基于依存树距离的语义角色识别方法*

王 鑫[1]，穗志方[1]，李 芸[2]
[1]北京大学 计算语言学研究所，北京 100871
[2]中国社会科学院 语言研究所，北京 100732
E-mail: wang-xin@pku.edu.cn; szf@pku.edu.cn; liyun@cass.org.cn

摘 要：在基于依存的语义角色标注研究中，大多数系统采用机器学习方法进行论元识别和分类。本文分析了依存树的特点，发现论元集中分布于依存树上的特定局部范围内，因此提出一种基于依存树距离的论元识别方法。该方法将候选论元限制在与目标动词的依存树距离不超过 3 的范围内，通过制订规则，提取目标动词的最佳候选论元集合。在 CoNLL2009 中文语料上采用正确的依存树，识别出了 98.5%的论元。在此基础上，结合基于机器学习的角色分类，系统 F 值达到 89.46%，比前人的方法 (81.68%)有了较为显著的提升。

关键词：论元识别；基于依存树距离的方法；语义角色标注

Semantic Role Identification Method Based on Dependency Tree Distance

Wang Xin[1], Sui Zhifang[1], Li Yun[2]
[1] Institute of Computational Linguistics, Peking University, Beijing, 100871
[2] Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732
E-mail: wang-xin@pku.edu.cn; szf@pku.edu.cn; liyun@cass.org.cn

Abstract: In research of semantic role labeling based on dependency, most systems apply machine learning to arguments identification and arguments classification. This paper analyses dependency tree characteristic, noticing that arguments distribute in specific area of dependency tree, so we propose a novel semantic role identification rule based on dependency tree distance. The maximal distance from candidate arguments to verb is limited to no more than three. We also obtain best candidate arguments related to verb. Making use of gold syntactic dependency tree, this method recognizes 98.5% of arguments on CoNLL 2009 Chinese dataset, combined with arguments classification based on machine learning, F measure reaches 89.46%, which achieves significant improvements over previous work (81.68%).

Keywords: argument identification; method based on dependency tree distance; semantic role labeling

1 引言

语义角色标注是浅层语义分析的一种重要手段，此项研究主要分为两个方面，基于短语结构的语义角色标注和基于依存的语义角色标注。论元识别和论元分类是标注过程中需要解决的主要问题，其可以通过两类方法得以实现：基于统计的机器学习方法和基于规则的方法。

在基于依存的语义角色标注研究中，现阶段主要的论元识别方法都是基于机器学习的。本文通过对依存树中论元节点的特征分析，发现大于 98%的论元节点到目标动词的依存树路径长度不超过 3，这说明论元集中分布于依存树上的一个局部范围内。充分利用这一特点，本文参考赵海等(2008)[1]的剪枝算法，提出一种基于依存树距离的论元识别方法，通过制订规则，提取依存树中由动词的儿子、父亲、兄弟、第一祖父以及父亲的兄弟节点构成的候选论元集。在此识别方法基础上，本文采用机器学习的方法进行论元分类，综合原句的特征以及由识别所得候选论元构成的骨干句的特征，为候选论元标注相应的角色。在 CoNLL2009 中文语料上，以正确的依存树为输入，系统的 F 值达到 89.46%，与前人的方法 81.68%(王步康等, 2010) [2]相比有很大改善。

*本文得到国家自然科学基金 60873156、61075067 以及国家社会科学基金 09BYY032 的支持。

2 相关研究

语义角色标注通常分为 4 个步骤，剪枝、识别、分类、后处理，而前三个步骤都是在完成广义分类任务，因为剪枝和识别本质都是区分候选对象是否会是论元。这种广义分类任务可以通过基于机器学习的方法和基于规则的方法来实现，不同系统的实现方法不同。

– 全过程不使用规则，完全使用基于机器学习的方法

Pradhan 等(2004)[3]基于短语结构句法树使用 SVM 分类器(Kudo and Matsumoto, 2000, 2001) [4, 5]进行论元识别和分类。Johansson 等(2008) [6]在语义依存分析任务中使用基于线性逻辑回归模型的 LIBLINEAR 分类器(Lin 等 2008[7])完成角色识别和分类。

