软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(2):222−232 [doi: 10.3724/SP.J.1001.2011.03844]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于特征组合的中文语义角色标注*

李世奇[1+], 赵铁军[1], 李晗静[1], 刘鹏远[2], 刘 水[1]

[1](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[2](北京大学 计算语言学研究所,北京 100871)

Chinese Semantic Role Labeling Based on Feature Combination

LI Shi-Qi[1+], ZHAO Tie-Jun[1], LI Han-Jing[1], LIU Peng-Yuan[2], LIU Shui[1]

[1](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

[2](Institute of Computational Linguistics, Peking University, Beijing 100871, China)

+ Corresponding author: E-mail: sqli@mtlab.hit.edu.cn

Li SQ, Zhao TJ, Li HJ, Liu PY, Liu S. Chinese semantic role labeling based on feature combination. *Journal of Software*, 2011,22(2):222−232. http://www.jos.org.cn/1000-9825/3844.htm

Abstract: This paper proposes a semantic role labeling (SRL) approach for the Chinese, based on feature combination and support vector machine (SVM). The approach takes the constituent as the labeling unit. First, this paper defines the basic feature set by selecting the high-performance features of existing parsing-based SRL systems. Then, a statistics-based method is proposed to construct a combined feature set derived from the basic feature set. According to the distribution of combining features in both positive and negative instances, the ratio of between-class to within-class distance is utilized as the measurement of classifying the performance the feature, and then choosing the combining features with high ratios into the combining feature set. Finally, the experimental results show that the feature combination method-based SRL achieved 91.81% *F-score* on Chinese PropBank (CPB) corpus, nearly 2% higher than the traditional method.

Key words: semantic role labeling; natural language processing; support vector machine; feature combination

摘 要: 提出一种基于特征组合和支持向量机(support vector machine,简称 SVM)的语义角色标注(semantic role labeling,简称 SRL)方法.该方法以句法成分作为基本标注单元,首先从当前基于句法分析的语义角色标注系统中选出高效特征,构成基本特征集合.然后提出一种基于统计的特征组合方法.该方法能够根据正反例中组合特征的分布状况,以类间距离和类内距离之比作为统计量来衡量组合特征对分类所产生的效果,保留分类效果较好的组合特征.最后,在 Chinese PropBank(CPB)语料上利用支持向量机进行分类实验,结果表明,引入该特征组合方法后,语义角色标注整体 *F* 值达 91.81%,提高了近 2%.

关键词: 语义角色标注;自然语言处理;支持向量机;特征组合

中图法分类号: TP391 **文献标识码**: A

浅层语义分析是近年来自然语言处理领域的研究热点之一,而语义角色标注(semantic role labeling,简称

* 基金项目: 国家自然科学基金(60736014, 60803094, 60773069, 60903063)
收稿时间: 2009-10-29; 修改时间: 2010-01-20; 定稿时间: 2010-03-11

SRL)是目前浅层语义分析所采用的主要形式.该技术能够广泛应用于自然语言领域中的各项任务,如信息抽取[1]、问答系统[2]、信息检索[3]、篇章理解[4]等,为其提供结构化的浅层语义信息.中文语义角色标注是指从汉语句子中识别出与目标谓词相关的语义角色(或称“论元”)并判断其类型,如图 1 所示.根据目标谓词和语义角色之间的约束关系,可以把语义角色分为若干个类型,如施事、受事、与事、时间、地点、工具等.中文计算语言学的发展以及中文自然语言处理中底层技术的逐渐成熟,如分词、词性标注、句法分析,都为中文语义角色标注奠定了基础.本文所要研究的内容正是基于 SVM(support vector machine)分类器和短语结构句法分析的中文语义角色标注.

语义角色标注的基本标注单元主要有词、短语和句法成分 3 种,其中,词标注单元主要用于基于依存句法分析语义角色标注系统,短语主要用于基于 Chunk 的语义角色标注系统,句法成分主要用于基于短语结构句法分析的语义角色标注系统.目前,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[5].因此,本文将句法成分,即以短语结构句法分析树中的节点,作为语义角色基本标注单元.另外,本文采用了 PropBank[6]中的语义角色标记方式,具体语义角色标注形式如图 1 所示.与其他标记形式相比,如 FrameNet[7]和 VerbNet[8]等,该形式能够有效地减少语义角色的类别、有效地减轻数据稀疏现象,更加有利于机器学习.这种标记形式把所有语义角色分成两类:一类是谓词所辖语义角色,标记为 ARG0~ARG5,ARG0 通常表示动作的施事,ARG1 通常表示动作的受事,ARG2~ARG5 根据谓语动词不同具有不同的语义含义;另一类是连接性语义角色,包括副词、时间、地点、方式、条件、目的等 13 个子类型,标记为 ARGM,另外还要附加其子类型标记,如图 1 中地点类连接性语义角色“在日本”被标记为 ARGM-LOC.

