

汉语句法规则的排歧与禁止规则*

陈刚 苑春法 黄昌宁
“智能技术与系统”国家重点实验室
清华大学计算机系

摘要：本文在依存语法的基础上，首先提出了汉语句法规则中二类歧义的概念。然后结合语义和上下文有关信息提出了解决第二类歧义的方法，最后生成无歧义的禁止规则。实验结果表明，禁止规则很好地解决了汉语句法规则中的第二类歧义问题，并且其规则的学习对于汉语句法规则的自动学习具有普遍意义，显示了很好的应用前景。

关键词：依存语法 一类歧义 二类歧义 禁止规则

Disambiguation of Chinese Grammar Rules and Forbidden Rules

Chen Gang, Yuan Chunfa and Huang Changning

State Key Laboratory of Intelligent Technology and System

Department of Computer Science & Technology, Tsinghua University, Beijing 100084

ABSTRACT: Based on dependency grammar, this paper first put forward the concept of class 2 ambiguity in Chinese grammar rules, then we disambiguate the class 2 ambiguity by using part-of-speech, semantic and contextual information together, at last forbidden rules were produced. Experimental results proved that forbidden rules deal perfectly with the class 2 ambiguity in Chinese grammar rules, and are of general significance in the learning and expression of Chinese grammar rules. All of that show there is a good application prospect for this method.

Keywords: Dependency Grammar, Class 1 Ambiguity, Class 2 Ambiguity, Forbidden Rules

1. 引言

在自然语言的句法分析系统中，句法规则起着很重要的作用。因此，句法规则的构造和描述问题一直是计算语言学界研究的一个热点。国外在英语方面已做了许多工作。如：H.H. Shih, S. J. Young 和 N.P.Waegner (1995)提出了一个计算机辅助句法构造系统 (CAGC) [1]。G.Carroll, E.Briscoe 和 C.Grover 等在广义短语结构语法 (GPSG) 框架上开发的句法规则自动构造工具 GDE[4][5]等。

在中文信息处理方面，清华大学的周明等在汉语依存体系的基础上运用基于变换的方法对汉语句法功能标注的进行了探讨[10]。清华大学的周强在其博士后的研究中设计了一个汉语概率型上下文无关语法 (PCFG) 的句法规则的自动构造工具[9]。清华大学的许伟在其硕士论文中进行了汉语句法规则的分类和构造的研究[7]。另外，香港中文大学 Angel S.Y.Tse 等也曾结合语言学方法与基于语料库的统计方法

* 自然科学基金资助项目

对汉语名词短语规则的获取做过有益的探讨[6]。

但是，以上的方法都有一个共同点那就是无论从知识的表示和规则的学习上多是基于词类的。而对于汉语，由于词类与句法功能之间不存在明显的直接映射，所以在词类的基础上建立句法规则，其歧义性是相当严重的。本文尝试首先将学到的基于词类的上下文无关的句法规则的歧义进行分类，然后针对其中的二类歧义，利用上下文和语义信息将规则排歧，并最终形成上下文有关的禁止规则。

本方法认为上下文无关规则的歧义是由于没有上下文知识和知识表示的颗粒度太粗引起的。所以本文在引入了上下文的同时，在禁止规则的表示上，从知识的最粗颗粒度（即词性）到语义，到最细颗粒度（即词形）又形成了层次表示。在这里，知识的表示被限定到了它不产生歧义的最粗颗粒度。

在下面几节里，第二节介绍了句法规则中的二类歧义问题以及禁止规则的定义；第三节重点论述在学习过程中语义和上下文知识的引入以及由此对第二类歧义排歧生成的禁止规则；第四节是实验结果与分析。

2. 句法规则中的二类歧义及禁止规则的定义

本文选择了依存语法作为句法规则自动获取的基础。为了获取上下文相关的禁止规则，我们还在依存语法中加入了词的位置信息。

在依存语法的体系下，词性层面的上下文无关的句法规则可以形式化地表示如下：

$$R_b = A, B, \text{FLAG}$$

其中，A，B是相邻两词的词性。FLAG ∈ {BACK, FRONT}分别表示A和B的依存方向是向后还是向前。如(vg ed BACK)表示一个动词在助词“的”的前面，则马上可以判定该动词是依存于“的”的。

