

基于规则的事件描述单元识别*

邢欣 郭浩 郭海旭 李茹

山西大学计算机与信息技术学院, 山西 太原 030006

Xingu.w@gmail.com

摘要: 本文主要讲述基于谓词识别及合并规则的事件描述单元的识别方法, 将完整汉语句子划分为各个事件描述单元。实验中所涉及到的训练集及测试集均来自新闻学术类 TCT 语料。利用由 14248 条句子中事件描述单元总结得到的规则, 对 3751 条句子进行事件描述单元自动识别。实验证明规则处理的方法可以实现对事件描述单元的有效识别。

关键字 事件描述单元 谓词 合并 规则

Events Describe-unit Identify Based on the Rules

Xing Xin, Guo Hao, Guo Hai-xu, Li Ru

School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

Abstract: The paper is about the identification method of events describe-unit based on the rules of predicate identification and consolidation, in order to divide the whole sentence into events describe-unite. The training set and test set involved in the experiment are from TCT Corpus of news academic. Through the rules summarized from the events describe-unite in 14248 sentences, 3751 sentences have been identified into events describe-unite automatically. The experiment shows that by the method of rules handling, events describe-unite can be identified effectively.

Keywords: events describe-unit, predicate, consolidation, rule

1 引言

汉语的完整句子的句法语义分析, 一直以来都是自然语言处理的一个难点。CIPS-ParsEval-2009 测评任务中, 将汉语句法语义分析分成了三个步骤: 事件描述单元识别, 事件描述结构分析和事件逻辑关系分析。通过第一个步骤事件描述单元的识别, 将复杂的整句划分出各个的事件描述单元, 为进一步的分析做准备。根据定义, 事件描述单元主要有两个部分: 事件描述小句和句首状语成分。其中事件描述小句在事件描述单元中占绝大多数, 其内部要包含完整的主谓结构。因此在识别事件描述单元的过程中, 谓词的识别尤为重要, 识别出小句中的谓词就可以大致确定该事件描述小句边界。除此之外, 根据语料库中实际出现的语言现象总结出一部分规则, 提高提高事件描述单元的识别正确率。

在确定事件描述单元的界限时, 我们采取了先分后合的策略, 先将每个整句按照标点符号划开, 再根据其是否拥有主谓结构或特定句型, 确定多个片语 (利用标点分割形成的句子碎片) 间是否需要合并。下文将详细介绍这种基于谓词识别与合并片语思路的事件描述单元识别方法。

* **基金项目:** 国家 863 高技术研究发展计划资助项目 (2006AA01Z142) 国家社会科学基金青年项目 (07CYY022)

2 基于规则的事件描述单元识别

2.1 语料分析

训练集语料的 171 篇文章是来自于约 48 万词规模的新闻学术类 TCT 语料，其中包括 14248 条句子，33467 个事件描述单元。其中占绝大多数的是事件描述小句，分为谓词是形容词和动词的两大类，还有部分事件描述单元为句首状语或单个名词、连词等。按照 TCT 汉语词类标记集中规定，动词分为动词 (v)，系动词 (vC)，助动词 (vM)，名动词 (vN)，补助动词 (vB)，双宾动词 (vSB)，兼语动词 (vJY)；形容词分为形容词 (a) 和副形词 (aD)。其中动词 (v)，系动词 (vC)，双宾动词 (vSB)，兼语动词 (vJY) 和形容词 (a) 可以作为谓词。如下表：

表 1. 事件描述小句中的谓词

谓词词性	事件描述小句举例
动词 (a)	[它/rNP 既/d 产生/v 生理/n 的/uJDE 快适感/n]
系动词 (vC)	[这/rN 也/d 是/vC 世界/n 上/f 最/d 早/a 的/uJDE 国立/b 医学/n 教育/n 机构/n]
双宾动词 (vSB)	[而是/c 迂回/v 曲折/a 地/uJDI 给/vSB 人/n 以/p 潜移默化/v 的/uJDE 感染/vN]
兼语动词 (vJY)	[使/vJY 人/n 感到/v 有所/v 领悟/v]
形容词 (a)	[由于/c 社会/n 性质/n 和/p 西欧/nS 各/rB 国/n 不同/a]

