文章编号：1003-0077(2015)01-0139-07

一种处理结构化输入输出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，山西 太原 030006）

摘 要：中文句法结构复杂，特征维数较高，目前已知最好的汉语句法分析效果与其他西方语言相比还有一定的差距。为进一步提高中文句法分析的效率和精度，该文提出一种采用二阶范数软间隔优化的结构化支持向量机（Structural Support Vector Machines，Structural SVMs）方法对基于短语结构的中文句法进行分析，通过构造结构化特征函数 $\psi(x,y)$，体现句法树的输入信息，并根据中文句子本身具有的强相关性，在所构造的 $\psi(x,y)$ 中增加中文句法分析树中父节点的信息，使 $\psi(x,y)$ 包含了更加丰富的结构信息。在宾州中文树库 PCTB 上的实验结果表明，该文方法与经典结构化支持向量机方法以及 Berkeley Parser 相比可取得较好的效果。

关键词：中文句法分析；加权上下文无关文法；结构化 SVM；二阶范数软间隔优化

中图分类号：TP391 **文献标识码**：A

A Chinese Parsing Method Based on Interdependent and Structured Input and Output Spaces

ZHAO Guorong, WANG Wenjian

(School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: Chinese syntax has complex structure and high dimension features, and the best known Chinese parsing performance is still inferior to that of other western languages. In order to improve the efficiency and accuracy of Chinese parsing, we propose a L2-norm soft margin optimization structural support vector machines (structural SVMs) approach. By constructing the structural function $\psi(x,y)$, the input information of syntactic tree can be mapped well. Since Chinese syntax has a strong correlation, we use father node of phrase structure trees to enrich the structure information of $\psi(x,y)$. The experiment results on the benchmark dataset of PCTB demonstrate that the proposed approach is effective and efficient compared with classical Structural SVMs and Berkeley Parser system.

Key words: Chinese parsing; weighted context-free grammars; Structural SVMs; L2-norm soft margin optimization

1 引言

句法分析是自然语言处理中一个重要的环节，句法分析的结果直接决定着信息检索、信息抽取、机器翻译和自动文摘等自然语言处理系统的最终性能，也是语用、语义等自然语言处理深层研究的基础。所谓句法分析是根据给定的语法[1]，自动地推导出句子所包含的句法单位和这些句法单位之间的关系，即句子的语法结构。句法分析要遵循某一语法体系，根据该体系的语法确定语法树的表示形式。目前，在句法分析中使用比较广泛的有短语结构语法和依存语法，前者（特别是上下文无关语法）应用最为广泛。

句法分析方法的研究大体分为两种途径：基于规则的方法[2]和基于统计的方法[3-4]。前者从汉语句子最本质的特征出发，规则相对稳定且易于表达汉语句子成分的构成规律，方法也较成熟，但是规则的获取过程十分繁琐，句法分析的歧义问题很难解决，常会出现不稳定以及规则冲突等问题。基于统计的句法分析方法具有效率高、鲁棒性好的优点，可以有效地消除语言现象中的各种歧义现象，但是在处理高维数据时效果不太理想。由 Vapnik 等人[5]提出的支持向量机（Support Vector Machine，

收稿日期：2013-05-06 定稿日期：2014-07-23

基金项目：国家自然科学基金(60975035,61273291)；山西省回国留学人员科研资助项目(2012-008)

SVM)具有简洁的数学形式、成熟的求解方法和良好的泛化能力,尤其处理高维数据具有较强的优势,在自然语言处理领域被广泛应用。但是传统的SVM训练效率偏低,在处理结构复杂的数据时有一定局限性,对输出的结果也不能从概率上进行解释。

Hofmann和Joachims等人在2004年[6]针对实际应用中数据具有比较复杂的结构,而且数据本身存在相互依赖关系(如队列结构、树形结构或网状结构等数据),改进了传统支持向量机,首次提出结构化支持向量机学习方法,并将其应用于英文句法分析中,取得了较好的效果。由于结构化支持向量机处理复杂结构这一特点,有望在未来10年在多个应用领域得到广泛应用[7]。

