ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.9, September 2006, pp.1876−1883
DOI: 10.1360/jos171876

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于词汇支配度的汉语依存分析模型*

刘 挺+, 马金山, 李 生

(哈尔滨工业大学 信息检索研究室,黑龙江 哈尔滨 150001)

Chinese Dependency Parsing Model Based on Lexical Governing Degree

LIU Ting+, MA Jin-Shan, LI Sheng

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86413683, Fax: +86-451-86413683, E-mail: tliu@ir-lab.org, http://ir-lab.org

Liu T, Ma JS, Li S. Chinese dependency parsing model based on lexical governing degree. *Journal of Software*, 2006,17(9):1876−1883. http://www.jos.org.cn/1000-9825/17/1876.htm

Abstract: Use of structural information and lexicalization are two of the main challenges facing syntactic analysis, and they are investigated in this paper. First, the probabilities of lexical dependencies are obtained by training a large-scale dependency treebank and used to build the lexical model. Second, the governing degree of words is introduced to utilize the structure information. The lexical method overcomes the weakness of POS dependencies in the past work; meanwhile the governing degree of words is helpful to distinguish the syntactic structures so some ill-formed structures are avoided. Finally, the paper shows a good experimental result of around 74% accuracy on the test set that consists of 4000 sentences.

Key words: dependency grammar; parsing; governing degree; dynamic programming

摘 要: 如何应用句法结构和词汇化是句法分析建模所面临的两个主要问题,汉语依存分析对这两方面做了初步的探索.首先通过对大规模依存树库的统计学习,获取其中的词汇依存信息,建立了一个词汇化的概率分析模型.然后引入词汇支配度的概念,以充分利用了句子中的结构信息.词汇化方法有效地弥补了以前工作中词性信息的粒度过粗问题.同时,词汇支配度增强了对句法结构的识别,有效地避免了非法结构的生成.在 4 000 句的测试集上,依存分析获得了约 74%的正确率.

关键词: 依存语法;句法分析;支配度;动态规划

中图法分类号: TP301 **文献标识码**: A

随着树库资源的丰富及统计方法的深入研究,句法分析的重点逐渐转向词汇化的分析方法.而在词汇关系的表达上,依存语法较之上下文无关的短语结构语法存在的一些优势,使得依存语法的研究重新获得重视[1].

依存语法是 20 世纪 30 年代提出的语法理论,用 5 条公理限定了其语法体系,包括汉语在内的各种语言都遵循这个语法体系,并在此基础上进行了适当的扩充,根据自己的问题进行相应的定义[2].与短语结构语法不同的是,依存语法没有非终结点,用一条依存弧直接连接两个词汇,分析后的句子形成一棵依存树,图 1 为一个依存

* Supported by the Key Project of National Natural Science Foundation of China under Grant No.60435020 (国家自然科学基金重点项目); the National Natural Science Foundation of China under Grant Nos.60575042, 60503072 (国家自然科学基金)

Received 2005-04-28; Accepted 2005-10-10

骨架结构树的分析实例.

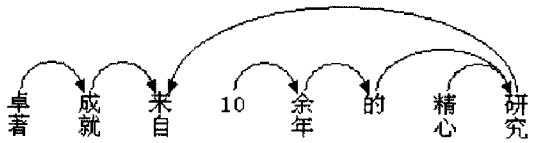

Fig.1 An example of bare-bones dependency tree

图 1 依存骨架分析树

在依存语法中,词与词之间通过依存和支配发生联系,核心词支配从属词,从属词依存于核心词.如图 1 所示,核心词"来自"支配左边的"成就"和右边的"研究".从图中能够看出,依存分析树中两个词汇之间直接发生联系,没有非终结点,没有短语结构树中的"扁平结构(flat structure)",对句法树的词汇化表达达到了最细的区分度.另外,依存树直接体现了句子中的词汇依存关系,这便于上层的应用及各种语言知识(如谓词论元结构)的获取.

