ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2017/11(07)-1114-08
doi: 10.3778/j.issn.1673-9418.1605034

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

# 融合多结构信息的中文句法分析方法*

赵国荣,王文剑+
山西大学 计算机与信息技术学院,太原 030006

# Method for Chinese Parsing Based on Fusion of Multiple Structural Information*

ZHAO Guorong, WANG Wenjian+
School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
+ Corresponding author: E-mail: wjwang@sxu.edu.cn

**ZHAO Guorong, WANG Wenjian. Method for Chinese parsing based on fusion of multiple structural information. Journal of Frontiers of Computer Science and Technology, 2017, 11(7): 1114-1121.**

**Abstract:** Syntactic parsing is a basic technology of natural language understanding, and it is the cornerstone of deep language understanding. At present, the parsing method is based on the hypothesis of context free grammar. In fact, the context has a strong correlation in phrase structure trees. If the structural information can be used, it can further improve the accuracy of the parser. This paper combines the multiple structural information in syntactic structure trees, the structural information (such as father node or left and right sister nodes) in the non-terminal node can strengthen grammar rules of context constraints. And then this paper uses the method of structural support vector machines (SSVMs) for Chinese parsing. The experimental results show that the method of multiple structural information fusion can resolve the structural ambiguity and improve the accuracy and *F*1 value.
**Key words:** structural support vector machines; context-free grammar; structure context correlation; Chinese parsing

**摘 要:** 句法分析是自然语言理解的一项基础技术,是迈向深层语言理解的基石。目前常用的句法分析方法的语法模型建立在上下文无关文法的假设上。事实上,短语结构树的节点之间具有很强的上下文相关性,充分利用结构信息,可进一步提高句法分析的准确性。融合了句法结构树中的多结构信息(在非终节点中增加

* The National Natural Science Foundation of China under Grant Nos. 61273291, 61503229 (国家自然科学基金); the Natural Science Foundation of Shanxi Province under Grant No. 2015021096 (山西省自然科学基金); the Science and Technology Innovation Project of Shanxi Province under Grant No. 2015110 (山西省高等学校科技创新项目).
Received 2016-04, Accepted 2016-06.
CNKI网络优先出版: 2016-06-23, http://www.cnki.net/kcms/detail/11.5602.TP.20160623.1139.004.html

父亲节点及左、右姐妹节点等标记)以加强语法规则的上下文约束,并采用结构化支持向量机的方法对句法进行了分析。实验表明,该融合多结构信息的句法分析方法可以消解结构歧义,提升句法分析精确率和$F1$值。

**关键词**:结构化支持向量机;上下文无关文法;结构上下文相关;中文句法分析

**文献标志码**:A **中图分类号**:TP391

## 1 引言

句法分析是自然语言处理的关键性问题之一。对句法分析进行可计算化处理,句法分析算法和语法模型是两个重要的元素,其中语法模型无论是使用统计的方法,还是使用单纯的规则,在进行句法分析时都需要建立一种模型。最早的语法模型是简单的上下文无关的语法模型(context-free grammar, CFG)[1]。但是CFG是在一些非常理想化的独立性假设的基础上建立的,它的规则的建立只和其孩子节点有关,因而这些假设忽略了句法树中其他许多隐含的信息。为了得到更好的基于短语结构的句法分析效果,一些算法的研究集中在挖掘短语结构树的上下文相关的信息上,通过增加丰富的结构信息和词汇信息等来提升句法分析的效果。

最具代表性的研究就是在概率上下文无关文法[2](probabilistic context-free grammar, PCFG)中增加结构上下文相关的策略。文献[3]尝试了祖先节点相关、父亲节点相关等几种结构上文相关的策略;文献[4]尝试了加入结构下文孩子节点相关的策略,构成结构下文相关的概率语法模型;文献[5]加入了每个短语节点的父亲节点和左、右姐妹节点的结构上下文信息,这些方法都对突破上下文无关语法研究中的独立性假设进行了尝试,都是对经典PCFG模型进行的优化。文献[6]采用机器学习方法——结构化支持向量机(structural support vector machine, SSVM)对基于短语结构的中文句法进行分析,语言模型采用的是上下文无关文法。本文的工作是尝试融合句法分析树中节点的结构信息,研究使用结构化支持向量机对中文句法进行分析时所产生的影响,实验证明它可以提高句法分析系统的精确率和$F1$值。