– 剪枝阶段使用规则，后续阶段使用机器学习方法

Xue 等(2004)[8]基于短语结构树使用启发式规则完成剪枝，使用最大熵分类器进行角色识别和分类。王步康等(2010)[2]也提出一种剪枝算法，即在依存树中，保留与谓词具有一定关系的节点，如父亲，儿子，孙子等，其他节点都被过滤掉，之后再用机器学习方法进行角色识别和分类。

– 将剪枝和识别合为一步，并用基于规则的方法完成，只在分类阶段使用机器学习技术

丁金涛等(2008)[9]使用规则，在 CoNLL2005 共享任务的 WSJ 测试集上，基于自动句法分析识别出了 97.17%的论元，在此基础上角色标注系统的 F 值达到了 77.84%，在基于单一句法分析的角色标注系统中处于领先位置。

基于机器学习的方法和基于规则的方法各有特点，基于机器学习的方法优点是需要的人工干预少，对研究者语言学背景要求少，但此方法的缺点在于对训练语料的依赖性强，易出现数据稀疏问题；对训练语料中未出现的实例，分类效果较差；系统时间效率较低等问题。

基于规则的方法在某种程度与基于机器学习的方法有着互补的关系，此方法中研究者可以根据丰富的语言学知识对规则进行细化，利于处理分类中的细节问题，在一定程度上缓解了数据稀疏问题。此外，由于不必需要大规模语料库支持也不必进行模型训练，其在时间性能方面也表现出了较强优势。然而，由于规则需要人工制定，如果待区分的类别较多，并且某些待区分对象间相似度较高，就极大增加了制定规则的难度以及规则本身的复杂度，因此在一定意义上，相比于多分类问题，其处理二分类问题时优势更为显著。

因此，如果可以找到规则与机器学习运用范围的最佳组合，就可以将两者优势相结合，充分发挥规则和统计各自的特点，取得良好的标注效果。对语义角色标注任务来说，剪枝与识别本质是二分类问题，在这两个阶段运用规则方法既可以充分发挥规则在时间性能上的优势，又不会因为需要区分的类别过多而使规则过于复杂。而对于论元分类，由于论元类别较多，机器学习方法则更具优势。因此，本文将规则与机器学习相结合，构建出了一个性能良好的角色标注系统。

3 基于依存树距离的论元识别

3.1 依存树距离对语义角色的影响分析

在现代依存语法理论(又称从属关系语法，配价语法)中，周国光（1994）[10]对依存语法进行了定义：“依存语法是一种结构语法，主要研究以谓词为中心而构句时由深层语义结构映现为表层句法结构的状况及条件，谓词与体词之间的同现关系，并据此划分谓词的词类”。因此，基于依存理论所构建的依存树，在表达词语间依赖关系的同时，强调动词在句子中的重要作用。从这个角度讲，在围绕动词展开的角色标注任务中，依存树相比短语结构树而言，具有明显的优势。在某种意义上，依存树上的某些特征可以直接决定词语间语义上支配关系的远近。例如，词语与目标动词的距离特征直接决定着这个词语是否会与动词有语义上的依赖关系，即是否会成为谓词的论元，距离特征在依存树中的作用要大于其在短语结构树中的作用，主要原因有以下两方面：

1）依存树中节点数量比短语结构树少（张育等2010[11]），依存树中节点都是句子中的词语，而短语结构树中除了词语节点外，还有句法成分节点，因此词语之间的距离包含了这些句法成分，距离特征对于词语间关系远近的决定作用会因此受到影响。依存树则不会存在此类问题。

2）依存树偏重于一种关系结构，是语义层面的表示，节点间距离是他们语义关系远近的一种形式表现。短语结构树主要体现的是句子的句法层次结构，节点间距离基于句法关系，对语义的指示程度相对较低。