另一方面我们统计了，训练集中事件描述单元按照标点划分的情况。其中，标点符号 wE (结句点号，包括句号、叹号、问号、省略号等)，wP (分隔点号，包括逗号、分号、破折号) 以及 wM (冒号) 可以作为事件单元块的分隔符。统计后发现，对于只含有一个标点的句子，95.75% 的句子以该标点作为分割，进行事件单元块划分；对于含有两个以上标点的句子，无论其标点数量多少，其以所有标点做完全分割的标记结果所占比例是最高的。即形如 [/w] [/w] ... [/w] 的标记结果在所有标记中比例最高，在 171 篇测试集中，其句数占到了 44.83%。部分标点个数及标记形式如下表统计：

表 2. 标点及标记形式统计

0 个标点				1 个标点			
标记形式个数	2	总句数	927	标记形式个数	8	总句数	2401
标记形式		句数		标记形式		句数	
[]		740		/w[]		60	
“无标记”		187		[/w]		2299	
				其他		42	
2 个标点				3 个标点			
标记形式个数	9	总句数	3723	标记形式个数	19	总句数	2991

标记形式	句数	标记形式	句数
[/w][/w]	1906	[/w/w/w]	631
[/w/w]	1795	[/w/w][/w]	438
其他	22	[/w][/w/w]	302
		[/w][/w][/w]	1508
		其他	112

标记形式说明:

1. 未标记表示该句子不包含事件单元块。如: 缪鲁/nP
2. “[]”表示该事件块并未以标点作为分割。如: [本/rB 报/n 渥太华/nS 二/m 月/qT 二/m 十/m 六/m 日/qT 电/n]
3. “[/w]”表示该事件块以标点作为分割。如: [医/n 和/cC 药/n 是/vC 密切/aD 相关/v 的/uJDE]。/wE
4. “[/w/w]”表示该事件块以标点作为分割,且其中仍含有一个标点,如: [约/d 在/p 公元前/t 200/m 年/qT , /wP 就/d 有/v 人/n 用/p 动物药/n 马宝/n 解救/v 食物中毒/vN 者/k]。/wE

除此之外,在训练集中还有少量特殊的事件描述单元,这些事件描述单元不以标点符号为边界,直接在词间划分。这些特殊的事件描述单元大部分考虑到了句法功能及语义信息,划分难度大,难以处理。下面罗列出部分含有特殊事件描述单元的句子:

例1: [由于/c] [其时/t] [天气/n 炎热/a] 又/c [阴/a 晴/a 多变/v] , /wP [忽而是/c 骄阳/n 似/vC 火/n]、/wD [使/vJY 人/n 口/n 干/a 唇/n 裂/v , /wP 汗流浹背/v] ; /wP [忽而/c 又/d 大雨/n 倾盆/v] , /wP [使/vJY 人/n 如/vC 落汤鸡/n] , /wP [全/a 身/n 泥浆/n]。/wE

例2: [看来/l] [找到/v 楔进点/n] , /wP [再/d 热/a 的/uJDE 市场/n 也/d 能/vM 挤/v 进去/vB]。/wE

例3: [在/p 日本/nS 和/cC 美国/nS , /wP 他们/rNP 的/uJDE 面包/n 是/vC 必须/d 加/v 维生素/n B/n 1/m 的/uJDE] , /wP 如果/c [不/dN 加/v 的话/u] [是/vC 违反/v 国家/n 的/uJDE 规定/n , /wP 是/vC 犯法/v 的/uJDE]。/wE

2.2 算法思想

经过上述分析后,我们确立了识别事件描述单元的基本思想——“先分割,后合并”。这个思想主要有两方面的考虑,一方面训练库中大量的事件描述单元是有以标点作为边界,只出现了370次两个事件描述单元间的划分没有任何标点形式;另一方面通过对语料的统计,完全按照标点划分事件描述单元的正确率超过50%。在这个基础上增加规则合并片语,能容易提高识别的正确率。在“分”的基础上再利用认为总结出的规则,对其进行“合”的操作,从而正确识别事件描述单元边界。

因此我们识别事件描述单元的大致思路为首先对句子按标点进行完全分割,然后根据合并规则再进行块的合并。表示如下:

1. 对给定句子进行从左至右扫描,获取词及词性;
2. 扫描标点个数及位置;

3. 若有 1 个标点，则从左至右进行完整分割；
4. 若有两个标点以上。扫描标点所在位置，并根据位置将句子分割为若干片语，并依次设置片语标记符
5. 依次实现各合并规则，设置合并标记符；
6. 根据合并标记符，若右合并，则置片语标记符(n) =片语标记符(n+1)；若左合并，则置片语标记符(n) =片语标记符(n-1)；递归。
7. 根据片语标记符进行事件描述单元的识别：片语标记符相同的片语，则表示同在一个事件描述单元内；否则，片语分属两个不同的时间描述单元。