但是在实际的规则学习中，很多直观上认为是词性层面的上下文无关规则，在真实的大规模语料库中经常碰到反例，一种反例是A，B的依存方向在真实文本中正好与二元词性规则相反，比如在词组“教育经费”学习到的二元词性规则(vg,ng,BACK)用于分析词组“学习知识”时所遇到的错误；而另一种反例是A，B在真实文本中出现时并没有依存关系，比如上下文无关规则(a, ng,BACK)，在词组“学vg好a技术ng”中所遇到的优先捆绑错误，这里我们把反例中的实际依存方向与规则相反时的歧义叫做第一类歧义，而把反例中的相邻词不存在实际的直接依存关系的歧义叫做第二类歧义。

本文重点是针对二类歧义规则的排歧。第二类歧义规则的出现，我们认为这是由于上下文知识和语义信息不足引起的，所以对于第二类歧义规则的排歧，应该从引入上下文和语义知识入手，生成新的上下文有关规则来指明何时该词性层面的上下文无关规则不应该使用。本文把这种规则叫做禁止规则，其形式化的表示成如下形式：

$$R_f = A, B, \text{FLAG } E;$$

R_f 表示在上下文环境E下，上下文无关规则：A，B，FLAG应该被禁用。其中，FLAG的意义同 R_b ，但是此时的A，B不仅可以是相邻词的词性，而且还可以是相邻词的语义类代码。E表示上下文信息，我们如下规定E的形式：

设A，B对应的两个词是 w_1 和 w_2 ，它们出现在句子中的横切面形式是... $w_{i2}w_{i1}w_1w_2w_{r1}w_{r2}$...，（即 w_1 与 w_2 相邻且 w_1 在 w_2 的前面），设 w_i 的词类是 $c_i(i=...,l2,l1,l2,r1,r2...)$ ，如果在 R_b 中有元素(c_1,c_2,FRONT)，在观察实例中 w_2 的主词是 w_n ， w_n 所对应的主词是 w_m ，规定：

当 $n \geq r1$ 时，E的形式为“+C”，表示 w_2 新的主词 w_n 的位置在 w_2 之后。C表示 w_n 的词类或语义类。

当 $n \leq 11$ 时, E 的形式为 “-C”, 表示 w_2 新的主词 w_n 的位置在 w_2 之前, C 表示 w_n 的词类或语义类。

因为我们是利用语义来解决歧义规则中的歧义, 而象助词, 感叹词这样的虚词只与句子的结构有关, 而与语义无关。所以当 w_n 是虚词时, 我们还要考虑 w_n 的所对应的主词 w_m 的信息。在这里, 我们是在 E 中加上了一项表示, 规定:

当 w_n 是虚词时, E 表示为如下形式:

$E = \text{LOC1 C LOC2 D};$

其中, $\text{LOC1}, \text{LOC2} \in \{+, -\}$ 。LOC1 和 C 定义如前所示。D 为 w_m 的词类或语义类, 当 $m \geq n$ 时, $\text{LOC2} = “+”$; 当 $m \leq n$ 时, $\text{LOC2} = “-”$ 。

当 $\text{FLAG} = \text{BACK}$ 时, 定义基本如上, 只不过 w_n, w_m 分别是 w_1 的主词和 w_1 的主词的主词, 而非 w_2 的主词和 w_2 的主词的主词。

总结一下, 禁止规则可以表示为如下几种形式之一:

$(c_1, c_2, \text{BACK}-c)$	$(c_1, c_2, \text{FRONT}-c)$
$(c_1, c_2, \text{BACK}-c, -d)$	$(c_1, c_2, \text{FRONT}-c, -d)$
$(c_1, c_2, \text{BACK}-c, +d)$	$(c_1, c_2, \text{FRONT}-c, +d)$
$(c_1, c_2, \text{BACK}+c)$	$(c_1, c_2, \text{FRONT}+c)$
$(c_1, c_2, \text{BACK}+c, -d)$	$(c_1, c_2, \text{FRONT}+c, -d)$
$(c_1, c_2, \text{BACK}+c, +d)$	$(c_1, c_2, \text{FRONT}+c, +d)$

3. 利用上下文和语义知识获取禁止规则

如前所说, 二类歧义的出现, 是由于无上下文信息和词性知识颗粒度太粗所引起的。所以, 二类歧义规则排歧的关键在于上下文有关信息和语义知识的引入。

3.1 对第二类歧义规则的上下文限定

第二类歧义规则是指规则所出现的相邻词对在语料库中有不存在实际依存关系的反例的词性层面的上下文无关规则。对于二类歧义规则的排歧我们首先要确定其上下文环境。

举例来说明学习的过程, 假设在依存关系树库中有如下的两个句子:

