第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·软件技术与数据库· 文章编号：1000—3428(2009)15—0082—03 文献标识码：A 中图分类号：TP311

基于依存关系的语义角色标注

汪红林，王红玲，周国栋
(江苏省计算机信息处理技术重点实验室苏州大学计算机科学与技术学院，苏州 215006)

摘 要：针对以句法成分为基本标注单元语义角色标注的瓶颈问题，描述一个以依存关系为标注单元的语义角色标注系统，经过依存关系分析、谓词标识、特征抽取、角色识别和角色分类，最终在CoNLL2008 SRL Shared Task自动依存分析的WSJ测试集取得了较好的结果，*F*1值达到了80.94%，结果证明其性能明显好于基于句法分析的SRL。

关键词：语义角色标注；依存分析；依存关系

Semantic Role Labeling Based on Dependency Relationship

WANG Hong-lin, WANG Hong-ling, ZHOU Guo-dong
(Jiangsu Provincial Key Lab for Computer Information Processing Technology, School of Computer Science & Technology, Soochow University, Suzhou 215006)

【Abstract】Aiming at the bottlenecks of syntactic tree-based Semantic Role Labeling(SRL), this paper explores dependency relationship-based semantic role labeling. By properly integrating dependency parsing, predicate identification, feature extraction, semantic role identification and semantic role classification, this system achieves the *F*1 measure of 80.94% on the WSJ portion of the CoNLL2008 SRL Shared Task, using automatic dependency parsing. Experimental results show that it is better than the *F*1 measure of syntactic tree-based semantic role labeling apparently.

【Key words】Semantic Role Labeling(SRL); dependency parsing; dependency relationship

1 概述

近年来，随着基于语义研究的广泛和深入，语义角色标注(Semantic Role Labeling, SRL)越来越受关注，已经成为自然语言处理的重要组成部分。给定一个句子，其任务是识别并标注句中每个目标谓词的所有充当语义角色的句法成分。进行语义角色标注的基础技术，如词性标注、句法分析、统计学习方法等目前已经比较成熟。同时语义角色标注在问答系统、信息抽取、机器翻译等领域有着广泛的应用。

基于特征向量的方法主要集中于特征工程和机器学习模型的研究。文献[1]最早进行基于句法分析的语义角色标注工作，提出了角色分类的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)，在基于手工句法分析的PropBank测试语料上所得的*F*1值为87%。文献[2]加入了内容词、命名实体、中心词词性和内容词词性等新特征并使用感应决策树学习进行实验，性能稍有提高。文献[3]使用最大熵分类器进行角色分类，还对加入的新特征(句法框架、谓词与当前句法间的距离)和组合特征(谓词+句法类型、谓词+中心词、语态+位置)进行了研究，在基于手工句法分析的PropBank测试语料上所得的*F*1值为88.51%。因为大多数的句法成分是不充当角色的，所以还提出了一个很有用的剪枝算法来过滤掉这些成分。

文献[4]使用SVM分类器进行角色分类，分成2步操作，首先角色识别，删除高概率为空角色的句法成分，然后将保留下来的句法成分作为分类阶段的输入，进行角色分类，并表明角色的识别和分类分开操作效果更好。在PropBank上，基于Chariniak自动句法分析的*F*1值为79%。文献[5]使用了最大熵分类器，在单一自动句法分析上报告了取得的最好的结果，*F*1值为77.13%。

和基于句法分析的SRL相比，基于依存分析的SRL系统相对较少。文献[6]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，提出了一种比较有效的剪枝算法，采用SVM分类器实现了角色的分类，给出了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*F*1值分别为84.6%, 79.8%。虽然使用的信息比基于句法分析的SRL少，但也取得了相似结果。国内做过这方面研究工作的人很少。目前，依存分析的方法有很多种，典型的是：使用CCG模型产生依存关系；采用机器学习的方法实现依存分析，并且能够移植到其他领域或语言，使SRL系统可移植性也更大。

2 系统描述

CoNLL2008提供的语料包含有名词性谓词和动词性谓词。本文对两者分开标注，前期试验表明，若放在一起标注，性能会比较差，究其原因是名词性谓词和动词性谓词性质不一样，充当它们角色的依存关系也有很大不同。例如：在依存树中，动词性谓词的角色除了其祖先节点外还有兄弟、孩子等，而名词性谓词的角色大多数集中在其祖先节点上。

基金项目：国家“863”计划基金资助项目(2006AA01Z147)；国家自然科学基金资助项目(60673041)；高等学校博士学科点专项科研基金资助项目(20060285008)

