利用配价信息的语义角色标注

袁里驰
(江西财经大学信息管理学院,江西南昌 330013)

摘 要: 语义角色标注是一种浅层语义分析.现有的汉语语义分析方法和语义角色标注体系没有结合汉语的特点并有效刻画出汉语的本质特性,导致目前汉语语义角色标注性能与英语相比相差较大.在汉语中,配价结构可以较好地刻画汉语句子的句法结构和语义构成关系,因此,我们在考察配价语法的基础上适当修改了语义角色标注体系并将谓词本身的配价信息融入语义角色标注.实验结果表明,配价信息的使用能够较大幅度提高动名词性谓词的语义角色标注性能:基于正确句法树和正确谓词识别,动词性谓词的 SRL 性能 F1 值达到 93.69%;名词性谓词的 SRL 性能 F1 值达到 79.23% ;均优于目前国内外的同类系统.

关键词: 配价结构;动词性谓词;名词性谓词;语义角色标注
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2017)10-2533-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.10.031

Semantic Role Labeling Utilizing Valence Information

YUAN Li-chi
(*School of Information Technology ,Jiangxi University of Finance and Economics ,Nanchang ,Jiangxi* 330013 ,*China*)

Abstract: Semantic roles labeling is a kind of shallow semantic analysis. Existing Chinese semantic analysis methods and semantic roles labeling systems do not effectively characterize Chinese essential features,and it causes the currently larger difference between Chinese SRL systems and English SRL systems. Valence structures can better characterize syntactic structures and semantic constitution relations of Chinese sentences,so we appropriately modified the semantic roles labeling systems and incorporated the valence information of predicates into semantic roles labeling. Experimental results show that proper use of valence information significantly improves the performance of semantic roles labeling system:the verbal SRL approach achieves the performance of 93.69% in F1-measure and the nominal SRL approach achieves the performance of 79.23% in F1-measure on golden parse trees and golden predicates,and all outperform the state-of-the-art SRL systems.

Key words: valence structure;verbal predicates;nominal predicates;semantic role labeling

1 引言

语义分析是自然语言处理的一个关键问题.作为目前的热点研究课题之一,语义角色标注是浅层语义分析的一种,其实质是在句子级别进行浅层的语义分析,即对于给定句子,对句中的每个谓词标注出句中相应的语义成标注分,并确定其相应的语义标记,包括核心语义角色(如施事者、受事者等)和附属语义角色(如地点、时间、方式、原因等).根据谓词类别的不同,可以将现有的 SRL(Semantic Roles Labeling)分为动词性谓词 SRL 和名词性谓词 SRL.语义角色标注已广泛应用于自动问答[1]、机器翻译[2]、信息检索[3]等领域,具有广泛的应用前景.目前大多数语义角色标注系统[4-15]采用统计学习的方法,基于统计的机器学习方法可以分为两类:基于特征向量的方法和基于树核函数的方法.

FrameNet[16],PropBank[17]等语料库的发布极大地推动了基于动词性谓词的英文语义角色标注的研究,语义角色标注任务越来越受到国际自然语言处理领域的关注,国际上先后举行了多次语义角色标注任务的评测.与 FrameNet 相比,PropBank 基于 Penn TreeBank[18]手工标注的句法分析结果进行标注,因此标注的结果几乎不受句法分析错误的影响,精确率较高.它

收稿日期:2015-07-29;修回日期:2015-09-29;责任编辑:郭游
基金项目:国家自然科学基金(No. 61562034, No. 61262035);江西省科技支撑计划(No. 20151BBE50082);江西省自然科学基金(No. 20142BAB207028)

几乎对 Penn 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. NomBank[19]语料库采用与 PropBank 一致的标注框架,进一步标注了 Penn TreeBank 中的名词性谓词及其语义角色. 由于中文 PropBank[20]和中文 NomBank[21]发布较晚,中文语义角色标注研究相对较少. Xue[4]等人利用中文 PropBank 和中文 NomBank 展开了中文动词性和名词性谓词的语义角色标注,在使用正确和自动句法树情况下,性能 F1 值分别取得了 91.3% 和 61.3%.

根据对句子的不同标注情况,语义角色标注系统可分为基于短语结构句法分析的语义角色标注、基于依存句法分析的语义角色标注和基于组块的语义角色标注,从整体效果上看,以句法成分为标注单元的语义角色标注要优于以词和短语为标注单元的方法.

