统计与规则相结合的汉语句法分析研究

李京葵　周明　黄昌宁
清华大学计算机科学与技术系，北京　100084

摘　要

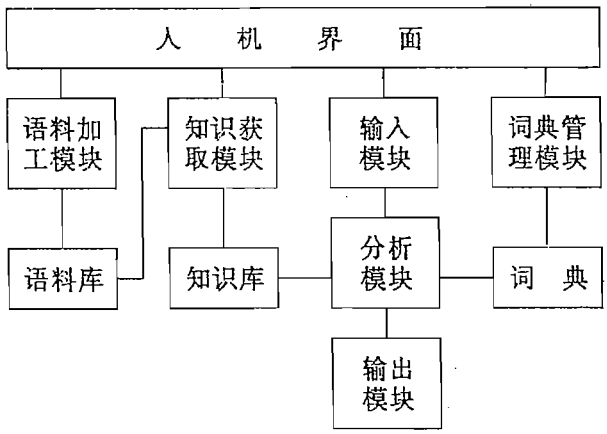
本文采用依存文法，设计了一个统计与规则相结合的汉语句法分析系统，其中，统计方法和规则方法共同用于依存网络的建立和剪枝过程，用语料库支持知识获取和分析结果评价。

关键词：自然语言处理　句法分析　依存网络　依存文法

1　引言

传统的基于规则的句法分析方法有如下缺陷：①规则是从封闭语料中总结出来的，对开放语料的处理不理想；②分析器鲁棒性较低，在遇见意外情况（如生词）时常常失败；③无法表示小颗粒度的知识，因此处理歧义的能力不强；④难以保证规则的一致性。近年来兴起的以语料库和统计方法为基础的句法分析方法的主要优点是：①对语言处理提供了较客观的度量；②对处理语言的不确定性问题有一定优势；③可以获取大量小粒度的知识，因此处理歧义的能力增强；④知识一致性好。其主要不足是：①在某些情况下难以避免时间和空间的组合爆炸；②难以表达语言的确定性现象。

规则方法已有成熟的经验，语料库和统计方法也在自然语言处理的许多方面取得了令人满意的结果，因此，我们考虑建立一个融和二者优点的弹性分析系统，其思想是：利用语料库支持各类知识和统计数据的抽取，并检验句法分析的结果；对简单语言现象沿用规则方法以保持较高效率；对规则难以处理的现象则探讨用统计方法解决。系统结构如下：



主要资源有：(1) 词典，含7万词条，每一词条标记了语义和句法类；(2)语料库：1300个典型简单句；(3) 知识库：存放从语料库获取的词与词，类与类之间的各种依存关系知识。分析步骤是：①输入一个已分词的句子；②用统计方法标注词性；③应用规则归并邻接词；④在知识库中检索每一词的搭配、传递和配价知识；⑤建立依存网络；⑥通过统计方法标注向上依存关系，简化依存网络；⑦应用规则对依存网络作进一步的简化；⑧对依存网络中的树进行评价并输出。

2　依存网络（DRN）的建立

2.1 语料库的标注

本系统定义了45种依存关系，标注时标出每个词的主词和对应的向上依存关系(UDR)，就可表达出该句的依存关系树（DRT）。标注形式见下表：

序号	词项	主词序号	向上依存关系	说　明
1	我	2	SUB	主语
2	是	0	GOV	全句的主词
3	他	4	DEP	'的'字结构
4	的	6	ATTA	定语
5	好	6	ATTA	定语
6	同学	2	OBJ	宾语

2.2 知识的获取

对每一词Wi，可从标注的语料库中获取关于具体词的三类低层知识：搭配模式，传递模式，配价模式，进一步归纳出关于词性类的三种高层知识[1]。若某词未在语料库中出现，低层知识库中就查不到以该词为索引的知识，可根据该词的语义句法类查高层知识库，只要与该词的语义句法类相同的其它词曾经在语料库中出现过，就会查到以该语义句法类为索引的知识，这样可以克服知识短缺现象，提高鲁棒性。

2.3 邻接词的归并

在汉语句子中，某些特定类型的相邻词之间往往存在确定的依存关系。邻接词归并就是使用简单规则测试相邻词的词性，在满足规则的相邻词间建立确定的UDR。例如规则

a + n + (~n) → <ATTA, 1, 2>　　a ←ATTA— n　~n

说明，当Word1词性为a（形容词），后面紧邻词Word2词性为n（名词），且Word 2后紧邻词（包括标点符号）词性不为n时，在Word1 和Word2之间建立依存关系"ATTA"（定语），其中Word2是Word1的主词，即Word2>>Word1。