句法分析的词汇化建模过程,首先要解决的就是词汇数据的稀疏问题.英文的句法分析通过对大规模树库 Penn Treebank 的统计学习,词汇信息已经被应用到短语结构语法的分析中,使英文的句法分析达到了较高的水平[3,4].Eisner 将 Penn Treebank 中的句子转换为依存结构,并引入结构规则,构造了一个英文依存骨架分析的概率模型[5].

在汉语方面,周强自动抽取句子中的结构优先关系,利用这些关系进行局部优先的短语结构分析[6].Bikel 采用了英语的 TAG 模型,通过非词汇化的片断树(elementary tree)结构建模,并在原来的基础上作了一些规则方面的修改,以适应汉语分析[7].Xiong 将语义知识引入句法分析中,缓解了训练数据不足的问题[8].在汉语的依存分析方面,Zhou 采用分块的思想,应用一些制定的语法规则,先对句子进行分块处理,找出关系固定的语块,然后再对整个句子进行依存分析[9].

以上工作均取得了较好的效果,为汉语的句法分析研究开辟了道路.但是,由于受树库资源的限制,多数工作还没有充分应用词汇化信息.本文以人工标注的 4 万句依存树库为训练集,统计训练集中的词汇依存信息,并进行适当的平滑,增加了模型对词汇特征的区分能力,同时在模型中引入了词汇支配度,以利用依存语法的结构特征.通过对 4 000 句的测试集进行评测,本方法取得了较好的依存分析效果,在汉语句法分析的词汇化方面进行了初步探索.

1 依存分析的概率模型

统计句法分析建模的目的是寻找一个概率评价函数,使其在给定的语法框架下能够评价某个句法分析结果是当前句子正确语法解释的概率.设 G 为某种语言的语法,S 为一个句子,T 为根据语法 G 生成 S 的所有可能分析结果所组成的集合.对于任意分析树 $t\in T$,句法分析的概率模型能够计算出 t 的条件概率 $p(t|S,G)$,且满足

$$\sum_{t\in T} p(t|S,G)=1.$$

1.1 基于词汇信息的统计建模

本文的工作是建立一个依存语法的概率分析模型,当输入一个线性的词序列 S,模型输出一棵结构化分析树 t,并使分析树 t 符合汉语的依存语法规范.其中,S 是经过分词和词性标注的句子,表示为

$$S=\{\langle w_1,t_1\rangle,\langle w_2,t_2\rangle,\ldots,\langle w_n,t_n\rangle\},$$

其中,$w_i(1\le i\le n)$是句子中第 i 个词,$t_i(1\le i\le n)$是第 i 个词的词性.

句法分析是一个处理语法歧义的问题.对于一个给定的句子,存在数量巨大的符合语法的分析树,我们用概率评价函数对每个分析结果进行评价,把消歧问题转化为一个最优化的过程,即给定句子 S,找出一棵概率最大的依存分析树 t^*,该过程可以表示为

$$t^*=\arg\max P(t|S)=\arg\max_{t\in T} P(t|w_1,w_2,\ldots,w_n) \quad (1)$$

为了计算分析树 t 的概率,我们对式(1)作了独立假设,即假设分析树中的各个依存弧相互独立,则整个依存

树的概率为 $n-1$ 条依存弧的概率乘积,即

$$P(t|S)=\prod_{k=1,\ldots,n-1}P(L_k|w_i,w_j) \tag{2}$$

其中,$L_k(1\le k\le n-1)$是构成 n 个节点分析树的依存弧,w_i 和 $w_j(j>i)$是两个词汇节点.除了两端的节点 w_i 和 w_j 之外,依存弧 L_k 还需要另外两个构成要素:*Direction* 和 *Distance*.*Direction* 是依存弧的方向,取 0 和 1 两个值,

$$Direction=\begin{cases}0, & \text{如果}w_j\text{依存于}w_i\\ 1, & \text{如果}w_i\text{依存于}w_j\end{cases}.$$