中文句子结构复杂,词类和句法成分之间的关系错综复杂,同一词类可担任多种句法成分并且无形态变化,还有兼类词等问题,实现中文句法分析要更困难一些。文献[8]直接引入经典的结构化支持向量机方法进行中文句法分析,结构特征函数的构造也仅仅是句法规则的抽取以及规则出现的频次统计,但所取得的实验结果说明了Structural SVMs进行中文句法分析的可行性和有效性。本文采用二阶范数软间隔优化形式的结构化支持向量机,同时由于中文句法结构具有较强相关性,在特征构造上引入短语结构树中父节点的信息,使特征函数的构成信息更加丰富。在宾州中文树库PCTB上进行中文句法分析实验,并和文献[8]方法以及Berkeley Parser[9]的实验结果进行了分析比较。

2 优化的结构化支持向量机学习模型

2.1 结构化支持向量机模型

结构化支持向量机是基于判别式的模型,结构化数据分析问题的目的是要构造出样本的输入与输出对之间的一个映射函数 $f: X \rightarrow Y$,在句法分析中 f 给定输入句子 X 到输出短语结构树 Y 的一个映射。构造函数 f 的一个重要任务是需要学习一个基于输入/输出对的判别式函数 $F: X \times Y \rightarrow \Re$,通过对输出变量的最大化,实现对输出结果的预测。设Structural SVMs的目标函数为[6,10]:

$$f(x;w) = \underset{y \in Y}{\operatorname{argmax}} F(x,y,w) \tag{1}$$

F 是基于输入/输出组合特征表示 $\psi(x,y)$ 的线性函数,如式(2)所示。

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

这里,w 是权向量,$\psi(x,y)$ 的形式取决于具体要解决问题。以句法分析为例,句子 x 的句法分析树 y 中的每一个结点对应一个规则 g_j,相应的规则对应一个权值 ω_j。对于一个句子 x 的有效句法树 y,每一个结点的权值 ω_j 的和作为这个句法树的分值,计算分值的函数为 $F(x,y;w) = \langle w, \psi(x,y) \rangle$。对于给定的句子 x,通过CKY(Cocke-Younger-Kasami)算法[1]找出符合文法的句法分析树集 Y,从中找出分值最大的 $F(x,y,w)$,$y \in Y$,即为句子的语法树。对于Structural SVMs,最关键的是特征函数 $\psi(x,y)$ 的构造。

2.2 二阶范数软间隔优化

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[5]:

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

(3)式可以等价转换为:

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

其中定义 $\delta \Psi_i(y) \equiv \Psi(x_i, y_i) - \Psi(x_i, y)$。

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:

$$SVM_0: \min_{w} \frac{1}{2} \| w \|^2 \tag{5}$$
$$\forall i, \forall y \in Y \setminus y_i: \langle w, \delta \psi(y) \rangle \geqslant 1$$

在式(5)中引入松弛变量的软间隔,以容忍部分噪声和离群点,并兼顾更多的训练点,而不只是靠近边界的那些点。松弛变量软间隔的形式可以是一阶的,也可以是二阶的,它的泛化性的好坏和实际数据有关,文献[8]使用一阶的形式,在本文使用二阶范数软间隔优化形式[11]:

$$SVM_2: \min_{w,\xi} \frac{1}{2} \| w \|^2 + \frac{C}{2n} \sum_{i=1}^{n} \xi_i^2 \tag{6}$$
$$\forall i, \forall y \in Y \setminus y_i: \langle w, \delta \psi_i(y) \rangle \geqslant 1 - \xi_i, s.t.\ \forall i, \xi_i \geqslant 0$$

这里常量 $C>0$,用来平衡训练错误最小化和间隔最大化。

在实际问题中常常把损失函数引入到约束条件中,因为偏离标准值大的点更应调整,即具有高损失的点更应受到惩罚。所以在式(6)的约束条件中引入了损失函数:

$$\Delta S\_SVM_2: \min_{w,\xi} \frac{1}{2} \| w \|^2 + \frac{C}{2n} \sum_{i=1}^{n} \xi_i^2 \tag{7}$$
$$\forall i, \forall y \in Y \setminus y_i: \langle w, \delta \psi_i(y) \rangle \geqslant 1 - \frac{\xi_i}{\sqrt{\Delta(y_i, y)}},$$

$$s.t.\ \forall i,\xi_i \geqslant 0$$

除了对松弛变量进行再调整外，还可针对间隔进行再调整，得到以下优化问题：

$$\Delta M\_SVM_2: \min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{2n}\sum_{i=1}^{n}\xi_i^2 \qquad (8)$$