现有的语义角色标注研究表明,虽然基于手工句法树,汉语语义角色标注的性能 F1 值达 91%,和英语差不多,但是基于自动句法树,汉语语义角色标注的性能 F1 值与英语相比则相差较大. 造成这种现象的主要原因是现有的汉语语义分析方法和语义角色标注体系[22]不适合汉语的特点,没有有效刻画出汉语的本质特性,导致目前汉语语义角色标注性能与英语相比相差较大. 在汉语中,配价结构[23,24]可以较好地刻画汉语句子的句法结构和语义构成关系,因此,我们有必要更系统广泛地考察和研究形式化语法理论,尤其是配价语法,并在此基础上建立语义角色标注体系和语义角色标注方法.

论文后续主要内容的安排如下:第一部分介绍配价语法和配价结构;第二部分讨论基于配价结构的语义角色标注体系和语义角色标注方法;第三部分描叙了用于后处理的统计模型;第四部分给出语义角色标注的实验结果及分析.

2 配价语法和配价结构

配价语法强调动词是句子的中心,所以是动词中心论. 动词根据它联系的动元(动词所联系的强制性语义成分即语义角色)的数量来分类,即动词的"价"分类,可分为一价动词、二价动词和三价动词三类. 动词的配价结构与动词的语义角色标注尤其是核心语义角色标注有关,因而配价语法和句子级的语义分析(特别是语义角色标注)有着紧密的联系. 现在,配价的研究已经不仅仅局限于动词,形容词和名词的配价也有很多人在研究. 比如说,形容词"年轻"和名词"姐姐"都是一价,分别需要支配一个名词词组,用于说明"谁年轻"和"谁的姐姐". 袁毓林受朱德熙对汉语动词的配价研究的直接影响,着手对汉语名词的配价研究[24]. 从配价的角度看,现代汉语名词可分为无价名词(或零价名词)和有价名词两大类,这是根据名词有无配价要求分类的. 有价名词又分为两类:一类是从谓词派生出来的,另一类不是从谓词派生出来的,它们往往包含一个降级述谓结构. 其中根据其支配能力又可以分为一价名词和二价名词两小类. 名词的配价结构与名词性谓词的语义角色标注尤其是内部角色标注有关.

3 基于配价结构的语义角色标注体系和语义角色标注方法

3.1 语义角色标注体系

随着配价语法和格语法等理论的提出和引入,从事语言研究的学者开始越来越多地关注语义角色问题,纷纷将这些理论应用于语言研究,提出了许多语义角色标注体系. 目前,绝大多数的语义角色标注研究都是基于 PropBank、NomBank 及以此为基准语料库的语义角色标注体系.

PropBank 是宾夕法尼亚大学在 Penn TreeBank 句法分析语料库基础上标注的语义角色标注语料库. PropBank 只对动词进行标注,相应地被称为动词性谓词. PropBank 只定义了 20 多种语义角色,其中核心语义角色为 Arg0 ~ 5 六种,Arg0 通常表示动作的施事,Arg1 通常表示动作的受事等,Arg2 ~ 5 由于动词不同会有不同的含义. 其余的语义角色为附加语义角色,用 ArgM 表示,在这些标记后附加其他标记来表示语义角色的具体类别,如 ArgM-LOC 表示地点,ArgM-TMP 表示时间等等. 与 PropBank 不同的是,NomBank 标注了 Penn TreeBank 中的名词性的谓词及其语义角色,参数的类别和表示与 PropBank 是相同的.

我们在 PropBank 语义角色标注体系的基础上,结合配价语法,增加了两种附加语义角色 ArgM-Tol、ArgM-Mat,分别表示工具、材料,并标注介词附属名词分别为动词性谓词的 VTol 配价、动作的受事名词(通常为宾语)的 NMat 配价. 如句子"天文学家用望远镜观察天空"、"工厂用大米生产白酒"在修改后的 PropBank 语义角色标注体系下的语义标注分别如图 1、图 2.