Distance 是节点 w_i 和 w_j 之间的距离,取 1,2,3 三个值,

$$Direction=\begin{cases}1, & \text{如果 } j-i=1\\ 2, & \text{如果 } j-i=2\\ 3, & \text{如果 } j-i>2\end{cases}.$$

距离是为了更为准确地描述依存关系所引入的信息,要求这个信息的粒度既不能太粗也不能太细:太粗则描述能力不强;太细则会产生较为严重的数据稀疏.汉语中的远距离依存搭配,多是由于中间有过多的状语或定语等修饰成分,这些搭配虽然在物理距离上变化较大,但依存关系比较稳定.所以,对相距为两个或两个以上词语的依存搭配,距离值 *Distance* 均取 3.

1.2 参数估计

通过对训练数据进行统计,采用极大似然估计计算依存弧 L_k 的概率

$$\tilde{P}(L_k|w_i,w_j)=\frac{C(L_k,w_i,w_j)}{C(w_i,w_j)} \tag{3}$$

其中,分子表示节点 w_i 和 w_j 构成依存弧 L_k 的次数;分母表示 w_i 和 w_j 在一个句子中以距离 *Distance* 出现的次数,既包括存在依存关系,也包括不存在依存关系的次数.

对词汇关系的稀疏问题,我们采用词性关系进行插值平滑.上式的词汇依存概率记为 $\tilde{P}_1$,将其中的 w_i 替换成该词的词性 t_i,得到词性依存概率记为 $\tilde{P}_2$,则平滑后依存弧的概率为

$$\tilde{P}=\lambda_1\tilde{P}_1+\lambda_2\tilde{P}_2+\lambda_3\xi,$$

其中,$\lambda_1+\lambda_2+\lambda_3=1$,由训练获得;$\xi$是常量,这里取 0.001.通过平滑,有效地缓解了词汇依存信息的数据稀疏问题.

2 结构信息的应用

在上一节的概率模型中,我们假定一个句子中各依存弧之间相互独立,这种假设简化了问题,却忽略了依存树中的结构信息.模型在分析时常出现如图 2(a)所示的错误.

(a) The parse with an error
(a) 错误的分析结果

(b) The correct parse
(b) 正确的分析结果

Fig.2
图 2

图 2(a)中的"推动"多支配了一个从属词,而"的"缺少从属词,我们将这种错误称为非法结构.产生非法结构的一个原因是词汇信息的过度拟合,如"推动"和"价格"是一个很强的搭配,在这里错误地构成了依存关系.显然,这种方法忽略了词汇的结构信息:"推动"所能支配的从属词上限以及"的"所能支配的从属词下限.图 2(b)为正确的分析结果,"推动"仅支配"上升","的"字支配"价格".

2.1 词汇支配度

在文献[5]中,Eisner 建立了一个面向说话者的概率依存分析模型.该模型利用依存语法中的结构信息,由父节点和最近的兄弟节点来预测当前节点.Gabriel 认为该模型存在两个不必要的生成无限长派生过程的问题:一是一个结点可能生成无限多的子结点;二是一个结点可能生成无限长的依存结构[10].其中第 2 个问题我们通过依存弧的距离参数进行了限制.为解决第 1 个问题及避免图 2(a)中的非法结构,我们引入了词汇支配度.

词汇支配度是用来描述一个词支配其从属词的能力,被支配度是指一个词依存于核心词的能力.如果把依存树看作有向图,则支配度相当于节点的入度,被支配度相当于节点的出度.根据依存语法,依存树中每个节点(根结点除外)有且只有一个父结点,所以每个结点的被支配度均为 1.但结点的支配度却不受此限制,一个词可以同时支配多个从属词.

在配价语法理论中,以"价"来描述一个动词所能携带的论元个数.如果一个动词不能支配任何论元,那么它就是零价动词,如"地震、刮风";如果一个动词能够支配一个论元,那么它就是一价动词,如"休息、游泳"等,以此类推.

本文引入的词汇支配度是对动词配价的扩展,这种扩展包括以下两个方面:

(1) 将动词扩展为所有的词;

(2) 将体词性的论元扩展为所有的从属词.