综合以上发现，本文提出了基于依存树距离规则的论元识别方法，充分利用依存树本身的特点进行语义角色标注。

3.2 基于依存树的剪枝方法

在基于依存的语义角色标注研究中，赵海等(2008)[1]提出一种剪枝规则：构建集合S,由依存树中目标动词到根节点上的节点组成（包括目标动词和根节点）。集合S中的元素以及依赖于集合中元素的节点就会被保留下来进入识别阶段。为了方便说明，本文称S中的节点为“主节点”。在赵海等(2008)[1]中，以上规则只覆盖剪枝过程，此后，系统还将依赖机器学习方法进行论元识别和分类。规则方法能否进一步放大范围来完成角色标注中的论元识别这一主要任务？

本文基线实验将赵海等(2008)[1]的剪枝算法直接用作论元识别的规则，结果表明，此方法的召回率较高R=99.3%，但是准确率很低(P=24.6%),这是因为保留了较多的非论元成分，保留的非论元数量是实际论元数量的3倍。因此，为提高论元识别的准确率，需要对此基线方法进行修改。

3.3 基于依存树距离的论元识别方法

在基线实验基础上，本文对经过识别阶段被标注为候选论元的词语特征以及语料中真正论元的特征进行了对比。表3-1统计了在训练集上使用基线识别方法所得的与目标动词不同距离的候选论元数目。表3-2统计了不同路径长度对应的真正论元数目，从中发现，真正的论元在与目标动词的距离特征上表现出了明显的聚集性：训练集的真实论元总计17547个，其中只有1个论元与目标动词的距离大于6，而当距离大于4时，论元的数目也急剧减少，这有力说明了依存树在表达句子语义方面的优势：依存树结构使句中核心词语间的距离变短，依存树上的论元分布的局部性更加明显。如图3-1所示，设目标动词是“鼓励”，真正的论元是“中国”、“企业家”和“投资”。在短语结构树中“鼓励”和三个论元的距离都是3，而且三个论元在树中分布的位置的局部性不明显。而在依存树中，目标动词与三个论元的距离都是1，而且在树状结构中三个论元都处于动词的下一层，表现出了极好的局部性特征。从表3-1和表3-2的比较中，我们受到启发，利用词语与目标动词的距离特征，将距离限定在一定的阈值之内，满足阈值条件下的词语才可以被选为候选论元进入分类阶段，进而有效地减少了非论元被识别为论元的数量，提高了识别阶段的准确率。

基于以上分析，本文提出了基于依存树距离的论元识别方法：提取从目标动词到根节点路径中与目标动词距离不大于L的节点构成集合S，集合S中的节点以及依赖于S中节点的节点构成候选论元。在此条件下，候选论元与目标动词的最长距离被限制为L+1。本文分别设置L=3、2、1进行实验，结果表明当L=2时，系统性能达到最优，此条件下，被识别为候选论元的节点包括动词的儿子、父亲、兄弟、第一祖先和父亲的兄弟。

表3-1 训练集中不同路径长度下对应的候选论元数目

与目标动词的距离D	1	2	3	4	5	6	7	8	9
论元数量	21840	11760	13479	8654	5463	3753	2021	1034	554
与目标动词的距离D	10	11	12	13	14	15	16	17	
论元数量	281	86	52	21	4	4	1	2	

表 3-2　训练集中不同路径长度下对应的真正的论元数目

与目标动词的距离 D	1	2	3	4	5	6	7
论元数量	14540	1892	869	178	41	26	1
距离不大于 D 的论元占总论元的比例	82.86%	93.64%	98.59%	99.61%	99.84%	99.99%	100%

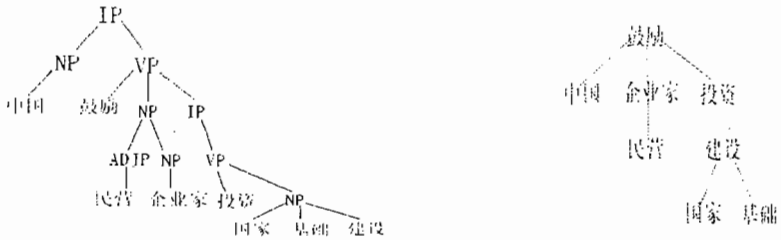

图 3-1　短语结构句法树与依存句法树的比较

4　基于机器学习的论元分类

在论元分类阶段，本文采用序列标注模型，以识别所得的候选论元为基本标注单元，选择了现阶段大多数角色标注系统所广泛使用的特征。表 4-1 列举了论元分类阶段的特征集合。由于论元

