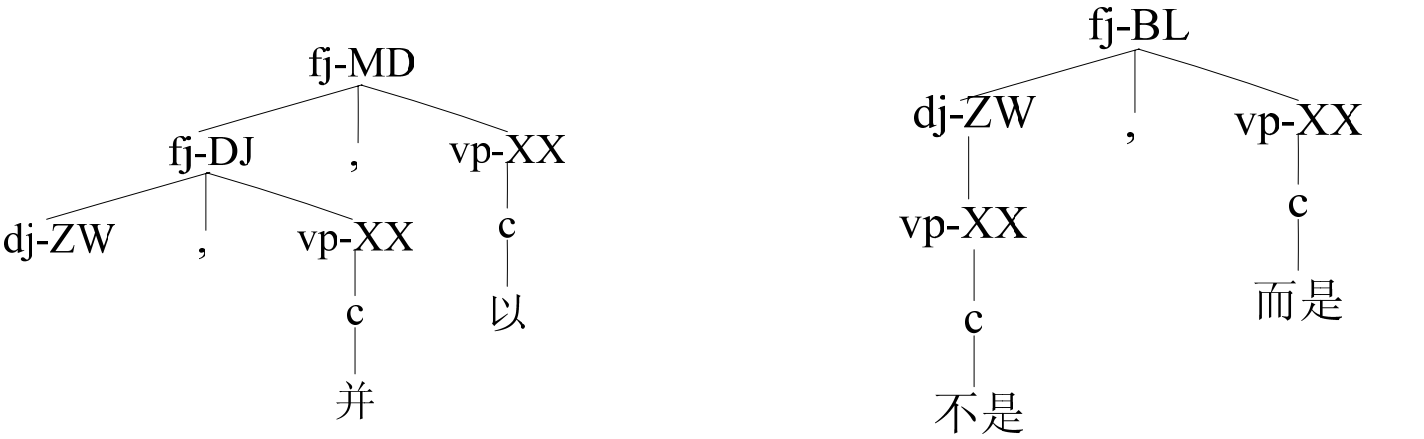
序号	标记符号	关系类型	序号	标记	关系类型	序号	标记	关系类型
1	BL	并列关系	5	YG	因果关系	9	ZE	转折关系
2	LG	连贯关系	6	MD	目的关系	10	JZ	解注复句
3	DJ	递进关系	7	JS	假设关系	11	LS	流水复句
4	XZ	选择关系	8	TJ	条件关系			

复句关系词的抽取与分类



- 他根据先秦有关理论原则，制订和推行了盐铁官营、酒类专卖，**并**首创均输和平准，**以**实现“民不益赋**而**天下用饶”（《史记·平准书》）的目的。

复句关系词的抽取与分类



fj-??

dj-??|fj-??|vp-?? ,|; dj-??|fj-??|vp-??

[vp-??|dj-??|c|d|l] [vp-??|dj-??|c|d|l]

图3 清华汉语树库中的复句关系标注

复句关系词的抽取与分类

表3 关系词及类别分布情况

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
l	407 / 187	12	22	0	1	5	0	4	2	3	358

表4 出现次数最多的复句关系词

副词（d）词	次数	连词（c）词	次数	连接语（l）词	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

实验结果

表5 是否为复句关系词识别准确率

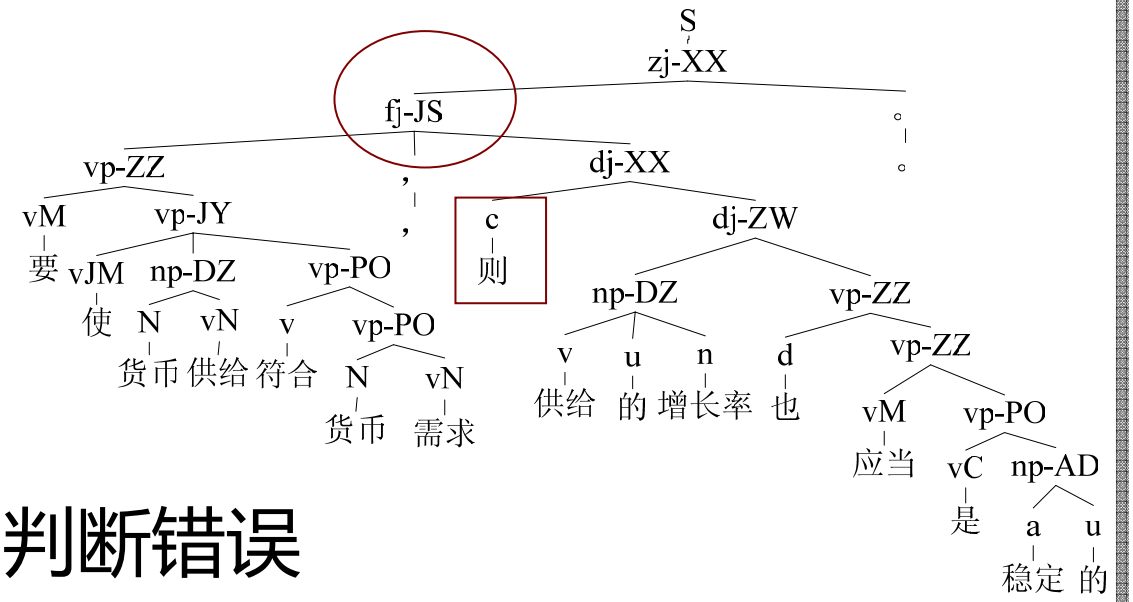
	自动句法树（不带功能标记）			自动句法树（带功能标记）		
特征	最大熵	决策树	贝叶斯	最大熵	决策树	贝叶斯
词汇	69.5	71.7	65.8	82.2	81.6	78.5
句法	90.8	90.7	90.6	94.8	95.0	93.5
词汇+句法	91.1	92.0	88.3	95.5	95.6	93.3
词汇+句法+位置	91.2	92.1	88.1	95.5	95.7	93.6

表6 复句关系词类别识别结果

		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
	F1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
	F1值	67.9	78.6	80.0	67.8	80.9	86.4	75.5	75.9	82.6	77.2

错误分析

– 抽取复句关系词错误



– 复句关系词类别判断错误

- 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。（并列关系错判为递进关系）

结论

主要进行汉语复句关系词的识别与分类

根据清华树库标注规则，利用算法抽取复句关系词及类别，形成实验数据

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F1值为77.2%