系统中使用的邻接词归并规则如下（共11条）：

(1) a + n + (-n) → <ATTA,1,2>　　a ←ATTA— n　-n

(2) a + usde → <DEP,1,2>　　a ←DEP— usde

(3) a + usdi → <DIP,1,2>　　a ←DIP— usdi

(4) d + v → <ADVA,1,2>　　d ←ADVA— v

(5) d + usdi → <DIP,1,2>　　d ←DIP— usdi

(6) v(-vc) + usdf → <COMP,2,1>　　v(-vc) —COMP→ usdf

(7) v(-vc) + vc → <COMP,2,1>　　v(-vc) —COMP→ vc

(8) usde + n + (-n&-cm) → <ATTA,1,2>　　usde ←ATTA— n　-n&-cm

(9) usdi + v → <ADVA,1,2>　　usdi ←ADVA— v

(10) usdf + d + (-d&-a) → <DEIP,2,1>　　usdf —DEIP→ d　-a&-d

(11) usdf + a + (-d&-a) → <DEIP,2,1>　　usdf —DEIP→ a　-a&-d

〖句1〗　1　2　3　4　5　6　7　8　9　10　11　12

她　上午　收　到　朋友　的　一　封　很　长　的　信　。

r　t　vg　vc　ng　usde　m　q　d　a　usde　ng　w

运用规则(7)(2)(8)，建立依存关系<COMP, 3, 4>，<DEP, 11, 10>，<ATTA, 12, 11>。

应用完这些规则后，没有主词的词称为活动结点，已有主词的词称为非活动结点，后者不再与其它词建立任何UDR，不需参与知识库的查找，分析效率可以得到很大提高。

2.4 依存网络的建立

建立DRN时，先以词汇为索引查低层知识库，若未查到，再以词性为索引查高层知识库，然后即可建立句子的依存网络[1]。句1 查询知识库后建立的DRN共17 个UDR如下：

向上依存关系	主词	从词	向上依存关系	主词	从词
0 (GOV)	0	3	5 (COMP)	3	4
3 (ATTA)	12	11	2 (OBJ)	3	12
0 (GOV)	0	6	24 (DEP)	6	5
24 (DEP)	11	10	24 (DEP)	11	5
4 (ADVA)	10	9	3 (ATTA)	7	6
1 (SUB)	3	1	3 (ATTA)	12	6
24 (DEP)	6	1	17 (LA)	8	7
24 (DEP)	11	1	3 (ATTA)	12	8
4 (ADVA)	3	2			

3 剪枝

3.1 统计方法剪枝

上述方法建立的DRN中，每一词可能有多个UDR，采用UDR自动标注方法根据统计数据得出它们出现的可能性，即可对DRN作一定程度的裁剪。

UDR自动标注就是标注每一词的UDR，系统采用动态规划的FB算法[2]，并得到每一UDR的可信度，反映了出现的可能性。目前对1300个句子进行了封闭测试，当返回一个标记和两个标记时，正确率分别为94.6%和96.2%。

对〖句1〗UDR自动标注结果如下，阈值取0.005时带“*”的UDR被剪去：

序号	UDR	主词	从词	可信度	序号	UDR	主词	从词	可信度
1	1	3	1	0.985577	9	3	7	6	0.499987
2	24	6	1	0.007212	10	3	12	6	0.499987
3	24	11	1	0.007212	11	17	8	7	1.000000
4	4	3	2	1.000000	12	3	12	8	1.000000
5	0	0	3	1.000000	13	4	10	9	1.000000
6	5	3	4	1.000000	14	24	11	10	1.000000
7	24	6	5	0.500000	15	3	12	11	1.000000
8	24	11	5	0.500000	16	2	3	12	1.000000
*	0	0	6	0.000025					

3.2 规则方法剪枝

3.2.1 句法图(Syntactic Graph)

DRN是一个以句中所有词为顶点，词与词之间的依存关系为边的有向图，用一系列三元组表示，构成一个句法图，每个元组<relation, head, modifier >表示一对词之间的依存关系，即图中的一条有向边，其中head是头结点即主词的序号，modifier是修饰结点即从词的序号，relation是modifier与head之间的一个UDR。

● 术语：

①入弧(in-arc)：句法图中指向某结点的弧称为该结点的入弧，相当于DRN中与该结点相对应的词的UDR。

②出弧(out-arc)：句法图中从某结点发出的弧称为该结点的出弧，相当于DRN中与该结点相对应的词的LDR。

③根结点(root-node)：句子的一棵句法树中无入弧的结点称为该句法树的根结点，相当于分析树的中心词。为便于分析，与DRN一样，我们也相应建立一个虚拟头结点，则根结点仅受该虚拟结点的支配。