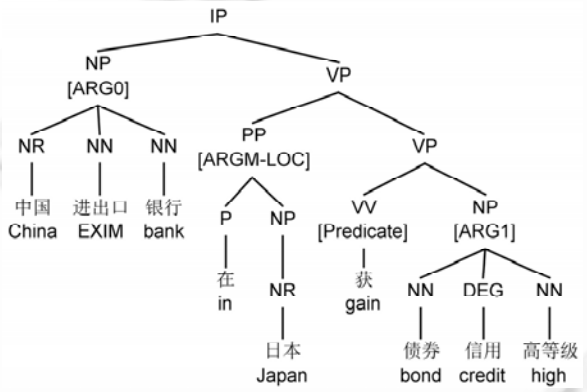


Fig.1 Phrase structure parsing based SRL

图 1 基于短语结构句法分析的语义角色标注

基于特征的有指导机器学习方法仍然是当前语义角色标注中的主导方法,基于有指导机器学习方法、以句法成分为单元的语义角色标注可分为两个子任务:一是语义角色识别,目标是从句子中抽取出所有充当语义角色的句法成分;二是语义角色分类,也就是判断语义角色识别阶段所得的语义角色的类型.另外,在进行语义角色识别之前通常还要进行一个重要的剪枝过程,用简单的方法过滤句法分析树中很多不可能成为语义角色的句法成分,以缩小候选范围,提高准确率[9].基于句法成分的语义角色标注通常还要将句法成分与其对应的谓词相结合,组成“谓词-论元”二元组,然后再将二元组转换成为特征向量作为学习和预测的样本,以补充谓词相关的语义信息.

目前,基于特征的语义角色标注方法中具有代表性的研究主要有:Gildea 等人[10]率先提出了基于概率统计模型的语义角色标注,并提出基于短语结构句法分析 SRL 系统的 7 个基本特征:谓词、句法类型、次范畴框架、路径、位置、语态和中心词.这 7 个特征被后续研究者称为标准特征;Chen 等人[11]以树邻接语法为基础,从句法分析树中挖掘出深层的特征,并用于语义角色标注.Pradhan 等人[12]在标准特征基础上又引入了名实体、中心词

词性、谓词类别、部分路径、时间指示词等 12 种新特征,并详细分析了这些特征各自在语义角色识别和分类中所起的作用.上述研究中尚未引入组合特征的概念,但是它们所提出的多种基本特征为组合特征的应用奠定了基础.

Xue 等人[9]首先在 Gildea 标准特征集合的基础上对组合特征进行了尝试.他们在语义角色标注中采用了“谓词+论元短语类型”和“谓词+论元短语中心词”和“谓词语态+论元位置”这 3 个组合特征,实验结果是,前两个组合特征能够显著提升语义角色标注效果.另外,他们还提出了现在被广泛采用基于启发式规则的句法成分剪枝方法.接着,Xue[13]发现了一系列新的特征,其中包括“谓词类别+论元短语中心词”和“谓词类别+论元短语类型”两个有效组合特征.Ding 等人[14]提出了一种层次化特征选择策略,以短语结构句法分析为基础定义了一些基本特征和几个组合特征模板,提出一种贪心选择算法,从中提取出最有效的特征集合.类似地,Zhao 等人[15]采用贪心特征选择算法,从大规模依存句法分析特征中提取出有效的特征集合,在 CoNLL-2009 语义依存分析评测中取得了较好的成绩.Boxwell 等人[16]提出了一种基于丰富特征的 SRL 方法,其中结合了组合范畴、短语结构和依存 3 种句法分析的特征.但多种句法分析在带来了丰富信息的同时,也带来了较大的噪声.

中文语义角色标注中组合特征的研究从方法上来讲与英文相类似,但语料库资源较少.具有代表性的研究主要有:Sun 等人[17]率先将英文中短语结构句法分析的基本特征集合移植到中文语义角色标注上,然后利用在宾州中文树库上训练的 Collins 句法分析器进行句法分析,并利用 SVM 分类器在手工标注的小规模语料上进行了实验.Xue 等人[18]在宾州中文树库的基础上建立了中文命题语料库(Chinese proposition bank,简称 CPB),并以此为依托采用了若干语言学特征和组合特征,利用有指导的机器学习方法,在正确句法分析基础上取得了高于 90%的 F 值.刘挺等人[19]定义了 19 个基本特征以及 12 个组合特征,其中还包括两个以上基本特征的组合,利用最大熵模型进行语义角色分类取得了较好的效果,但并未详细给出组合特征的构造方法及其产生的性能提升.王红玲[20]研究通过向标准特征集中逐个加入新的特征,其中包括 12 个组合特征,来分析各个特征的作用,证实了将一些特定的特征组合成新特征时,能够增强原特征集的表达能力并由此提高系统性能.车万翔[21]提出了一种语法驱动的卷积树核方法,用于语义角色标注.此外,刘怀军等人[22]和丁金涛等人[23]探索了中文语义角色标注的特征工程中组合和优化问题.

综上所述,目前在基于特征的语义角色标注中,组合特征的研究并不充分,只有少数几个被证实切实有效的组合特征被高频地使用,仍有许多能够提高系统性能的组合特征有待发掘,而缺乏快速、有效的组合特征选择方法是目前的一个主要障碍.因此,本文提出了一种基于统计的二元特征组合方法.该方法能够以基本特征和语料库资源为基础,高效地筛选出有助于语义角色识别和分类的组合特征,进而提高基于特征的语义角色标注的整体性能.本文首先总结现有的语义角色标注系统中证实有效特征作为基本特征,以此为基础构造组合特征.接着,以类间距离和类内距离的比值作为统计量来估计组合特征对于语义角色识别和分类所产生的效果.然后,在 CPB 标准语料库上进行实验,给出在语义角色识别和分类过程中 TOP-N 的组合特征,并在基本特征和组合特征共同组成的特征集合上采用 SVM 分类器进行语义角色识别和分类,证明本文方法的有效性.本文最后一节对该方法进行总结.