$\forall i, \forall y \in Y \setminus y: \langle w, \delta\psi_i(y)\rangle \geqslant \sqrt{\Delta(y_i,y)} - \xi_i, s.t. \forall i, \xi_i \geqslant 0$

优化问题(6)、(7)、(8)可分别通过相对应的对偶形式进行求解。

在解决二次凸优化问题中，最大的困难就是遇到规模很大的约束条件，用常规的优化方法很难求解[12]。而对于句法分析，每个可能的句法分析树 y 都会对应一个约束，因而会产生规模非常大的约束条件，甚至可能达到指数级数量的约束，但是考虑到大间隔问题的特殊结构，实际上在指数级数量的约束集合中只需要非常少的约束即可求解问题，依据这个特性可以将指数级的约束数量减少为多项式数量集。最大间隔方法通过构造初始优化问题的一个嵌套序列的不断紧密的松弛，从而可保证产生一个足够精确的近似解。

3　基于优化的结构化支持向量机中文句法分析方法

句法分析的任务是对于给定的输入句子，生成一棵与之对应的句法树。句法分析的过程主要有两步：首先判断进行句法分析的句子在句法上是否正确，然后对于句法正确的句子，输出其短语结构树。

本文采用短语结构的句法分析，将文法限定为乔姆斯基范式的形式[1]，所有的语法规则为：$n_l[A \rightarrow BC]$或者 $n_l[A \rightarrow \alpha]$的形式，这里 A、B、C 是非终结符，α 是终结符，并且每一条规则都对应相应的权值 w_l，并将其称之为加权上下文无关文法。假定 x 为给定的句子，针对 x 的分析结果为若干个句法树用 Y 表示，最佳分析树设为 $h(x)$，$rules(y)$ 代表每棵句法树 y 中所有的语法规则的集合，加权上下文无关文法模型的形式如式(9)所示。

$$h(x) = \underset{y\in Y}{\operatorname{argmax}}\left\{\sum_{nl\in rules(y)} w_l\right\} \qquad (9)$$

加权上下文无关文法模型中引入结构化特征函数 $\psi(x,y)$，则

$$\langle w, \psi(x,y)\rangle = \sum_{nl\in rules(y)} w_l \qquad (10)$$

其中 $\psi(x,y)$表示规则及规则出现的次数，w 表示统计学习中训练得到的权值。

3.1　结构化特征函数 $\psi(x,y)$ 的构造

在结构化支持向量机方法中，特征函数 $\psi(x,y)$ 的构造是重点和难点，在实际应用中，$\psi(x,y)$ 构造的合适与否直接影响结构化支持向量机方法的有效性。

中文句子具有很强的上下文相关性，对于词语的序列来说，出现在序列前或后的词语都会对确定句子序列有所帮助。所以本文在构造特征函数时，加入了短语结构树中父节点信息，以丰富特征函数的结构信息。图 1 为中文句法分析的输入输出示例，图 2 为本文增加父节点信息后针对中文句法分析构造的 $\psi(x,y)$。

