第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)04—0161—03 文献标识码：A 中图分类号：TP18

基于依存树的中文语义角色标注

安强强，张 蕾
(西北大学信息科学与技术学院，西安 710127)

摘 要：现有中文语义角色标注主要集中在基于短语结构句法树的标注。基于此，提出一种基于依存树的中文语义角色标注方法。将中文句子转化为标准的依存树，作为实验数据集，特征选取时结合知网，将语义信息引入特征集，以提高系统的召回率，并采用最大熵分类器进行实验，获得 90.68%的 *F* 值。结果表明，在标准的句法树上，当基于依存关系的标注系统中加入新特征时，该中文语义角色标注取得了比基于句法成分标注更好的成绩。

关键词：最大熵分类器；语义角色标注；依存树

Chinese Semantic Role Labeling Based on Dependency Trees

AN Qiang-qiang, ZHANG Lei
(College of Information Science & Technology, Northwest University, Xi'an 710127)

【Abstract】 Current Chinese semantic role labeling mainly focuses on using phrase structure trees. This paper presents an approach of Chinese semantic role labeling method which is based on dependency trees. Chinese sentences are converted into gold dependency trees which are divided into training and testing set. By using maximum entropy classifier and adding the first sememe of word concept to the feature set, the system gets an *F*-score of 90.68%. Results show that dependency-based system adding new features performs better than constituent-based system on gold standard parses.

【Key words】 maximum entropy classifier; semantic role labeling; dependency trees

1 概述

语义分析就是根据句子的句法结构和句中每个实词的词义，推导出能够反映句子意义的某种形式化表示。计算语言学对语言的分析一直以来追求“全面”和“深层”的目标，但在复杂语言现象下，这种思想难免收效甚微。与之相对，浅层分析采用“片面”和“浅层”的理念，在满足应用的前提下，为解决复杂语言现象提供了一条新的途径，而语义角色标注成为当前浅层语义分析的主要手段[1]。

语义角色就是谓词与它的参数之间的语义关系。语义角色标注就是将词语序列分组，并按照语义角色对它们进行分类。它对问答系统、机器翻译、自动文摘、信息抽取等系统性能的提高，有着重要的作用。它并不对整个句子进行详细的语义分析，而只是标注句子中的一些成份为给定谓词的语义角色，这些成份作为此谓词的参数被赋予一定的含义。

对于语义角色标注，国际上在 2004 年~2008 年举行过 5 次评测，分别为 Senseval-3、SemEval2007、CoNLL 会议主办的 SRL Shared Task 2004, 2005, 2008。

汉语语义角色标注的研究刚刚起步，使用的资源主要是文献[2]在宾州中文树库的基础上建成的中文命题库(CPB)，在其中进行了语义角色的自动标注，并使用了谓词的类提高了系统的性能。文献[3]运用支持向量机的方法进行了浅层语义标注的实验，并比较了中文实验结果与英语的实验结果。文献[4]针对中文的特点，在英文语义角色标注特征的基础上，提出了一些新的特征和组合特征，并在 CPB 语料数据上使用最大熵分类器进行了实验。文献[5]以宾州中文树库为基础，选取了 5 种主要的语义角色，采用了两阶段的分类方法，取得了较好的结果。文献[6]鉴于当前数据稀疏的问题，采用了基于知网的回退模型，很好地改善了标注的准确率。文献[7]通过整合主动学习与半监督学习，在小规模标注样本环境中取得了良好的学习效果，文献[8]将角色分类阶段分为 3 个子任务，提高了分类的准确率。标注的基本单元可以是句法成分、短语、词或者依存关系等。以上介绍的实验主要是基于短语结构语法的，而在英文的语义角色标注中，已经有学者利用依存句法进行标注实验，CoNLL-2008 的评测就是利用系统自动生成的依存句法树进行语义角色标注。

文献[9]的实验表明，相比基于句法成分的英文语义角色标注，基于依存关系的标注对词汇的依赖性较弱，鲁棒性较高。这对于当前中文语料库较少、数据稀疏等问题，有着重要的意义。

2 中文依存句法树库的建立

2.1 知网

概念是人类对客观世界认识的结果，在本质上都是符号化的实体，它表示的是客观世界中的事物及其含义。在这个客观世界里，一切事物都在特定的时间和空间内不停地运动和变化，它们通常是从一种状态变化到另一种状态，并通常由其属性值的改变来体现。