## 2 结构化支持向量机方法简介

在现实世界中,需要处理的大部分数据(如网状结构数据、队列结构或树形结构等)都比较复杂,而且数据之间相互依赖,具有特定的结构化关系,传统的支持向量机[7]已经不适合处理这些复杂的数据。为了解决传统支持向量机在处理复杂数据时的难题,Hofmann和Joachims等人首次提出了结构化支持向量机[8-9],它可以根据不同的应用领域设计不同的结构化特征函数去拟合数据,从而有效地处理结构化数据。

结构化支持向量机是一种基于判别式的学习模型,使用结构化支持向量机的关键是构造出样本的输入与输出对之间的一个映射函数 $f: x \to y$ 。当使用结构化支持向量机进行句法分析时, $f: x \to y$ 中 $f$ 表示的意思是输入句子$X$到输出短语结构树 $Y$ 的一个映射。在构造函数 $f$ 时,关键任务是需要学习一个基于输入/输出对的判别式函数 $F: X \times Y \to \Re$ ,通过使输出变量最大化的方法,实现对输出结果预测的目的。结构化支持向量机的目标函数[10-11]为:

$$f(x;w) = \arg\max_{y \in Y} F(x,y;w) \tag{1}$$

$F$是基于输入/输出组合特征表示 $\psi(x,y)$ 的线性函数:

$$F(x,y,w) = \langle w, \psi(x,y) \rangle \tag{2}$$

式(1)中带参数 $w$ 的函数 $f$ ,假设它的经验风险为0,可以写成一个非线性约束的形式[8]:

$$\forall i \{1,2,\cdots,n\}: \max_{y \in Y \setminus y_i} \{\langle w, \psi(x_i,y) \rangle\} \leqslant \{\langle w, \psi(x_i,y_i) \rangle\} \tag{3}$$

式(3)可以等价转换为:

$$\forall i,\ \forall y \in Y \setminus y_i: \langle w, \delta\psi_i(y) \rangle \geqslant 0 \tag{4}$$

定义 $\delta\psi_i(y) \equiv \psi(x_i,y_i) - \psi(x_i,y)$ 。

采用最大间隔法可以将式(4)转化为一个凸二次规划形式的最优化问题[10]:

$$\mathrm{SVM}_0: \min_w \frac{1}{2}\|w\|^2$$
$$\forall i, \forall y \in Y \setminus y_i: \langle w, \delta\psi(y) \rangle \geqslant 1 \tag{5}$$

为了容忍部分噪声和离群点,同时兼顾除靠近边界

之外更多的训练点，在式(5)中引入松弛变量的软间隔，本文采用一阶范数 $\xi$ 的形式[10]：

$$\mathrm{SVM}_1: \min_{w,\xi} \frac{1}{2}\|w\|^2+\frac{C}{n}\sum_{i=1}^{n}\xi_i, \text{ s.t. } \forall i,\xi_i\geqslant 0$$
$$\forall i,\forall y\in Y\backslash y_i: \langle w,\delta\psi_i(y)\rangle\geqslant 1-\xi_i \tag{6}$$

## 3 融合多结构信息的语法模型构造

文献[6]使用结构化支持向量机进行中文句法分析时，将文法限定为乔姆斯基范式的形式[2]，其语法规则为：

$$n_i[A\rightarrow BC] \tag{7}$$
$$n_i[A\rightarrow \alpha] \tag{8}$$

这里 $A$、$B$、$C$ 是非终结符，$\alpha$ 是终结符。设 $x$ 是需要进行句法分析的句子，$Y$ 是针对 $x$ 分析出的若干个句法树的集合。假设最佳分析树为 $h(x)$，每棵句法树 $y$ 中所有的语法规则的集合用 $rules(y)$ 表示，每一个语法规则所对应的权值参数为 $w_l$，文献[6]使用的上下文无关文法模型为：