输入 x：秘鲁侨胞举行新年晚会

$f: x \rightarrow y$

输出 y：

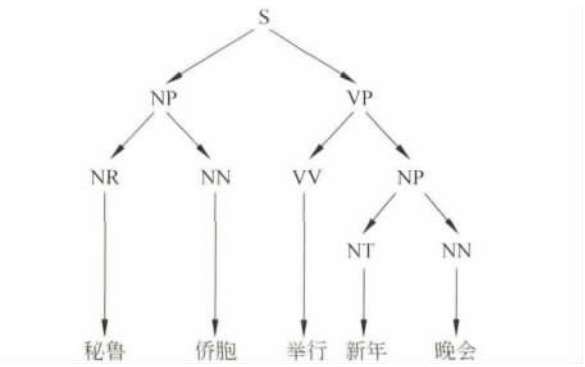

图 1　句法分析输入输出示例图

$$\psi(x,y)=\begin{bmatrix} 1 & S\rightarrow NP\hat{}S\ VP\hat{}S \\ 1 & NP\hat{}S\rightarrow NRNN \\ 1 & NP\hat{}VP\rightarrow NTNN \\ 1 & VP\hat{}S\rightarrow VVNP\hat{}VP \\ \vdots & \vdots \\ 1 & NR\rightarrow 秘鲁 \\ 1 & NN\rightarrow 侨胞 \\ 1 & VV\rightarrow 举行 \\ 1 & NT\rightarrow 新年 \\ 1 & NN\rightarrow 晚会 \\ 0 & VV\rightarrow 召开 \end{bmatrix}$$

图 2　构造的 $\psi(x,y)$

3.2　基于结构化输入输出的中文句法分析算法

首先使用求解结构化支持向量机中文句法分析任务的近似优化算法对训练数据进行训练，得到相应的权值。对于测试数据，首先使用 CKY 算法，对每一条输入的句子分析出符合语法规则的句法树集[13]，然后使用加权上下文无关文法模型，将训练得到的权值带入式(9)，求出最佳分析树。

基于结构化输入输出的中文句法分析算法的主要步骤如下：

step1：输入训练样本$(x_1,y_1),\cdots(x_n,y_n)$，设置参数$C$，$\varepsilon$

step2：初始化工作集S_i为null，$i=1,\ldots,n$.

step3：设定损失函数

SVM_2：$H(y)\equiv(1-\langle\delta\psi_i(y),w\rangle)\Delta(y_i,y)$

ΔS_SVM_2：$H(y)\equiv(1-\langle\delta\psi_i(y),w\rangle)\sqrt{\Delta(y_i,y)}$

ΔM_SVM_2：$H(y)\equiv\sqrt{\Delta(y_i,y)}-\langle\delta\psi_i(y),w\rangle$

其中$w\equiv\sum_j\sum_{y'\in S_j}\alpha_{jy'}\delta\psi_j(y')$

step4：计算出$\hat{y}=\arg\max_{y\in y}H(y)$

step5：计算得出$\xi_i=\max\{0,\max_{y\in S_i}H(y)\}$

如果$H(\hat{y})>\xi_i+\varepsilon$，则$S_i\leftarrow S_i\cup\{\hat{y}\}$，

在S上进行二次优化更新α_s，

$S=U_iS_i$

如果S_i不发生变化，则结束；否则，继续返回step3进行优化。

算法开始时，约束集Si为空。之后在每次迭代中，从可能的指数级的约束集合中寻找样本集中最违反约束条件的x_i所对应的输出$\hat{y}$，不断更新约束集S_i，直到达到约束集不再发生变化，在该算法收敛之前，仅仅多项式数量的约束被加入到了约束集S_i中，因此本算法可以大大减少约束项的数目。

4 实验结果分析

4.1 评价指标

对句法分析模型的评价是句法分析研究的一项重要内容，在句法分析中，PARSEVAL句法分析评价体系被认为是一种粒度适中较为理想的评价方法，在句法分析系统中被广泛使用[1]，它主要由精确率、召回率两部分组成。

对于一组需要分析的句子，假设语料库中对这组句子标注的所有成分的集合为目标集，句法分析系统实际分析出的句子成分为分析集，分析集和目标集的交集为共有集，即正确识别出的句子成分的数量。精确率是衡量句法分析系统分析的所有成分中正确的成分的比例，召回率是衡量句法分析系统分析的所有正确成分在实际成分中的比例，F_1用来协调精确率和召回率。它们分别定义如式(11)～(13)所示。