1 基于特征组合的语义角色标注方法

本文选择支持向量机(support vector machine,简称 SVM)模型作为语义角色标注的分类器.SVM 是目前基于有指导机器学习的语义角色标注领域中最常用、性能最好的判别模型之一[12,24],而且在自然语言处理其他任务上也有成功的应用.SVM 方法具有适于高维特征空间、适于小规模样本、适于非线性问题、推广能力强等优点.其不足之处是,在处理大规模训练样本时,模型较为复杂,训练和测试所需时间较长.语义角色标注包含语义角色识别和语义角色分类两个子任务,语义角色识别是一个二值分类任务,语义角色分类是一个多值分类任务.而 SVM 是一种典型的二值分类器,因此在采用 SVM 进行语义角色标注时,需要针对这两个子任务分别进行学习.在语义角色识别中,可直接采用 SVM 模型进行学习和预测;在语义角色分类中,我们采用一对多的方法处理该多值分类问题,即训练与语义角色类别个数相同的 SVM 分类器,针对各个语义角色类别进行判断.下面首

先介绍 SVM 的基本方法,接下来对剪枝方法进行后处理,然后定义一些基本特征,并提出一种统计方法,在基本特征的基础上构造出组合特征,最后在基本特征和组合特征组成的特征集合上用 SVM 方法进行学习和预测.在使用 SVM 分类器时,核函数的选择对于模型的训练和预测性能有很大的影响,本文中采用自然语言处理中 3 种常用的核函数,分别是:

(1) 线性核函数(linear kernel function):$K(x_i,x_j)=(x_i\cdot x_j)$.

(2) 多项式核函数(polynomial kernel function):$K(x_i,x_j)=(s(x_i\cdot x_j)+c)^d$.

(3) 径向基函数核函数(RBF kernel function):$K(x_i,x_j)=\exp(-\gamma|x_i-x_j|^2)$.

1.1 剪枝方法及后处理

以句法成分为标注单元的语义角色标注,首先需要一种简单的剪枝预处理方法来过滤句法分析树中一些不可能成为语义角色的句法成分,保留尽量少的候选句法成分,以提高准确性.在目前的剪枝方法中,最有效的是 Xue 在文献[9]中提出的基于启发式规则的方法.该方法可描述为以下 3 个步骤:

(1) 获得目标谓词所在句子的句法分析树,然后将谓词所在的节点设置为当前节点.

(2) 抽取出当前节点的所有兄弟节点放入语义角色候选集合.如果某兄弟节点为介词短语,则将该节点的直接儿子节点也放入语义角色候选集合.

(3) 将当前节点的父节点设置成当前节点,然后重复上述抽取过程,直至达到子句根节点为止.

实际应用中发现,在使用该剪枝方法抽取出的候选语义角色中仍包含较多的冗余句法成分.本文在上述剪枝方法的基础上,通过对语料库的统计分析引入一种后处理方法,能够减轻冗余,进一步缩小候选语义角色的范围.我们在 CPB 语料上统计了语料库中语义角色与其对应的短语类型的共现情况,通过对训练语料中正反例的对比得出,正确语义角色对应的短语类型在全部语义角色对应短语类型中所占比率为 45.3%(24/53),且正例中 24 种正确短语类型对应的语义角色数量在全部语义角色中所占比率为 84.16%;相应地,正确语义角色的父节点短语类型在全部语义角色父节点短语类型中所占比率为 63.6%(14/22),且正例中 14 种正确父节点短语类型对应的语义角色数量在全部语义角色中所占比率为 93.65%.因此,根据上述信息,本文提取了正例中出现的 24 个短语类型和 14 个父节点短语类型作为判断条件(见表 1),进一步将从上述基于启发式的剪枝方法获得的句法成分中短语类型及其父节点短语类型不在类型集合中的句法成分过滤掉,这样可以快速而有效地进一步减少候选语义角色的数量.

Table 1 All types of constituents and their parent nodes in the positive instances

表 1 正例中出现的短语类型和父节点短语类型

Phrase type	Parent's phase type
NP, ADVP, PP, IP, QP, LCP, VP, DP, DVP, CP, PRN, UCP, VV, LST, NN, DNP, CLP, AD, VA, NR, NT, CD, PN, VCD	VP, IP, NP, PP, CP, QP, VSB, VRD, LCP, DVP, UCP, DNP, ADVP, VPT

该方法本质上是基于这样的假设:语料规模足够大,所有可能作为语义角色的短语类型都出现在语料库中.事实上,本文采用的 CPB 语料库规模较大,共有近 10 万个语义角色正例,而句法成分的短语类型仅有 50 多种(包括词性标注类型).在这种情况下,任何一种合理的语义角色短语类型在近 10 万个正例中一次都不出现的概率很小.因此,该假设在本文所采用的语料库上是合理的,实验表明该方法也是可行的.

1.2 基本特征选择

在基于机器学习方法的语义角色标注中,特征选择是关键.本文总结了现有中英文文献中所出现的有效语义角色特征,构造了谓词、句法成分和上下文 3 大类共 26 个基本特征集合.表 2 中详述了这些特征的名称、类型、符号表示、含义描述以及实例.该基本特征集合在语义角色识别和语义角色分类子任务中都会被采用.