$$h(x)=\arg\max_{y\in Y}\left\{\sum_{n_l\in rules(y)} w_l\right\} \tag{9}$$

但是，在实际生活中自然语言具有很强的上下文相关性，上下文无关语法表现能力有限，当遇到结构依存的问题时就显得能力有限了。

上下文无关文法对句法树中的结构以及词汇等信息利用不足，无法描写句法树结构上隐藏的许多信息，如每个短语节点的父节点或(和)左、右姐妹节点的信息。文献[5]成功地将上下文相关信息(即父节点或(和)左、右姐妹节点的信息)加注到每个短语节点(即非终节点)上，使用概率上下文无关文法进行句法分析，并取得很好的效果。故本文也尝试将这些信息增加到使用结构化支持向量机进行句法分析的方法中，从而提升句法分析器的精度。假设将单纯地增加“父亲”、“左妹”或“右妹”信息称为一阶标注；那么增加“父亲+左妹”、“父亲+右妹”或“左妹+右妹”就是二阶标注；增加“父亲+左妹+右妹”为三阶标注。因为只是在非终节点上增加上下文相关的结构信息，所以语法规则(7)(8)的形式要发生变化。以语法规则(7)的形式变化为例，在每一个非终节点后用括号注明相关结构信息范畴。

一阶标注中增加父亲信息后，规则(7)的形式变换为：

$$n_i[A(parent)\rightarrow B(A)C(A)] \tag{10}$$

增加左妹信息后，规则(7)的形式变换为：

$$n_i[A(Lsister)\rightarrow B(0)C(B)] \tag{11}$$

增加右妹信息后，规则(7)的形式变换为：

$$n_i[A(Rsister)\rightarrow B(C)C(0)] \tag{12}$$

二阶标注中增加父亲+左妹信息后，规则(7)的形式变换为：

$$n_i[A(parent-Lsister)\rightarrow B(A-0)C(A-B)] \tag{13}$$

增加父亲+右妹信息后，规则(7)的形式变换为：

$$n_i[A(parent-Rsister)\rightarrow B(A-C)C(A-0)] \tag{14}$$

增加左妹+右妹信息后，规则(7)的形式变换为：

$$n_i[A(Lsister-Rsister)\rightarrow B(0-C)C(B-0)] \tag{15}$$

三阶标注增加父亲+左妹+右妹信息后，规则(7)的形式变换为：

$$n_i[A(parent-Lsister-Rsister)\rightarrow B(A-0-C)C(A-B-0)] \tag{16}$$

语法规则(8)和规则(7)箭头左部的变化一样，因为箭头右边是终结符，所以不发生变化。简单地以增加父亲节点信息为例，短语的结构受到上层短语的制约。比如做主语的NP短语(NP位于S之下)和做宾语的NP短语(NP位于VP之下)的内部结构明显不同，这样可以快速帮助分析器抉择，减少不必要的子树生成。

## 4 结构化支持向量机特征函数的构建

在结构化支持向量机中，关键任务是特征函数 $\psi(x,y)$ 的构造，在不同的领域需要构造不同的特征函数，从而和实际数据达到较好的拟合。因而特征函数构造合适与否会直接影响结构化支持向量机方法的有效性。图1是短语结构树的输入输出示例，图2为其在学习时构造的 $\psi(x,y)$。

以在每个非终节点增加“父亲”、“父亲+左妹”和“父亲+左妹+右妹”节点为例，短语结构树以及构造的相对应的特征函数 $\psi(x,y)$ 变换后的示例如图3所示。

在使用结构化支持向量机进行句法分析时，在

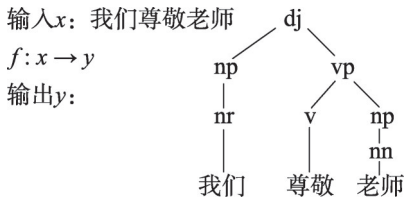



Fig.1 A sample of input and output without structural information

**图1 未增加结构信息的输入输出示例**

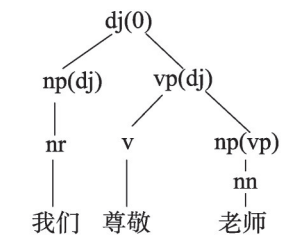
$$\psi(x,y)=\begin{bmatrix}1 & \text{dj} \to \text{np vp}\\ 1 & \text{np} \to \text{nr}\\ 1 & \text{vp} \to \text{v np}\\ 0 & \text{vp} \to \text{v pp}\\ 1 & \text{np} \to \text{nn}\\ \cdots & \cdots\cdots\cdots\cdots\\ 1 & \text{nr} \to \text{我们}\\ 1 & \text{nn} \to \text{老师}\\ 1 & \text{v} \to \text{尊敬}\\ 0 & \text{nr} \to \text{他们}\end{bmatrix}$$