在知网中，概念是由词表示的概念标识符，一个词有多种语义，就对应有多个不同的概念。知网认为，任何一个事物都一定包含着多种属性，事物之间的异或同是由属性决定

基金项目：陕西省教育厅专项科研基金资助项目(HD01302)
作者简介：安强强(1983－)，男，硕士研究生，主研方向：人工智能，自然语言理解；张 蕾，教授
收稿日期：2009-09-06 **E-mail：**zhlei@nwu.edu.cn

的，没有了属性就没有了事物。属性和它的宿主之间的关系是固定的，即有什么样的宿主就有什么样的属性，反之亦然。而属性必有一定的属性值体现。

知网正是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

2.2 语料资源

Chinese Proposition Bank(CPB)是 Upenn 基于 Penn Chinese Treebank(PCT)标注的汉语浅层语义标注资源，在 PCT 句法分析树的对应句法成分中加入了语义信息。PCT 的标注数据主要来自新华新闻专线、Sinorama 新闻杂志和香港新闻。CPB 包含 20 多个语义角色，相同语义角色对于不同谓语动词有不同的语义含义。其中核心的语义角色为 Arg0-56 种，其余的语义角色为附加语义角色，用前缀 ArgM 表示，后面跟一些附加标记来表示这些参数的语义类别，如 ArgM-LOC 表示地点、ArgM-TMP 表示时间等。

2.3 中文依存句法树库

依存语法是一种充分利用句子中词汇信息的语法体系，它的核心思想是，句子中不同的成分(词)之间是不平等的，存在着支配与被支配，从属与被从属的关系。传统的句法分析把句子分为主语、谓语、宾语等语法结构，而依存语法首先关注的是句子的动词，然后再探寻句子中其他成分与动词的关系。

依存语法认为，词之间的关系是有方向的，通常是一个词支配另一个词，这种支配与被支配的关系就称作依存关系。依存关系既可以是句中词与词之间的句法关系，也可以是语义关系。这为依存句法关系转化为依存语义关系提供了极大的方便。

语义角色标注建立在句法分析的基础之上，由于中文自动句法分析的准确率不高，因此使用标准的树库来进行语义角色标注实验。

依据英文语义角色标注实验的习惯，按照文章的数量将语料分为训练集与测试集，而没有采用按照动词的分类方法。这样可以确保测试集中含有未训练过的动词。为了与基于短语结构句法树的标注相比较，采用了文献[3]所选取的实验数据集。

由于没有大型的中文依存句法树库，因此参照 CoNLL-2008 中的英文树库转换方法，利用 CoNLL-2008 共享的支持中文树库转换的工具包，将 CPB 5.0 的前 1 652 个句子的短语结构树转换为依存树，并对其中的错误进行手工校正。然后，利用基于知网概念体系的词义标注系统进行预处理，自动标注词语的词义。

3 基于依存树的中文语义角色标注

语义角色标注一般分为 2 个步骤：

(1)识别：确定成分是否是谓词的参数。

类似在基于短语结构树的识别中采用的方法，用一个二值分类器来判断给定的依存节点是否是谓词的参数。

(2)分类：对已确定的谓词参数进行分类，即确定成分是什么类型的参数。

本文以依存关系作为标注的基本单元，并与该单元所对应的语义角色类型组成学习实例，最后使用最大熵分类器对这些实例进行自动学习，从而可以对新的实例进行预测。为了解决数据稀疏问题，提高召回率，在特征集中加入了词语的概念首义原等新特征。由于最大熵分类器的效率很高，因此本文把角色识别和分类通过一步实现，句中的每个词都作为候选分类对象，不属于任何语义角色的词被标为空角色。

3.1 最大熵分类器

近年来，最大熵模型被广泛地应用于自然语言处理中。最大熵模型是最大熵分类器的理论基础，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最大熵模型有以下 3 类特点：不要求具有条件独立的特征，较为容易地对多类分类问题进行建模，训练效率较高[10]。