表 4-1　论元分类阶段的特征集

基于目标动词的特征	
目标动词及其词性	目标动词的首字和尾字
目标动词的子类框架	目标动词的类别
目标动词的字数	
基于当前词语的特征	
当前词语及其词性	当前词语相对于目标动词的位置
当前词语的首字和尾字	当前词语的字数
基于词语上下文的特征	
原句中当前词语前后一个词及其词性	原句中以当前词语为中心的大小为 1 的窗口内二元文法的词语特征、词性特征
原句中目标动词前后一个词及其词性	骨干句中当前词语前后两个词及其词性
原句中以目标动词为中心的大小为 1 的窗口内二元文法的词语特征、词性特征	骨干句中以当前词语为中心的大小为 2 的窗口内二元文法的词语特征、词性特征
骨干句中目标动词前后两个词及其词性	骨干句中以目标动词为中心的大小为 2 的窗口内二元文法的词语特征、词性特征
基于依存关系的特征	
中心词及其词性	依存关系
基于当前词语和目标动词之间的联系	
依存树中当前词语到目标动词的路径	路径长度
原句中当前词语到目标动词的直线距离长度	骨干句中当前词语到目标动词的直线距离长度
组合特征	
当前词语-位置-目标动词	当前词语词性-位置-目标动词
目标动词首字和尾字-位置-目标动词字数	当前词语-位置-目标动词类别

识别阶段删除了大量的非论元成分，被标注为候选论元的词语会构成一个新的句子（本文称之为"骨干句"）。对于候选论元来说，其在骨干句中的语境与其在原句中语境有很大不同，因此对于和语境相关的特征，如表 4-1 中基于词语上下文的特征以及基于当前词语与目标动词之间关系的特征，我们从原句以及识别后的"骨干句"中分别提取了相应的特征。

5 后处理

为了解决一个句子中出现多个相同核心论元的问题，本文提出了基于距离的后处理方法。从 3.3 的观察中可以得出结论，绝大多数论元被限制在以目标动词为中心的一定范围内，从某种意义上讲，与目标动词距离近的节点，有更高的概率成为论元。因此，如果多个候选论元被同时标注为核心角色 Ai，则可以首先比较这些节点在依存树上与目标动词的距离，距离近的候选论元优先获得此角色，其他候选论元则标注为空。如果基于依存树的路径长度相同，则可以比较候选论元与目标动词在原句中的直线距离，较近的一个被标注为核心论元。

6 数据与实验结果分析

本文选用 CoNLL 2009 Closed Challenge 提供的中文训练集语料进行模型训练，使用开发集进行系统测试。系统基于正确的依存树进行实验，在角色分类阶段，选用了随机梯度 CRF 软件包[1]，借助此工具本文较快获得了分类时的最优特征集，并取得了较好的角色标注结果。

6.1 基线识别方法

本文将赵海等(2008)[1]中的剪枝规则放大作用范围来完成论元识别任务，如表 6-1 所示，识别阶段召回率较高(R=99.3%)，但准确率很低(P=24.6%)。因此增强对候选论元的约束，减少被错误识别为候选论元的词语数是十分必要的。表 6-2 对比了基线识别方法基础上的角色标注与王步康等(2010)[2]的角色标注结果。两个实验采用了相同的数据集和系统输入，结果表明，本文基线角色标注结果在 F 值上相比王步康等(2010)[2]已经取得了大幅提高（7.3%）。

表 6-1 基线识别方法的识别结果

	真正的论元	真正的非论元
标注为候选论元	26466	80757
标注为非论元	185	222349

表 6-2 基线识别方法基础上的角色标注结果与前人工作的对比

	P(%)	R(%)	F(%)
王步康等	88.29	76	81.68
基线识别方法	90.17	87.67	88.91

6.2 基于依存树距离的论元识别方法

表 6-3 表示了基于依存树距离的识别方法中距离对于角色标注系统的影响，其中 L 采用了 3.3 中的定义，即集合 S 中的主节点与目标动词的距离不超过 L，结果表明，L=2 时系统性能达到最优，这说明利用依存树上节点与目标动词的距离特征来对主节点进行约束，进而限制候选论元到目标动词的距离对于取得良好的角色标注性能有着重要意义。表 6-4 表示了 L=2 条件下识别阶段的结果，召回率为 98.3%，相比基线实验，进入分类阶段的候选论元数减少了 38345（占基线条件下候选论元总数的 35.76%），有力证明了依存树距离特征对于筛选候选论元的积极意义。