$$\text{精确率}=\frac{\text{共有集元素个数之和}}{\text{分析集元素个数之和}} \tag{11}$$

$$\text{召回率}=\frac{\text{共有集元素个数之和}}{\text{目标集元素个数之和}} \tag{12}$$

$$F_1=\frac{2\times\text{精确率}\times\text{召回率}}{\text{精确率}+\text{召回率}} \tag{13}$$

4.2 实验语料

本文使用PCTB宾州中文树库语料，从它的1 500个文档中提取出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，从训练集中抽取150句用来进行封闭测试，除训练集之外的150句用来进行开放测试，称为语料1，进行简单句的实验；然后又提取300句复杂句(平均句长为26)，其中250句用来训练，50句用来进行开放测试，称为语料2，进行复杂句的对比实验。本实验采用线性核[14] k=u*v，其中惩罚参数C=1.0，参数ε=0.01。

4.3 对实验语料的处理

在做本实验时，本文将从宾州中文树库选出来的2 000个单句进行实验处理，将句法树上原有的空语类、指同索引和功能标记一概删除[15]。

例如，下面的例句A要删除转换成B的形式：

A

```
<S ID=9359>
((IP-HLN (NP-SBJ (NP-PN (NR 秘鲁))
                 (NP (NN 侨胞)))
         (VP (VV 举行)
             (NP-OBJ (NT 新年)
                     (NN 晚会)))))
</S>
```

B

```
((S (NP (NR 秘鲁)
        (NN 侨胞))
    (VP (VV 举行)
        (NP (NT 新年)
            (NN 晚会)))
))
```

4.4 两种损失函数实验结果的比较

采用语料1先进行简单句的实验，分别使用

SVM_2、ΔS_SVM_2、ΔM_SVM_2 在加注父节点的情况下对中文句法进行分析，并选用 0-1 损失和 F_1 损失进行实验，实验结果如表 1 与表 2 所示。

表 1　采用 0-1 损失的三种方法的实验结果比较

方法	封闭测试/%			开放测试/%			训练时间 CPU(s)
	精确率	召回率	F_1	精确率	召回率	F_1	
SVM_2	87.1	86.4	86.7	83.1	82.7	82.9	309
ΔS_SVM_2	89.1	88.5	88.8	83.2	82.4	82.8	656
ΔM_SVM_2	86.3	85.8	86.1	84.1	83.4	83.7	461

表 2　采用 F_1 损失的三种方法的实验结果比较

方法	封闭测试/%			开放测试/%			训练时间 CPU(s)
	精确率	召回率	F_1	精确率	召回率	F_1	
SVM_2	84.8	83.9	84.4	82.7	82.1	82.4	666
ΔS_SVM_2	87.7	86.9	87.3	83.9	83.1	83.5	991
ΔM_SVM_2	84.7	84.1	84.3	83.1	82.3	82.7	945

从表 1 与 2 可以看出，采用 0-1 损失的 SVM_2、ΔS_SVM_2、ΔM_SVM_2 三种方法的训练时间均比使用 F_1 损失的三种方法耗费的时间少，而且在 0-1 损失下 SVM_2 训练时间是最少的。由于测试数据需要的时间非常短，消耗的时间基本差不多，所以没有对三种方法的测试时间进行对比。

在封闭测试中，三种方法各自相比均是采用 0-1 损失的效果好。在开放测试中，SVM_2 和 ΔM_SVM_2 方法采用 0-1 损失的开放测试的结果比使用 F_1 损失的效果好，ΔS_SVM_2 方法是采用 F_1 损失的开放测试结果比使用 0-1 损失的效果好。总体来看，使用三种方法对中文句法进行分析，即是采用 0-1 损失比采用 F_1 损失总的效果要好。

在表 1 采用 0-1 损失的方法比较中，SVM_2 方法在封闭测试和开放测试中 F_1 值均排在第二，使用的训练时间是最少的；ΔS_SVM_2 方法封闭测试中 F_1 值排在第一，开放测试中 F_1 值却排在第三，但是和 SVM_2 方法开放测试的 F_1 值非常接近，然而使用的训练时间最多；ΔM_SVM_2 方法是 F_1 值在封闭测试排在第三，开放测试的 F_1 值排在第一，使用训练时间在其它两种方法之间。所以综合来看，使用 SVM_2 方法测试的效果居中，比较稳定，训练时间用时最短。