Table 2 Basic features and their descriptions

表 2 基本特征及其描述

Category	ID	Feature	Description
Predicate (Take the node with label "Predicate" in Fig.1 for example)	*p*1	Predicate	The predicate word, e.g. "获"
	*p*2	Predicate POS	The part-of-speech (POS) of predicate word, "VV"
	*p*3	Subcategorization	Phrase type pattern of all children of the predicate's parent, e.g. "VP→VV+NP"
	*p*4	Voice	Voice of the predicate, active or passive
Constituent (Take the node with label "ARG0" in Fig.1 for example)	*a*1	Position	Relative position of the argument to the predicate, before or after
	*a*2	Phrase type	Phrase type of the argument, e.g. "NP"
	*a*3	Head word	Head word of the argument phrase, e.g. "银行"
	*a*4	Head word POS	POS tag of the head word, e.g. "NN"
	*a*5	First word	The first word of the argument phrase, e.g. "中国"
	*a*6	First word POS	The POS of the first word of the argument phrase, e.g. "NR"
	*a*7	Last word	The last word of the argument phrase, e.g. "银行"
	*a*8	Last word POS	The POS of the last word of the argument phrase, e.g. "NN"
	*a*9	Temporal cue words	Whether the argument contains temporal keywords or not
Context (Take the two nodes with label "Predicate" and "ARG0" in Fig.1 for example)	*c*1	Parent's Type	Phrase type of the argument's parent node, e.g. "IP"
	*c*2	Parent's head	Head word of the argument's parent node, e.g. "获"
	*c*3	Parent's head POS	POS tag of the "Parent's head" (c2) feature "VV"
	*c*4	Left sibling type	Phrase type of the argument's left sibling, e.g. NONE
	*c*5	Left sibling head	Head word of the argument's left sibling, e.g. NONE
	*c*6	Left sibling POS	POS tag of the "Left sibling head" (c5) feature, e.g. NONE
	*c*7	Right sibling type	Phrase type of the argument's right sibling, e.g. "VP"
	*c*8	Right sibling head	Head word of the argument's right sibling, e.g. "获"
	*c*9	Right sibling POS	POS tag of the "Right sibling head" (c8) feature, e.g. "VV"
	*c*10	Path	Path of phrase type from the argument to the predicate in constituent parse tree, e.g. "NP↑IP↓VP↓PP"
	*c*11	Partial path	Argument branch of the "Phrase type" (a2) feature, e.g. "NP↑IP"
	*c*12	Layer difference	Layer difference between the argument and the least common ancestor (LCA) of the predicate and the argument, e.g. "1"
	*c*13	Syntactic frame	Positions of the NPs surrounding the predicate, e.g. "Cur_v_NP"

1.3 基于统计方法的组合特征选择

文献[9,12]等都发现,加入由基本特征组成的组合特征能够有效地提高语义角色标注的性能,目前大多数语义角色标注系统也基本都会采用组合特征.但是,由于特征组合的方式较多,采用不当的组合特征非但不会提高系统的性能,反而会大大增加特征空间的维数,提高运算的复杂性.目前,对于组合特征的方法研究并不充分,大多数都是针对语言现象通过人工定义得到.本文采用一种基于统计的方法定义了一种新的统计方法对组合特征进行筛选,根据各个组合特征在相应语料中分布情况高效地发掘出对于分类有帮助的潜在组合特征,构造出组合特征集合.

设训练集合 D 由正例和反例两部分组成:$D=\{D_{pos},D_{neg}\}$.这两部分都是从 CPB 语料中获得的,正例可直接从标注语料中抽出,反例的构造方法在语义角色识别和语义角色分类过程中有所不同.这样做是为了使训练反例更接近预测时的真实反例情况,本质上是由于在语义角色识别和语义角色分类两个阶段候选实例的生成机制有所不同.在语义角色识别中是随机地从候选句法成分中选择不是语义角色的成分作为反例,正、反例数量相当;语义角色分类阶段对于每个待识别的语义角色类型都要构造一组正反例训练语料,对于某一语义角色类型,其正例同样可直接从标注预料中获得,反例则是由除该类型之外其余语义角色类型的正例所组成.因此,反例数量通常要比正例多.

假设共有 N 个基本特征,每次仅针对两个不同基本特征 f_a 和 $f_b(a\neq b,a,b\in[0,1,\ldots,N])$,将这两个特征组合构造一个新特征 f_{ab},并将其作为集合中第 N+1 个特征也就是 f_{N+1},再将这个新特征的值也写入正反例训练数据 D 中形成 $D'=\{D'_{pos},D'_{neg}\}$.然后在特征向量化的过程中,分别将正例和反例中的特征先后进行向量化,这样才能保证正反例特征的维数之间具有统计性差异.在特征向量化之后的正反例训练数据集可表示为

$$D'_{pos}=\{(x,y)\mid x\in D',y=1\},\ D'_{neg}=\{(x,y)\mid x\in D',y=-1\},$$

其中,x 代表待分类句法成分所代表的特征向量,采用稀疏特征表示方法,即仅使用值为 1 特征的维数来表示特征向量,值为 0 特征不显式表示,y 代表该候选句法成分所属的语义角色类型.在引入新特征的数据集 D'中,我们首先计算每个特征在正例和反例中的样本均值,即

$$MeanPos(f_i)=\frac{\sum_{x\in D'_{pos}} x(i)}{|D'_{pos}|},\ 1\leqslant i\leqslant N+1,$$

$$MeanNeg(f_i)=\frac{\sum_{x\in D'_{neg}} x(i)}{|D'_{neg}|},\ 1\leqslant i\leqslant N+1,$$