做这种扩展是面向我们所要解决的问题.句法分析要求解析句子中的所有词汇关系.而除了动词外,其他词在支配从属词的时候也都表现出一定的规律性,如介词总是支配其右侧的一个名词成分,支配度常为 1;助词"的"总是支配左侧的一个或多个词,支配度大于或等于 1.

另外,配价理论中动词的论元只为名词或代词等体词性词语,我们将词汇所支配的从属词扩充到所有的成分,支配度不但是只包括所支配的论元,而且对一些修饰成分、附加成分也计算在内.这两点较之单纯的动词配价更适合句法分析.

2.2 融合词汇支配度的统计模型

我们用 $D(w_i)$表示词 w_i 的支配度,如图 1 中 D(来自)=2,D(研究)=2.一些词汇的支配度受词性的影响较小,比如动词、及物动词的支配度要大于不及物动词的支配度,即支配度对词汇更为敏感.另外,不同的词其支配度并不完全一致,同一个词在不同的上下文中支配度的变化较大.于是我们统计词汇一级的支配度,并采用均值和方差对词汇支配度进行描述.均值 $E(D_i)$记录了 w_i 在训练数据中支配度的平均值,而方差 σ_i 则反映了词 w_i 的支配度变化情况.如果方差较小,则说明该词的支配度比较稳定,对句子的分析就更有用;如果方差较大,则该词的支配度变化很大,对模型的效果影响较小.

另外,词在支配从属词的时候表现出一定的方向性:有的倾向于支配前面的词,有的倾向于支配后面的词,有些词对前后的词都可以支配.为了更准确地描述支配度,我们把支配度分为左向支配度和右向支配度,分别表示为 $D_L(w_i)$和 $D_R(w_i)$,则图 1 中 D_L(来自)=1,D_R(来自)=1,D_L(研究)=2,D_R(研究)=0.

同时,我们也将均值分为左均值 $E_L(D_i)$和右均值 $E_R(D_i)$,方差分为左方差 σ_{Li} 和右方差 σ_{Ri},并通过均值和方差划定每个词语支配度范围 C_L 和 C_R:

$$E_L(D_i)-(1+\xi)\sigma_{Li}\le C_L\le E_L(D_i)+(1+\xi)\sigma_{Li},$$

$$E_R(D_i)-(1+\xi)\sigma_{Ri}\le C_R\le E_R(D_i)+(1+\xi)\sigma_{Ri}.$$

其中,ξ 是调整因子,我们取经验值为 0.1.

在将词汇支配度信息融合进模型的过程中,我们将依存树分解成不同的依存子结构.这里称一个词汇节点及其支配的所有从属词为一个依存子结构,用 H_i 表示以 w_i 为核心词的依存子结构.对图 1 中的"来自"和"研究"两个词,其依存子结构如图 3 所示.

(a) A dependent structure of "来自"
(a) "来自"的依存子结构

(b) A dependent structure of "研究"
(b) "研究"的依存子结构

Fig.3
图 3

依存子结构的概率为在左向支配度和右向支配度分别符合限定范围的条件下,各依存弧的概率乘积.设 H_i 的左侧有 u 条依存弧,右侧有 v 条依存弧,则依存子结构 H_i 的概率为

$$P(H_i)=P(L_{1i},L_{2i},...,L_{ui}|u\in C_L)P(L_{i1},L_{i2},...,L_{iv}|v\in C_R)$$
$$=\prod_{1\le j\le u}P(L_{ji}|u\in C_L)\prod_{1\le k\le v}P(L_{ik}|v\in C_R) \quad (4)$$

式(4)也作了类似于式(2)的独立假设.在式(4)中,将依存树中的多元结构信息通过词汇支配度转化为一元信息,表示为在给定词汇支配度的约束下其有向依存弧的概率.然后,将每个词的依存子结构替换式(1)中的依存弧,我们得到一个新的依存分析概率模型:

$$t^*=\arg\max_{t\in T}P(H_1,H_2,...,H_{n-1}|w_1,w_2,...,w_n)=\arg\max_{t\in T}\prod_{1\le i\le n}P(H_i) \quad (5)$$

其中,n 是句子中节点的个数,同时也是依存子结构的个数.