采用语料 2 进行复杂句的实验，从对表 1～2 的分析中可以看到采用 0-1 损失比采用 F_1 损失的实验效果好，故针对复杂句的实验采用 0-1 损失，SVM_2、ΔS_SVM_2、ΔM_SVM_2 三种方法的结果比较如表 3 所示。

表 3　采用 0-1 损失复杂句实验结果比较

方法	开放测试/%			训练时间 CPU(s)
	精确率	召回率	F_1	
SVM_2	53.9	47.1	50.3	103
ΔS_SVM_2	50.4	44.3	47.2	106
ΔM_SVM_2	45	40	42.4	96

相对于简单句而言，复杂句不仅是句子的长度增加了，而且句子的结构也变得更加复杂，句法分析树的深度也相应增加。即使采用结构化支持向量机的方法对复杂句进行句法分析，效果也不太理想。但仅是对 SVM_2、ΔS_SVM_2、ΔM_SVM_2 三种方法进行比较的话，在对复杂句的分析中，SVM_2 方法的实验结果更好一些。当针对复杂句时，约束条件越少，反而效果更好一些。

4.5　不同模型的实验结果比较

文献[8]采用的是经典的结构化支持向量机方法对中文句法进行分析，特征函数的构造也较简单。Berkeley Parser 是一个纯粹的基于 PCFG[16] 方法的句法分析器，目前比较成熟，应用也比较广泛，而且 PCFG 方法可以看成加权上下文无关文法的一个特例。所以将这两个模型与本文提出的方法进行

比较，以验证本文所提方法的有效性。实验语料采用的是语料1。

在0-1损失的情况下，Berkeley Parser、文献[8]方法与SVM_2（增加了父节点信息）模型实验结果比较如表3所示。

表4 三种模型的实验结果比较

方法	封闭测试/%			开放测试/%			训练时间 CPU(s)
	精确率	召回率	F_1	精确率	召回率	F_1	
Berkeley Parser	87.1	86.7	86.9	82.9	82.5	82.7	530
文献[7]	83.7	82.1	82.9	80.6	79.8	80.2	251
SVM_2	87	86.4	86.7	83.1	82.7	82.9	309

从表4可以看出，SVM_2（增加了父节点信息）方法在封闭测试上的F_1值略低于Berkeley Parser，但是在开放测试F_1值又比Berkeley Parser略好。文献[8]在三个指标上的值均比较差，其原因是它的特征函数的构造比较简单，ε的值采用的是绝对值的形式，而本文SVM_2方法在特征函数的构造上增加了父节点的信息，ε值采用平方项的形式更适合中文句法分析应用。但是从表中可以看到本文和文献[8]所使用的训练时间比Berkeley Parser少很多。

5 结束语

句法分析在信息处理中处于重要地位，它的效果的好坏，直接影响机器翻译、自动文摘、信息获取等的处理效果，因而对句法分析的研究有很重要的意义。本文采用结构化支持向量机的方法对中文句法进行分析，主要针对单句进行了实验分析，也对复杂的句子进行简单的实验分析，从结果上看，取得了令人满意的效果，也验证了使用结构化支持向量机对中文句法进行分析的可行性。由于结构化支持向量机在中文信息处理中应用还处于探索阶段，很多问题还需要继续进行深入的探讨和研究。

参考文献

[1] Manning C D, Schutze H. Foundations of statistical natural language processing [M]. London: the MIT Press, 1999.

[2] 冯志伟. 基于短语结构语法的自动句法分析方法[J]. 当代语言学, 2000, 2(2): 84-98.

[3] 马金山. 基于统计方法的汉语依存句法分析研究[D]. 哈尔滨：哈尔滨工业大学, 2007.

[4] 吴伟成，周俊生，曲维光. 基于统计学习模型的句法分析方法综述[J]. 中文信息学报. 2013, 27(3): 9-19.

[5] Vapnik V. Statictical Learning Theory [M]. New York: Wiley, 1998.