其中,$x(i)$代表中 x 中第 i 个元素的值,也就是 x 中特征 f_i 的值在特征空间中的维数.然后将该特征在正反例集合中样本均值之差的平方作为该特征正反例的类间距离 $B\text{-}Distance(f_i)$,将该特征在正反例样本中的样本方差之和作为该特征正反例的类内距离 $W\text{-}Distance(f_i)$.

$$B\text{-}Distance(f_i)=(MeanPos(f_i)-MeanNeg(f_i))^2,$$

$$W\text{-}Distance(f_i)=S^2_{pos}(f_i)+S^2_{neg}(f_i),$$

其中,$S_{pos}(f_i)$和 $S_{neg}(f_i)$分别代表特征 f_i 的取值在正例和反例中的样本标准差.事实上,$B\text{-}Distance(f_i)$描述了特征 f_i 正例样本中心和反例样本中心之间的距离.如果 $B\text{-}Distance(f_i)$较小,则说明特征 f_i 的正例样本中心和反例样本中心相距较近,反之亦然.$W\text{-}Distance(f_i)$是特征 f_i 正例样本方差和反例样本方差之和.如果 $W\text{-}Distance(f_i)$较小,则说明该特征 f_i 的所有正例样本距正例样本中心与所有反例样本距反例样本中心的总和较小,也就是说,正反例样本与其对应的中心相比较为集中,反之亦然.本文受 Fisher 线性判别模型的启发,采用类间距离与类内距离的比值作为判断组合特征优劣性的依据,比值越大,说明特征 f_i 对于类别区分的作用越明显.因此,定义统计量 $G(f_i)$:

$$G(f_i)=\frac{B\text{-}Distance(f_i)}{W\text{-}Distance(f_i)}.$$

为了比较组合特征的有效性,我们还需要对统计量 $G(f_i)$进行标准化,计算其标准化后的值 $Z\text{-}score$ 作为特征的最终得分.$Z\text{-}score(G(f_i))$(简记作 Z_i)能够有效衡量某个样本与样本均值之间相差多少个标准差,但是该值依赖于总体分布的均值和方差.本文中采取一种简化的形式,用样本均值和样本方差来代替总体分布的均值和方差:

$$Z_i=\frac{G(f_i)-\overline{G(f_i)}}{S_G},$$

其中,$\overline{G(f_i)}$ 为统计量 $G(f_i)$的样本均值,S_G 为 $G(f_i)$的样本标准差:

$$\overline{G(f_i)}=\frac{\sum_{i=1}^{N+1}G(f_i)}{N+1},\ 1\leqslant i\leqslant N+1,$$

$$S_G=\sqrt{\frac{\sum_{i=1}^{N+1}(G(f_i)-\overline{G(f_i)})^2}{N}},\ 1\leqslant i\leqslant N+1.$$

计算出基本特征 f_a,f_b 和组合特征 f_{ab} 所对应的 Z_a,Z_b 和 Z_{ab} 之后,我们定义组合特征的 Z 提高值 I_{ab} 为组合特征的 Z 值与两个基本特征之中得分较高的 Z 值之差:

$$I_{ab}=Z_{ab}-\mathrm{Max}(Z_a,Z_b).$$

最后,通过设定阈值的方法,保留 Z 值超过某阈值 Z_{th},并且 I 值超过某阈值 I_{th} 的组合特征,将其加入到特征集合中.这样,不但有效地过滤掉了对于正反例样本间均值无明显差异的组合特征,保留了适于语义角色标注问题的特征,同时,较小的计算量也保证了对于处理大规模组合特征时的速度,为接下来基于 SVM 模型的学习和分类提供切实有效的组合特征.需要注意的是,本文在利用 SVM 分类器处理语义角色识别和语义角色分类两个子任务以及在处理语义角色分类中各个语义类别的分类任务时,都会采用不同的组合特征集合.由于这些任务的训练正反例完全不同,因此,在语料上根据上述统计方法获得的组合特征也不尽相同.

2 实验结果和分析

2.1 实验数据及评测指标

本文采用语料库为 CPB1.0.该语料库形式上与英文中的 PropBank 类似,是目前中文语义角色标注研究中的通用标准语料库.它是在宾州中文树库基础上手工标注了其中的语义角色信息,共包含 760 篇文章、10 364 个句子、4 854 个谓词以及 92 959 个语义角色[18].我们将其中前 100 篇作为测试语料,将其余 660 篇作为训练语料,采用 Cornell 大学开发的 SVM-light 工具包[25]作为 SVM 分类器.为了验证和分析本文提出的特征组合方法在语义角色识别和语义角色分类这两个 SRL 子任务上的不同效果,我们将针对这两个子任务分别进行评价.为了评价本文所提出的特征组合方法的有效性,本实验在 CPB 中标注的句法分析结果上进行,语义角色识别阶段的输入实例为在测试语料中正确标记的短语结构句法树基础上,经过剪枝所获得的全部候选句法成分;语义角色分类阶段的输入实例为语义角色识别阶段自动识别所得到的句法成分.两个步骤之间衔接紧密,无须人工干预.因此,整个语义角色标注过程仅以正确标记的短语结构句法分析作为输入.我们把仅采用基本特征作为特征集合的系统作为 Baseline 系统,将同时采用基本特征和组合特征作为特征集合的系统记为 Combined 系统.评价指标为 SRL 中广泛使用的精确率(precision,简称 *P*)、召回率(recall,简称 *R*)和 *F* 值(*F-score*,简称 *F*).