(1) 好 a 人 ng 一生 ng 平安 a

(2) 我们 rp 应该 u 努力 wg 学 vg 好 a 技术 ng

因为它们都在树库中, 所以我们知道它们的正确结构, 现在我们来描述学习过程 (只考虑两个句子中加下划线的形容词): 当首先看到句子(1)时, 能学习到 “形容词在名词前面时, 可以做名词的修饰语” 这样的知识, 记为规则

(a)(a,ng,BACK),

(a)就是前面讲的上下文无关规则。在试图用规则(a)分析(2)时遇到了矛盾: (2)中的形容词 “好” 不是依存于其后的名词 “技术” 而是依存于其前的动词 “学”。也就是 (2) 中的 a (好) 和 ng(技术)不存在依存关系。这时(a)规则变成了一条第二类歧义规则。为了解决歧义, 一个简单的方法是把由主词确定的环境记忆下来, 即学习到知识 “当一个形容词在一个名词前面但它前面有动词时, 规则(a)应该禁用”, 这样的知识我们简记为

(b) (a,ng,BACK-vg)

(b)规则对(a)在上下文环境中作了限制。它属于禁止规则集, 在这里, 我们记录的是(a)规则的从词在歧义情况下所对应的新的主词的词性。如果该新的主词是实词, 我们认为该上下文限制已经够了。如果是虚词, 我们还要记录(a)规则中主词在歧义情况下所对应的新的主词词性。这主要是出于以后语义排歧

考虑。因为虚词只于句子的结构有关，而于语义无关。如在“参观 vg 图书馆 ng 的 ed 人们 ng”中学到的规则（vg,ng,FRONT）在遇到的歧义情况“参观 vg 图书馆 ng 的 ed 大厅 ng”时，光靠 ng 所对应的新主词“的”是解决不了歧义的，必须在歧义规则所对应的禁止规则的上下文中包括“的”所对应的主词“大厅”。

3.3 语义知识的运用

假设在依存关系树库中又出现第三个句子：

(3) 政府 ng 给予 vg 特殊 a 补助 ng

如果语料库中没有句子(3)的话，上下文无关规则(a)和禁止规则(b)在判定形容词和名词的关系的问题上已经足够了。但是用(b)分析句子(3)时又提出了难题，在同样的上下文中，形容词和名词是否有依存关系无法确定。显然，该冲突的出现也是由于句子(2)和句子(3)在词类层面上同形结构的存在。我们必须在更细的知识颗粒度上对其进行排歧。

对于同形结构：学 vg 好 a 技术 ng；给予 vg 特殊 a 补助 ng，直观上，只要我们有这样的知识：“好”可以和“学”搭配，而“特殊”不能和“给予”搭配，就能将歧义规则排歧。于是在规则学习时就将歧义结构标注上了语义类。如上例，在《词林》中可以找到能区分同形结构中“学”和“给予”两词最抽象的语义类为 H 和 J；能区分“好”和“特殊”的最抽象的语义类分别是 ED03 和 ED04；“技术”和“补助”分别是 D 和 H；这样同形结构被区分开了，规则歧义也就解决了，我们得到了一条无歧义的禁止规则：ED04, D BACK-H。它表示歧义规则：(a,ng,BACK) 在当形容词的语义类为 ED04，名词语义类为 D，并且其前有一个动词的语义类为 H 时才被禁止使用，其他情况下是无歧义的。

又比如另一条出现频度较高的词类层面上有歧义的规则：vg,ng,FRONT+ed+ng，假设它在树库中的实例是“参观图书馆的大厅”和“参观图书馆的人们”。显然要解决这个歧义只需要让机器知道“参观”可以和“大厅”搭配，而不能和“人们”搭配就可以了。于是我们通过语义类的标注得到了能区分“大厅”和“人们”的最抽象的语义类，它们分别为 B 和 A，那么我们在知道了正确依存关系的树库中就可学到一条无歧义的禁止规则：vg,ng,FRONT+ed+B。

3.6 禁止规则集合自动学习算法

在实验中，我们禁止规则学习所用到的树库是一个有 967 句、25293 个词次的依存关系树库。

在学习过程中，设在词性层面上上下文无关规则集合为 B 而禁止规则集合为 F，同时我们用了一个辅助的集合 S，它表示在词性层面上的禁止规则集合（显然，其中的规则有一些仍是有歧义的）。另外，为了便于义类标注，我们把 S 集合中每一条规则 s_i 的同形结构记录在一个新的集合 A_i 中。我们禁止规则学习算法的步骤如下：