3.2 语义角色标注方法

3.2.1 动词性谓词角色剪枝

在句法短语结构树中,仅有极少部分短语与目标谓词之间存在语义关系. 我们剪枝的基本思想是在 Xue[4]提出的角色剪枝算法的基础上将句法树中当前短语至目标谓词的路径作为剪枝的依据,由于在句法短语结构树中标注了每个短语的中心词(一般情况下,中心词只有一个词,对于并列名词短语,中心词则为两个名词),如果当前短语到目标动词性谓词的路径只包含当前短语中心词和目标动词性谓词中心词(对于当前短语的父亲结点是介词短语,则路径中可包含父亲

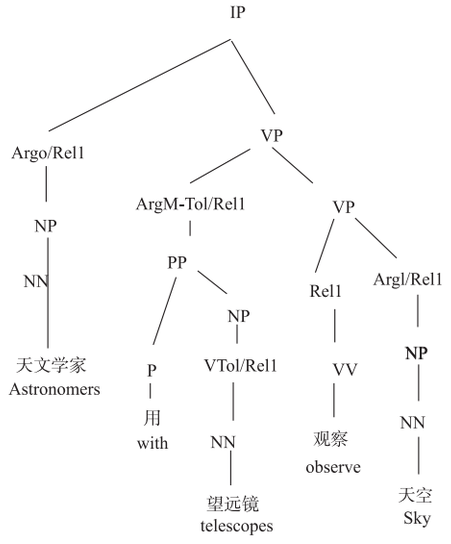

图1　动词性谓词“观察”及其语义角色

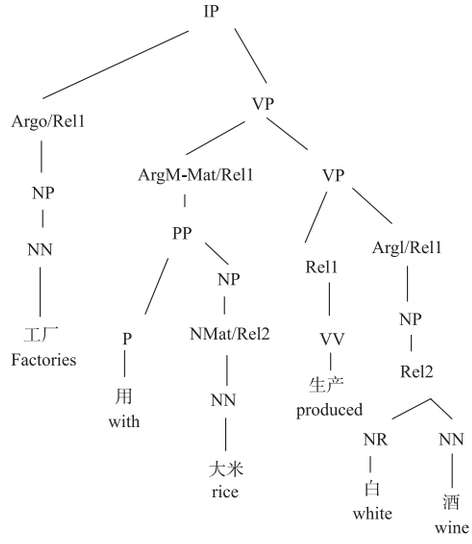

图2　动词性谓词“生产”、名词性谓词“酒”及其语义角色

结点的中心词介词),则保留当前短语为进入角色识别阶段的标注单元.

3.2.2　动词性谓词角色识别和角色分类

对每个目标谓词及担当其语义角色的候选短语集合,采用二元分类器进一步过滤掉不大可能担当语义角色的短语. 对于角色分类,则为每个担当语义角色的候选短语采用多元分类器标注其最可能的语义角色类别. 在语义角色标注系统中广泛应用的基准特征集[4,5]基础上,增加动词性谓词本身的配价数作为基准特征,对于介词短语,则增加名词短语的中心词类作为基准特征. 表 1 列出了本文所使用的基准特征集,表 1 同时结合例句“天文学家用望远镜观察天空”给出了在图 1 中,当前谓词为 VV(观察),句法成分为 PP(用望远镜)时对应的各特征值.

表 1　动词性谓词角色识别和角色分类特征

特征	说明
b1	谓词:谓词本身及其配价数.(观察/2)
b2	成分类型:句法成分的不同类型表现了该句法成分成为不同角色的趋势.(PP)
b3	成分中心词及其词性.(用,P,NTol)
b4	路径:句法树中句法成分至当前谓词的路径. 每个语义角色都处于谓词祖先结点的特定的句法域,该特征表达了这样的信息.(PP < VP > VP > VV)
b5	位置:句法成分相对于谓词的位置,有两种可能:左/left 和右/right.(left)

3.2.3　名词性谓词角色剪枝、角色识别和角色分类

根据担当语义角色的短语在句法树中与谓词之间的句法结构关系,可以将名词性谓词语义角色分为两大类:1)内部角色,指那些与谓词共同位于某名词短语内部的角色,这些角色与谓词之间的路径不存在 VP(动词短语)结点;2)外部角色,指那些与谓词之间路径存在 VP 结点的角色,通常存在一个支持性动词将外部角色与谓词相关联. 对外部角色,如果当前短语到目标名词性谓词的路径只包含当前短语中心词、支持性动词及目标名词性谓词中心词;对内部角色的剪枝,如果当前短语到目标名词性谓词的路径只包含当前短语中心词和目标名词性谓词中心词(或对于当前短语的父亲结点是介词短语,则路径中可包含父亲结点的中心词介词、支持性动词,主要指被标注为支持性动词的 ArgM-Mat 语义角色的句法成份;或对于当前短语未标注为支持性动词语义角色的句法成份,则路径中可包含支持性动词)则保留当前短语为进入角色识别阶段的标注单元. 在李军辉[3]提出的名词性谓词特征集基础上,增加内部角色识别和分类的配价特征 ai,即谓词的名词词性及作为名词时的配价数;增加外部角色识别和分类的配价特征 ao,即谓词的动词词性及作为动词时的配价数. 例如图 2 中,“白”、“大米”均为名词性谓词“酒”的内部语义角色,“酒”作为名词时的配价数为 2. 而在句子“中国工商银行向中国石油公司提供贷款”中,“中国工商银行”、“中国石油公司” 均为名词性谓词“贷款”的外部语义角色,名词性谓词“贷款”作为动词时的配价数为 2.