由式(4)和式(5),我们得到了一个融合词汇支配度的依存句法分析模型.

3 搜索最优路径

本文的搜索问题可归结为在有向图中找出满足依存语法的最优依存树,但符合语法的分析结果随句长的增加呈指数级增长,所以,高效的搜索过程至关重要.

依存结构的生成具有递归特性,同一结构同时存在于多个依存树中.对于这一点,文献[10]给出了形式化的详细描述.利用句法分析的这种特性,我们采用动态规划的思想,按照自底向上的过程,设计了一个较为高效的搜索算法,在 $O(n^3)$ 时间内,从全部分析结果中找出概率值最高的路径,具体细节可参考文献[11].

在以式(5)作为评价函数进行搜索时,由当前节点的支配度和依存弧的概率共同制约依存树的生成.其中,节点当前支配度情况在搜索过程中动态获得.

4 实验及分析

统计模型词汇化的基础是建立类似于 Penn Treebank 的大规模树库.目前,我们已手工建立了容纳 46 000 个句子的汉语依存树库*,全部句子取自《人民日报》语料库.我们将树库中的 40 000 句作为训练集,2 000 句作为开发集,4 000 句作为测试集.由于句子的长度对句法分析的准确率影响很大,我们统计了测试集中句子的长度分布,并根据不同词长的句子分别进行评价,以便于和其他工作进行比较,句子的长度分布见表 1.

Table 1 Average length in test set (words per sentence)

表 1 测试集的句子长度分布(以词为单位)

Length	≤10	≤15	≤20	≤30	≤50
Number	472	1357	2321	3796	4000
Average length	7.9	11.2	14.0	18.3	19.0

在评价指标上,依存语法与短语结构语法存在一些不同:短语结构语法的评价单元是句法成分(constitute),同一个句子的不同短语结构树含有的句法成分数量可能不同,所以短语结构语法用准确率和召回率进行评价;依存语法对句子中的词汇进行两两解析,分析结果及标准测试集中全部依存弧的数量均为 $n-1$(n 为句子长度),

* 该依存树库已在网上免费发布,可通过网址 http://ir-lab.org 获取.

准确率与召回率相等,所以通常只用准确率一个指标进行评价.本文测试了依存弧的准确率和全句核心词的准确率.

这里,将模型分析的结果称为S_1,测试集称为S_2,则依存关系的准确率定义为S_1中正确的依存弧占S_2中全部依存弧的比率,句子核心词的准确率为S_1中正确识别的句子核心词占S_2中句子核心词的比率.

我们将第 1.1 节中基于词汇依存关系的方法称为模型 1,第 2.2 节中融入词汇支配度的方法称为模型 2,同时以文献[11]中的方法作为 Baseline,该方法统计树库中的词性依存信息,并利用不同距离的词性关系进行插值平滑,未考虑句子的结构信息.根据不同长度的句子,我们对这 3 个模型分别进行测试,测试结果见表 2~表 4,同时,将 3 个模型的依存弧准确率结果显示在图 4 中,以便于比较.

Table 2 Parsing results in Baseline model

表 2 Baseline 的依存分析结果

Length	≤10	≤15	≤20	≤30	≤50
Dependency accuracy (%)	82.7	76.6	72.7	69.8	69.3
Root accuracy (%)	88.1	80.6	76.1	71.2	70.4

Table 3 Parsing results in model 1 (lexical model)

表 3 模型 1 的依存分析结果(词汇化模型)

Length	≤10	≤15	≤20	≤30	≤50
Dependency accuracy (%)	84.8	79.3	76.0	73.2	72.7
Root accuracy (%)	90.8	84.3	80.1	75.6	75.0

