基于依存关系的语义角色标注

汪红林 丁金涛 王红玲 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: 064227065055@suda.edu.cn

摘 要：本文利用依存关系进行语义角色的标注，在CONLL2008提供的shared task语料库上进行训练和测试。经过相关剪枝算法处理以后，使用最大熵分类器进行学习和分类，在手工句法分析基础上取得的F1值为:84.42%(Labeled)和92.58%（Unlabeled），在基于MaltParser的自动句法分析上取得的F1值为：81.15 %（Labeled）和88.73 %（Unlabeled），在基于MSTParser的自动句法分析上取得的F1值为：80.81%（Labeled）和88.47%（Unlabeled）。

关键字：语义角色标注，依存关系，最大熵分类器

Dependency Tree-based Semantic Role Labeling

Wang Honglin Ding Jintao Wang Hongling Zhou Guodong

(School of Computer Science & Technology, Soochow University, Suzhou, China, 215006)

(Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou, China, 215006)

E-mail: 064227065055@suda.edu.cn

Abstract: This paper explores dependency tree-based semantic role labeling on CONLL2008 shared task. We adopt maximum entropy classifier to learn and classify after proper pruning. Evaluation on the gold-standard dependency trees shows that our system achieves 84.42% (Labeled) and 92.58%（Unlabeled）in F-measure. It also shows that our system achieves 80.81%(Labeled) and 88.47%(Unlabeled) in F-measure using automatic Maltparer, which has the performance of 81.15 %（Labeled） and 88.73 %（Unlabeled） in F-measure.

Key words: Semantic Role Labeling, Dependency Tree, Maximum Entropy Classifier

1 引言

近年来，基于语义的研究越来越广泛深入。对自然语言进行浅层语义分析[10]逐渐兴起，它成为自然语言处理应用的重要组成之一。作为其具体实现，当前的语义角色标注是一项定义完整的任务，它有着充实的工作内容和可比较的评测。

语义角色标注的目的是为了表示出句子中一系列单词组成的短语结构或者是某个单词所充当的语义角色。角色有很多种，不同的语料库角色的表示符号不尽相同。目前支持语义角色标注的英文语料库有 FrameNet[8]、PropBank[10]和 NomBank[11]等。在 PropBank 语料库中，核心角色有 A0~A5，其他辅助角色有 ARGM-LOC 、ARGM-TMP 、ARGM-MNR 等等。每种不同的角色都表示不同的语义，如 A0 表示实施者，A1 表示受事者等等。

基金资助："863"国家高技术研究发展计划(the "863" National High-Tech Research and Development of China under Grant No.2006AA01Z147)；国家自然科学基金(the National Natural Science Foundation of China under Grant No.60673041)；高等学校博士学科点专项科研基金(the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20060285008)。.

作者简介：汪红林（1985-），男，硕士研究生，主要研究方向：自然语言处理；丁金涛（1979-），男，硕士研究生，主要研究方向：自然语言处理；王红玲（1975-），女，博士研究生，主要研究方向：中文信息处理；周国栋（1967-），男，博士生导师，主要研究方向：自然语言处理、信息抽取、网络挖掘等

大多数语言都是以谓词为中心来进行角色标注。标注方法建立在以下三种标注单元基础上：句法成分、短语和单词。最早进行语义角色标注相关工作的是 Gildea and Jurafsky, 2002[10]。在其基础上随后有很多人尝试使用不同的特征、机器学习算法等方法加以改进，比较典型的有：Hacioglu 等, 2004[1]; Pradhan 等, 2003[2],;Xue 等 2005[4]; Surdeanu 等, 2005[5]; Hacioglu 等, 2003[6]; Yih S, Toutanova K 等，2006[7]; 刘挺 等,2007[12]。