作者简介：汪红林(1985－)，男，硕士研究生，主研方向：自然语言处理；王红玲，博士研究生；周国栋，教授、博士生导师

收稿日期：2009-04-10 **E-mail：**064227065055@suda.edu.cn

系统主要分成 3 个部分：依存关系的构造，谓词的标识，角色标注(识别和分类)。整个流程如图 1 所示。

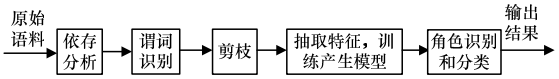

图 1　系统结构流程

2.1　依存分析实例

在语义角色标注中，谓词既可以是动词也可以是名词。现给定一个实例(1)，其中有 2 个谓词，分别为名词 evidence 和动词 remain，现只标注出以名词 evidence 为谓词的角色标注。

Meanwhile, [AM-MNR overall][A2 evidence [A1 on the economy]] remains fairly clouded. (1)

图 2 和图 3 分别说明了实例(1)的依存关系和其对应的依存树结构。

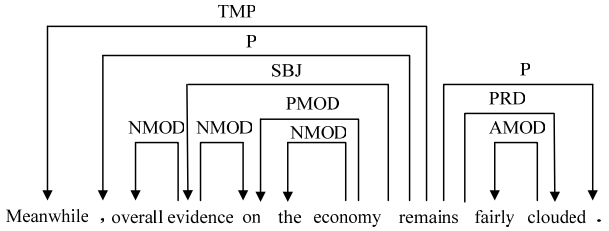

图 2　实例(1)对应的依存关系

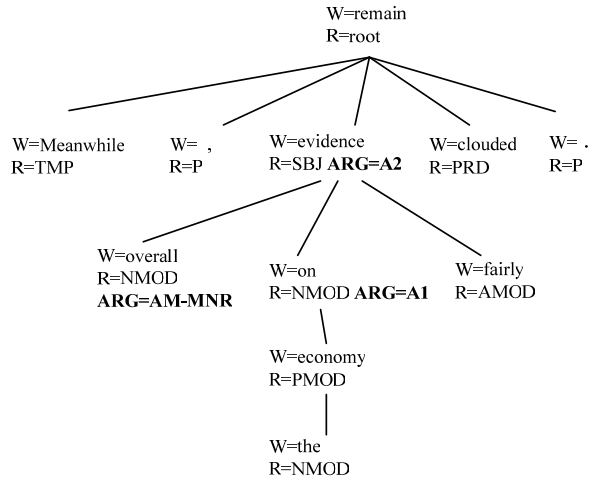

图 3　实例(1)对应的依存树

在图 3 中，用黑体字表示各依存关系承担的角色，W 表示单词，R 表示依存关系。

2.2　谓词标识和依存关系构造

CoNLL2008 shared task 测试语料不包含谓词及词义，需自行标识，对于动词性谓词及其词义的标识，参照 CoNLL2005 shared task 语料(谓词及词义已发布)，将相应句子中 05 语料里出现的谓词及其词义在 08 语料里全部标识出来，准确率达到 98.67%。对于名词性谓词标识，根据词表，判断每个名词是否有充当动词的可能性，若有，则进行标识，反之放弃标识。准确率达到 92.82%，并把所有的词义都设为“01”，由于多数词义是“01”，因此这种词义标识方法对结果影响不大。

目前有很多 parser 工具能够产生依存关系，CoNLL2008 shared task 测试集仅提供了 MaltParser 的结果，除此之外，本文还利用 MSTParser 生成了依存关系，2 个工具对比如表 1 所示。

表 1　MaltParser 和 MSTParser 性能比较　(%)

句法分析器	LAS	UAS	LA
MaltParser	85.50	88.41	90.41
MSTParser	87.01	89.72	91.75

由表 1 知，MSTParser 性能更好。

2.3　剪枝和预处理

本文对依存关系节点树进行了剪枝，在依存树上，只保留与谓词有一定关系的节点进行特征的抽取，采用的剪枝算法有 3 种：(1)Hacioglu 算法。保留谓词的父亲、孩子、孙子、兄弟、兄弟的孩子、兄弟的孙子节点。(2)新 Hacioglu 算法。在 Hacioglu 算法基础上，加上谓词祖父、祖父的孩子、祖父的父亲节点。(3)Xue 算法。从当前谓词一直走向根节点，保留经过的所有节点及它们的孩子节点。

在二元识别和分类后，保留了所有非空角色实例和小于某阈值的空角色实例，相关情况如表 2 所示。

表 2　对谓词处理的相关情况的说明

谓词类别	剪枝算法	阈值	误剪率/(%)	
			训练集	测试集(Gold)
动词性	Xue+Hacioglu	0.9	0.7	0.7
名词性	新 Hacioglu	0.95	4.9	43.5

