基于依存关系的中文谓词标注研究*

袁晓虹　王步康　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: 20074227065077@suda.edu.cn

摘　要：谓词标注是语义角色标注中的重要一步，它的性能直接影响到语义角色标注的性能。本文实现了一个基于依存关系的中文谓词分析平台，使用最大熵分类器在 CoNLL'2008 和 CoNLL'2009 评测数据上进行了系统实验，对各种词法、语法和语义特征及其组合进行了测试，以得到系统最好性能。同时，与基于依存关系的英文谓词标注进行了分析比较。

关键字：谓词标注；语义角色标注；依存关系；最大熵分类器

Predicate Labeling for Dependency-Based Chinese Semantic Role Labeling

Xiaohong Yuan　Bukang Wang　Hongling Wang　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: 20074227065077@suda.edu.cn, Phn: +86-0512-65165848

Abstract: Predicate labeling plays a critical role in semantic role labeling (SRL). This paper explores dependency-based predicate analysis in Chinese SRL using a maximum entropy classifier. In particular, various kinds of lexical, syntactic and semantic features are incorporated to improve the performance with systematic evaluation on CoNLL'2008 and CoNLL'2009 shared tasks .Moreover, we compare the performance of predicate analysis between Chinese dependency-based SRL and English dependency-based SRL.

Key words: Predicate Labeling; Semantic Role Labeling; Dependency Relationship; Maximum Entropy Classifier

1. 引言

谓词标注就是根据句子的短语句法分析树或依存关系树上的结构，以及各个短语或词的词性等各种词法、语法和语义特征，识别出句子的谓词，并分析谓词的词义。所以，谓词标注可分为两个子任务：谓词识别和词义标识。所谓谓词识别（PI, Predicate Identification）是识别出句子中的谓语动词或名词，词义标识（PC， Predicate Classification）是在前者所识别出的谓词的基础上进行词义的识别。一直以来谓词标注并没有成为一项研究的热点，但是随着语义角色标注（Semantic Role Labeling，SRL）的广泛应用，谓词标注的重要性日益凸显，作为语义角色标注

* 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：袁晓虹（1985-），女，硕士研究生，主要研究方向：自然语言处理；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；周国栋（1967-），男，教授，博士生导师，研究方向：自然语言处理

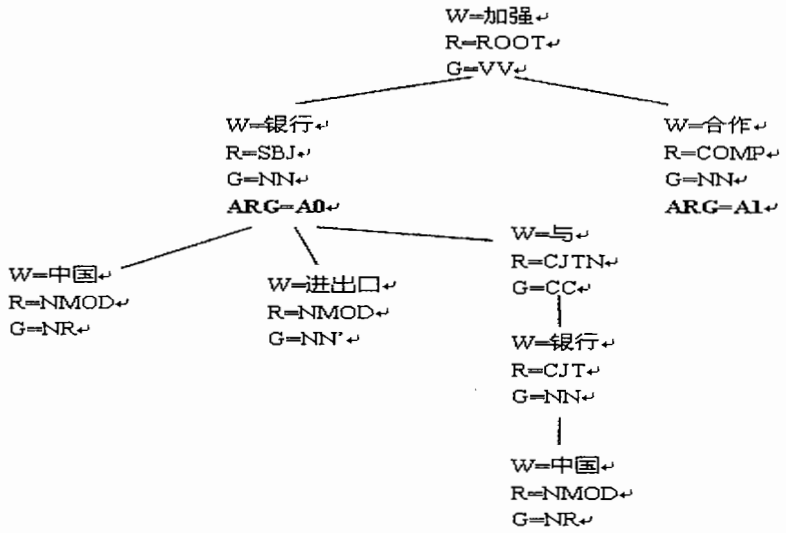
* 通讯作者：hlwang@suda.edu.cn

的前提，谓词标注的结果直接影响语义角色标注的性能。

谓词标注是语义角色标注的一个关键问题。通常谓词标注是在短语结构句法分析或依存关系分析基础上进行的，使用规则或者统计的方法识别句子中的谓词成分及词义。

下面给出了一个CoNLL2009 shared task♥提供的语料实例(1)，实例（1）所构成的依存关系树如图（1）所示：

中国进出口银行与中国银行加强（.01）合作。



图（1） 中文上的依存关系树

图（1）中，用黑体字表示各依存关系承担的角色，W 表示单词，R 表示语法依存关系，G 表示词性，ARG 表示语义角色。在实例（1）中只有一个谓词：加强，(.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。

文章第 2 部分简单介绍了谓词标识的相关研究。第 3 部分介绍基于依存关系的中文谓词标注系统的实现、特征选择以及实验结果分析。最后第 4 部分对本文进行总结并对后期工作进行展望。