④无歧义边：若句法图中某一结点有唯一入弧，则称该结点的入弧为无歧义边。

【性质】

①在任一句法树中，任一结点有且仅有一个支配结点，即任一结点有且仅有一条入弧。

②在任一句法树中，每个结点必须且仅出现一次。

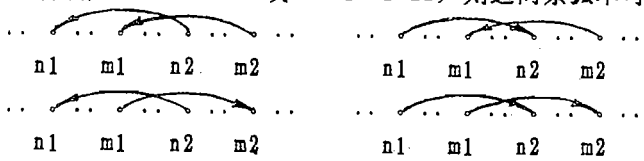
③在任一句法树中，具有相同修饰结点的元组不可共现。

④无歧义边必在每一棵句法树中出现。

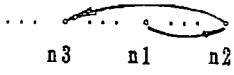
● 弧（元组）的共现问题：

【性质】在任一句法树中（文中提到的共现均指在同一句法树中共现）

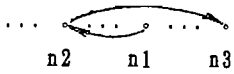
①、交叉弧不得共现。设句法图中有两弧，n1、n2为一弧的两端点，m1、m2为另一弧的两端点，若满足n1<m1<n2<m2或m1<n1<m2<n2，则这两条弧不可共现。如图所示：



②、在句中位于其头结点右侧的结点不可支配位于其头结点左侧的任一结点。设有弧＜r1 n2, n1＞和＜r1', n3, n2＞且n1<n2，若n3＜n1，则这两条弧不可共现，如图所示：



③、在句中位于其头结点左侧的结点不可支配位于其头结点右侧的任一结点，设有弧＜r , n2, n1＞和＜r1', n3, n2＞，且n1>n2，若n3＞n1，则这两条弧不可共现，如图所示：



④、指向同一结点的弧不可共现。

3.2.2 规则方法剪枝算法

根据句法图性质，无歧义边必在每棵句法树中出现，若句法图中有与无歧义边不可共现的边，则这些边必不可在任一句法树中出现，可将它们删去。这样，只要找到一些特殊的无歧义边，删去与它们不可共现的边，即可大大简化DRN。

无歧义边的确定需要依据依存公理和简单的语法规则：

①、若对应助词“的”、“地”、“得”的结点仅有一条出弧，则该弧为无歧义边，理由：助词“的”、“地”、“得”至少有一个LDR（分别为DEP, DIP, DEIP）。

②、若虚拟头结点仅有一条出弧指向结点i，则该弧为无歧义边，结点i即为根结点。

③、若结点i仅受一个结点j的支配，i与j之间存在一条或多条弧，则这些弧从j 出发指向i的一组弧，从整体上可以看作一组无歧义边。

3.2.3 辅助的语法规则

①、任何词不能与动词建立PSUB（介词结构主语）和POBJ（介词结构宾语）的UDR。

②、任何词不能与介词建立SUB（主语）和OBJ（宾语）的UDR。

③、语气词仅与句子中心词建立UDR ZYQA，不与其它词存在依存关系。

④、一个词至多只有一个主语和一个宾语，若结点i 有唯一入弧SUB或OBJ，且其支配结点为j，则删去从j出发的其他同名的弧。

3.3 两种方法的有效结合

剪枝过程中，如何合理安排各种剪枝策略的顺序，是提高算法效率和剪枝效果的关键。统计方法中UDR自动标注采用FB算法为线性的时间复杂度，而规则方法大部分具有（N2 ）的时间复杂度，仅有个别简单规则为线性复杂度，因此从时间复杂度的角度看，应先进行统计方法剪枝，后进行规则方法剪枝。

从剪枝的效果看，先用统计方法利于规则方法中确定更多的无歧义边，剪枝效果较好；而先用规则方法，UDR自动标注时标记减少，有利于提高标记可信度的计算精度。由于DRN中边数随结点数呈指数增长，因此结合以上两方面的考虑，本系统首先采用部分线性复杂度的简单规则剪枝，其次用统计方法剪枝，最后用弧的共现性质剪枝。

经过统计与规则相结合的方法的剪枝，〖句1〗的依存网络简化为如下结果：

序号	UDR	主词	从词	可信度	序号	UDR	主词	从词	可信度
1	1	3	1	0.985577	8	17	8	7	1.000000
2	4	3	2	1.000000	9	3	12	8	1.000000
3	0	0	3	1.000000	10	4	10	9	1.000000
4	5	3	4	1.000000	11	24	11	10	1.000000
5	24	6	5	0.500000	12	3	12	11	1.000000
6	3	7	6	0.499987	13	2	3	12	1.000000
7	3	12	6	0.499987					