3.2 特征选择

参照文献[11]在英文语义角色标注中选取的特征，并根据中文的特点加以修改。结合知网信息，加入词语概念首义原等新特征，以测试这些特征在中文依存树库上的标注效果。

下面对部分特征以图 1 为例来说明。

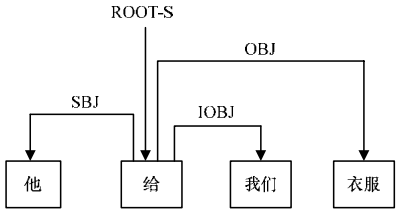

图 1 依存树的例子

3.2.1 单一特征

首先根据汉语语言学知识及依存树的特点选取特征。

依存关系：与父节点的依存关系。如图 1 所示，“他”的依存关系为“SBJ”，“SBJ”对于“他”的语义分类具有重要的指示作用。

词：词语本身。

中心词：词语的父节点词。中心词含有丰富的信息，如果一个名词短语的中心词为“现在”，它很可能是一个时间角色。类似的还有介词“于”、“在”等。介词短语中的名词短语的中心词也含有很重要的预测信息。例如，“在西安”，“在早上”，由于“在”的宾语既可以是时间名词，又可以是地点名词，因此，在介词短语中，宾语的中心词也是一个重要的特征。

实体名词概念首义原：概念首义原为词语概念的第一个义原。在知网中，词义称为概念。例如：“教师”的概念定义为“人: HostOf={职位}, domain={教育}, {教:agent={~}}”，概念的首义原为“人”。

首词及其词性：所有子孙节点中第一个词。如图 1 所示，“给”的“首词”为“他”。

尾词及其词性：所有子孙节点中最后一个词。如图 1 所示，“给”的“尾词”为“衣服”。

词性模式：在子孙节点中，除去“首词”、“尾词”后，剩余的词性标记构成一个“词性集合”。由集合的性质可知，集合中的元素没有顺序、不可重复。词性模式由首词、词性集合与尾词组成。

词性路径，关系路径：分别为从谓词到一个词的路径上的词性标记或依存关系构成的序列。方向有“上”、“下”、“左”、“右”。相对于位置，路径携带更多的信息，如图 1 所示，从“给”到“他”的关系路径为“SBJ 左下”。

上位路径：从一个词到它与谓词的第一个公共祖先节点的路径。

路径长度：从起始节点到终止节点的路径长度。

位置：一个词相对于谓词的位置，值为“左”或“右”。

某一特定语义角色的节点通常出现在特定的位置，例如，大部分附属角色出现在谓词前面。

谓词：动词。在 CPB 中，不同的动词具有不同的角色框架定义、不同的句法结构变化。

谓词子节点的词性序列：在测试集中出现相同谓词子节点的概率很小，所以用词性代替词语，提高系统的泛化能力。

谓词子节点的概念首义原序列：由于词性泛化范围太大，因此引入词语概念首义原来提高准确率。

谓词子节点的依存关系序列：即子类框架。如图 1 所示，“给”的子类框架是 SBJ+IOBJ+OBJ。

谓词兄弟节点的依存关系序列：有些句子有多个动词，非根节点动词具有兄弟节点。

谓词的类别信息：文献[2]的分类结果。由于本实验所采用的训练数据与测试数据是根据文章的数目而不是动词实例来划分的，因此，必定会有一些词出现在测试集中，而没有在训练集中出现。每一个动词的语义角色都有特定的定义，幸运的是，许多动词有着相似的参数结构，文献[2]对此进行了分类。

并列动词的主语：如果并列的动词只有一个主语，则特征值为 true，反之为 false。这个特征是为了解决并列动词的主语不确定性问题。例如：在“我唱歌，他跳舞。”中，并列的动词共有 2 个主语；在“我一边唱歌，一边跳舞。”中，并列的动词共有 1 个主语。

3.2.2　组合特征

由于最大熵分类器不能自动地对特征进行组合，因此，使用上述一些特征的组合来构造组合特征，经过实验验证后，去掉那些降低系统性能的，剩余的作为最终系统的特征。以下组合特征为实验验证后的有效特征：

词+依存关系+中心词

实体名词概念首义原+依存关系+中心词

词+谓词家族关系

(谓词家族关系：一个词相对于谓词的关系。值为“孩子”、“子孙”、“双亲”、“祖先”、“自身”、“兄弟”、“无关系”。)