2.2 语义角色识别结果

在语义角色识别之前,本文首先采用了 Xue[9]中提出的剪枝方法来减少候选语义角色的数量,在此基础上提出一种统计后处理方法,进一步对候选语义角色进行筛选.首先,我们在测试语料上对该后处理方法进行了评价.我们采用两个指标对剪枝效果进行评价:一是剪枝召回率,为剪枝后保留的正确语义角色数量与总正确语义角色数量之比;二是剪枝效率,表示正确剪掉的短语数与总短语数之比.测试语料中共包含 12 270 个语义角色和 38 0623 个短语,测试结果详见表 3.

Table 3 Experimental results of the post-processing method for pruning

表 3 剪枝后处理方法实验结果

Method	Number of recalled semantic roles	Number of correctly pruned phrases	Pruning recall (%)	Pruning efficiency (%)
Xue	12 017	344 447	97.94	90.50
Our method	12 012	352 421	97.90	92.59

从表 3 中可以看出,本文中的后处理方法剪枝召回率没有明显的下降,但剪枝效率却有明显的提高,达 2%.剪枝效率的提高能够减少后续分类的样本数量,进而降低语义角色分类的错误率.为了进一步评价该剪枝后处理方法在实际应用的有效性,我们还在第 2.3 节中对该后处理方法对语义角色标注整体性能的影响进行了描述.在语义角色识别实验中,我们分别设定统计量 Z 值的阈值 Z_{th} 以及统计量 I 值的阈值 I_{th} 为 0.5 和 0.1,也就是保留组合特征中 Z 值大于 Z_{th} 并且 I 值大于 I_{th} 的组合特征,加入到组合特征集合,其中,Z 值前 20 名组合特征,即对于语义角色识别较为有效的组合特征,见表 4.

Table 4 Top-20 combining features for semantic role recognition on *Z-score*

表 4 语义角色识别中 *Z* 值 Top-20 的组合特征

Rank	Combined feature ID	Rank	Combined feature ID
1	*c*1+*c*11	11	*p*1+*c*12
2	*c*1+*c*12	12	*c*12+*c*13
3	*c*1+*c*10	13	*c*7+*c*12
4	*p*1+*a*3	14	*a*1+*c*12
5	*p*1+*a*7	15	*c*9+*c*10
6	*c*6+*c*12	16	*c*8+*c*10
7	*c*4+*c*12	17	*p*4+*a*2
8	*c*8+*c*12	18	*c*5+*c*10
9	*c*5+*c*12	19	*a*4+*c*2
10	*c*9+*c*12	20	*a*8+*c*2

然后,分别采用基于线性核、多项式核和 RBF 核的支持向量机进行语义角色识别的分类实验.其中,多项式核的参数 s,c,d 根据经验值分别取值 1,1,3;RBF 核的参数 γ 通过在一个独立样本集上采用交叉验证方法获得,取值为 0.03.在测试集上的实验结果见表 5.

Table 5 Results of semantic role recognition

表 5 语义角色识别结果

Kernel	Baseline			Combined		
	P (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)
Linear	92.52	99.39	95.83	93.01	99.16	95.99
Polynomial	93.78	98.83	96.24	93.68	99.17	96.35
RBF	93.82	99.06	96.37	93.81	99.11	96.39

从表 5 中可以看出,基于 RBF 核的组合特征系统取得的 F 值最高,为 96.39%.这些组合特征对基于 3 种核函数的语义角色识别都有所提高,但并不显著;在线性核和多项式核上的 F 值提高幅度相对较大,也仅为 0.16%和 0.11%;在 RBF 核上提高幅度最小,其原因可能是在该问题上 RBF 核函数对于特征的组合能力强于线性核和多项式核,因此对于组合特征的加入不如两者敏感.可见,在基于多特征 SVM 分类器的语义角色识别问题上,RBF 核所表现的性能要优于线性核和多项式核,但总体来说,特征组合的方法对语义角色识别任务的帮助不大.

2.3 语义角色分类结果

CPB 语料中共有 18 种类型的语义角色,本文主要处理其中最主要的 5 类:施事(Arg0)、受事(Arg1)、副词(ArgM-ADV)、地点(ArgM-LOC)和时间(ArgM-TMP).对于其余较为稀疏的语义类型,可采用独立的基于启发式规则或基于机器学习的方法有针对地进行处理.实验方法与语义角色识别类似,先根据统计量 Z 和 I 的值来选择组合特征,在各语义角色类型中,Z 值前 10 名的组合特征见表 6.语义角色分类是多值分类,因此采用一对多的方法需要训练 5 个独立的 SVM 分类器,这里,我们均采用在上一步中取得较好性能的 RBF 核函数,参数 γ 同样在每个类别的独立样本上采用交叉验证方法获得.在测试集上的实验结果见表 7.其中,带星号的数字表示该值的提升具有统计显著性($p<0.05$).