本文的方法是基于依存关系（R-by-R）进行语义角色标注，以谓词为中心，通过挖掘句子中单词或者短语之间所存在的关系，如：SBJ、NMOD、NN、PMOD 等等，采用基于特征向量的方法来训练和预测，最终得出句子中各依存关系的角色。较早使用此方法的是 Hacioglu 等, 2004[1]，在其构造的语料库 Depbank（由 Penn TreeBank 和 PropBank 转化而来）上的准确率（precision）为 85.6%，在 CONLL2004 shared task 语料库上的准确率为 84.9%。

本文余下的内容如下安排：第 2 部分主要介绍了依存关系的基本概念，并给出实例说明；第 3 部分对系统做了系统描述，介绍了系统的处理过程；第 4 部分是对实验结果和分析；第 5 部分进行了总结，并对未来的工作做了简述。

2 依存关系

依存文法分析主要是通过分析句子内词语之间的依存关系，以此来揭示其句法结构。Robinson 提出了依存语法中关于依存关系的 4 条公理：

(1) 一个句子只有一个成分是独立的；

(2) 其它成分直接依存于某一成分；

(3) 任何一个成分都不能依存两个或以上的成分；

(4) 如果A 成分直接依存于B，而C 成分处于A，B 之间，那么C 或者直接依存于A，或者直接依存于B，或者直接依存于A 和B 之间的某一部分。

除此之外，一般依存语法还遵循一条公理：中心成分左右两边的成分不发生依存关系。

在语义角色标注中，以动词性谓词为中心进行标注的角色是不可以嵌套的，而以名词性谓词为中心进行标注的角色是可以嵌套的。给定一个实例(1)，其中有两个谓词，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 1 和图 2 中分别说明了(1)的依存关系和其对应的依存树结构。