4　后处理

在语义角色标注的后处理阶段,除根据语义角色标注本身的一些固有约束进行后处理以外,还根据句法分析与语义角色标注的联合学习结果选择合理的语义角色组合. 传统的自然语言处理任务(词性标注、句

法分析、语义分析、信息抽取等)通常是按序进行的,即后一项任务在前一项任务的基础上进行,例如语义角色分析通常要基于句法分析的结果. 按序执行任务并非唯一可选方案,如果能够实现两个或多个连续任务的联合学习,任务之间能够彼此相互利用信息,从而从中获利. 联合学习的一个潜在应用是句法分析和融合配价结构的语义信息标注及分析.

我们的基本思想是:在句法分析的过程中,每当形成一条新的产生式 $p\rightarrow c_1,c_2,\cdots,c_n$ 时,(其中 p 为祖先结点,$c_1,c_2,\cdots,c_n$ 为子结点.)进行配价结构等语义信息标注及分析. 同时将标注的语义信息融入产生式的概率计算. 设 P 为非终结符,H 表示中心成分,L_1 表示左边修饰成分,R_1 表示右边修饰成分. hw,lw,rw 均是成分的核心词,ht,lt,rt 分别是它们的词性,$P(h)$ 表示句法树上当前核心词 h 所依赖的上层核心词. 进一步假设,首先由 P 产生核心成分 H,然后以 H 为中心分别独立地产生左右两边的所有修饰成分. 这样,在我们的句法分析模型中,每一条文法规则写成如下形式:

$$\begin{aligned}&P(ht,hw\mid P(h))\rightarrow L_m(lt_m,lw_m)\cdots L_1(lt_1,lw_1)\\&\cdot H(ht,hw\mid P(h))R_1(rt_1,rw_1)\cdots R_n(rt_n,rw_n)\end{aligned}\tag{1}$$

形如(1)式的文法规则的概率为:

$$\begin{aligned}&P_h(H\mid(ht,hw),P(h))\\&\cdot\prod_{i=1}^{m+1}P_i(L_i(lt_i,lw_i)\mid L_{i-1}(lt_{i-1},lw_{i-1}),\cdots,\\&L_1(lt_1,lw_1),(ht,hw),P(h))\\&\cdot\prod_{i=1}^{n+1}P_i(R_i(rt_i,rw_i)\mid R_{i-1}(rt_{i-1},rw_{i-1}),\cdots,\\&R_1(rt_1.rw_1),(ht,hw),P(h))\end{aligned}\tag{2}$$

其中,L_{m+1} 和 R_{n+1} 分别为左右两边的停止符号. (2)式中的概率

$$P_i(R_i(rt_i,rw_i)\mid R_{i-1}(rt_{i-1},rw_{i-1}),\cdots,R_1(rt_1.rw_1),(ht,hw),P(h))$$

可分解为两个概率

$$P_i(rt_i\mid rt_{i-1},rt_{i-2},\cdots,rt_1,ht,rw_i)\tag{3}$$

$$P_i(rw_i\mid rw_{i-1},rw_{i-2},\cdots,rw_1,hw,P(h))\tag{4}$$

的乘积,记 $S(rw_i)$ 表示词 $rw_{i-1},rw_{i-2},\cdots,rw_1,P(h)$ 中与当前词 rw_i 有语义搭配关系的词(由句子分析树标注的配价结构确定),则有:

$$\begin{aligned}&P_i(rw_i\mid rw_{i-1},rw_{i-2},\cdots,rw_1,hw,P(h))\\&=P_i(rw_i\mid hw,\Delta_r(i-1),S(rw_i))\end{aligned}\tag{5}$$

再假定 $hw,S(rw_i)$ 关于 rw_i 条件独立有:

$$\begin{aligned}&P_i(rw_i\mid hw,\Delta_r(i-1),S(rw_i))\\&=\frac{P_i(rw_i\mid hw,\Delta_r(i-1))\cdot P_i(rw_i\mid S(rw_i))}{P_i(rw_i)}\end{aligned}\tag{6}$$

式(6)中概率 $\frac{P_i(rw_i\mid S(rw_i))}{P_i(rw_i)}=\frac{P_i(rw_i,S(rw_i))}{P_i(rw_i)\cdot P_i(S(rw_i))}$ 即为 $rw_i,S(rw_i)$ 间的互信息,因而整个式(6)概率意义十分明确,符合语言现象.

