基于树核函数的中文语义角色标注研究 *

王步康[1]，王红玲[2]，袁晓虹[3]，周国栋[4]

（苏州大学计算机科学与技术学院,江苏 苏州 215006;

江苏省计算机信息处理技术重点实验室,江苏 苏州 215006）

摘 要：目前使用特征方法进行语义角色标注研究已经遇到发展瓶颈，性能难以进一步提高；而基于核函数的方法可以充分利用特征方法无法表示的结构化信息，有进一步研究的空间。本文使用 SVM 提供的卷积树核函数构造了一个中文语义角色标注系统，该系统以依存关系作为标注单元进行中文语义角色标注。本文重点描述了通过不同的裁剪方法来获得依存树的结构化信息，裁剪后的依存树分别为最短路径树和最小树。在中文 PropBank 和 NomBank 的转换语料上的实验结果表明：使用最小树能得到系统的最佳性能，在动词性谓词和名词性谓词上分别获得 82.87，76.40 的 F1 值。

关键词：语义角色标注，树核，依存关系

Tree Kernel-Based Semantic Role Labeling in Chinese Language

Wang Bukang[1], Wang Hongling[2], Yuan Xiaohong[3], Zhou Guodong[4]

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China;

Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou 215006, China)

E-mail: 20094227014@suda.edu.cn

Abstract: Currently, it is hard to further improve the performance of feature-based semantic role labeling because the features' limits. The kernel-based method can represent the structural information better than the feature-based method, so it has great value to the future research. This paper implements a Chinese dependency-based semantic role labeling system, uses the convolution tree kernel of SVM. In this paper, we focus on how to properly express the structural representation between predicates and arguments on dependency tree and let the input tree contain less noise information. We explore two methods to prune the dependency tree: Shortest Path Tree(SPT) and Minimum Tree (MT). The experiments on the transferred corpuses from Chinese PropBank and Chinese NomBank show our system achieves the best performance by using the minimum tree. It achieves 82.87 in labeled F1 on verbal predicates and 76.40 in labeled F1 on nominal predicates.

Key words: Semantic Role Labeling, Tree Kernel, Dependency Relationship

1. 引言

语义角色标注(Semantic Role Labeling, SRL)是浅层语义分析的一种实现方式，是目前自然语言处理领域中的热点研究课题之一，其在问答系统、信息抽取、机器翻译等领域有着广泛的应用。语义角色标注采用“谓词-角色”的结构形式，标注出句子中给定谓语的语义角色，每个语义角色被赋予一定的语义含义，如施事、受事、工具或附加语等。当前语义角色标注通常是在句法分析的基础上进行的，根据使用的句法分析不同，可分为基于短语结构句法分析的 SRL 和基于依存关系的句法分析的 SRL。另外，常用的语义角色标注方法又可分为基于特征的方法和基于核

* 基金资助：国家自然科学基金(60673041, 60873150)；国家教育部博士点基金(200802850006)；江苏省自然科学基金(BK2008160)；江苏省高校自然科学重大基础研究项目（08KJA520002）。

作者简介：王步康（1987-），男，硕士研究生，主要研究方向：自然语言处理；E-mail：20094227014@suda.edu.cn；王红玲（1975-），女，讲师，通讯作者，主要研究方向：自然语言处理；袁晓虹（1985-），女，硕士研究生，主要研究方向：自然语言处理；周国栋（1967-），男，教授，博士生导师，研究方向：自然语言处理

函数的方法。目前使用基于特征的方法已经遇到发展瓶颈，很难找出更有效特征，性能难以进一步提高等。而基于核函数的方法目前的研究还较少，有进一步研究的空间。

通过分析两种句法分析的结果发现，表示依存关系的句法树和短语结构的句法树存在一些区别：在短语结构句法树上，每个词为叶子节点且顺序排列，非叶子节点能够表示句法信息（如NP表示名词性短语）或词性的标记；而在依存树上，所有节点都是句子中的词，词与词之间并不是顺序排列，使用依存树的结构来描述他们之间的依存关系。因此我们认为在研究使用依存关系的SRL时，结构化信息的使用尤为重要，如王等[1]采用基于特征向量的方法研究依存关系SRL中，所选择的20个特征中，有9个用来表示结构化信息，而表示句法和词法的特征仅有4个。

然而，在使用特征向量表示结构化信息时，当结构化信息转化为平面特征向量时，可能会丢失部分有效信息。例如，即使依存树中两条非常相似的路径，可能因为某一中间节点的差异会被当成截然不同的特征。与基于特征向量的方法不同，基于树核函数的方法以结构树为处理对象，通过直接计算两个离散对象（如句法树）之间的相似度来进行分类，这使得基于核函数的方法可以充分利用特征方法无法表示的结构化信息，因此近年一些研究人员开始研究和使用该方法（如Moschitti等[2][3]和Zhang等[4]）进行语义角色标注。

本文在中文依存句法分析的基础上，使用树核函数的方法，进行语义角色标注研究。本文后续组织结构如下：第二部分回顾了SRL的相关工作；第三部分重点论述了依存树的构建裁剪方法；第四部分为相关的实验、实验结果和性能分析；最后是全文总结和将来的工作方向。

2. 相关工作

Moschitti等[2]首次提出在语义角色分类中应用卷积树核。在短语结构句法树上，他们选择包含谓词-论元的句法分析树的子结构作为谓词-论元特征 (predicate-arguments feature ，PAF)空间，并在PAF空间中定义卷积树核，使用CoNLL2005英文的训练语料和测试语料（WSJ）F1值为69.80。由于PAF核不利于角色识别，为了改进系统性能，Mochitti等[3]又提出了一个改进的PAF核 (MPAF，iMproved PAF)来进行语义角色标注。在MPAF核中，一个句法成分的根节点附加一个“-B”符号，从而，新的核能够区分路径特征与句法成分结构特征的边界线，在相同的语料上F1值提高到70.61。

Zhang等[4]一文中利用句法结构的近似匹配和句法分析树中节点的近似匹配设计了句法驱动卷积树核，通过引入简化的产生式（通过相似子结构匹配）和节点特征变种（通过相似节点匹配）来放宽子树出现的条件，从而保留了基本的语言学约束以及原始产生式的语义。使用句法驱动卷积树核和基于特征多项式核构建混合核后在CoNLL2005语料上取得78.13的F1值。

而到目前为止，还未有文献报告利用树核函数的方法研究基于依存关系的中文SRL。王等[1]使用基于特征向量的方法对基于依存关系的动词性谓词作了相关研究，选择了7个基础特征和13个扩展特征，最终使用CTB转换语料取得了84.30的F1值。

3. 依存树构建

对于基于树核函数的方法，有效的表示结构化信息尤为重要。目前大部分基于树核方法的SRL都是在短语结构句法树上进行的。依存树不同于短语结构句法树，依存树的节点上包含很多词法信息（词性，单词等），而树核函数的输入树结构中，每个节点只可描述一种信息（如依存关系）。为此我们通过构造依存关系树来获得结构化信息。下面是中文NomBank中的一个例

句：

他希望今后两国的经贸科技合作与交流进一步扩大和发展 。　　例句（1）

图 1 给出了以往研究中常用的例句（1）的依存树，图中 W 表示单词，R 表示依存关系，G 表示词性，黑体字分别表示谓词，及各个谓词所对应的角色。

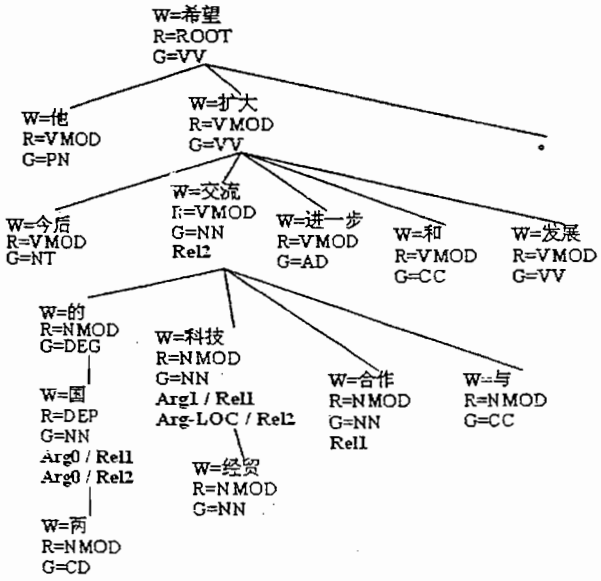


图 1　谓词"合作"和"交流"的语义角色标注实例

为了能够表示依存树上的词法特征且符合树核函数的输入结构，本文在图 1 的基础上进行修改，即对每个依存树节点进行扩充，将“词性-单词”作为每个节点的孩子链，以描述更多的节点信息（如图 2 所示），这样的树称为“完全依存树”。

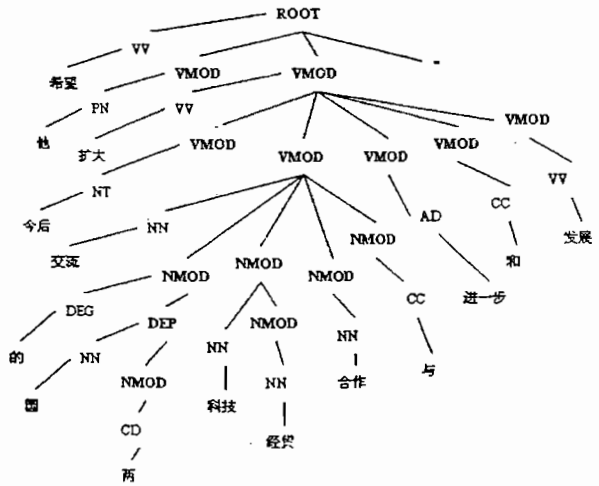


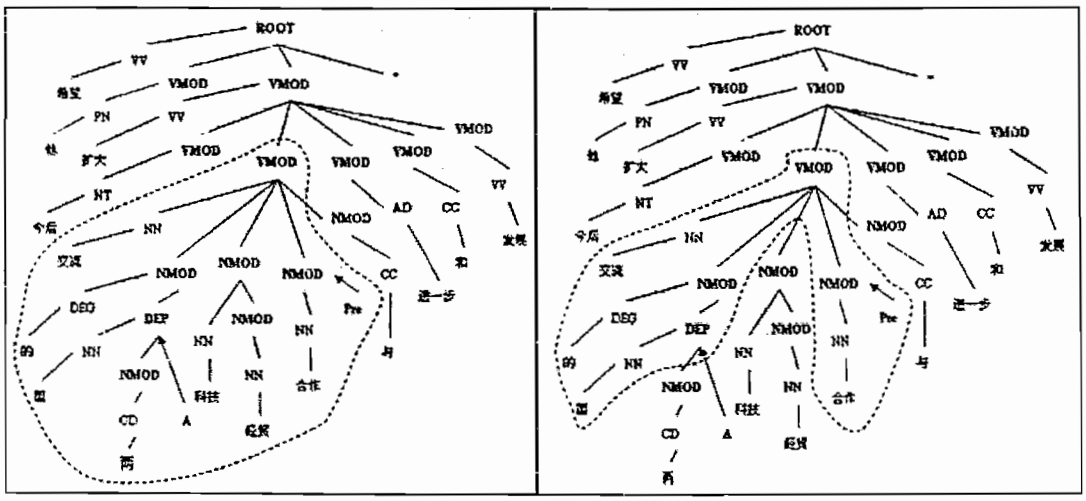
图 2　例句 1 对应的完全依存树

从图 2 中可以看出完全依存树包含了丰富的结构化信息，但是对于 SRL 而言，所处理的对象不仅仅是角色正例，还有大量的角色反例，而反例中两个谓词和候选角色在依存树中的位置往往较远，使得它们的完全树相当复杂，导致卷积树核函数在计算两棵树之间的相似度时要消耗更多的时间。因此尽管原始的完全依存树包含了丰富的结构化信息，由于其规模过于庞大，且包含

了太多的与语义角色无关的噪音，并不适合于基于卷积树核函数的 SRL。

Moschitti 等在短语结构句法树上基于树核方法的 SRL 使用的 PAF 空间实质上是公共节点树（CT），即：指谓词和候选角色（或成分）向上找到两者共同的深度最深的祖先结点，保留该祖先结点的所有孩子结点，去除完全树中的其他部分。通过分析发现这种 CT 树并不适合于依存关系树，依然会包含很多噪音而影响系统结果。本文在 CT 树的基础上进一步裁剪，通过分析依存树上谓词与角色的关系，主要采用了两种裁剪策略，以例句（1）中谓词“合作”和候选角色“国”为例，方法如下：

（1）最短路径树（Shortest Path Tree ，SPT）：在公共结点树的基础上，只包含依存树中谓词和候选角色词之间的部分，具体如图 3(a)虚线部分所示。



（a）例句 1 对应的最短路径包含树（SPT）　　（b）例句 1 对应的最小树（MT）

图 3 依存树的裁剪方法

（2）最小树（Minimum T ree ，MT）：对最短路径树进一步裁剪，只保留谓词和候选角色的若干直接祖先节点，向上回溯到它们的深度最深的公共结点为止，具体如图 3(b)虚线部分所示。

4. 基于树核函数的 SRL

本文构建的基于树核函数的 SRL 系统，以依存关系作为标注单元，其标注过程分三个步骤：预处理、语义角色识别和语义角色分类。

4.1 卷积树核

本文采用的树核函数为 Collins 和 Duffy[5]定义的卷积树核函数(Convolution T ree Kernel ，CTK)，CTK 通过枚举两棵句法树之间的相同子树的数目来计算它们之间的相似度，并且能够有效的在高维空间计算两个向量之间的点积。其公式为：

$$K(T_1,T_2)=\langle\Phi(T_1),\Phi(T_2)\rangle=\sum_i(\Phi_i(T_1)\cdot\Phi_i(T_2))=\sum_{n_1\in N_1}\sum_{n_2\in N_2}\sum_i I_i(n_1)*I_i(n_2) \quad (1)$$

其中，N_1 和 N_2 分别是树 T_1 和树 T_2 全部结点的集合，指示函数 $I_i(n)$的值为 1 ，当且仅当存在一棵以 n 为根节点的类型为 i 的子树，否则值为 0。Collins 和 Duffy 指出： $K(T_1,T_2)$是树结构

上的一个卷积核的实例，并可以通过下面的递归定义在 $O(|N_1|\times|N_2|)$的时间内计算出，其中 $\Delta(n_1,n_2)=\sum_i I_i(n_1)*I_i(n_2)$，用来计算以 n_1 和 n_2 为根结点的两棵子树之间的相似度：

(1) 如果 n1 和 n2 处的产生式规则不同，则有 $\Delta(n_1,n_2)=0$；

(2) 否则，如果 n1 和 n2 子节点相同并且都是叶子节点，则有 $\Delta(n_1,n_2)=\mu$；

(3) 否则，$\Delta(n_1,n_2)=\mu\prod_{j=1}^{nc(n_1)}(1+\Delta(ch(n_1,j),ch(n_2,j)))$。

其中 $nc(n_1)$ 是节点 n_1 儿子的个数，ch(n,j)是节点 n 的第 j^{th} 个儿子，$\mu(0<\mu<1)$ 是衰减因子，树的规模越大，则会乘上更多的 μ，因此可以控制核函数的值不会随着树的规模变大而急剧变大。

4.2 实验语料

根据目标谓词的词性，SRL 一般分为动词性谓词 SRL 和名词性谓词 SRL。为了便于比较，对于动词性谓词 SRL 语料，本文选择使用王等[1]的 CTB 转换语料，基本语料库是 Chinese TreeBank5.0，标注信息来源于 PropBank1.0，使用 Penn2Malt 工具将基于短语结构的句法树库转换成依存关系树库。实验选取 CTB 中的前 760 篇文档（chtb_001.fid 到 chtb_931.fid），共 10364 个句子，其中（chtb_100.fid 到 chtb_931.fid）中 9127 个句子作为训练语料，共有谓词 32387 个；（chtb_001.fid 到 chtb_099.fid）中共 1238 个句子作为测试语料，共有谓词 4793 个。

而名词性谓词 SRL 语料，选择使用中文 NomBank（对应于中文 PropBank2.0 和中文 TreeBank5.1），同样使用 Penn2Malt 转换成依存关系树库，参照 Xue[6]的实验数据划分，取 NomBank 转换语料中的 648 个文件（chtb_081.fid-chtb_899.fid）作为训练集，40 个文件（chtb_041.fid-chtb_080.fid）作为开发集，72 个文件（chtb_001.fid-chtb_040.fid 和 chtb_900.fid-chtb_931.fid）作为测试集。其中，训练集、开发集和测试集所包含的名词性谓词数分别为 8642、731 和 1124。

4.3 预处理

预处理阶段主要对依存关系树的节点进行过滤，过滤掉依存树上最不可能承担谓词角色的关系结点，不予以标注，以有效地减少输入到分类器中的实例个数，尤其是减少负例的数量。Hacioglu[7]提出了一种简单的过滤算法：在依存树中，仅考虑与谓词具有以下关系的结点：父亲、孩子、孙子、兄弟、兄弟的孩子、兄弟的孙子结点，其他结点都被过滤掉。本文根据中文依存关系树结构的特点扩展了 Hacioglu 剪枝方法，增加了与谓词具有以下关系的结点：保留了谓词结点的祖父结点、祖父的孩子结点等。系统使用该改进的 Hacioglu 算法后，经过统计两个训练集的实例大大减少（均超过 50%），同时对动词性谓词语料过滤的正例不足 1%，而对名词性谓词语料不足 1.5%。

4.4 实验结果与分析

在角色识别和角色分类中，我们使用SVM Light工具包♥中树核函数作为分类模型，特别地，由于SVM Li ght分类器本质上是一个二元分类器，所以在角色分类阶段采用一对多方法(one v s. others)将其重新包装为多元分类器，最终分类结果取得分最大的那个类别。在训练时，训练参数C值大小设置为 4.0。

♥ SVM-LIGHT-TK. http://dit.unitn.it/~moschitt

针对上述两种裁剪后的依存树结构化信息表示方法，用 4.2 节所述的动词性谓词语料和名词谓词语料分别进行训练和测试，最终的 SRL 性能如表 1 所示。

表 1　基于树核函数的名词性谓词 SRL 性能

	方法	准确率 P（%）	召回率 R（%）	F1 值
动词性谓词 SRL	SPT 方法	90.01	76.52	82.72
	MT 方法	89.69	77.01	82.87
	王等[1]特征向量方法	88.00	80.89	84.30
名词性谓词 SRL	SPT 方法	86.84	66.37	75.23
	MT 方法	86.60	68.37	76.41
	袁等[8]特征向量方法	71.37	86.20	78.09

从表 1 中可以看出，对于两种谓词的 SRL，MT 方法的性能均略高于 SPT 方法，这主要因为，SPT 方法最大程度的保留了谓词和候选角色之间的结构化信息，所以系统能够更加准确的判断两树的相似度，在准确率上略高；相反，由于比较更加精确，识别的角色也就较少，导致了召回率的损失；MT 方法尽可能去除了依存树上的冗余信息，保留了尽可能少的有效结点，删除了很多的噪音，从而较之 SPT 方法，召回率有了较大的提高，虽然准确率有所降低，但总体的 F1 值却得到提高，由此也说明在使用树核函数进行研究时，树的结构化信息表示尤为重要。

从表 1 中还可以看出，对于两种不同的结构化信息表示方法，动词性谓词 SRL 的性能均明显高于名词性谓词 SRL。名词性谓词 SRL 性能低的主要原因如下：首先，虽然语料规模相当，但是名词性谓词的标注实例还是远远低于动词性谓词的标注实例，即正例较少。其次，名词性谓词的角色识别更加困难。即使某个名词为动词的派生词，该名词的所有修饰成分也不一定是该名词的语义角色，这使得名词性谓词的 SRL 要复杂。

相比于基于特征向量的中文 SRL，本文所使用的基于树核函数的方法性能均略低。首先，动词性谓词 SRL 性能相比于王等[1]（使用相同语料和相同的评测方法）基于特征的方法取得性能 F1 值 84.30，低了 1.43；名词性谓词 SRL 相比于袁等[8]（使用相同语料和相同的评测方法）基于特征的方法 F1 值低了 1.68。主要原因在于，本文所使用的依存树的结构化信息表示方法还不能很好的描述句子的句法特征，以至于召回率较低，从而影响 F1 值。

为了进一步与基于特征向量方法的 SRL 比较，对于名词性谓词 SRL，我们分别取原始划分语料中训练语料的 1/4、1/2、3/4 作为新的训练语料，并分别在树核函数的 SRL 系统和袁等[8]的基于特征向量的系统上进行实验。并画出两种方法的性能随语料大小变化的趋势图，如图 4 所示。

从图 4(a)中可以看出，随着训练语料的增加，两种方法的总体性能 F1 值都逐渐增加，但增加趋势减缓，并且基于特征向量的方法性能变化大于基于树核函数的方法，说明基于特征向量的方法对语料库的依赖性更大；图 4(b)显示了准确率随训练集变化而变化的趋势，从中我们可以发现，基于树核函数的方法在准确率上一直高于基于特征向量的方法，并且训练语料的缩减对其影响不大，而基于特征向量的方法的准确率对语料库大小比较敏感，语料库减小会使准确率迅速降低；图 4(c)显示了两种方法的召回率随训练集变化的变化趋势，从中可以发现，基于特征向量的方法随语料库的增加，召回率的增加趋势变缓，而基于树核函数的方法召回率的变化几乎呈线性变化，说明随着训练集的增加，采用基于树核的方法能够获得更多分类正确的角色数。

总的来说，在使用基于树核函数的方法时，我们应该更多的考虑，如何构造更有效的结构，来识别出更多正确的角色。

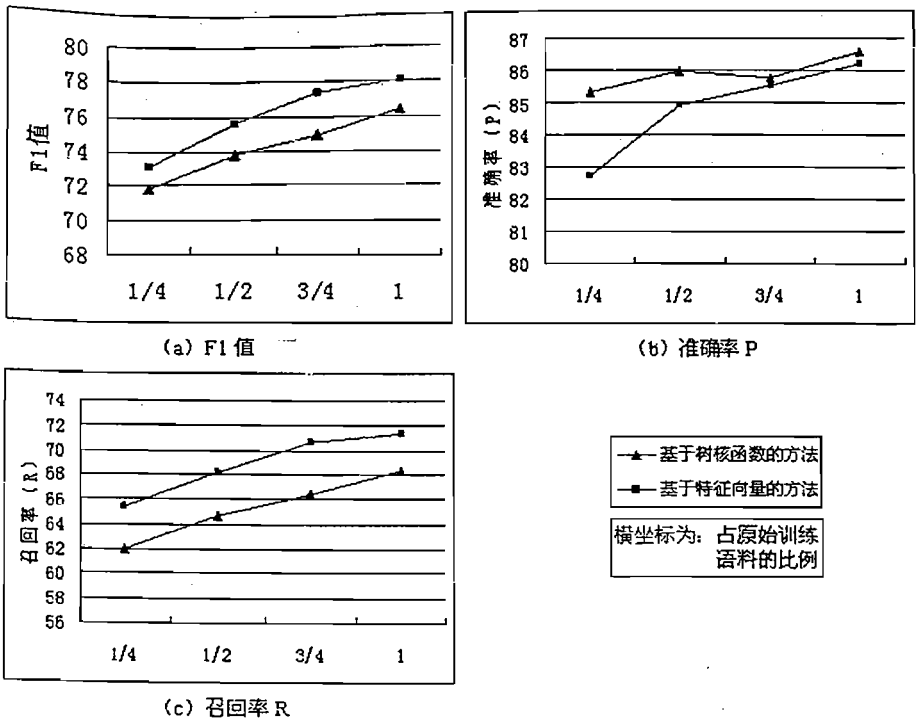


图 4 系统性能随训练集变化的趋势图

5. 总结与展望

本文探索了应用树核函数在依存关系进行中文语义角色标注，重点探讨了依存树的结构化信息表示方法，并在两种语料上分别进行了实验，取得了不错的效果。同时本文也将实验结果与基于特征向量的 SRL 进行了详细比较与分析，具有一定的研究参考价值。

虽然本文的实验结果性能略低于基于特征的方法，但特征方法中要获取新的平面特征提高性能已经十分困难，而在基于树核函数的方法中，还可以进一步探索结构化信息的表达方式以及树核函数的计算方法，因此使用树核方法进行语义角色标注具有很大的研究空间。

参 考 文 献

[1] 王步康, 王红玲, 袁晓虹, 周国栋: 基于依存句法分析的中文语义角色标注. 中文信息学报[J]. 2010, 24(1): 25-29.

[2] Moschitti A. A Study on Convolution Kernels for Shallow Semantic Parsing [C]. ACL'2004, 2004, 335–342.

[3] Moschitti A, P ighin D, and B asili R . Tree K ernel Engineering i n S emantic R ole Labeling S ystems [C]. EACL'2006, 2006, 49-56.

[4] Zhang M., Che W. X., Zhou G. D, Aw A.T.Semantic Role Labeling using a Grammar-driven Convolution Tree Kernel [J]. IEEE Transaction on Audio, Speech and Language Processing. 2008, 16(7): 1315-1329.

[5] Collins M, Duffy N. Convolution kernels for natural language [C]. Proceedings of NIPS-2001. 2001.

[6] Xue N. Labeling Chinese predicates with semantic roles[C]. Computational Linguistics, 2008, 34(2):225-255.

[7] Hacioglu.K. Semantic Role Labeling Using Dependency Trees [C].In Proceedings of CoNLL-2004, US, 2004.

[8] 袁晓虹, 王红玲, 王步康, 周国栋: 基于依存关系的中文名词性谓词语义角色标注研究[J]. 计算机应用与软件, 已录用.