可以看出，Xue 的剪枝算法是很有效的，但是却不能适用于名词性谓词，因为产生的特征实例太大，无法训练生成模型文件。名词性谓词更多的角色是分布在其祖先节点上，而新 Hacioglu 只保留了长度不超过 3 层的祖先节点，所以误剪率比较高。

在预处理阶段，在生成的特征文件中，用空角色代替了出现次数少于 200 的角色，如：A5, AM-PRD, C-AM-ADV, R-A2 等，因为这些角色会对模型产生误导作用。

2.4　所选的特征

本文抽取语义角色标注中常用的 8 个特征作为基本特征，参考 Hacioglu 等的特征，然后加入了其他的增强特征，如表 3 所示。用图 2 中的依存树作为实例，设当前关系节点是 on，特征值为空的用“_”表示。

表 3　所选特征及特征说明

	特征名称	特征说明
基本特征	谓词原型	当前谓词的词性(evidence)
	谓词词性	当前谓词的词性(NN)
	谓词语态	谓词的主动语态或者被动语态，有主动或者被动 2 种(_)
	子类框架	当前谓词节点的所有孩子节点的依存关系链(SBJ-NMOD-NMOD)
	路径	从当前节点走向当前谓词，将途经的节点的依存关系用“->”链接起来(NMOD->SBJ)
	位置	当前节点的中心词相对于当前谓词的前后顺序值分别是“before”或者“after”或者“equal”(equal)
	依存关系	当前节点所对应的依存关系(NMOD)
	中心词	当前节点的父亲节点所对应的单词本身(evidence)
增强特征	谓词的孩子的依存关系链	当前谓词的所有孩子节点的依存关系组成的链(NMOD-NMOD)
	谓词的兄弟的词性链	当前谓词的所有兄弟节点的词性组成的链(RB-,-JJ-.)
	谓词的兄弟的依存关系链	当前谓词的所有兄弟的依存关系组成的链(TMP-P-PRD-P)
	谓词依存关系	当前谓词节点的依存关系(NMOD)
	谓词的孩子的词性链	当前谓词的所有孩子节点的词性组成的链(NN)
	家族成员	剪枝后剩下的节点，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，当前关系节点与当前谓词的家族关系，如 father, child, siblings 等(child)
	依赖词	指当前节点本身单词(on)
	中心词词性	中心词的词性(NN)
	依赖词的词性	当前节点单词的词性(IN)
	谓词+中心词	当前谓词原型+中心词(evidence+ evidence)
	中心词+当前关系	中心词+当前节点依存关系(evidence+ NMOD)

2.5 分类器

本文选用的是最大熵分类器。其基本思想是为所有已知的因素建立模型，而把所有未知的因素排除在外。也就是说，要找到这样一个概率分布，它满足所有已知的事实，且不受任何未知因素的影响。

最大熵模型的一个最显著的特点是其不要求具有条件独立的特征。因此，可以相对任意地加入对最终分类有用的特征或者删除有误导倾向的特征，而不用顾及它们之间的相互影响。另外，最大熵模型能够较为容易地对多类分类问题进行建模，并且给各个类别输出一个相对客观的概率值结果，可以根据这些概率值来进行相关的特征实例处理。最后，最大熵分类器的训练速度快也是一个很大的优点。

在实验中，采用的最大熵原型是 maxent-2.4.0，在此基础上进行了相关的修改，使输出符合系统的要求，并且把参数 cutoff 和 interation 分别设为 2 和 100。

3 实验结果与分析

本文使用的是 CoNLL2008 shared task 提供的 WSJ 语料，数据来源于 PropBank 和 Nombank，其中训练集有 39 280 句，开发集有 1 335 句，训练集有 2 400 句，并采用其提供的 eval08.pl 脚本进行评测，评测指标为准确率、召回率和 *F*1 值，在基于自动依存分析的基础系统平台上达到的 *P*, *R*, *F*1 值为 84.31%, 78.64%, 81.38%。

3.1 特征表现

在手工依存分析上利用基本特征首先建立一个基准系统，然后依次加入其他特征，性能变化如表 4 所示。

表 4 基准系统上分别加入单个特征后的性能 (%)