算法的实现主要依靠了对谓词识别及片语合并的规则，规则的正确和细致程度，直接影响到了最后的识别结果。下一章节，将详细讲述我们所总结出的规则。

3 事件描述单元识别规则

3.1 规则介绍

根据训练库中 171 篇文章的人工总结，我们共总结出 10 条谓语识别规则和 25 条片语合并规则。在这些规则的基础上，对汉语句子的时间描述单元的识别。下文将详细介绍谓语规则及合并规则。

3.1.1 谓词识别规则

正如上文所述，依据 TCT 词类标记规范中的定义，在训练集中只有词性为 v、vC、vBS、vJY 和 a 能作为谓词。但是并不意味着只要有这些词出现，就一定是作谓语成分。例 4 中的事件描述单元中只有第一个动词“看”和“是”是谓语动词，其他的动词“戴”、“蹬”、“看”、“睡”都不是谓语动词。

例 4：[你/rNP 看/v ， /wP 杨秀珍/nP 身上/s 穿/v 的/uJDE 老太太衫/n ， /wP 手/n 上/f 戴/v 的/uJDE 金/b 戒指/n ， /wP 脚/n 上/f 蹬/v 的/uJDE 皮鞋/n ， /wP 看/v 的/uJDE 电视机/n ， /wP 睡/v 的/uJDE 席梦思床/n ……/wE 都/d 是/vC 孩子/n 们/k 买/v 的/uJDE]

因此我们根据 171 篇文章总结了 10 条谓语识别规则，来判断片语中出现的动词及形容词是否作为谓语成分。下面是我们总结出的谓语识别规则：

- RULE1. 结构 v+uJDE 判断 v 为非谓语动词。
- RULE2. 结构 uJDE+v 判断 v 为非谓语动词。
- RULE3. 结构 v+uA+uJDE 判断 v 为非谓语动词。
- RULE4. 结构 v+f 判断 v 为非谓语动词。
- RULE5. 结构 v+fT 判断 v 为非谓语动词。
- RULE6. 结构 v+nT 判断 v 为非谓语动词。
- RULE7. 助动词 vM，名动词 vN，补助动词 vB 判断为非谓语动词。
- RULE8. 所有不违反规则 1-7 的动词判断为谓语动词。
- RULE9. 在没有谓语动词的情况下判断有无谓语形容词，句尾结构“uJDE+ (d) +a”判断为非谓语形容词。

RULE10. 若句尾是形容词 (a) 或形容词之后加“极、多、透”组成的述补结构, 判断 a 为谓
语形容词。

3.1.2 片语合并规则

在确定了谓词后, 就需要考虑到多个片语间的合并, 即事件描述单元的边界设置了。训练
库的丰富语料, 是规则的主要来源。通过观察已有的语言现象, 找出事件描述单元划分的规律,
总结出我们所需要的合并规则。理想的情况下, 每一个汉语句子都应该找到符合要求的若干规则
进行事件描述单元的识别, 且这些规则是互不干扰的。但在制定规则的过程中, 很难多做到全面
的考虑, 总会不断地出现与现有规则冲突的新现象。所以规则的制定是一个不断跟进的过程, 目
前我们共总结了 25 条规则。部分规则如下表所示:

RULE1. 谓词判断。若分句中不含谓词, 右合并, 递归。如: [藏族/nR 人民/n 的/uJDE 祖
先/n , /wP 在/p 公元前/t 几/m 个/qN 世纪/nT , /wP 已/d 认识/v 到/vB 某些/rB
动物/n 、 /wD 植物/n 、 /wD 矿物/n 有/v 治疗/v 疾病/n 的/uJDE 作用/n]。/wE

RULE2. 若小句首尾依次如: “介词/p……nt/f/fs/ns”, 右合并。如: [藏族/nR 人民/n 的
/uJDE 祖先/n , /wP 在/p 公元前/t 几/m 个/qN 世纪/nT , /wP 已/d 认识/v 到/vB
某些/rB 动物/n 、 /wD 植物/n 、 /wD 矿物/n 有/v 治疗/v 疾病/n 的/uJDE 作用
/n]。/wE

RULE3. 若分句尾为 fT (时间方位词), 右合并。如: [50/m 年代/nT 以来/fT , /wP
对/p 医史/n 、 /wD 古籍/n 文献/n 的/uJDE 发掘/vN 、 /wD 整理/vN 和/cC 出版
/vN ; /wP 以/p 现代/b 科学/n 方法/n 进行/v 中药/n 研究/vN 和/cC 临床/b 研
究/vN , /wP 已/d 取得/v 有/v 价值/n 的/uJDE 成果/n]。/wE

RULE4. “介词/p……qT”结构前后没有谓词, 右合并。如: [在/p 雷达/n 发明/vN 之
前/f , /wP 利用/v 脉冲/n 无线电/n 装置/n 测量/v 电离层/n 高度/n 的/uJDE 工
作/n 已/d 进行/v 多/m 年/qT]。/wE

.....