4 分析树的生成和评价

4.1 E矩阵

● E矩阵的概念

E矩阵即排斥矩阵(Exclusion Matrix)，记录句法图中两两弧（三元组）间的排斥（不可共现）关系，E矩阵为N×N对称阵，N为图中三元组个数，三元组序号从1至N，则

$$E(i,j)=\begin{cases}1 & \text{（第i个与第j个三元组不可共现）}\\ 0 & \text{（第i个与第j个三元组可以共现）}\end{cases}$$

● E矩阵的生成步骤:

①将E阵初始化全0;

②对句法图中每一对元组i和j根据弧的共现特性判断：若i与j不可共现则置E(i,j)=1。

4.2 分析树的生成

根据E矩阵，找出句法图中的所有句法树，对句1经上述算法后共找到两棵句法树，形式如下:

```
GOV[收]─ SUB[她]
       ├ ADVA[上午]
       ├ COMP[到]
       └ OBJ[信]─ ATTA[封]─ LA[一]─ ATTA[的]─ DEP[朋友]
                └ ATTA[的]─ DEP[长]─ ADVA[很]
GOV[收]─ SUB[她]
       ├ ADVA[上午]
       ├ COMP[到]
       └ OBJ[信]─ ATTA[的]─ DEP[朋友]
                ├ ATTA[封]─ LA[一]
                └ ATTA[的]─ DEP[长]─ ADVA[很]
```

4.3 评价函数

经过多次剪枝，从句法图仍有可能找出多棵分析树，需要对每棵树进行评价，输出其中评价值最高的一棵或几棵分析树作为分析结果，目前采用的评价函数如下:

$$SCORE = \prod SCORE(i)$$

其中SCORE(i)为第i个词UDR的可信度，该可信度取自FB算法UDR自动标注部分。

5 实验结果与讨论

对1300句进行了封闭试验，对 100句进行了开放试验，结果统计如下:

	封闭测试	开放测试
无分析树	20 句	2 句
第一选正确	1000 句	30 句
第二选正确	127 句	24 句
有多棵树且前两棵均不正确	8 句	44 句
只有1棵或2棵树且均不正确	45 句	0 句
二选正确率	93.91%	54%

目前分析器还有一些不足，以下是影响分析效果的主要因素及改进设想：

① 语料库标注的一致性：语料库标注不一致会增加歧义现象。

② 语料库的规模：影响知识库中知识的全面性，DRN建立和UDR自动标注精度。

③ 依存关系集的划分：依存关系划分过细，种类过多会导致歧义现象增加，分析效果不理想，适当合并依存关系种类可改善分析效果，并可减少语料库标注不一致的现象。

④ 词性集的选取：本系统使用[2]的词性标注系统，开放测试采用其小标注集（共41类），因种类过少，知识过粗，检索知识库后建立的DRN复杂度极高，很难有效剪枝，分析树总数随句中词数剧增。若使用划分更细、带有较多语法信息的词性集（如[2]的大标注集），细化高层知识库，并添加邻接词归并规则和剪枝规则，可以改善分析效果。

⑤ 知识获取：现有知识库中词与词建立UDR时的位置信息仅记录了两词的前后关系，如果将它细分，增加紧邻和远离的信息，则可在建立DRN时减少一些盲目性。

⑥ 改进或选用新的UDR自动标注算法，使之能够区分具有不同主词的同名UDR。

通过这几个方面的改进，分析器的分析效果（特别是开放语料的分析正确率）将会有明显的提高。

参考文献

[1]周明，黄昌宁，《统计与规则并举的汉语句法分析模型》，计算机研究与发展（待发表）。

[2]白栓虎，《基于统计的汉语语料库词性自动标注的研究与实现》，清华大学计算机系硕士论文，1992。

[3]吴升，《基于语料库的汉语句法的研究与实现》，清华大学计算机系硕士论文，1992。

[4]张敏，《语料库，知识获取与汉语依存分析》，清华大学计算机系硕士论文，1993。

[5]Jung Seo, Robert F.Simmons, Syntatic Graphs: AReprosentationforthe Union of All Ambiguous Parse Trees, Computational Linguistics, Vol15, No1,Mar 1989

A Study on Chinese Parsing Based on Statistics and Rules

Li Jingkui, Zhou Ming, Huang Changning

Dept. of Computer Science, Tsinghua University, Beijing, P.R.China

Abstract

In thefield of natural language processing, it's ofgreat significance to take advantage of both rule based approach and statistically based approach. In this paper, the author designs and implements a new Chinese dependency parsing system which combines corpus, rule, and statistically based approach.The rules and statistical datas are used to build and simplify the DRN (dependency relation net) .The corpus is used for knowledge acquisition and scoring of parse trees.The close test and open test show that this method is perspective.

Key Words: Natural language processing, Chinese parsing Dependency relation net, Dependency grammar