Fig.2 Structural feature function $\psi(x,y)$

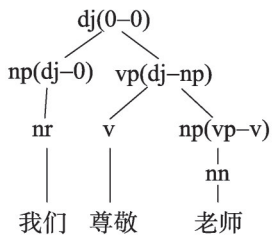
**图2 结构化特征函数 $\psi(x,y)$**

学习过程中要从树库自动抽取语法规则并进行统计，从而对模型进行分析。$r_j$ 表示句子 $x$ 的句法分析树 $y$ 中每一个节点所对应的规则，$a_j$ 表示规则 $r_j$ 出现的次数，$w_j$ 表示每条规则相对应的权值。$x$ 表示一个句子，$y$ 表示其对应的有效句法树，$w_j$ 表示每一个节点的权值，其和作为这个句法树的分值，$F(x,y,w)=\langle w,\psi(x,y)\rangle$ 为计算分值的函数。特征函数 $\psi(x,y)$ 的构造就是由树库中出现的规则及其次数组成。对于给定的句子 $x$，通过乔姆斯基算法[2]（Cocke-Younger-Kasami，CKY）找出符合文法的句法分析树集 $Y$，再从句法分析树集中找出分值最大的 $F(x,y,w)$，$y \in Y$，即为所求句子的语法树。

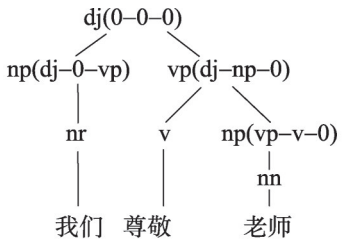
也就是在式（9）$h(x)=\arg\max_{y\in Y}\left\{\sum_{n_l\in rules(y)} w_l\right\}$ 中引入特征函数 $\psi(x,y)$，则 $\langle w,\psi(x,y)\rangle=\sum_{n_l\in rules(y)} w_l$。其中 $w$ 表示特征函数通过统计学习训练得到的权值；$\psi(x,y)$ 表示通过训练得到的规则及规则出现的次数。




(a) Phrase structure tree after adding "father" nodes
(a) 增加"父亲"节点的短语结构树




(b) Phrase structure tree after adding "father and left sister" nodes
(b) 增加"父亲+左妹"节点的短语结构树




(c) Phrase structure tree after adding "father, left and right sisters" nodes
(c) 增加"父亲+左妹+右妹"节点的短语结构树

$$\begin{bmatrix}1 & \text{dj(0)} \to \text{np(dj) vp(dj)}\\ 1 & \text{np(dj)} \to \text{nr}\\ 1 & \text{vp(dj)} \to \text{v np(vp)}\\ 0 & \text{vp(dj)} \to \text{v pp(vp)}\\ 1 & \text{np(vp)} \to \text{nn}\\ \cdots & \cdots\cdots\cdots\cdots\\ 1 & \text{nr} \to \text{我们}\\ 1 & \text{nn} \to \text{老师}\\ 1 & \text{v} \to \text{尊敬}\\ 0 & \text{nr} \to \text{他们}\end{bmatrix}$$

(d) Structural feature function after adding "father" nodes
(d) 增加"父亲"节点后构造的特征函数

$$\begin{bmatrix}1 & \text{dj(0-0)} \to \text{np(dj-0) vp(dj-np)}\\ 1 & \text{np(dj-0)} \to \text{nr}\\ 1 & \text{vp(dj-np)} \to \text{v np(vp-v)}\\ 0 & \text{vp(dj-np)} \to \text{v pp(vp-v)}\\ 1 & \text{np(vp-v)} \to \text{nn}\\ \cdots & \cdots\cdots\cdots\cdots\\ 1 & \text{nr} \to \text{我们}\\ 1 & \text{nn} \to \text{老师}\\ 1 & \text{v} \to \text{尊敬}\\ 0 & \text{nr} \to \text{他们}\end{bmatrix}$$