Table 4 Parsing results in model 2 (lexical model with governing degree)

表 4 模型 2 的依存分析结果(引入词汇支配度)

Length	≤10	≤15	≤20	≤30	≤50
Dependency accuracy (%)	86.1	80.7	77.4	74.4	73.9
Root accuracy (%)	90.9	84.5	80.4	76.6	76.0

表 2 是 Baseline 的分析结果.Baseline 中的方法是基于词性依存关系的统计模型,该模型只考虑词性信息,虽然能够正确分析大部分依存关系,但是由于词性信息的粒度过粗,句法分析的效果较易达到上限,之后增大训练数据的规模对分析的作用不再明显.这一点可以从与文献[11]的实验结果对比中得出**.在引入词汇信息之后,模型 1 的依存分析准确率较之 Baseline 有了显著提高,这是由于词汇信息对依存关系的计算更为准确,有效地解决了词性信息的粒度过粗问题.但是,由于词汇所产生的过度拟合问题使依存树中产生了许多非法结构,在模型 2 中引入词汇支配度之后,有效避免了模型 1 中出现的非法结构,并进一步利用了句子的结构规律,使分析的准确率得到进一步提高.

从图 4 能够看到:随着句子长度的增加,依存分析的准确率下降,Baseline 模型下降得较为明显;而引入词汇信息之后,由于词汇搭配能够有效解决长距离的依存问题,一定程度上缓解了下降的趋势.但由于词汇数据的稀疏,长句子较短句子在准确率方面还是有较大差距.所以,我们下一步将对词汇化模型进行改进,重点解决词汇数据的稀疏问题,同时探索结构信息,以改进模型的质量.

近几年,汉语的句法分析水平逐渐得到提高.依存分析方面,Zhou 在平均词长为 7.34 的新加坡小学课本的测试数据上获得了 90.5%的准确率,在词长为 14.52 的人民日报数据上,准确率为 67.7%[9];Cheng 在平均句长为 5.7 词的 CKIP 树库上进行分析,文学类文本的准确率达到了 87%,新闻类文本的准确率为 74%[12].短语结构分析的多数工作是在 Upenn 的中文树库上进行,Bikel 在新华语料上对评测集中的全部文本进行测试,获得了 73.9%的 F 值[7];Xiong 对测试集中长度小于 40 的句子进行评测,F 值达到了 79.4%[8].但是,由于目前的汉语句法分析还没有完全统一的测试集,许多工作是在不同的数据集上进行的,而不同语料的规模、文本类型以及句子的长度对分析结果的影响很大,这些差别给句法分析结果的比较造成了一定困难.另外,由于评价标准不一样,不同

** 文献[11]中测试集的平均词长为 9,依存弧准确率为 80.5%.

的语法体系之间也很难直接比较,这些都是今后需要解决的问题.

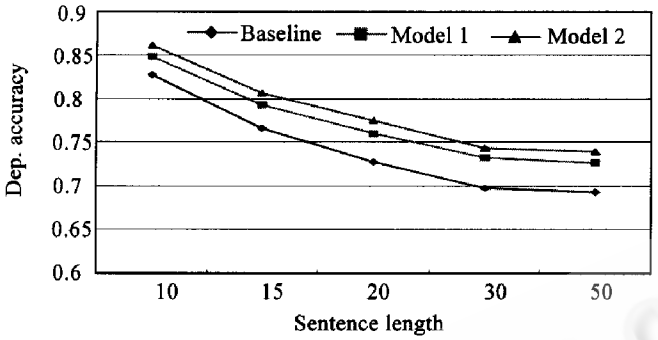

Fig.4 The parsing results for three models

图 4 3 种模型的依存弧准确率比较

5 结束语

依存语法是句法分析领域一个逐渐被重视的研究热点,凭借其语法简洁、易于标注等特点,被越来越广泛地应用于各种语言的分析中.本文在前人工作的基础上,借鉴短语结构语法分析的成果,并参照一些依存分析方面的工作,建立了一个汉语的词汇化依存分析模型.