Table 6 Top-10 combining features for semantic role classification on *Z*-score

表 6 语义角色分类中 *Z* 值 Top-10 的组合特征

Rank	Arg0	Arg1	ArgM-ADV	ArgM-LOC	ArgM-TMP
1	*p*1+*c*12	*p*1+*a*3	*p*4+*a*2	*a*2+*a*6	*a*3+*a*9
2	*c*12+*c*13	*p*1+*a*7	*p*1+*a*3	*a*5+*c*4	*a*7+*a*9
3	*c*7+*c*12	*p*4+*a*1	*p*1+*a*7	*p*1+*a*3	*p*1+*a*3
4	*a*1+*c*12	*p*1+*c*11	*a*6+*c*1	*p*1+*a*7	*p*1+*a*7
5	*c*9+*c*10	*a*1+*c*5	*a*4+*c*1	*a*5+*c*6	*p*3+*a*3
6	*c*8+*c*10	*a*1+*c*2	*a*8+*c*1	*a*5+*a*9	*p*3+*a*7
7	*p*4+*a*2	*p*1+*c*10	*c*6+*c*11	*a*5+*c*9	*p*2+*a*3
8	*c*5+*c*10	*c*5+*c*8	*c*6+*c*10	*a*5+*c*12	*p*2+*a*7
9	*a*4+*c*2	*c*5+*c*6	*a*9+*c*10	*a*5+*c*7	*p*3+*a*4
10	*a*8+*c*2	*a*1+*a*3	*c*5+*c*11	*a*5+*c*1	*p*3+*a*8

Table 7 Results of semantic role classification

表 7 语义角色分类结果

Argument type	Baseline			Combined		
	P (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)
Arg0	91.57	95.64	93.56	91.85	96.76	94.24*
Arg1	90.19	85.39	87.72	95.94	87.11	91.31*
ArgM-ADV	98.10	95.79	96.93	97.95	96.43	97.18
ArgM-LOC	85.00	87.46	86.21	90.49	84.26	87.26*
ArgM-TMP	93.50	88.62	90.99	97.05	89.61	93.18*
All Args	91.73	90.63	91.18	94.60	91.75	93.15*

从表 7 中可以看出,引入特征组合后,语义角色分类的整体 F 值大幅度提升,由原来的 91.18%提高到 93.15%,增加了近 2%.而且,各个语义角色类型的分类精度也都有所提高,其中,Arg1 类型语义角色的 F 值提高最

为明显,达 3.59%;唯有 ArgM-ADV 类型语义角色的 F 值提高不显著,并且在 Baseline 系统上,其 F 值就已高达 96.93.可见,基础特征对于这一类型的语义角色识别就可以取得较好的效果.

2.4 语义角色标注整体结果

将上述语义角色标注和语义角色分类两个步骤整合,就形成了一个完整的语义角色标注器.接下来我们评价各个阶段的方法对语义角色标注整体性能的影响.除了 Baseline 系统和 Combined 系统以外,我们定义了另外 5 个分支系统,其中两个为 Baseline*和 Combined*系统,分别表示不采用本文提出的剪枝后处理方法的 Baseline 系统和 Combined 系统,其余 3 个为 ComTop-5,ComTop-10 和 ComTop-20,分别表示仅采用 top-5,top-10 和 top-20 组合特征的系统.这些系统之间性能比较的结果见表 8.

Table 8 Overall results of semantic role labeling

表 8 语义角色标注整体实验结果

	P (%)	R (%)	F (%)
Baseline*	88.03	93.22	90.55
Baseline	88.39	93.31	90.78
ComTop-5	90.04	93.28	91.63
Combined*	89.97	93.37	91.64
ComTop-10	90.15	93.42	91.76
ComTop-20	90.15	93.50	91.79
Combined	90.17	93.51	91.81

通过比较表 8 中第 1 行、第 2 行以及第 4 行、第 7 行,也就是 Baseline*和 Baseline 系统以及 Combined 和 Combined*系统的性能可以看出,本文提出的剪枝后处理方法能够提升语义角色标注的整体性能.另外我们发现,随着采用组合特征的增多,系统 F 值会逐渐升高,但是增幅逐渐越小.比如,采用 top-10 组合特征与采用 top-5 相比,组合特征提升了 0.13 个百分点,而采用 top-20 组合特征与采用 top-10 相比,提升仅为 0.03%.另外,考虑到组合特征数增加时系统复杂性的增加,结合系统性能来看,本文提出的组合特征方法中采用 top-10 组合特征是较为合理的选择.此外,为进一步验证本文提出方法的有效性,我们将 Baseline 和 Combined 系统与另外两个基于 CPB 语料库的语义角色标注的相关工作进行了比较.这两个系统都采用 CPB 语料前 99 篇作为测试语料,其余 661 篇用作为测试语料,且都以正确的短语结构句法分析结果作为输入,其实验设置与本文方法基本一致.本文分别引用了文献[18,21]中所描述的方法和实验结果,见表 8 中 XUE 和 CHE 两项,其中,XUE 系统采用最大熵模型,定义了 9 个基本特征和 2 个组合特征,利用最大熵模型进行学习和测试;CHE 系统采用了 9 个基本特征、10 个扩展特征和 15 个组合特征,利用基于多项式核的 SVM 分类器进行语义角色标注,比较结果见表 9.