4 实验及实验结果分析

测评中我们只实现了部分规则 (5 条谓词规则和 13 条规则), 实验结果并不理想。下文将详
述分析这些规则在测试过程中发现的问题。

4.1 单一规则对于句子的影响

为了能够更好的了解所总结规则对句子事件描述单元识别的贡献程度, 单独将每条规则对
测试集进行了测试, 并用覆盖率这一指标来表征该规则的适用广度。覆盖率计算公式如下:

$$\text{覆盖率} = \text{规则适用句数} / \text{总句数}$$

在统计中我们选取了 13 条规则, 其中覆盖率最高的规则分别是 RULE1 (覆盖率为 45.39%)
以及 RULE14 (覆盖率为 15.87%)。而有 8 条规则的覆盖率不到 1%。其中 RULE8 的覆盖率为 0.00%,
表明测试集中并没有出现相应的语言现象。

对于规则的执行效果, 我们用贡献度这一指标来对其进行表征。其计算公式如下:

$$\text{贡献度} = (\text{使用该规则的正确数} - \text{不使用该规则的正确数}) / \text{该规则适用句数} * 100$$

通过统计我们看到, 在已实现的 13 条规则中, 5 条规则的贡献度为正值, 3 条规则的贡献

度为 0，而 5 条规则的贡献度为负值。那就意味着 13 条规则中，有 23%规则对于标记结果没有影响，而 38%的规则对于标记结果产生的负面影响。其中，RULE10 的贡献度为-100，如果抛开适用规则语料的稀疏问题的话，那么我们则有理由怀疑这条规则的制定是否正确。规则覆盖率及贡献度统计如下表所示：

表 3. 规则覆盖率及贡献度

规则	规则适用句数	覆盖率(%)	不使用该规则		单独使用该规则		贡献度
			正确	错误	正确	错误	
RULE1	1702	45.39	669	1033	686	1016	1.00
RULE2	150	4.00	47	103	56	94	6.00
RULE3	47	1.25	19	28	10	37	-19.15
RULE4	99	2.64	36	63	30	69	-6.06
RULE5	224	5.97	112	112	126	98	6.25
RULE6	13	0.35	7	6	5	8	-15.38
RULE7	1	0.03	1	0	1	0	0.00
RULE8	0	0.00	0	0	0	0	0.00
RULE9	1	0.03	1	0	1	0	0.00
RULE10	3	0.08	3	0	0	3	-100.00
RULE11	11	0.29	7	4	5	6	-18.18
RULE12	35	0.93	2	33	17	18	42.86
RULE13	19	0.51	1	18	3	16	10.53

4.2 规则间的约束关系

从规则角度出发，我们利用规则的覆盖率来表征该规则适用的广度。而从句子角度出发，我们则需要对每个句子适用规则的个数进行统计。统计结果如下：

表 4. 句子适用规则统计

适用规则数	句子数	比例 (%)	适用规则数	句子数	比例 (%)
适用 4 条规则	1	0.00	适用 3 条规则	54	1.44
适用 2 条规则	557	14.8	适用 1 条规则	1535	40.9
适用 0 条规则	1603	42.7			

统计结果中，适用不同规则数最多的 4 条规则，只适用 1 条规则的句子为 1535 句，达 40.9。而不适用任何规则的句子为 1603 条。这就意味着有 42.7%的句子没有被我们所设定的 14 条规则影响到。由此可见，目前我们规则的数量和覆盖率上还是远远不够的

同时我们注意到，有 16%的句子同时适用于多条规则。那么，不同规则对同一句子的执行是否会因不同规则间执行方法、执行顺序以及执行判别条件的不同而对句子产生不同的影响呢？即规则间是否存在相互的约束关系。如下是 RULE1 和 RULE5 共同适用的实例：