1. 统计树库中所有依存关系树的所有树边，就得到了词性层面上的上下文无关规则集合 B；

2. 对于 B 中的每一个元素 $b_i=(c_1,c_2,FLAG)$ (c_1,c_2 表示词性；FLAG 为 FRONT 或 BACK 之一) 扫描树库的所有依存关系树，考察树中的所有的相邻词对 (w_1,w_2) ，如果 (w_1,w_2) 满足 b_i 的形式，但它们不存在的依存关系时则向集合 S 中添加元素 $(c_1,c_2,FLAG,e_i)$ ，其中 e_i 的定义见上一节，它由 c_1 或 c_2 (FLAG='FRONT' 时为 c_2 ,FLAG='BACK' 时为 c_1) 的真正主词所决定，并且当该主词是虚词时，添加该主词所依存的主词；

3. 对 S 中的每一元素 $s_i=(c_1,c_2,FLAG,e_i)$ 再次扫描树库的所有依存关系树，考察树中的所有的相邻词对 (w_1,w_2) ，如果 (w_1,w_2) 和其上下文环境满足 s_i 的形式，则将该词对和规则要求的上下文环境以及正确的依存关系一起记录在 s_i 规则所对应集合 A_i 中。显然， A_i 就是相对于词性层面上的禁止规则 s_i 的所有同形结构所组成的集合；

4. 对 S 中的每一元素 s_i ，考查它所对应的 A_i 。若 A_i 中所有同形结构依存方向不满足 FLAG 的形式，

这说明上下文 e_i 已经可以使得 s_i 不出现歧义了。将规则 s_i 写入禁止规则集合 F ；若依存方向有满足 FLAG 的出现，则说明规则需要用语义排歧，于是执行下面步骤：

5. 为 A_i 中的每一个元素的每个词标上唯一的语义类。这里，我们标注的全是小类；重新考查 A_i 中的每一个元素，首先将它们分成依存方向满足和不满足 FLAG 的两个集合，这里我们把它们分别设为集合 T 和 H ，同时，我们把通过对 T 和 H 的比较后生成的义类一级的禁止规则设为 F_i ，其形式为 $F_i=(j, k \text{ FLAG } r)$ ；

6. 比较规则中 c_1, c_2 和 e_i 对应的实词（方便起见，设为 c_3 ）在 T 和 H 集合中的各自的词性、词形和语义类的出现，以 c_1 为例，当：

i. T 和 H 中对应于 c_1 位置的词形相同，则取 $j=c_1$ ；

ii. T 和 H 中对应于 c_1 位置的词形不同但语义类相同，则取 $j=c_1$ 在 H 集合所对应的词形。这说明只有在最细的知识层面上才能对该歧义规则排歧；

iii. T 和 H 中对应于 c_1 位置的词形不同且语义类也不同，则取 j 为能 H 中能区分开 T 和 H 的最粗的语义类代码。例如，若 c_1 在 T 和 H 中对应的两个语义类分别是 $Gb09$ 和 $Gc17$ ，则取 H 中能区分开 T 和 H 的最粗的语义类代码为 Gc 。因为本文的语义类代码是依照《词林》的分类体系，而《词林》的分类体系是一种层次结构。所以语义类代码的粗细程度就是用它的位数表示的；

如法炮制，确定出 F_i 中的 k, r 。这里要注意的是由于 T 和 H 集合中的元素不止一个，所以 F_i 并不一定唯一，它表示了歧义规则应该被禁止的所有情况。