2. 相关研究

传统的基于短语结构句法分析的语义角色标注研究论文大多基于手工标注的谓词，对谓词标注的研究大都是基于英文依存句法分析，而对中文的谓词标注研究还未展开。不过，随着越来越多中文语义角色标注的研究，中文谓词标注研究也将日益受到关注。

在英文上的谓词标注有些采用基于某种特定的规则的方法，如Llu'ıs[1]等提出比较简单且典型的规则识别方法：对动词而言，由于词性和原型足以正确标识出大多数动词谓词，所以，除了助动词和被动语态中的动词外，他将所有动词都当作是谓词；对名词而言，将训练集中出现过的名词性谓词作为参考，如该名词出现在测试集中，则作为谓词。这种规则实现简单但存在明显的缺点，主要是对于名词性谓词的处理不够严谨，会漏标和错标很多名词，影响系统性能。

谓词标注的另一种方法是基于统计的方法，如 Ciaramita[2]等在CoNLL2008 语料中抽取了分

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

词原型、分词语态、谓词词性、孩子节点数等7种典型特征，并在谓词的识别和词义标识中采用同样的特征，谓词识别在测试集▼得到的F1的值为84.87%。

Che[3]等在谓词识别时增加了更多的特征，并做了大量的实验，最后得出了效果明显的特征集，例如：单词原型、中心词、中心词词性、单词原型+中心词和一些有关兄弟结点的特征等。而在词义标识时采用了与谓词识别不同的特征集，这个特征集同样是选取了效果明显的特征。谓词识别和词义标识的性能得到了显著提高，在测试集上得到的F1值分别为84.87%/78.94%。

同小组成员汪[4]等同样采用基于统计的方法，使用了30多个特征，除了相同的7个基本特征之外，在谓词的识别与词义标识阶段采用不完全相同的特征，对系统进行不断的测试，最后在测试集得出谓词识别（PI）和基于自动谓词识别的词义标识（PC）的最佳性能，分别为：89.8%、82.1%。

由于中文的谓词分析还未展开，因此没有可比的系统，主要是因为现在基于中文的语义角色标注相关研究比较少，同时缺少基于依存句法的中文手工标注语料，以及中文句子结构的复杂性。这些都成为中文语义角色标注任务面临的重要问题。

3. 系统实现

3.1 系统概述

如前，我们也把谓词标注分为两个子任务：谓词识别和词义标识。后者是在前者的基础上进行。同时，我们研究了在正确的谓词识别基础上进行词义的标识，以正确的谓词作为输入可以帮助我们评估谓词识别阶段的性能对词义标识阶段的影响。

为了和同类实验进行比较，我们搭建的平台使用中文依存句法分析，针对语料中的动词性谓词进行标注。谓词识别时首先对所给语料进行预处理，过滤掉非动词的依存关系，然后进行特征抽取。系统采用最大熵分类器[5]，评测使用信息检索中常用的准确率(precision)、召回率(recall)和 F-Score 来评价系统的性能。

3.2 语料资源

和其他基于统计的自然语言处理任务一样，谓词标注需要好的语料资源，为了能够在中文的谓词标注中有所对比，本文采用了两种语料资源。

一种语料是将Chinese PropBank 1.0（CPB）[6]的语料通过Penn2Malt工具转换为CoNLL2008中所使用的语料的形式，以下简称转换语料。CPB是Upenn基于Penn Chinese Treebank 标注的汉语浅层语义标注资源，在Penn Chinese Treebank句法分析树的对应句法成分中加入了语义信息。Penn Chinese Treebank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，相同的语义角色对于不同目标动词有不同的语义含义。CPB基于 Penn Chinese Treebank 手工标注的句法分析结果，准确率较高。它几乎对 Penn Chinese Treebank 中的每个动词及其语义角色进行了标注，因此覆盖范围更广，可学习性更强。我们的实验选取了前760篇文章（chtb_001.fid—chtb_931.fid），共9622个句子。其中训练集(chtb_100.fid-chtb_931.fid)包含句子数8384句，有32387个谓词，测试集(chtb_001.fid-chtb_099.fid)包含句子数1238句，

▼本小节的测试集均指 CoNLL2008 shared task 提供的WSJ测试集

有4793个谓词。

另一种语料是CoNLL2009 shared task提供的，其中训练集包含22277个句子，其中谓词数有102813个，开发集[♥]包含1762个句子，其中含有的谓词数为8103个。其数据来源是LDC发布的CTB6.0的一个子集，数据中的语义信息来自LDC发布的Chinese PropBank2.0。