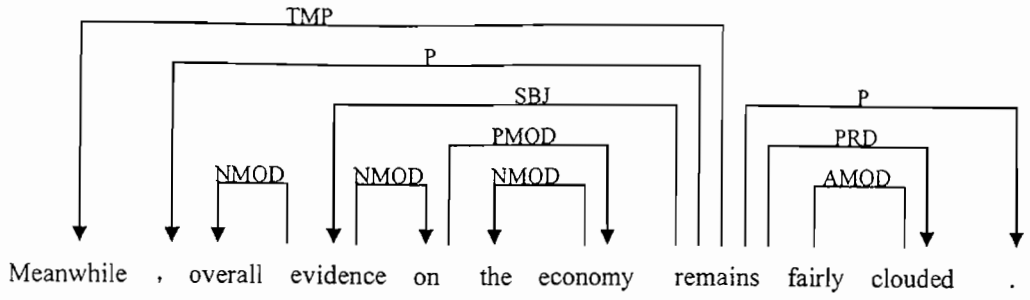

图 1 实例(1)对应的依存关系

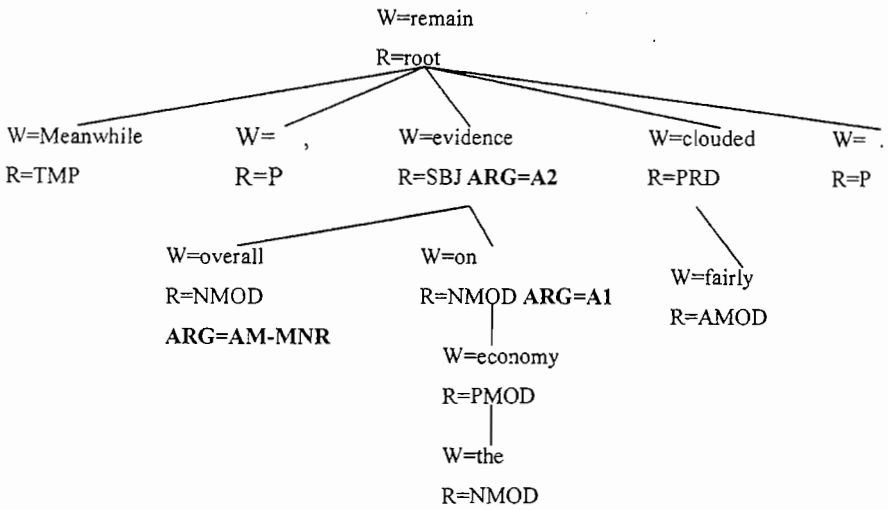

图 2 实例（1）对应的依存树

图 2 中，用黑体字表示各依存关系承担的角色。

3 系统描述

本系统是建立在类似于图 2 的依存树基础上，首先对语料库进行信息的抽取，使每个句子都能转化为一棵依存树，然后通过剪枝算法过滤掉一部分认为不可能承担角色的结点，接着对树中剩下的每个结点抽取特征，最后使用最大熵分类器进行角色的分类。

3.1 剪枝和处理

通常的语义角色标注分为 4 个步骤：剪枝（Pruning）、识别（Identification）、分类（Classification）和后处理（Post-processing）。本文首先对依存关系树进行了剪枝，只保留与谓词有一定关系的结点进行特征的抽取，采用的剪枝算法是在 Kadri Hacioglu[1]基础上进行扩充的，增加了更多的结点。譬如：谓词的祖父结点，谓词的祖父的孩子结点等等。凡不包含在此的结点都舍弃。此剪枝算法对已知角色的误剪率分别为：训练集 4.71%，开发集 5.24%。

另外，在对抽取出来的特征文件也进行了处理，删除了那些高概率空角色的特征实例，因为

大量的空角色的实例会对结果产生不利影响，实验中阈值设为 0.9999，即概率值大于 0.9999 的空角色实例不参与训练和测试，认为它们的角色为空。经统计，训练集删除了 118872 个实例，开发集删除了 4196 个实例。

3.2 所选的特征

对每个句子都构造依存树，然后在依存树上抽取相关特征，首先抽取的仍然是语义角色标注中常用的 7 个基本特征，然后加入了 9 个其他的特征，见表 1。用图 2 中的依存树作为实例，设当前关系结点是 on，特征值为空的用“_”表示。

表 1 所选特征及特征说明

特征名称	特征说明
与谓词本身有关的特征有：	
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，有主动或者被动两种(_)
谓词依存关系	当前谓词结点的依存关系 (NMOD)
子类框架	当前谓词结点的所有孩子结点的依存关系链 (SBJ-NMOD-NMOD)
谓词的孩子的词性链	当前谓词的所有孩子结点的词性组成的链 (NN)
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 (NMOD-NMOD)
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 (RB-,-JJ-.)
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(TMP-P-PRD-P)
与关系结点有关的特征：	
路径	从当前结点走向当前谓词，将途经的结点的依存关系用“->“链接起来。(NMOD->SBJ)
中心词	当前结点的父亲结点所对应的单词本身 (evidence)
位置	当前结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	当前结点所对应的依存关系 (NMOD)
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 （child）
依赖词	指当前结点本身单词 (on)
中心词词性	中心词的词性 (NN)
依赖词的词性	当前结点单词的词性 (IN)

3.3 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, 人们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。

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。

4 试验结果

本文语料库使用 CoNLL2008 shared task[▾]提供的语料，其中谓词不但有动词而且还有名词。训练集有 39280 句，开发集有 1335 句。抽取全部角色（包含空和非空角色）和非空角色（只含非空角色）的特征进行训练预测，所产生的实例个数见表 2：

表 2 分别对名词性谓词和动词性谓词抽取特征的实例个数

谓词词性	训练集		测试集	
	全部角色	非空角色	全部角色	非空角色
动词	1213892	227735	43397	7968
名词	1189882	173456	41663	6057