(e) Structural feature function after adding "father and left sister" nodes
(e) 增加"父亲+左妹"节点后构造的特征函数

$$\begin{bmatrix}1 & \text{dj(0-0-0)} \to \text{np(dj-0-vp) vp(dj-np-0)}\\ 1 & \text{np(dj-0-vp)} \to \text{nr}\\ 1 & \text{vp(dj-np-0)} \to \text{v np(vp-v-0)}\\ 0 & \text{vp(dj-np-0)} \to \text{v pp(vp-v-0)}\\ 1 & \text{np(vp-v-0)} \to \text{nn}\\ \cdots & \cdots\cdots\cdots\cdots\\ 1 & \text{nr} \to \text{我们}\\ 1 & \text{nn} \to \text{老师}\\ 1 & \text{v} \to \text{尊敬}\\ 0 & \text{nr} \to \text{他们}\end{bmatrix}$$

(f) Structural feature function after adding "father, left and right sister" nodes
(f) 增加"父亲+左妹+右妹"节点后构造的特征函数

Fig.3 Structural feature function after adding structural information in non-terminal node

**图3 在每个非终节点增加结构信息后的结构化特征函数**

## 5 实验及分析

本文为了测试结构化支持向量机在融合多结构特征后对中文句法分析精度的影响,进行了一组对比实验,并对结果进行了分析。

### 5.1 语料的预处理

#### 5.1.1 语料1预处理

本文的实验语料1来自北京大学计算语言学研究所公开的北大微型树库的A语料[12]。该语料来自汉英机器翻译研究的测试题库,它句型多样,句子较短,不同短语组合的分布也很广,便于进行自动分析处理。

该语料一共有1 434句,表1是对实验语料集的情况统计[12],表2为实验语料举例。

选取表2中887句单句作为实验语料1,句长最长为19,最短为3,其平均句长为7.01;抽取其中787句作为训练语料,100句用作开放测试的语料。

在进行实验之前,需要对北大树库语料进行改写,比如:

[dj 厂长/n [vp [vbar 宣布/v 了/u] [np 委员/n 名单/n]]]

改写后格式为:

```
( (s (n 厂长)
    (vp (vbar (v 宣布)
              (u 了))
        (np (n 委员)
            (n 名单)))
))
```

#### 5.1.2 语料2预处理

本文的实验语料2采用文献[6]的语料,该语料来自PCTB宾州中文树库语料,从1 500个文档中提取2 000条(句长小于等于12词)单句,其中的1 850句用来进行训练,剩下的150句用来进行开放测试。

在进行本实验前,同样需要对从宾州中文树库选出来的2 000个单句进行预处理[13-14],将句法树上原有的空语类、指同索引和功能标记一概删除[5]。

例如,下面例句A转换成B的形式:

A

```
<S ID=3620>
( (IP-HLN (NP-SBJ (NP-PN (NR 马来西亚))
                  (NP (NN 副总理)))
          (VP (VV 结束)
              (IP-OBJ (NP-SBJ (-NONE- *PRO*))
                      (VP (VV 访)
                          (NP-PN-OBJ (NR 华)))))) )
</S>
```

B

```
( (S (NP (NR 马来西亚)
         (NN 副总理))
     (VP (VV 结束)
         (VP (VV 访)
             (NR 华)) )))
```

### 5.2 评价指标

本文使用PARSEVAL评价体系[2]作为句法分析

Table 1 Statistics of experimental data

**表1 实验语料情况统计**

| 总句数 | 总词数 | 总字数 | 平均句长 | 简单句子 | | 复杂句子 | |
|---|---|---|---|---|---|---|---|
| | | | | 比例/% | 平均句长 | 比例/% | 平均句长 |
| 1 434 | 11 821 | 17 058 | 8.243 | 97.768 | 7.869 | 2.232 | 24.656 |