3.3 特征选择

假设实例（1）所构成的依存树中，当前单词为“加强”，现将各特征列举如下，特征如果不存在就用NULL代替。

3.3.1 谓词识别阶段

经过实验，谓词识别阶段我们取本阶段在本系统中取得最好性能的特征。

使用转换语料时所抽取的特征：

单词本身：当前单词本身，实例（1）中本特征为：加强。

单词词性：当前单词的词性，在转换语料中的词性为手工标注的（Gold），实例（1）中本特征为：VV。

使用CoNLL2009 shared task 提供的语料所抽取的特征：

单词本身：同上。

单词词性：这里取Gold词性，实例（1）中本特征为：VV。

依存关系：指当前单词与其父亲单词的关系，实例（1）中本特征为：Root。

3.3.2 词义标识阶段

类似于语义角色标注的分类阶段，在谓词识别的基础上，进行词义标识需要更多的特征，由于最大熵分类器不能自动地对特征进行组合，因此，我们使用了一些特征的组合来构造组合特征。

使用转换语料时添加如下特征：

依存关系：指当前单词与其父亲单词的关系，实例（1）中本特征为：Root。

当前结点中心词本身：当前结点的中心词，实例（1）中本特征为：NULL。

孩子词性链：指当前结点的孩子结点的词性链，实例（1）中本特征为：NN+NN。

孩子词性链（N）：指当前结点的孩子结点的词性链（遇到重复词性则只能用一次），实例（1）中本特征为：NN。

孩子依存关系：指当前结点的孩子结点的依存关系链，实例（1）中本特征为：SBJ+COMP。

孩子依存关系（N）：指当前结点的孩子结点的依存关系链（遇到重复关系则只能用一次），实例（1）中本特征为：SBJ+COMP。

由于CoNLL2009语料与转换语料并不完全相同，因此对于CoNLL09语料我们在词义识别阶段减少了特征“当前结点中心词本身”，增加了如下特征及组合特征：

当前节点中心词词性：当前结点中心词的词性，实例（1）中本特征为：NULL。

单词本身+中心词本身，实例（1）中本特征为：加强+NULL。

单词本身+依存关系，实例（1）中本特征为：加强+ROOT。

依存关系+单词词性：ROOT+VV。

[♥]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

3.4 实验结果与分析

3.4.1 系统结果

系统使用上节给出的特征及组合特征，得到结果如表 1 所示。表 1 给出了系统在两种不同的语料上取得的性能。从测试结果可以看出，该系统在转换语料上的谓词识别（PI）性能及词义标识（PC）性能均比在 CoNLL2009 share task 提供的语料库上的结果好，主要原因还是在于语料之间的差异，具体原因分析在下节展开。

表 2 给出在谓词正确时进行词义识别的结果，在此我们使用正确率（Accuracy=词义标注正确的词的个数/谓词总数）来描述系统性能。从表中可看出，基于正确谓词，使用同样的特征，在两种语料库上的词义识别正确率非常接近，而在表 1 中两者在 PC 上的 P 值相差约 3.6 个百分点，这充分说明了谓词识别对词义的识别有着很大的影响。

表 1 谓词标注性能

	P(%)	R(%)	F(%)
转换语料			
谓词识别（PI）	99.96	94.11	96.95
词义标识（PC）	95.17	89.61	92.31
CoNLL2009 shared task 提供的语料			
谓词识别（PI）	96.54	94.75	95.64
词义标识（PC）	91.51	89.82	90.66

表 2 在正确谓词识别上的词义标识结果

使用语料	Accuracy（%）
转换语料	94.82
CoNLL2009 shared task 提供的语料	94.97

3.4.2 结果分析

由于目前中文谓词标注没有可比的系统，为分析在这两种语料上的性能差异，我们对两种语料做了谓词相关情况统计，结果如表 3 所示。从统计结果看，由于 CoNLL2009 的语料规模远远大于转换语料，因此两个语料在训练集中的谓词个数相差很大。从这点上来说，基于 CoNLL09 语料的结果可信度更高。

表 1 中转换语料的谓词识别（PI）性能及词义标识（PC）性能均高于在 CoNLL2009 share task 提供的语料库上的性能。究其原因，我们可以从表 3 给出的数据来分析，在转换语料的训练集中谓词的词性均为动词（VV、VA、VC 和 VE），这个情况在该语料的测试集上也是如此，所以可以充当谓词的成分其词性全部为动词，并且在训练集和测试集中只有（117、10）个动词不是谓词，只占动词总数的 0.36%、0.21%，这两个特点使得在此语料中谓词识别的准确率极高，达到 99.96%。而 CoNLL2009 语料的训练集中谓词的词性除了动词外，还有其他词性，如（NN/MSP/AD/DER/SB）。另外，我们从表中“词性为动词但不是谓词”这一行来看，在训练集和开发集中分别有（4578、317）个词性为动词的词不是谓词，占动词总数的 4.26%、3.76%，这个比例与转换语料相比是相当高的，这也使得谓词在分类时受到负例的影响比较大，最终导致分