[6] Tsochantaridis I, Hofmann T, Joachims T, et al. Support Vector Machine Learning for Interdependent and Structured Output Spaces[C]//Proceedings of the twenty-first International Conference on Machine Learning, 2004: 104-112.

[7] Dietterich G H, Domingos P, Getoor L. Structured Machine Learning: the next ten years [J], Machine Learning, 2008, 73(1): 3-23.

[8] 王文剑，王亚贝. 基于结构化支持向量机的中文句法分析[J]. 山西大学学报（自然科学版）. 2011, 1: 66-72.

[9] http://code.google.com/p/berkeleyparser/

[10] T Joachims, T Hofmann, Yisong Yue, et al. Predicting Structured Objects with Support Vector Machines[J], Communications of the ACM, Research Highlight, November, 2009, 52(11): 97-104.

[11] Tsochantaridis I, Joachims T, Hofmann T, et al. Large Margin Methods for Structured and Interdependent Output Variables [J]. Journal of Machine Learning Research, 2005, 9: 1453-1484.

[12] Joachims T, Finley T, Chun-Nam Yu. Cutting-Plane Training of Structural SVMs [J]. Machine Learning, 2009, 77(1): 27-59.

[13] Eugene C, Mark J. Coarse-to-fine n-best parsing and MaxEnt Discriminative reranking[C]//Proceedings of the 43rd Annual Meeting of the ACL, 2005: 173-180.

[14] Nello C, John S T. An Introduction to SVM and Other Kernel-based Learning Methods [M]. 北京：电子工业出版社, 2004.

[15] 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.

[16] Collins M J. A new statistical parser based on bigram lexical Dependencies [C]//Proceedings of ACL, 1996: 184-191.

赵国荣(1979—)，博士研究生，讲师，主要研究领域为自然语言处理、机器学习等。
E-mail：zhaogr@sxu.edu.cn

王文剑(1968—)，博士，教授、博士生导师，主要研究领域为神经网络、支持向量机、机器学习理论、环境计算等。
E-mail：wjwang@sxu.edu.cn

(上接第 138 页)

[8] Coltuc, Dinu. Improved embedding for prediction-based reversible watermarking[J]. IEEE Transactions on Information Forensics and Security, September 2011, 6: 873-882.

[9] **刘志杰. 基于自然语言的文本可恢复水印研究**[D]. **湖南：湖南大学**, 2010.

[10] **姜传贤，陈孝威. 鲁棒可逆文本水印算法**[J]. **计算机辅助设计与图形学学报**，2010，22(5)：879-885.

[11] Topkara M, Topkara M, Mercan, et al. The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions[C]//Proceedings of ACM Multimedia and Security Workshop (MMSEC'07), 2006: 164-174.

[12] **甘灿，孙星明，刘玉玲，等. 一种改进的基于同义词替换的中文文本信息隐藏方法**[J]. **东南大学学报**，2007, 37, Sup(Ⅰ)：137-140.

[13] **张宇，刘挺，陈毅恒，等. 自然语言文本水印**[J]. **中文信息学报**，2005, 19(1)：56-62.

[14] Zheng Xueling, Huang Liusheng, Chen Zhili, et al. Hiding Information by Context-Based Synonym Substitution[C]//Proceedings of Digital Watermarking - 8th International Workshop Proceedings, Guildford, United kingdom, 2009: 162-168.

[15] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[C]//Proceedings of the Coling 2010:Demonstrations, Beijing, China, 2010: 13-16.

[16] Sougou Labs. SougouR[DB/OL]. http://www.sogou.com/labs/dl/r.html, 2011

[17] **向华，曹汉强，伍凯宁等. 一种基于混沌调制的零水印算法**[J]. **中国图象图形学报**，2006, 11(5)：720-724.

费文斌(1986—)，硕士，主要研究领域为通信网络与信息安全技术。
E-mail：feiwenbin111@sina.com

王静(1989—)，硕士，主要研究领域为通信网络与信息安全技术。
E-mail：842098130@qq.com

唐向宏(1962—)，博士、教授，主要研究领域为数字水印、图像修复、多载波通信等。
E-mail：tangxh@hdu.edu.cn