Table 2 Samples of experimental data

**表2 实验语料句型举例**

| 句型 | 示例 | 句子数量 |
|---|---|---|
| 短语 | [np [np 大学/n [np [np 管理/v 信息/n] 系统/n]] 的/u [vp 试制/v 和/c 应用/v]] | 61 |
| 单句(dj) | [dj 厂长/n [vp [vbar 宣布/v 了/u] [np 委员/n 名单/n]]] | 887 |
| 复句(fj) | [fj [vp 没有/v 计算机/n] [dj [np 复杂/a 的/u 计算/v] [vp 不/d [vp 可能/v 完成/v]]]] | 3 |
| 整句(zj) | [zj [dj 窗户/n [vp 是/v [dj 谁/r 开/v] 的/u]] ?/w] | 483 |

注:复句、单句和短语句尾有标点符号的就被标注为整句。

模型的评价指标,选取其中的精确率(Precision,*Pre*)、召回率(Recall,*Rec*)以及*F*1值(*Pre*和*Rec*的调和平均值)对结果进行评价。

精确率表示所有句法分析结果中所有正确的成分比例;召回率表示句法分析结果中正确的成分占所有句法实际成分的比例;$F1=2\times Pre\times Rec/(Pre+Rec)$。

### 5.3 实验分析

实验使用的句法分析器是从网上公开下载的SVMstruct-cfg(http://www.cs.cornell.edu/tj/svm-light/svm_struct.htm)。使用经典结构化支持向量机$SVM_1$方法,并与文献[6]中$SVM_2$方法以及经典的概率上下文无关文法PCFG[2]在语料1和语料2上进行了实验对比分析。这里的PCFG采用和文献[5]相同的算法,即规则的概率估计采用最简单的相对频率法。结构化支持向量机选取的核函数为线性核,其中惩罚参数$C=1.0$,参数$\varepsilon=0.01$。在文献[6]中,在采用$SVM_2$方法进行句法分析时,曾对选取*F*1损失函数和0-1损失函数进行实验对比,从实验结果中发现采用0-1损失函数要比*F*1损失函数的效果好,故本文在进行结构化支持向量机的实验时,都选取的是0-1损失函数。实验结果只采用开放测试的结果,结构化支持向量机的测试时间极短,可以忽略不计,故只对训练时间进行对比。对比实验结果如表3、表4所示。其中,Time表示模型在当前语料下的训练时间。

从表3、表4开放测试的实验结果可以看出:一阶标注、二阶标注、三阶标注的*F*1值均高于未进行标注的模型。它们之间在精确率上是三阶标注高于二阶标注,二阶标注高于一阶标注。在召回率上有高有低,出现了三阶标注的*F*1值低于二阶标注的情况。这是因为产生了数据稀疏的问题,当增加的结构信息越多时,句法分析的性能反而有下降的情况。同时,从表3、表4中可以看到,当增加一阶标注时,*F*1

Table 3 Comparison of experimental results of adding structural information in Corpus1

**表3 北大微型树库(语料1)上增加各种结构信息的对比实验结果**

| 模型 | $SVM_1$ | | | | $SVM_2$ | | | | PCFG | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Pre*/% | *Rec*/% | *F*1/% | Time/s | *Pre*/% | *Rec*/% | *F*1/% | Time/s | *Pre*/% | *Rec*/% | *F*1/% | Time/s |
| 基线 | 74.1 | 71.2 | 72.6 | 22.6 | 75.7 | 71.5 | **73.5** | 36.1 | 73.3 | 70.6 | 71.6 | 10.7 |
| +父亲 | 80.5 | 77.5 | **78.9** | 24.7 | 80.9 | 76.3 | 78.5 | 38.3 | 78.4 | 75.3 | 76.8 | 13.4 |
| +左妹 | 80.8 | 78.1 | **79.4** | 25.5 | 80.4 | 74.5 | 77.4 | 39.3 | 78.3 | 74.9 | 76.6 | 12.7 |
| +右妹 | 80.2 | 74.2 | 77.1 | 26.3 | 80.6 | 75.3 | **77.9** | 40.6 | 77.1 | 75.7 | 76.4 | 13.8 |
| +父亲+左妹 | 81.6 | 75.3 | 78.3 | 34.2 | 82.5 | 76.9 | **79.6** | 40.5 | 79.8 | 76.4 | 78.1 | 14.3 |
| +父亲+右妹 | 81.8 | 78.7 | **80.3** | 34.8 | 82.1 | 75.2 | 78.5 | 41.9 | 79.3 | 77.1 | 78.2 | 14.5 |
| +左妹+右妹 | 80.8 | 75.8 | 78.2 | 34.5 | 82.4 | 76.3 | **79.2** | 41.3 | 77.6 | 79.2 | 78.4 | 15.1 |
| +父亲+左妹+右妹 | 85.1 | 73.3 | 78.8 | 37.9 | 85.3 | 74.1 | **79.3** | 47.6 | 76.3 | 80.5 | 78.3 | 19.9 |