我们建立了一个大规模的汉语依存树库,统计分析树中词汇层次的依存信息,并根据依存树结构的特点引入了词汇支配度的概念,将结构信息有效地融合到统计模型中,建立了一个新的词汇化依存分析模型.在搜索算法上,采用动态规划的方法,在 $O(n^3)$的时间内搜索出全局最优的分析结果.

下一步的工作我们将重点解决词汇数据的稀疏问题以及对词汇支配度进行细化,即在应用支配度的同时,进一步考虑每个依存子结构中从属词的属性,以更充分地利用句法的结构信息.

References:

[1] Nasr A, Rambow O. A simple string-rewriting formalism for dependency grammar. In: Proc. of the Workshop on Recent Advances in Dependency Grammar. Barcelona: Association for Computational Linguistics, 2004. 17−24.

[2] Liu WQ, Wang MH, Zhong YX. On study of hierarchy structure dependency relations in Chinese. Journal of Chinese Information Processing, 1996,10(2):32−46 (in Chinese with English abstract).

[3] Charniak E. A maximum-entropy-inspired parser. In: Proc. of the 1st Conf. of the North American Chapter of the Association for Computational Linguistics. Seattle: Association for Computational Linguistics, 2000. 132−139.

[4] Collins M. Head-Driven statistical models for natural language parsing [Ph.D. Thesis]. Pennsylvania: University of Pennsylvania, 1999.

[5] Eisner J. Three new probabilistic models for dependency parsing: An exploration. In: Proc. of the COLING. Copenhagen: Association for Computational Linguistics, 1996. 340−345.

[6] Zhou Q, Huang CN. An improved approach for Chinese parsing based on local preference information. Journal of Software, 1999,10(1):1−6 (in Chinese with English abstract).

[7] Bikel DM, Chiang D. Two statistical parsing models applied to the Chinese treebank. In: Proc. of the 2nd Chinese Language Processing Workshop. Hong Kong: Association for Computational Linguistics, 2000. 1−6.

[8] Xiong DY, Li SL, Liu Q, Lin SX, Qian YL. Parsing the Penn Chinese treebank with semantic knowledge. In: Dale R, Wong KF, eds. Proc. of the IJCNLP 2005. Jeju Island: Springer-Verlag, 2005. 70−81.

[9] Zhou M. A block-based dependency parser for unrestricted Chinese text. In: Proc. of the 2nd Chinese Language Processing Workshop Attached to ACL-2000. Hong Kong: Association for Computational Linguistics, 2000. 78−84.

[10] Infante-Lopez G, Rijke M, Sima'an K. A general probabilistic model for dependency parsing. In: Blockeel H, Denecker M, eds. Proc. of the 14th Dutch-Belgian Artificial Intelligence Conf. BNAIC-02. Leuven: BNVKI, Dutch and the Belgian AI Association, 2002. 139−146.

[11] Ma JS, Zhang Y, Liu T, Li S. A statistical dependency parser of Chinese under small training data. In: Proc. of the Workshop on Beyond Shallow Analyses-Formalisms and Statistical Modeling for Deep Analyses, IJCNLP-04. Sanya: Asia Federation of Natural Language Processing, 2004. 1−5.

[12] Cheng YC, Asahara M, Matsumoto Y. Deterministic dependency structure analyzer for Chinese. In: Proc. of the IJCNLP-04. Sanya: Asia Federation of Natural Language Processing, 2004. 135−140.

附中文参考文献:

[2] 刘伟权,王明会,钟义信.建立现代汉语依存关系的层次体系.中文信息学报,1996,10(2):32−46.

[6] 周强,黄昌宁.基于局部优先的汉语句法分析方法.软件学报,1999,10(1):1−6.

刘挺(1972－),男,黑龙江哈尔滨人,博士,教授,CCF 高级会员,主要研究领域为自然语言处理,信息检索.

马金山(1974－),男,博士生,主要研究领域为句法分析.

李生(1943－),男,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.