可以说,目前词汇化的上下文无关文法所做的独立性假设与语言现象不相符合,既不适合于英文,更加不适合于中文. 我们的句法分析模型用条件独立性假设取代了中心词驱动句法分析模型中的独立性假设. 通过对 Collins 模型[25]的规则进行分解和修改,将标注的语义信息融入句法分析统计模型,提高句法分析和语义角色标注的性能.

5 实验结果及分析

试验数据取自中文 PropBank2.0 和中文 NomBank1.0. CTB 是由语言数据联盟(LDC)公开发布的一个语料库,为汉语句法分析研究提供了一个公共的训练、测试平台. PropBank2.0 是宾夕法尼亚大学在 Penn TreeBank5.1 句法分析语料库基础上标注了动词性谓词的语义角色标注语料库. 而中文 NomBank1.0 是为了弥补 PropBank 仅以动词为谓词的局限而开发的,它标注了 Penn TreeBank 5.1 中的名词性谓词及其语义角色. 为了在训练集、开发集和测试集中平衡各种语料来源,参照 Xue[1]的实验数据划分,分别取中文 PropBank2.0 和 NomBank1.0 中的各 648 个文件(chtb_081.fid-chtb_899.fid)共 1296 个文件用作训练集,各 40 个文件(chtb_041.fid-chtb_080.fid)共 80 个文件用作开发集,各 72 个文件(chtb_001.fid-chtb_040.fid 和 chtb_900.fid-chtb_931.fid)共 144 个文件用作测试集. 其中,训练集、开发集和测试集所包含的动词性谓词数分别为 31361,2060 和 3599;训练集、开发集和测试集所包含的名词性谓词数分别为 8642,731 和 1124. 在本文的所有实验中使用 SVM 分类器,SVM 分类器使用多项式核函数,模型的参数都是从训练集中采用极大似然法估计出来的;训练参数的调整设置均在开发集上进行;而模型和语义角色标注方法的性能评测在测试集上进行.

测试的结果采取了常用的 3 个评测指标,即精确率 P、召回率 R、综合指标 $F1$ 值. 其定义如下:

精确率(Precision)用来衡量语义角色标注系统分类器预测的语义角色总数中正确标注的语义角色的比例.

召回率(Recall)用来衡量语义角色标注系统分析出的所有正确语义角色在测试数据中的语义角色总数中的比例.

综合指标:$F1=(P\times R\times 2)/(P+R)$.

表 2 比较了在使用正确/自动句法树和正确/自动

动词性谓词,并使用配价词典获取动词配价数的情况下,各类角色的识别性能. 其中第二列为基于正确句法树和正确动词性谓词条件下取得的性能 F1 值;第三列为基于自动句法分析树和自动谓词条件下取得的性能 F1 值.

表 2　各类动词性谓词语义角色的标注性能

类型	正确	自动	比例(%)
Arg0	94.61	63.75	26.51
Arg1	95.68	74.13	33.08
Arg2	87.82	61.47	4.0
Arg3	65.40	42.37	0.31
Arg4	67.53	48.73	0.09
ArgM-Tol	93.17	70.81	1.64
ArgM-Mat	92.84	69.49	1.38
ArgM-ADV	95.73	77.24	17.10
ArgM-BNF	84.38	67.36	0.25
ArgM-CND	89.13	45.21	0.23
ArgM-DIR	82.53	57.91	0.3
ArgM-DIS	65.68	46.74	1.15
ArgM-EXT	40.23	25.36	0.18
ArgM-LOC	88.67	63.48	3.2
ArgM-MNR	92.15	59.86	2.2
ArgM-TMP	93.47	62.64	7.5
ArgM-PRP	87.15	54.06	0.68
ArgM-TPC	20.58	11.26	0.2
所有类型	93.69	68.87	100.00

本文还对利用统计学习的方法从语义角色标注语料中获取动词配价数并对将动词配价数作为基准特征的动词性谓词的语义角色标注进行了实验,其结果如表 3 所示.