例 5: [不过/c , /wP 由于/p 合理/aD 预期/v 的/uJDE 作用/n] , /wP [的确/d 会/vM 使/vJY 宏观/n 政策/n 的/uJDE 有效性/n 显著/aD 小于/v 凯恩斯/nP 模型/n 的/uJDE 期望值/n] 。/wE

对适用 2 条规则的 557 句语料进行规则的分配统计, 统计如下:

表 5. 适用 2 条规则的句子统计

RULE1	RULE11	6	RULE2	RULE12	1
	RULE12	23		RULE13	1
	RULE13	11		RULE5	3
	RULE2	68	RULE4	RULE12	1
	RULE3	30			
	RULE4	64	RULE5	RULE13	1
	RULE5	109			
	RULE6	9			

4.3 规则的串行执行

在规则的串行执行中, 不同的规则执行顺序是否会对标记结果产生影响? 利用上述统计, 我们对 RULE1 和 RULE5 共同适用的 109 句, 进行了改变规则执行顺序的实验, 试验结果如下:

表 6. 改变规则顺序结果

RULE1 和 RULE5	正确	错误
先执行 RULE1, 后执行 RULE5	60	49
先执行 RULE5, 后执行 RULE1	43	66
影响度 (%)	15.59	

由试验结果可知, 不同的规则执行顺序对于标记的结果是有影响的, 而面对不同的规则匹配对, 其影响的程度也是不同的。因此我们利用规则的权重权重来进行量化, 量化值越高的规则则优先执行。定义权重的计算公式:

$$\text{权重} = \text{覆盖率} * \text{贡献度}$$

规则权重计算结果如下:

表 7. 规则权重

规则	权重	规则	权重	规则	权重	规则	权重
RULE1	45.33	RULE5	37.33	RULE9	0.00	RULE13	5.33
RULE2	24.00	RULE6	-5.33	RULE10	-8.00		
RULE3	-24.00	RULE7	0.00	RULE11	-5.33		
RULE4	-16.00	RULE8	0.00	RULE12	40.00		

在此基础上，我们利用规则的权重对原有算法进行优化，首先去掉权重为负值的规则，然后根据权重来调整规则的执行顺序。试验结果如下：

表 8. 改变规则顺序结果

权重	规则执行顺序	正确句数	正确率 (%)
高->低	RULE1->RULE12->RULE5->RULE2->RULE13	2231	59.49
低->高	RULE13->RULE2->RULE5->RULE12->RULE1	2211	58.96

由表 8 中的结果可以看出，规则的顺序虽然对事件描述单元识别的结果有所影响，但幅度有限。一个原因可能是规则本身设定的问题，另一个原因可能是规则的串行执行并不适用于事件描述单元的识别。这还需要进一步的分析实验才能得出最后的结果。

4.4 结果分析

经过局部规则的添加和调整，在正式测评中我们的正确率只达到了 60.70%。这样的结果并不尽如人意，但是也在预料之中。在没有完整的观察并统计训练库所包含语言现象的情况下，规则的制定是缺少说服力的。另一个重要的方面就是没有对谓词充分考虑。如上文所述，谓词包括动词和形容词两个部分，而实验中对谓词规则缺少了对谓语形容词的判别，影响到形容词作为谓词的事件描述单元识别。

在进一步总结训练库的基础上得出更多更细致的规则后，相信结果应该还有很大的提高空间。

5 结束语

汉语句子的完全句法语义分析是复杂的，本次测评中提出的事件描述单元的概念将整句分成了若干小的单元，使得句法分析的可行性提高，为下一步的句法语义分析提供了很好的基础。但是事件描述单元的定义还是模糊的，在训练库中的许多事件描述单元的划分都依据了句法功能等因素，使得在只有词与词性信息的情况下识别难度增大。例如状语成分何时需要切开，事件描述小句缺失主语的情况，可能都需要更明确的定义。尽管我们初步的实验得到的结果并不是那么理想，但也让我们看到了利用规则方法解决事件描述单元划界问题的问题和继续下一步研究的方向。

致谢：我们来自刘开瑛教授牵头的基于 CFN 句法语义研究团队，参赛者有李茹教授和四个学生（郭海旭、邢欣、郭浩、温锋瑞），在评测中，受到刘开瑛教授的悉心指导，实现了任务四和任务五；另外，在实验中，得到了 HIT 提供的问句语料，在此，一并向他们表示诚挚的感谢。

参考文献

- [1] 朱德熙 《语法讲义》商务印书馆 1982 P55-P77.
- [2] 周强 句法树标注规范 清华大学