最后使用 CoNLL2008 提供的评测程序 eval08.pl 来评价系统的性能，结果见表 3：

表 3 在 CoNLL2008 开发集语料上的结果

		Precision	recall	F1
Gold	Labeled	86.49 %	82.45 %	84.42 %
	Unlabeled	94.85 %	90.41 %	92.58 %
MaltParser	Labeled	83.78 %	78.69 %	81.15 %
	Unlabeled	91.60 %	86.04 %	88.73 %
MSTParser[•]	Labeled	83.32 %	78.46 %	80.81 %
	Unlabeled	91.21 %	85.89 %	88.47 %

本试验中：

- MaltParser 分析器产生的中心词和依存关系已经在 CoNLL2008 shared task 公布，统计表明准确率为 84.1%，而 MSTParser 分析器产生的的中心词和依存关系是由工具产生的，准确率 83.5%，此处值得注意的是，由于 MSTParser 分析器训练时间太长，所以只用了四分之一的训练集进行训练（大约 1 万句），若用全部的训练集进行训练的话准确率会更高。
- 在后两者自动句法分析中，会产生明显的错误结果，如:生成的依存树会包含多个根结点，在训练集中有 725 句，开发集中有 21 句，这就会导致有的特征（如：路径）无法抽取，用

[*] http://maxent.sourceforge.net/
[▾] http://www.yr-bcn.es/conll2008/
[•] http://sourceforge.net/projects/mstparser

符号“-”替代。

- 与 Kadri Hacioglu[1]（CoNLL2004 开发集上的 P/R/F1：84.9%/75.2%/79.8%）相比，性能提高比较明显，若在同一语料上只以动词性谓词为中心进行角色标注，与其他以句法成份为基本单位的系统（如：丁金涛等[13]）相比，结果差别不是很大。

5 结论及展望

本文是基于依存关系的语义角色标注，是一种全新的方法，和以前的语义角色标注方法大不相同，使用的语料库是最新的。未来改进空间很大，拟从以下几方面继续开展工作：

将句法分析和语义角色标注结合起来，因为自动句法分析中，句法分析的正确与否对依存关系的产生有很大的影响，从而对语义角色标注的最终性能影响较大。

有效特征的选取，本实验所抽取的特征较少，还可以抽取更多的相关特征来进行训练和测试，譬如：当前谓词单词本身，当前结点和当前谓词的最近共同祖先结点，当前结点到最近共同祖先的路径，当前谓词到最近共同祖先的路径等等，并且可以尝试不同特征之间的组合来提高性能。

尝试使用 SVM 分类器取代最大熵分类器， PradhanS, 2003[2]等表明 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] Kadri Hacioglu. Semantic Role Labeling Using Dependency Trees . In Proc. of CoNLL-2004, 2004.

[2] Hacioglu K., Pradhan S., Ward W., et al. Shallow semantic parsing using Support Vector Machines[R]. TR-CSLR-2003-1, 2003a.

[3] Zhou Ming. A Block-based Robust Dependency Parser for Unrestricted Chinese Text[C]. 微软中国研究院2000 年论文集. 北京：微软中国研究院, 2000.

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

[5] Surdeanu M., Turmo J.. Semantic role labeling using complete syntactic analysis[C]. In Proc. of CoNLL-2005, 2005, 221-224.

[6] Hacioglu K., Pradhan S., Ward W., et al. Shallow semantic parsing using Support Vector Machines[R]. TR-CSLR-2003-1, 2003a.

[7] Yih S, Toutanova K. Automatic semantic role labeling. Microsoft Research[R], 2006

[8] Baker C.F., Fillmore C.J., Lowe J.B.. The Berkeley FrameNet project[C]. In Proc. of COLING-ACL-1998, 1998, 86-90.

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

[10] Palmer M., Gildea D., Kingsbury P.. The Proposition Bank: An annotated corpus of semantic roles[J]. Computational Linguistics, 2005, 31(1): 71-106.

[11] F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In Proc. of COLING-ACL.

[12] 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.

[13] 丁金涛等. 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