比较表 2、表 3 的语义角色标注测试结果,可以发现:对比使用配价词典获取动词配价数和利用统计学习的方法从语义角色标注语料中获取动词配价数两种语义角色标注方法,系统的整体性能提升不明显,只有 Arg0、Arg1、Arg2 等核心语义角色的标注性能有所提高.

本文在语义角色标注系统中广泛应用的基准特征集[4,5]基础上,增加动词性谓词本身的配价数作为基准特征,对于介词短语,则增加名词短语的中心词类作为基准特征. 因而将参考文献[4,5]所提出的语义角色标注方法作为 baseline 并与本文方法进行比较. 表 4 为基于正确句法树和正确动词性谓词条件下,几种动词性谓词语义角色标注方法的测试结果;表 5 为基于自动句法分析树和自动动词性谓词条件下,几种动词性谓词语义角色标注方法的测试结果.

表 3　各类动词性谓词语义角色的标注性能(统计方法)

类型	正确	自动	比例(%)
Arg0	94.73	64.07	26.51
Arg1	95.84	74.48	33.08
Arg2	87.96	61.72	4.0
Arg3	65.57	42.65	0.31
Arg4	67.82	49.14	0.09
ArgM-Tol	93.17	70.81	1.64
ArgM-Mat	92.84	69.49	1.38
ArgM-ADV	95.70	77.14	17.10
ArgM-BNF	84.38	67.36	0.25
ArgM-CND	89.13	45.21	0.23
ArgM-DIR	82.53	57.91	0.3
ArgM-DIS	65.68	46.74	1.15
ArgM-EXT	40.23	25.36	0.18
ArgM-LOC	88.67	63.48	3.2
ArgM-MNR	92.11	59.72	2.2
ArgM-TMP	93.47	62.64	7.5
ArgM-PRP	87.15	54.06	0.68
ArgM-TPC	20.58	11.26	0.2
所有类型	93.77	69.04	100.00

表 4　动词性谓词语义角色标注方法的测试结果:基于正确句法树和正确谓词

标注方法	召回率(%)	精确率(%)	*F*1(%)
Baseline 1[4]	91.0	93.0	92.0
Baseline 2[5]	93.23	92.33	92.78
本文方法	92.84	94.56	93.69

表 5　动词性谓词语义角色标注方法的测试结果:基于自动句法树和自动谓词

标注方法	召回率(%)	精确率(%)	F1(%)
Baseline 1[4]	60.3	74.8	66.8
Baseline 2[5]	64.57	72.71	68.4
本文方法	62.69	76.40	68.87

从表 4,表 5 与有关语义角色标注方法测试结果对比可以看出:谓词配价信息的使用,系统的整体性能 F1 值得到较大提高,达到 93.69%,尤其对于数量上占较大比例的语义角色,例如 Arg0、Arg1、ArgM-ADV 、ArgM-MNR、ArgM-TMP 和 ArgM-LOC 等类别,性能提升更为明显. 在句法分析与语义角色标注的联合学习过程中,既进行配价结构等语义信息标注及分析,又将标注的语义信息融入产生式的概率计算,自动句法分析与语义

角色标注的性能都得到很大的提高.

表6,表7给出基于正确句法树/正确谓词和基于正确句法树/自动谓词识别的名词性谓词语义角色标注结果,

其中表6给出基于正确句法树下,分别基于正确谓词和自动谓词识别的情况下,各类角色的识别性能.

表6 主要名词性谓词语义角色在测试集上性能

类型	正确 F1(%)	自动 F1(%)	比例(%)
Arg0	81.02	76.91	28.17
Arg1	85.94	79.57	36.26
Arg2	84.35	78.43	3.92
ArgM-ADV	60.06	55.26	3.43
ArgM-MNR	70.35	59.14	11.86
ArgM-LOC	70.37	58.73	7.06
ArgM-TMP	69.86	46.53	5.34
所有类型	79.23	71.50	96.04

表7 名词性谓词语义角色标注结果

方法	召回率(%)	精确率(%)	F1(%)
正确谓词(本文方法)	72.65	87.13	79.23
自动谓词(本文方法)	64.72	79.86	71.50
正确谓词[6]	68.4	77.51	72.67
自动谓词[6]	65.36	74.69	69.72
正确谓词[4]	66.1	73.4	69.6
正确谓词[15]	68.05	79.97	73.53

从表7可以看出,名词性谓词语义角色标注结果远低于动词性谓词的语义角色标注结果,说明名词性谓词的角色识别更加困难,具有非常大的挑战性. 表7还列出了其它几种语义角色标注方法在相同实验语料下的测试结果,与动词性谓词的语义角色标注结果对比,本文提出的名词性谓词语义角色标注方法性能F1值提升更为明显.