6.3 后处理

表 6-5 列出了测试集上同一语义角色在一个句子中出现多次的数量分布情况。从中可以发现，

[1] http://leon.bottou.org/projects/sgd

后处理之前核心论元的重复出现次数总计 332，经过后处理，消除了核心论元重复出现的情况。表 6-6 是采用基于依存树距离的识别方法并设置 L=2 时，后处理前后系统的性能对比，F 值提高了 0.1%，证明了后处理方法的有效性。

表 6-3　基于依存树距离的识别方法中距离 L 对于角色标注系统的影响

	P(%)	R(%)	F(%)
基线实验	90.17	87.67	88.91
L=3	90.43	87.92	89.16
L=2	90.47	88.27	89.36
L=1	92.49	84.12	88.11

表 6-4　基于依存树距离的识别方法中 L=2 条件下的识别结果

	真正的论元	真正的非论元
标注为候选论元	26198	42680
标注为非论元	453	260426

表 6-5　重复出现的语义角色数量统计

角色	A0	A1	A2	ADV	C-A0	C-A1	CND	DIR	DIS
标准语料中重复出现的数目	53	5	0	676	2	1	1	1	77
后处理前	201	112	19	697	0	0	1	0	62
后处理后	0	0	0	697	0	0	1	0	62

表 6-6　后处理前后的语义角色标注性能比较

	P(%)	R(%)	F(%)
后处理前	90.47	88.27	89.36
后处理后	91.17	87.81	89.46

7　总结

本文提出了一种基于依存树距离的论元识别方法，由于依存树结构有利于缩短论元与目标动词的距离，使论元分布的局部性更显著，本文充分利用此种局部性特征，制订规则将距离特征作为判定候选论元的重要条件，实现了基于规则的论元识别。结合基于机器学习的论元分类，基于正确的依存句法分析结果，本文角色标注系统 F 值达到 89.46%，相比前人工作取得了较大改进。

参 考 文 献

[1] Hai Zhao, Chunyu Kit. Parsing syntactic and semantic dependencies with two single-stage maximum entropy models[C]. // Proceedings of the 12th CoNLL-2008, Manchester, August 2008: 203-207.

[2] 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[J]. 中文信息学报, 2010, 24(1): 25-29, 47.

[3] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, et al. Shallow Semantic Parsing Using Support Vector Machines[C]// Proceedings of NAACL-HLT 04. 2004.

[4] Taku Kudo and Yuji Matsumoto. Use of support vector learning for chunk identification [C]. // Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000: 142-144.

[5] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines[C]. // In Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001).

[6] Richard Johansson, Pierre Nugues. Dependency-based syntactic-semantic analysis with PropBank and NomBank[C]. //Proceedings of the 12th CoNLL-2008, Manchester, August 2008: 183-187.

[7] Chih-Jen Lin, Ruby C.Weng, S. Sathiya Keerthi. Trust region Newton method for large-scale logistic regression[C]. //Proceedings of the 24 th International Conference on Machine Learning, Corvallis, OR, 2007.

[8] Nianwen Xue, Palmer M. Calibrating features for semantic role labeling[C]//Proceedings of EMNLP, Barcelona, Spain, 2004: 88-94.

[9] 丁金涛, 周国栋, 王红玲, 朱巧明. 语义角色标注中有效的识别论元算法研究[J]. 计算机工程与应用, 2008, 44(18), 153-156.

[10] 周国光. 汉语配价语法论略 [J]. 南京师范大学学报: 社科版, 1994(4): 103-106, 121.

[11] 张育, 王红玲, 周国栋. 基于两种句法分析的语义角色标注比较研究 [J]. 计算机应用与软件, 2010, 27(8): 565-573.