7. F_i 确定后，将 F_i 加入 F 即可。

8. 重复执行步骤 2 到 7，直到集合 F 不再增长为止。

4. 实验结果与分析

i. 下表是语料库中出现频率最高的前 5 位的出现第二类歧义的词性层面的上下文无关规则所生成的禁止规则的学习结果：（表 1）

歧义规则	禁止规则
vg,ng,FRONT	Vg, ng, FRONT -pz : vg, ng, FRONT -s Vg, ng, FRONT -ed -Bn; Ie, Cb, FRONT -Dj Ih10, Dk, FRONT + Hg03; Hi37, Dd05, FRONT -社会 Jd07, Di, FRONT -Dn; H, ngd, FRONT -ng
ng,ng,BACK	Ng, ng, BACK+pg; ng, A, BACK+vg Ng, D, BACK+Hi; Di02, Gb, BACK-Ih A, D, BACK+和; Dm, Di, BACK+Aa
vg,a,FRONT	Vg, Ed, FRONT+A; vg, ED01, FRONT+D Vg, Ed, FRONT+Ca; vg, Ed, FRONT+H Vg, Ef, FRONT+Ed01; vg, Ee, FRONT+Ed
pz,ng,FRONT	Pz, ng, FRONT+ng; pz, Dj, FRONT+I Pz, Dd, FRONT-Hj02; pz, Dj, FRONT+Ih
cm,ng,FRONT	Cm, Db, FRONT+ed+ng; cm, B, FRONT+B Cm+B, FRONT+De; Kb09+ng, FRONT+vg

表 1. 由 5 位的出现第二类歧义的词性层面的上下文无关规则所生成的禁止规则

ii. 表 2 是禁止规则收敛性的实验结果。为了验证禁止规则集学习过程的收敛性，我们按照 1/8、2/8、...、15/8 等 8 个比例抽取了 8 个不同规模的训练集，作为规则集的学习资源。其结果如下（表 2）

训练集规模	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%	87.5%	100%
规则集合								
词性层面上上下文无关规则集 B	331	527	651	828	1174	1203	1263	1284
禁止规则集 F	335	432	487	566	694	712	732	741

表 2. 禁止规则收敛性实验结果

在图 1 中可以更直观地看到禁止规则的收敛性:

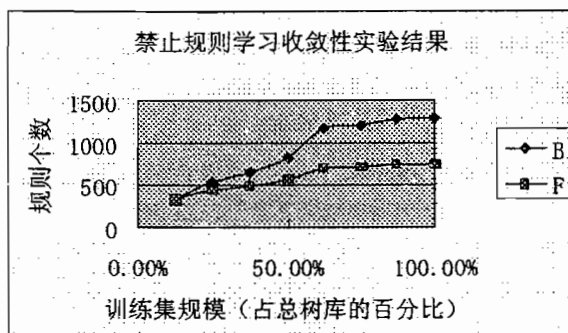


图 1. 禁止规则学习收敛性实验结果

从学习结果表 1 我们可以看出, 加入了语义和上下文的禁止规则虽然在形式上没有词性层面的上下文无关规则直观, 可是它们很好地解决了规则的歧义问题。我们在研究中发现仅靠词性的形式表示的规则歧义性是很严重的, 而仅以义类表示的规则数量又过于庞大。本文思路是将两者有机的结合起来, 并找到适当的语义和上下文知识的切入点。其方法对于汉语句法规则的自动学习具有普遍意义。

从表 2 和图 1 也可以看出, 本文所用的禁止规则学习算法学到的禁止规则集也是收敛的 (从图 1 可以直观看到, 在训练集规模占树库 60% 左右时开始收敛) 显示了很好的应用前景。

参考文献:

- [1] H-H. Shih, S. J. Young, N.P. Waegner, An inference approach to grammar construction, Computer Speech and Language, 1995.
- [2] Eric Brill, Automatic Grammar Induction and Parsing Free Text, A Transformation Based Approach, Proceeding of the 31th Meeting of the ACL, 1993
- [3] J. Carroll, E. Briscoe, & C. Grover, A development environment for large natural language grammars, Technical Report, Computer Laboratory, Cambridge University, England, 1991.
- [4] E. Briscoe, C. Grover, B. Boguraev, & J. Carroll, A formalism and environment for the development of a large grammar of English, Proceedings of the 10th International Joint Conference on Artificial Intelligence, Milan, Italy (1987).
- [5] E. Charniak & G. Carroll, Context-Sensitive Statistics For Improved Grammatical Language Models, Proc. of AAAI-94, 1994.
- [6] Angel S.Y. Tse, Kan-Fai Wong, Boon Toh Low, Wenjie Li, and Vincent Lum, CNP3 - Chinese Noun Phrase Partial Parser. Proc. of ICC'96, June 4-7, Singapore, 1996.
- [7] 许伟, 句法-语义一体化的汉语句法分析研究, 清华大学硕士学位论文, 1997.
- [8] 苑春法、许伟、黄昌宁, 汉语语义关联网研究, 《语言工程》, 1997.
- [9] 周强, 汉语句法规则的自动获取及其应用, 清华大学博士后出站报告, 1998.
- [10] 周明、潘海华, 基于变换的汉语句法功能标注探讨, 《中文信息学报》, Vol. 11, No 4, 1997.