Table 4 Comparison of experimental results of adding structural information in Corpus2

**表4 宾州中文树库(语料2)上增加各种结构信息的对比实验结果**

| 模型 | $SVM_1$ | | | | $SVM_2$ | | | | PCFG | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Pre*/% | *Rec*/% | *F*1/% | Time/s | *Pre*/% | *Rec*/% | *F*1/% | Time/s | *Pre*/% | *Rec*/% | *F*1/% | Time/s |
| 基线 | 76.4 | 74.1 | 75.3 | 198 | 77.1 | 75.6 | **76.3** | 263 | 70.5 | 77.8 | 74.0 | 79 |
| +父亲 | 80.6 | 79.8 | 80.2 | 251 | 83.1 | 82.7 | **82.9** | 309 | 77.5 | 82.5 | 79.9 | 103 |
| +左妹 | 81.2 | 78.3 | 79.7 | 249 | 82.4 | 79.8 | **81.1** | 311 | 78.9 | 80.3 | 79.6 | 107 |
| +右妹 | 79.8 | 80.9 | 80.3 | 256 | 82.9 | 80.6 | **81.7** | 317 | 77.8 | 81.4 | 79.4 | 109 |
| +父亲+左妹 | 82.5 | 78.7 | 80.6 | 260 | 83.3 | 80.4 | **81.8** | 321 | 79.2 | 81.7 | 80.4 | 127 |
| +父亲+右妹 | 81.9 | 80.4 | 81.1 | 267 | 83.8 | 80.3 | **82.0** | 323 | 79.1 | 81.9 | 80.5 | 131 |
| +左妹+右妹 | 82.7 | 81.3 | **82.5** | 259 | 83.5 | 80.1 | 81.7 | 319 | 79.5 | 82.2 | 80.8 | 135 |
| +父亲+左妹+右妹 | 83.9 | 77.4 | 80.9 | 274 | 84.1 | 78.2 | **81.0** | 326 | 77.3 | 83.1 | 80.1 | 148 |

值有明显的升高,但是增加为二阶标注和三阶标注,$F1$值的增加就不太明显。另外,随着结构信息的增加$F1$值会提高,但是需要的训练时间也在不断增加,而且语料规模越大,训练消耗的时间也越来越多。从$F1$值的对比来看,总的情况是$SVM_2$方法 > $SVM_1$方法 > PCFG方法,但其中也有$SVM_1$方法 > $SVM_2$方法的情况;从语料的训练时间对比来说,PCFG方法 < $SVM_1$方法 < $SVM_2$方法;因而从算法的$F1$值和训练时间双重考虑的话,增加了一阶、二阶、三阶标注后,$SVM_1$方法要好于$SVM_2$方法和PCFG方法。

## 6 结束语

中文句法相比较西文来说其结构更加复杂,具有较强的上下文相关性,在进行句法分析时难度更大。本文使用结构化支持向量机的方法并融合多结构信息对中文句法进行分析,丰富了结构化特征函数的形式。同时,本文使用了两种语料,并对3种句法分析方法在这两种语料库上的实验进行了对比分析,说明增加了结构信息可以在一定程度上提高句法分析的精度。由于对结构化支持向量机在中文信息处理中应用的研究还比较粗浅,在以后很多问题处理中还需要继续进行深入的探讨。

## References:

[1] Charniak E. Statistical parsing with a context-free grammar and word statistics[C]//Proceedings of the 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence, Providence, USA, Jul 27-31, 1997. Menlo Park, USA: AAAI, 1997: 598-603.