Table 9 Comparison of results with other systems

表 9 与其他系统比较实验结果

	P (%)	R (%)	F (%)
XUE	90.4	90.3	90.3
Baseline	88.39	93.31	90.78
CHE	92.68	89.97	91.31
ComTop-10	90.15	93.42	91.76
Combined	90.17	93.51	91.81

从表 9 的实验结果中可以看出,本文提出的组合特征方法 F 值比上述两个系统均有所提高,达到了目前该领域内的先进水平,也验证了本文提出的特征组合方法在中文语义角色标注上的整体有效性.

3 结束语

本文提出了一种基于特征组合和 SVM 分类器的中文语义角色标注方法.该方法以句法成分为语义角色的基本单元,以短语结构句法分析结果为基础,综合了现有的语义角色标注系统所采用的特征,共有谓词、句法成分、上下文 3 大类 26 个基本特征.然后采用一种基于统计的组合特征选择方法,根据各个特征在语料库中的分布状况,利用类间距离和类内距离之比标准化后的值作为衡量标准,根据设定阈值快速、有效地筛选出适于语

义角色识别和语义角色分类的组合特征,然后将这些组合特征和基本特征整合成为分类特征集合,构造相应的特征向量,利用 SVM 分类器进行学习和预测.最后,在 CPB 标准语料库上的实验结果表明了该方法的有效性,整体的 F 值达到了 91.81%.在语义角色识别和语义角色分类两个子任务上,F 值均有所提高,语义角色分类阶段的 F 值提高较为明显,提高幅度将近 2%.综上所述,本文提出了一种基于特征组合的中文语义角色标注方法.该方法能够根据语料库中的特征分布情况快速、有效地构造利于分类的组合特征,在基本特征和组合特征共同构成的特征集合上,利用 SVM 分类模型提高了中文语义角色标注的性能.

References:

[1] Surdeanu M, Harabagiu S, Williams J, Aarseth P. Using predicate-argument structures for information extraction. In: Hinrichs EW, Roth D, eds. Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg: ACL, 2003, 8−15. [doi: 10.3115/1075096.1075098]

[2] Shen D, Lapata M. Using semantic roles to improve question answering. In: Eisner J, ed. Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL). Stroudsburg: ACL, 2007. 12−21.

[3] Bilotti MW, Ogilvie P, Callan J, Nyberg E. Structured retrieval for question answering. In: Kraaij W, de Vries AP, Clarke CLA, Fuhr N, Kando N, eds. Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM, 2007. 351−358. [doi: 10.1145/1277741.1277802]

[4] Fillmore CJ, Baker CF. Frame semantics for text understanding. In: Proc. of the WordNet and Other Lexical Resources Workshop (NACCL). Stroudsburg: ACL, 2001. 59−63.

[5] Carreras X, Màrquez L. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In: Dagan I, Gildea D, eds. Proc. of the CoNLL. Stroudsburg: ACL, 2005. 152−164. [doi: 10.1162/0891201053630264]

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

[7] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. In: Boitet C, Whitelock P, eds. Proc. of the ACL-Coling. Stroudsburg: ACL, 1998. 86−90. [doi: 10.3115/ 980451.980860]

[8] Schuler KK. VerbNet: A broad-coverage, comprehensive verb lexicon [Ph.D. Thesis]. Philadelphia: University of Pennsylvania, 2005.

[9] Xue N, Palmer M. Calibrating features for semantic role labeling. In: Lin D, Wu D, eds. Proc. of the EMNLP. Stroudsburg: ACL, 2004. 88−94.

[10] Gildea D, Jurafsky D. Automatic labeling of semantic roles. Computational Linguistics, 2002,28(3):245−288. [doi: 10.1162/089120102760275983]

[11] Chen J, Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments. In: Lin D, Wu D, eds. Proc. of the EMNLP. 2004. 41−48. [doi: 10.3115/1119355.1119361]

[12] Pradhan S, Hacioglu K, Krugler V, Ward W, Martin JH, Jurafsky D. Support vector learning for semantic argument classification. Machine Learning Journal, 2005,60(3):11−39. [doi: 10.1007/s10994-005-0912]

[13] Xue N. Labeling Chinese predicates with semantic roles. Computational Linguistics, 2008,34(2):225−255. [doi: 10.1162/coli.2008.34.2.225]

[14] Ding W, Chang B. Improving Chinese semantic role classification with hierarchical feature selection strategy. In: Lapata M, Ng HT, eds. Proc. of the EMNLP. Stroudsburg: ACL, 2008. 324−323.

[15] Zhao H, Chen WL, Kit C. Semantic dependency parsing of NomBank and PropBank—An efficient integrated approach via a large-scale feature selection. In: Koehn P, Mihalcea R, eds. Proc. of the EMNLP. Stroudsburg: ACL, 2009. 30−39.

[16] Boxwell SA, Dennis Mehay D, Brew C. Brutus: A semantic role labeling system incorporating CCG, CFG, and dependency features. In: Su KY, ed. Proc. of the ACL-IJCNLP. Stroudsburg: ACL, 2009. 37−45.

[17] Sun H, Jurafsky D. Shallow semantic parsing of Chinese. In: Hirschberg JB, Dumais S, