基于依存句法分析的中文语义角色标注*

王步康　王红玲*　袁晓虹　周国栋

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

E-mail: hlwang@suda.edu.cn

摘　要：依存句法是句法分析的一种，相比于短语结构句法分析，依存句法具有更简洁的表达方式。本文采用英文语义角色标注的研究方法，实现了一个基于中文依存句法分析的语义角色标注系统。该系统针对中文依存关系树，采用有效的剪枝算法和特征，使用最大熵分类器进行语义角色的识别和分类。系统使用了两种不同的语料，一种是由标准短语结构句法分析（CTB5.0）转换而来，另一种是 CoNLL2009 公布的中文评测语料。系统分别在两种语料的正确谓词和自动谓词的基础上进行实验，在正确谓词上取得的 F1 值分别为 84.30 和 81.68，在自动谓词上的 F1 值为 81.02 和 81.33。

关键字：语义角色标注，依存关系，最大熵分类器

Chinese Dependency-Based Semantic Role Labeling

Bukang Wang　Hongling Wang　Xiaohong Yuan　Guodong Zhou

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

Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou, 215006, China

E-mail: hlwang@suda.edu.cn, Phn: +86-0512-65243192

Abstract: Dependency representations are more simple and intuitive than constituent representations. This paper implements an Chinese dependency-based semantic role labeling (SRL) by using the similar methods used in English SRL. In the system, effective pruning algorithm and useful features are adopted for Chinese dependency tree, semantic role identification and classification are exploited by a maximum entropy classifier. Two different corpora are used in our system, one is transferred from constituent-based corpus (CTB5.0), and another is Chinese dataset provided by CoNLL 2009 shared task. Based on the two datasets, the system achieves 84.30 and 81.68 respectively in labeled F1 for gold predicates, 81.02 and 81.33 for automatic predicates.

Key words: Semantic Role Labeling, Dependency Relations, maximum entropy classifier

1. 引言

当前根据采用的句法分析结果，自动语义角色标注（Semantic Role Labeling, SRL）可分为：基于短语结构句法分析的语义角色标注和基于依存结构句法分析的语义角色标注。针对前者的研究已较为成熟，并取得了很好的性能，然而伴随此方法的发展带来的瓶颈问题也日渐突出，

* 基金资助：国家 863 计划(2006AA01Z147);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，本科，专业：计算机科学；王红玲（1975-），女，博士研究生，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

* 通讯作者

如局部模型的机器学习方法很难有更大进展，语料的稀疏问题严重，更有效的特征难以抽取等等，导致了性能无法进一步提高。因此近两年来基于依存句法的语义角色标注开始受到重视，尤其是CoNLL2008 shared task[1]和CoNLL2009 shared task[2]都将基于依存关系的SRL作为评测主题，更加推进了基于依存句法的语义角色标注的发展。

依存结构句法分析相比于短语结构句法分析，它表达的句法结构是单词与单词之间的依赖关系。从理论上分析，依存句法中的句法-语义接口更简单、更直观，并提供了更透明的谓词-论元关系表达。因此在基于短语结构句法分析的语义角色标注系统遭遇到发展瓶颈后，研究基于依存结构句法分析的语义角色标注更具有现实意义。

本文采用英文语义角色标注的研究方法，使用中文依存句法分析，构建了一个中文语义角色标注系统。文章第2部分简述了基于依存关系的SRL的相关工作。第3部分介绍了基于依存句法的中文语义角色标注系统，重点描述构建系统的各个步骤，基础特征和扩展特征。第4部分给出了各个扩展特征的表现，并对实验结果进行了分析和比较。最后第5部分对本文进行了总结，并对后期工作进行了展望。

2. 相关研究

和基于短语结构句法分析的SRL相比，基于依存分析的SRL研究相对较少。在英文方面，Hacioglu等[3]首次采用基于依存分析的方法来实现语义角色标注，所使用的依存树是由句法树转化而来，采用SVM分类器实现了角色的分类，提出了12个特征（依存关系，位置，中心词，依赖词等），并且表明谓词相关信息的重组对性能影响很大。最终在基于手工依存分析语料库Depbank和CoNLL2004 shared task语料库上的F1值分别为84.6和79.8。而最新的基于依存关系的SRL研究出现在CoNLL2008评测中，代表作是Johansson等[4] [5]的工作，在文中详细分析比较了两种SRL系统在PropBank语料上的性能，文章的贡献在于分别使用基于部分短语的（segment-based）和基于依存关系（dependency-based）的衡量标准来公平的比较代表当前最好性能的两类SRL系统的性能。他们实现的基于依存句法的SRL系统在上述两项衡量标准下F值分别为77.97（WSJ+Brown）和84.29（CoNLL2008测试集）。

到目前为止，还未有文献报告基于依存句法分析的中文语义角色标注研究。正在进行的CoNLL 2009 shared task是在CoNLL2008 shared task的基础上，进行包括中文在内的多种语言的依存句法和语义的联合分析，最新结果显示在中文上使用CoNLL2009评测语料达到的最好SRL系统性能是80.66（Labeled F1值）。在基于短语结构句法分析的SRL方面，代表作是Xue等[6][7]的研究，其主要工作是比较和分析了中文和英文语义角色标注的性能以及影响因素，在Chinese Propbank上的实验结果表明：基于手工标注句法树的SRL系统F1值可达91.3%；基于单一自动标注句法树，F值大幅降为61.3%。这说明基于手工分析的中文语义角色标注的系统结果基本与英文的相当，甚至稍微高出一点；但对于自动产生的句法树，结果要比英文的差得多。

3 中文语义角色标注系统

3.1 语料资源

尽管目前针对中文的依存句法分析研究很多，但是尚未出现通用的大规模标注的中文依存关系语料库。当然也没有大规模标注的基于依存关系的语义角色标注语料。因此要进行基于依存句

法的自动语义分析研究，首先要解决语料库的来源问题。

为了便于评测比较，系统使用了两种语料资源。一种是转换语料（后文简称 CTB 转换语料），获得方法类似于 CoNLL2008 shared task 的语料获得，基本语料库是 Chinese TreeBank5.0，标注信息来源于 PropBank1.0，并且只针对动词性谓词标注语义角色。借助 MaltParser*工具将基于短语结构的句法树库转换成依存关系树库，并使用 Penn2Malt[1]工具将语料转换成 CoNLL2008 标注的格式。实验选取 CTB 中的前 760 篇文档（chtb_001.fid 到 chtb_931.fid），共 10,364 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。

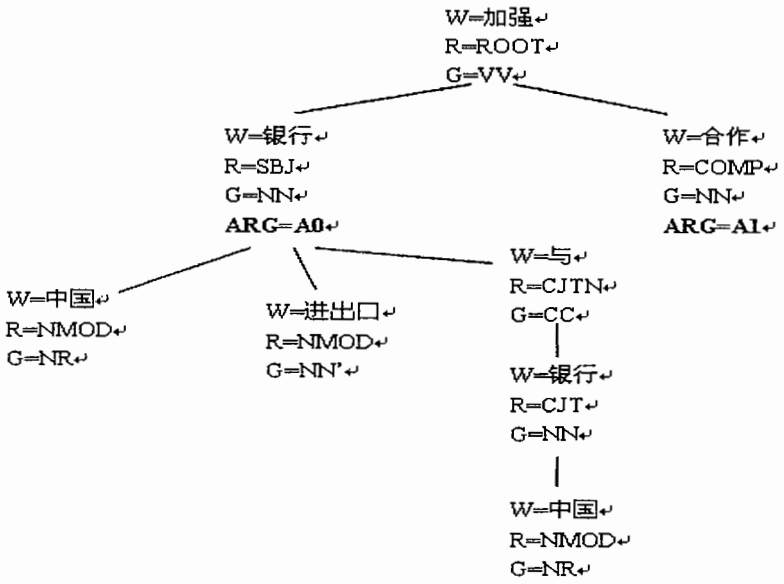

图 1 中文依存关系树实例图

另一种语料是 CoNLL2009 share task 提供的，其中训练集有句子 22277 句，谓词 102813 个，使用开发集*作为测试语料，共有句子 1762 句，含有的谓词数为 8103 个。该语料是 CTB6.0 的一个子集，语义信息则源自 Chinese PropBank 2.0。

例句（1）给出了转换语料中的一个例子，图 1 给出了例句的中文依存关系树图，图中 ARG 表示谓词角色，W 表示单词，R 表示依存关系，G 表示词性。

中国进出口银行与中国银行加强（.01）合作。　　（例句 1）

在例句（1）中只有一个谓词：加强，其中（.01）表示这个谓词的词义，该词义是 Chinese PropBank 中谓词“加强”的框架语义中对应的词义项编号。

3.2 标注步骤

本文构建的基于依存关系的中文 SRL 系统其标注过程分成了四个部分（如图 2 所示）：谓词标注（Predicate Labeling）、预处理（Pre-processing），语义角色识别（Semantic Role Identification）、语义角色分类（Semantic Role Classification）。

* http://w3.msi.vxu.se/~nivre/research/MaltParser.html

*由于 CoNLL2009 shared task 提供的测试集还没有公布 Gold 标注，所以我们用开发集充当测试集

其中谓词标注是识别出句子中的动词性谓语，并为它们分配词义。在传统的基于短语结构句法分析的SRL系统中通常不执行这步，默认谓词已识别正确。由于CoNLL2008要求进行谓词标注，所以在此我们也对系统在自动谓词标注下进行实验。所采用的自动谓词标注使用基于统计的方法实现（具体实现方法另文介绍），在CTB转换语料和CoNLL2009语料上的谓词识别的F1的值分别为96.95和95.64。另外为方便比较，在本系统中只采用谓词识别的结果，不为它们分配词义。

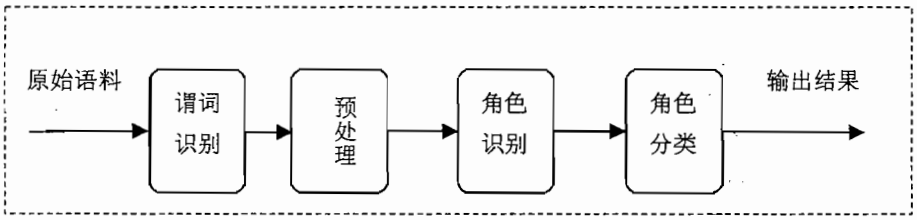

图2 系统标注过程

3.3 预处理

在预处理阶段主要对依存关系树进行剪枝，删除依存树上最不可能承担谓词角色的关系结点，以消除不必要的结构化信息，有效地减少输入到分类器中的实例个数。

此前Hacioglu[3]提出了一种简单的剪枝算法，其方法是：在依存树中，保留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。该算法主要针对英文句法树且谓词为动词性谓语。

在仔细分析中文依存关系树结构的基础上，我们扩展了Hacioglu剪枝方法，增加了与谓词具有以下关系的结点，即保留了谓词结点的祖父结点、祖父的孩子结点，祖父的父亲结点等。系统使用该改进的Hacioglu算法后，经过统计进入分类器的训练实例大大减少（减少约76.9%），同时误剪率不足1%。

3.4 基础特征

特征一直是决定语义角色标注系统性能的重要因素。类似于基于短语结构句法分析的系统，参照Gildea[8]等选取的7个基本特征（谓词、句法类型、子类框架、分析树路径、位置、语态和中心词），我们选取以下作为系统基础特征。假设例句（1）对应依存树（图1）中当前结点为“W=银行”，当前谓词为“加强”，现将各特征列举如下：

谓词原型：当前谓词的原型。（加强）

谓词词性：当前谓词的词性。（VV）

子类框架：当前谓词结点的所有孩子结点的依存关系链。（ROOT-SBJ-COMP）

路径：句法树上当前结点到谓词的路径，即途经结点的依存关系。（SBJ->ROOT）

位置：当前结点的中心词相对于当前谓词的前后顺序。（before）

依存关系：当前结点所对应的依存关系。（SBJ）

中心词：当前结点的父亲结点所对应的单词本身。（加强）

3.5 扩展特征

借鉴基于英文依存关系的语义角色标注系统，我们在系统上加入了以下扩展特征。

假设例句（1）对应依存树（图 1）中当前结点为“W=加强”，当前谓词为“加强”，现将各特征列举如下，特征如果不存在使用 NULL 代替。

谓词的孩子的依存关系链：当前谓词的所有的孩子结点的依存关系组成的链。（SBJ-COMP）

谓词的孩子的词性链：当前谓词的所的孩子结点的词性组成的链。（NN-NN）

谓词的兄弟的依存关系链：当前谓词的所有的兄弟的依存关系组成的链。（NULL）

谓词的兄弟的词性链：当前谓词的所有兄弟结点的词性组成的链。（NULL）

依赖词：指当前结点本身单词。（加强）

中心词词性：中心词的词性。（NULL）

依赖词的词性：当前结点单词的词性。（VV）

家族成员：剪枝后剩下的结点，几乎都是与谓词在同一个家族树中，此特征说明了在此家族树中，当前关系结点与当前谓词的家族关系，如：father，child，siblings等等。（myself）

谓词+中心词：当前谓词原型+中心词。（加强+NULL）

当前关系+中心词：当前结点依存关系+中心词 。（ROOT+NULL）

谓词原型+路径：当前谓词原型+路径。（加强+ ROOT）

依存关系+依存关系前一个词：当前依存关系的类型+依存关系前一个词。（ROOT+银行）

依存关系+依存关系后一个词：当前依存关系的类型+依存关系后一个词。（ROOT+合作）

4 实验结果与分析

实验采用最大熵分离器，其原型是开源软件 maxent-2.4.0[*]，并在此基础上进行了相关的修改，使输出符合系统的要求，参数 cutoff 和 interation 分别设为 2 和 100。评测时采用 CoNLL2008 Share Task 提供的评测程序 eval08.pl[1]，仍使用 Precision、Recall 和 Labeled F1 对最终系统的性能进行评价。

4.1 特征表现

实验时我们首先建立一个基于基础特征的系统，称为基础系统；然后把扩展特征单独加入基础系统中，得到每个特征的表现（如表 1 所示）。

从表 1 中可以看出，与依存关系有关的特征（依赖词和依存关系）对系统性能提高较明显。特别加入依赖词特征后，系统在两种语料上的性能分别提高了 5.56%和 1.92%，效果最明显。特征“依存关系+依存关系前一个词”和“依存关系+依存关系后一个词”表达了当前依存关系的上下文特征，在分别加入这两个特征后，系统性能也有提高。这首先说明依存关系对系统性能贡献很大，反过来也说明系统对依存关系的依赖较强，系统受到依存句法性能的影响。另外从 Wang 等[9]的研究可发现，在基于短语结构句法分析的 SRL 系统中，中心词特征对系统性能贡献很大。而实际上，依存句法中的依赖词就相当于短语结构句法分析中的成分中心词，因此它们的作用也是类似的，我们的实验也证明了这一点。而我们在此提出的中心词概念与前者的中心词概念有所不同，与之有关的特征也基本未起作用。

与谓词孩子和兄弟结点有关的特征（谓词孩子的词性链、依存关系链和兄弟结点的词性链、

[*] http://maxent.sourceforge.net/

依存关系链等）加入系统后，系统性能略有下降。使用这些特征的本意是想表达与谓词相关的上下文信息，但由于依存关系本身已包含了一些这样的信息，因此这些特征的作用不大。

表 1 每个扩展特征和组合特征单独加在基础系统上的结果

特征	CTB 转换语料			CoNLL2009 语料		
	P(%)	R(%)	F1	P(%)	R(%)	F1
基础系统	79.35	73.72	76.43	84.87	74.26	79.21
+谓词的孩子的词性链	79.16	74.16	**76.58**	84.68	74.39	79.20
+谓词的孩子的依存关系链	79.20	73.55	76.27	84.59	74.34	79.14
+谓词的兄弟的词性链	79.23	72.66	75.80	84.95	74.07	79.14
+谓词的兄弟的依存关系链	78.94	73.55	76.15	85.00	74.12	79.19
+依赖词	84.91	79.61	**82.17**	86.79	75.66	**80.85**
+中心词词性	79.50	73.53	76.40	85.05	74.15	**79.23**
+依赖词的词性	79.41	73.85	**76.53**	84.87	74.12	79.13
+家族成员	79.87	74.54	**77.11**	85.04	74.28	**79.30**
+谓词+中心词	80.28	73.09	**76.52**	86.31	73.00	79.10
+当前关系+中心词	79.15	73.69	76.32	85.17	74.45	**79.45**
+谓词原型+路径	79.43	74.27	**76.76**	85.72	74.58	**79.76**
+依存关系+依存关系前一个词	80.89	75.71	**78.21**	85.17	74.24	**79.33**
+依存关系+依存关系后一个词	83.19	77.78	**80.39**	85.35	74.44	**79.52**

4.2 系统结果

在基础系统上添加了全部扩展特征以后，得到了系统在两种语料上的性能，结果如表 2 所示。为评测谓词标注对系统性能的影响，我们分别在标准谓词和自动谓词上进行了实验。

表2 系统结果

	P(%)	R(%)	F1
CTB 转换语料			
标准谓词	88.00	80.89	84.30
自动谓词	86.39	76.29	81.02
CoNLL 2009 中文语料			
标准谓词	88.29	76.00	81.68
自动谓词	86.03	77.13	81.33

从表 2 中可以明显看出系统在使用标准谓词的两个语料上获得的性能有所差别，F 值分别为 84.30 和 81.68 ，系统在 CTB 转换语料上的性能比 CoNLL2009 语料上高了 2.62。同时系统在自动谓词上的性能都有所降低，说明谓词标注也是影响系统性能的一个重要因素。相比较而言，转换语料上由谓词标注带来的性能降低比较明显，准确率和召回率均降低，造成整个 F1 值下降了 3.28%；而在 CoNLL2009 语料上，准确率降了 2.26%，召回率反而有所上升，因此整个系统的 F1 值下降不明显，这说明在 CoNLL2009 语料上系统受谓词标注性能的影响较小，出现这种情况原因可能在于标注语料库（CTB5.0 与 CTB6.0，CPB1.0 与 CPB2.0）之间的差别，详细原因有待

进一步分析。但总的来说，由于 CoNLL2009 的语料数据量大，因此结果更加可信。

相比于 CoNLL2009 公布的中文 SRL 的系统性能（使用主办方提供的测试集得到 F1 值为 80.66），我们的系统在基于自动谓词的 CoNLL09 中文语料上的性能与之基本相当。另外相比于基于短语结构句法分析的中文 SRL 系统（手工标注，标准谓词，F 值为 92.75），两者的性能差距很大，这与中文依存句法分析结果不完善有关。因此现阶段，中文依存句法分析性能是影响中文依存语义分析的关键因素。

相比于英文基于标准依存分析和标准谓词标注的 SRL 系统，如 Johansson 等[5]其 F1 值为 85.52，中文的 SRL 系统性能略微有所下降。这其中的主要的原因可能在于中英文标准依存关系的来源上。由于目前两种语言都没有大规模手工标注的依存关系语料库，因此本文中所指的中英文标准依存关系，均由短语结构句法分析转换得来。对于英文的转换，CoNLL2008 的主办方经过了精心处理，因此转换结果较为可靠；而对于中文的转换，我们的转换语料只是使用了 MaltParser 工具进行，该工具是一个针对多语言的句法分析器，因此对中文的许多特征语言现象不可能做很多的特殊处理，因此转换结果存在一定的误差，也就影响了后续的 SRL 性能。

5 结论及展望

本文使用英文语义角色标注的方法，首次实现了一个基于依存句法的中文语义角色标注系统。相比于传统的基于短语结构句法分析的中文语义角色标注，该系统使用依存句法分析结果构建相应句法分析树，并在此树上抽取特征，进行角色的识别和分类，得到了系统结果。由于这方面的研究还未开展，同时也缺乏可靠的手工标注的语料库和统一的评价标准，因此无法详细评价系统性能，在此我们只是报告一个初步的实验结果，起到抛砖引玉的作用。

后续的工作中，我们将在这个方向展开进一步的研究，包括如何选取更为丰富、有效的特征提高系统性能，使用自动依存句法分析的进行中文 SRL，以及如何进行句法分析和语义分析联合学习等内容。

参 考 文 献

[1] CoNLL 2008, http://www.yr-bcn.es/conll2008/, [EB].

[2] CoNLL 2009, http://ufal.mff.cuni.cz/conll2009-st/,[EB].

[3] Kadri Hacioglu. Semantic Role Labeling Using Dependency Trees.In Proc.of CoNLL-2004, Boston,MA, USA, 2004

[4] Johansson R. and Nugues P.. Dependency-based semantic role labeling of PropBank[C]. In Proceedings of EMNLP-2008. 2008.

[5] Johansson R. and Nugues P.. Dependency-based syntactic-semantic analysis with PropBank and NomBank[C]. In Proceedings of CoNLL-2008.23-24 Aug. 2008.

[6] Xue Nianwen., Palmer M.. Automatic semantic role labeling for Chinese verbs[C]. In Proc. of IJCAI-2005, 2005.

[7] Xue Nianwen, Palmer M. Calibrating features for semantic role labeling[C]. In Proc. of EMNLP-2004, 2004.

[8] Gildea D, Jurafsky D. Automatic labeling of semantic roles[J]. Computational Linguistics, 2002, 28(3):245-288.

[9] Wang HongLing, Zhou GuoDong, Zhu QiaoMing and Qian PeiDe. Exploring various features in semantic role labeling. In Proceedings of ALPIT'2008. 23-25 July 2008.