6 结论

a. 现有的汉语语义分析方法和语义角色标注体系不适合汉语的特点,没有有效刻画出汉语的本质特性,导致目前汉语语义角色标注性能与英语相比相差较大. 在汉语中,配价结构可以较好地刻画汉语句子的句法结构和语义构成关系,因此,我们有必要更系统广泛地考察和研究形式化语法理论,尤其是配价语法,并在此基础上建立语义角色标注体系和语义角色标注方法.

b. 提出了一种句法分析与语义角色标注的联合学习模型,两者能够彼此相互利用信息:在句法分析的过程中,进行语义信息标注及分析. 同时将标注的语义信息融入产生式的概率计算. 语义角色标注实验已验证了联合学习模型对提高语义角色标注性能的有效性.

参考文献

[1] S Narayanan, S Harabagiu. Question answering based on semantic structures[A]. Proceedings of the 20th international conference on Computational Linguistics[C]. Stroudsburg, PA, USA, 2004. 693 – 701.

[2] D K Wu, P Fung. Can semantic role labeling improve SMT? [A]. Proceedings of the 13th Annual Meeting of the European Association for Machine Translation[C]. Barcelona, 2009. 218 – 225.

[3] M Surdeanu, S Harabagiu, J Williams, P Aarseth. Using predicate-argument structures for information extraction [A]. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics[C]. Sapporo, 2003. 8 – 15.

[4] XUE Nian-wen. Labeling Chinese predicates with semantic roles[J]. Computational Linguistics, 2008, 34(2): 225 – 255.

[5] 李军辉. 中文句法语义分析及其联合学习机制研究[D]. 苏州,苏州大学,2010. 64 – 103.
Li Jun-Hui. Research on joint syntactic and semantic parsing for Chinese[D]. Suzhou, Soochow University, 2010. 64 – 103. (in Chinese)

[6] 李军辉,周国栋,朱巧明,钱培德. 中文名词性谓词语义角色标注[J]. 软件学报,2011,22(8):1725 – 1737.
Li Jun-hui, Zhou Guo-dong, Zhu Qiao-ming, Qian Pei-de. Semantic role labeling in Chinese language for nominal predicates[J]. Journal of Software, 2011, 22(8): 1725 – 1737. (in Chinese)

[7] 吴方磊,李军辉,朱巧明,李培峰. 基于树核函数的中文语义角色分类研究[J]. 中文信息学报,2011,25(3):51 – 58.
Wu Fang-lei, Li Jun-hui, Zhu Qiao-ming, Li Pei-feng. Tree kernel-based semantic role classification in Chinese language[J]. Journal of Chinese Information Processing, 2011, 25(3): 51 – 58. (in Chinese)

[8] 李业刚,孙福振,李鉴柏,吕新宇. 语义角色标注研究综述[J]. 山东理工大学学报(自然科学版),2011,25(6):19 – 24.
Li Ye-gang, Sun Fu-zhen, Li Jian-bai, Lu Xin-yu. A survey on semantic role labeling[J]. Journal of Shandong University of Technology (Natural Science Edition), 2011, 25(6): 19 – 24. (in Chinese)

[9] 庄成龙,钱龙华,周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报,2009,23(1):3 – 8.
Zhuang Cheng-long, Qian Long-hua, Zhou Guo-don. Research on tree kernel-based entity semantic relation extrac-

tion[J]. Journal of Chinese Information Processing, 2009, 23(1):3 - 8. (in Chinese)

[10] 王红玲,袁晓虹,王步康,周国栋. 依存关系上的中文名词性谓词识别研究[J]. 计算机工程与应用,2011,47(20):113 - 116.
Wang Hong-ling, Yuan Xiao-hong, Wang Bu-kang, Zhou Guo-dong. Nominal predicate identification for dependency-based Chinese semantic role labeling[J]. Computer Engineering and Applications, 2011, 47(20):113 - 116. (in Chinese)