[2] Mallning C D, Schutze H. Foundations of statistical natural-language processing[M]. Cambridge, USA: MIT Press, 1999.

[3] Zhang Hao, Liu Qun, Bai Shuo. Structural context conditioned probabilistic parsing of chinese[C]//Proceedings of the 1st Students' Workshop on Computational Linguistics, Beijing, Aug 20-23, 2002. Beijing: Chinese Information Processing Society of China, 2002: 46-51.

[4] Chen Gong, Luo Senlin, Chen Kaijiang, et al. Method for layered Chinese parsing based on subsidiary context and lexical information[J]. Journal of Chinese Information Processing, 2012, 26(1): 9-15.

[5] Huang Changning, Li Yumei, Zhou Qiang. Implicit information of treebank[J]. Journal of Chinese Linguistics, 2012 (15): 149-160.

[6] Zhao Guorong, Wang Wenjian. A Chinese parsing method based on interdependent and structured input and output spaes[J]. Journal of Chinese Information Processing, 2015, 29(1): 139-145.

[7] Vapnik V. Statictical learning theory[M]. New York: John Wiley & Sons, Inc, 1998.

[8] Joachims T, Finley T, Yu C N J. Cutting-plane training of structural SVMs[J]. Machine Learning, 2009, 77(1): 27-59.

[9] Tsochantaridis I, Hofmann T, Joachims T, et al. Support vector machine learning for interdependent and structured output spaces[C]//Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, Jul 4-8, 2004. New York: ACM , 2004: 104-112.

[10] Tsochantaridis I, Joachims T, Hofmann T, et al. Large margin methods for structured and interdependent output variables[J]. Journal of Machine Learning Research, 2005, 6(2): 1453-1484.

[11] Joachims T, Hofmann T, Yue Yisong, et al. Predicting structured objects with support vector machines[J]. Communications of the ACM, 2009, 52(11): 97-104.

[12] Zhou Qiang, Zhang Wei, Yu Shiwen. Building a chinese treebank[J]. Journal of Chinese Information Processing, 1997, 11(4): 42-51.

[13] Johnson M. PCFG models of linguistic tree representations [J]. Computational Linguistics, 2002, 24(4): 613-632.

[14] CollinsM J. A new statistical parser based on bigram lexical dependencies[C]//Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, Santa Cruz, USA, Jun 24-27, 1996. Stroudsburg, USA: ACL, 1996: 184-191.

## 附中文参考文献:

[3] 张浩, 刘群, 白硕. 结构上下文相关的概率句法分析[C]//第一届学生计算语言学研讨会, 北京, 2002. 北京: 中国中文信息学会, 2002: 46-51.

[4] 陈功, 罗森林, 陈开江, 等. 结合结构下文及词汇信息的汉语句法分析方法[J]. 中文信息学报, 2012, 26(1): 9-15.

[5] 黄昌宁, 李玉梅, 周强. 树库的隐含信息[J]. 中国语言学报, 2012(15): 149-160.

[6] 赵国荣, 王文剑. 一种处理结构化输入输出的中文句法分析方法[J]. 中文信息学报, 2015, 29(1): 139-145.

[12] 周强, 张伟, 俞士汶. 汉语树库的构建[J]. 中文信息学报, 1997, 11(4): 42-51.

ZHAO Guorong was born in 1979. She is a Ph.D. candidate and associate research librarian at Shanxi University. Her research interests include Chinese information processing and machine learning, etc.
赵国荣(1979—),女,山西大同人,山西大学博士研究生、副研究馆员,主要研究领域为中文信息处理,机器学习等。

WANG Wenjian was born in 1968. She received the Ph.D. degree from Institute for Information and System Science, Xi'an Jiaotong University in 2004. Now she is a professor and Ph.D. supervisor at School of Computer and Information Technology, Shanxi University, and the senior member of CCF. Her research interests include data mining and machine learning theory, etc.
王文剑(1968—),女,山西太原人,2004年于西安交通大学获得博士学位,现为山西大学计算机与信息技术学院教授、博士生导师,CCF高级会员,主要研究领域为数据挖掘,机器学习等。