词+关系路径

中心词+关系路径

中心词概念首义原+关系路径

中心词+上位关系路径

中心词+词性模式

谓词子节点的依存关系序列+谓词家族关系

词性路径+关系路径

谓词子节点的词性序列+谓词子节点的依存关系序列

4　实验结果与分析

本文将 CPB 中前 1 652 个句子转化为标准的依存树，并进行了词义自动标注预处理，利用这部分作为训练集与测试集，采用最大熵分类器，在特征集中加入词语的概念首义原等新特征，获得了 90.68%的 *F* 值。

实验详细结果见表 1。

表 1　标注结果　(%)

系统	准确率	召回率	*F* 值
本系统	89.61	91.78	90.68
基于句法成分的系统	89.73	91.26	90.49

从表 1 可以看出，在标准的句法分析树上，本系统取得了比基于句法成分标注更高的 *F* 值，主要有以下几点因素：

(1)与基于句法成分的语义角色标注系统相比较，基于依存关系的系统对词汇的依赖性较弱，受小语料库的影响较小。

(2)采用了较多的、有效的单一特征与组合特征。

(3)由于测试集与训练集中出现相同的参数中心词的情况不多，因此大部分系统引入词性来解决这个问题。但在很多情况下，中心词词性、与谓词的关系相同，但语义角色不同。如：吃 食堂，吃 米饭，“食堂”与“米饭”都为实体名词，这 2 个短语都为动宾关系，通过词性差异无法区分；为此引入词语的概念首义原来提高系统的召回率。例如，“米饭”的概念首义原为“食品”，“食堂”的概念首义原为“场所”，当“吃”的宾语首义原为“食品”时，语义关系为 ArgM-LOC，当宾语首义原为“场所”时，语义关系为 Arg1。

(4)参数的中心词大部分为实体名词，由于实体名词一词多义的现象较少，即使是多义词，义项较少，因此，实体名词的词义标注准确率很高，极少部分标注错误的词义对系统的整体性能影响较小。

对分类错误的语义角色进行分析，发现主要是由于某些动词没有在训练集中出现，而且动词参数的句法结构又不符合一般规律。例如，主语通常为 Arg0，宾语通常为 Arg1，而“水装满了木桶。”、“一个房间睡五个人。”等句子的参数结构却与此不同。

5　结束语

本文选取了小部分语料作为训练集与测试集，在标准的依存句法树上，使用最大熵机器学习算法进行语义角色的自动标注实验，取得了较好的成绩。下一步将把 CPB 中的短语结构树全部自动转化为依存树，选取更加符合中文特点的特征，添加剪枝和后处理等处理步骤，以提高系统的性能。

参考文献

[1] 陈耀东, 王　挺, 陈火旺. 浅层语义分析研究[J]. 计算机研究与发展, 2008, 45(zl): 321-325.

[2] Xue Nianwen, Palmer M. Automatic Semantic Role Labeling for Chinese Verbs[C]//Proc. of IJCAI’05. Edinburgh, UK: [s. n.], 2005.

[3] Sun Honglin, Jurafsky D. Shallow Semantic Parsing of Chinese[C]//Proc. of NAACL’04. Boston, USA: [s. n.], 2004.

[4] 刘怀军, 车万翔, 刘　挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, 21(1): 79-84.

[5] 连乐新, 胡仁龙, 杨翠丽, 等. 基于中文宾州树库的浅层语义分析[J]. 计算机应用研究, 2008, 25(3): 674-676.

[6] Wang Xia. Semantic Role Labeling in Chinese Using HowNet[J]. Language and Linguistics, 2008, 9(2): 449-461.

[7] Chen Yaodong, Wang Ting, Chen Huowang, et al. Semantic Role Labeling of Chinese Using Transductive SVM and Semantic Heuristics[C]//Proc. of IJCNLP’08. Hyderabad, India: [s. n.], 2008.

[8] Ding Weiwei, Chang Baobao. Improving Chinese Semantic Role Classification with Hierarchical Feature Selection Strategy[C]//Proc. of EMNLP’08. Honolulu, USA: [s. n.], 2008.

[9] Johansson R, Nugues P. The Effect of Syntactic Representation on Semantic Role Labeling[C]//Proc. of the 22nd International Conference on Computational Linguistics. Manchester, UK: [s. n.], 2008.

[10] 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.

[11] Che Wanxiang, Li Zhenghua, Hu Yuxuan, et al. A Cascaded Syntactic and Semantic Dependency Parsing System[C]//Proc. of CoNLL’08. Manchester, UK: [s. n.], 2008.

编辑　任吉慧