类器预测出的谓词的准确率明显降低。

表 3 两种语料库的谓词个数统计

谓词数 \ 语料	转换语料		CoNLL2009 share task 提供的语料	
	训练集	测试集	训练集	开发集
VV	27079	4110	84053	6829
VA	2733	329	8241	494
VC	1762	223	6786	485
VE	813	131	3726	295
NN/ MSP/ AD/ DER/SB	0/0/0/0/0	0/0/0/0/0	1/1/3/1/1	0/0/0/0/0
非词性相关的统计				
非第一词义	2488	378	10382	756
谓词总数	32378	4793	102813	8103
词性为动词但不是谓词	117	10	4578	317

另外与汪[4]等在英文上的谓词标注结果（PI：94.1/85.9/89.8，PD：93.8/73.0/82.1）相比，由于中文树库中动词数量占的比例很大，同时结构性信息又多于英文树库，因此中文谓词的识别率很高。在英文中词性这一特征的作用虽然很重要，但仍然不是占主要地位的。而在中文中词性的作用非常明显，可以说是起决定作用的，在实验中我们发现，即使只用词性这一个特征，也可以使系统的F值达到 96.91%。因此系统在谓词识别阶段最终只采用了两个特征得到表 1 中的性能，我们的大量特征实验证明，其他特征的加入反而会降低系统性能。

在词义标识（PC）方面，由于中英文谓词的词义标识方法类似，都是基于框架语义标注的，同时标注信息都是框架语义中的具体语义项的编号，如编号 01、02 等。但从表 4 中“非（.01）词义”这行的结果看，在我们使用的两种语料中都有多于 90%的词义为.01；而英文谓词的词义相比于中文谓词的词义较分散，根据对CoNLL2008 shared task*（请将该标注放到文章首次出现conll2008 的地方）提供的WSJ训练集进行统计，其中谓词数有 185138 个，词义为.01 的有 155143 个，占谓词总数的 83.8%，这说明尽管.01 词义也占多数，但没有中文占的比例高，因此中文的词义识别性能较英文的高。

4. 结论与展望

本文在Wang（2008）[7]等人的工作基础上进行了扩充比较，使用基于统计的方法在不同的语料上进行中文的谓词标注研究。实验结果表明，在中文的谓词识别中，决定谓词成分的关键特征是词性，但是在词义标识中需要引入更多的特征及组合特征才能达到好的结果。谓词标注作为语义角色标注的一项子任务，其重要性已经得到广泛关注，并且随着多语言语义角色标注的研究，多语言谓词标注工程也会随之展开，相信会取得更好的结果。

* http://www.yr-bcn.es/conll2008/

未来我们将在自动依存句法分析的结果上进行相关的实验，并将实验结果应用于中文语义角色标注。根据目前进行的小部分实验结果，我们发现谓词标注的准确率依然很高，但召回率很低，有很多应该是谓词的成分被错误的排除，因此我们将采用各种手段提高系统性能，使之达到实用的目的。

参 考 文 献

[1] Xavier Llu´ıs,Llu´ıs M`arquez . A Joint Model for Parsing Syntactic and Semantic Dependencies. Proceedings of the 12th Conference on Computational Natural Language Learning, pages 188–192, Manchester, August 2008.

[2] Massimiliano Ciaramita, Giuseppe Attardi, Felice Dell'Orletta. DeSRL: A Linear-Time Semantic RoleLabeling System. Proceedings of the 12th Conference on Computational Natural Language Learning, pages 258–262. Manchester, August 2008.

[3] Wanxiang Che, Zhenghua Li, Yuxuan Hu and so on. A Cascaded Syntactic and Semantic Dependency Parsing System. Proceedings of the 12th Conference on Computational Natural Language Learning, pages 238–242, Manchester, August 2008.

[4] 汪红林,王红玲,周国栋. 语义分析中谓词标识的特征工程(已被《计算机工程与应用》录用未发).

[5] N. Kwon , M. Fleischman , E. Hovy. Senseval automatic labeling of semantic roles using Maximum Ent- ropy models [A] . Senseval23 : Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text [C] . Barcelona , Spain : Association for Computational Linguistics , 2004 , 129 132

[6] N. Xue , M. Palmer . Annotating the Propositions in the Penn Chinese Treebank [A] . In : Proceedings of the Second SIGHAN Workshop on Chinese Language Processing [C] . Sapporo , J apan : 2003 , 47 54.

[7] Hongling Wang， Honglin Wang, Guodong Zhou. Dependency Tree-based SRL with Proper Pruning and Extensive Feature Engineering. Proceedings of the 12th Conference on Computational Natural Language Learning, pages 253–257, Manchester, August 2008.