[11] 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 - 30.
Wang Bu-kang, Wang Hong-ling, Yuan Xiao-hong, Zhou Guo-dong. Chinese dependency parse based semantic role labeling[J]. Journal of Chinese Information Processing, 2010, 24(1):25 - 30. (in Chinese)

[12] 尹晓丽. 通用语义角色自动标注研究[J]. 长春工业大学学报(自然科学版),2012,33(2):171 - 175.
Yin Xiao-li. An automatic general semantic role labeling system[J]. Journal of Changchun University of Technology(Natural Science Edition), 2012, 33(2):171 - 175. (in Chinese)

[13] 王智强,李茹,阴志洲,刘海静,李双红. 基于依存特征的汉语框架语义角色自动标注[J]. 中文信息学报,2013,27(2):34 - 40.
Wang Zhi-qiang, Li Ru, Yin Zhi-zhou, Liu Hai-jing, Li Shuang-hong. Automatic labeling of Chinese frame semantic roles based on dependency features[J]. Journal of Chinese Information Processing, 2013, 27(2):34 - 40. (in Chinese)

[14] 张秀龙,李新德,戴先中. 基于组块分析的路径自然语言语义角色标注方法[J]. 东南大学学报(自然科学版),2012,42(s1):127 - 131.
Zhang Xiu-long, Li Xin-de, Dai Xian-zhong. Semantic role labeling method for route natural language based on chunk parsing[J]. Journal of Southeast University (Natural Science Edition), 2012, 42(s1):127 - 131. (in Chinese)

[15] 徐靖,李军辉,朱巧明,李培峰. 基于短语和依存句法结构的中文语义角色标注[J]. 计算机工程,2011,37(24):169 - 172.
Xu Jing, Li Jun-hui, Zhu Qiao-ming, Li Pei-feng. Chinese semantic role labeling based on phrase and dependency syntactic structure[J]. Computer Engineering, 2011, 37(24):169 - 172. (in Chinese)

[16] C F Biker, C J Fillmore, J B Lowe. The Berkeley FrameNet project[A]. Proceedings of the 17th international conference on Computational linguistics[C]. Montreal, 1998. 86 - 90.

[17] M Palmer, D Gildea, P Kingsbury. The proposition bank: an annotated corpus of semantic roles[J]. Computational Linguistics, 2005, 31(1):71 - 106.

[18] M P Marcus, M A Marcinkiewicz, B Santorini. Building a large annotated corpus of English: the Penn treebank[J]. Computational Linguistics, 1993, 19(2):313 - 330.

[19] A Meyers, R Reeves, C Macleod, R Szekely, V Zielinska, B Young, R Grishman. Annotating noun argument structure for NomBank[A]. Proceedings of the International Conference on Language Resources and Evaluation[C]. Lisbon, 2004. 803 - 806.

[20] N W Xue, M Palmer. Annotating the propositions in the Penn Chinese treebank[A]. Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, 2003. 47 - 54.

[21] Xue NW. Annotating the predicate-argument structure of Chinese nominalizations[A]. Proceedings of the International Conference on Language Resources and Evaluation[C]. Genoa: 2006. 1382 - 1387.

[22] 刘金凤. 面向自然语言处理的汉语句子语义知识库构建研究[D]. 烟台,鲁东大学,2009:27 - 41.
Liu Jin-feng. Research on the construction of a NLP-oriented Chinese sentence semantic knowledge database[D]. Yantai, Ludong University, 2009:27 - 41. (in Chinese)

[23] 袁里驰. 基于配价结构和语义依存关系的句法分析统计模型[J]. 电子学报,2013,41(10):2029 - 2034.
Yuan Li-chi. A statistical parsing model based on valence Structure and semantic dependency[J]. Acta Electronica Sinica, 2013, 41(10):2029 - 2034. (in Chinese)

[24] 袁毓林. 汉语配价语法研究[M]. 北京,商务印书馆,2010. 55 - 170.
Yuan Yu-lin. The Study of Chinese Valence Grammars[M]. Beijing: Commercial Press, 2010. 55 - 170. (in Chinese)

[25] Collins M. Head-Driven Statistical Models for Natural Language Parsing[J]. Computational Linguistics, 2003, 29(4):589 - 637.

作者简介

袁里驰　男,博士,1973 年 5 月出生于湖南邵阳,江西财经大学信息管理学院计算机科学与技术系副教授,硕士生导师. 研究方向为自然语言处理.

E-mail:yuanlichi@sohu.com