加入的特征	准确率	召回率	*F*1 值
Gold Baseline	84.31	78.64	81.38
+ 家族成员	84.70	78.87	81.68
+ 依赖词	86.74	83.01	84.84
+ 中心词词性	84.44	78.55	81.37
+ 依赖词的词性	84.42	78.33	81.47
+ 谓词的孩子词性链	84.35	78.73	81.47
+ 谓词的孩子依存关系链	84.75	78.97	81.76
+ 谓词的兄弟依存关系链	84.29	78.52	81.30
+ 谓词的兄弟词性链	83.75	78.32	80.95
+ 谓词本身	84.03	78.83	81.34
+ 谓词原型+中心词	83.30	78.94	81.30
+ 依存关系+中心词	84.66	79.37	81.93

可以看出，加入依赖词特征后，性能提高了 3.5%，效果最明显。因为依存关系是指依赖词和中心词之间的关系，所以依赖词对依存分析的影响必然很大，最终会对性能影响很大。

而加入谓词的兄弟词性链和依存关系链特征后，性能略有下降，因为充当谓词角色的兄弟节点不多，所以此特征效果不大，反而会带来噪音。加入谓词原型+中心词后，性能也略有下降，因为谓词和中心词两者之间几乎没有任何关系，是相互独立的，所以会起到反作用。

3.2 最佳系统

添加了上述的所有特征以后，在手工依存分析和自动依存分析上效果得到明显加强，结果如表 5 所示。

表 5 在 CoNLL2008wsj 测试语料上的结果 (%)

使用的句法分析器	准确率	召回率	*F*1 值
标准句法 Gold	87.24	83.86	85.52
MaltParser	76.17	72.02	74.04
MSTParser	82.39	79.54	80.94

在本实验中，MSTParser 的结果比 MaltParser 的结果好，是因为 MSTParser 预测的依存关系准确率更高。由表 1 可以看出，其准确率高了近 1.5%，对后期的角色标注必然会产生很大影响。参加 CoNLL2008 shared task 开放测试的还有 Vickrey(Stanford), Meza-Ruiz(Edinburgh), Zhang(DFKI 2), Li (HIT-ICR)，他们的 *F*1 值分别为 77.38, 75.72, 73.08, 70.32。其中，前两者使用的是 Gold 的依存关系，结果相对较好，后两者使用的依存分析器精确率分别为 88.14%和 87.42%，仍然比笔者所使用的依存分析器的效果好。即使如此，由表 5 可以看出，本文的 SRL 结果仍然是最好的。另外，和基于依存分析的 SRL 结果(在 CoNLL2004 shared task 测试集上 *F*1 为 79.8%)相比，结果稍好，因为笔者在其基础上采用更多的特征，剪枝算法也更加精确。与基于句法分析的结果相比，结果依然稍好，因为依存分析更能反映出句子中单词和短语之间的修饰关系，对句子的结构分析得更加清晰。

4 结束语

本文描述了一个基于依存分析的语义角色标注系统。这是一种全新的方法，和以往的语义角色标注方法大不相同，使用的语料库是最新的。首先对语料进行谓词及词义的识别，然后对不可能承担角色的依存关系进行过滤和简单预处理，如用空角色代替出现次数少的角色，并且使用了一些有用的特征及其简单组合，最后使用最大熵分类器来对角色进行识别和分类，取得了很好的结果。但是未来的改进空间还是很大，拟从以下几方面继续开展工作：(1)剪枝算法优化，尤其是对名词性谓词的剪枝算法作调整，降低误剪率。(2)有效特征以及特征组合的选取，如当前谓词单词本身、当前节点和当前谓词的最近共同祖先节点等，并且可以尝试不同特征之间的组合来提高性能。(3)尝试使用 SVM 分类器取代最大熵分类器，文献[4]表明 SVM 分类器能取得很好的性能，并且微调 SVM 分类器的参数，使性能达到最佳。

参考文献

[1] Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles[J]. Computational Linguistics, 2002, 28(3): 245-288.

[2] Surdeanu M, Harabagiu S, Williams J, et al. Using Predicate-argument Structures for Information Extraction[C]//Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. Tokyo, Japan: [s. n.], 2003.

[3] Xue Nianwen, Palmer M. Calibrating Features for Semantic Role Labeling[C]//Proc. of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: [s. n.], 2004.

[4] Pradhan S, Ward W, Hacioglu K, et al. Shallow Semantic Parsing Using Support Vector Machines[C]//Proc. of NAACL-HLT'04. Boston, Mass, USA: [s. n.], 2004.

[5] 刘 挺, 车万翔, 李 生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

[6] Hacioglu K. Semantic Role Labeling Using Dependency Trees[C]// Proc. of CoNLL-2004. Boston, Mass, USA: [s. n.], 2004.

编辑 顾逸斐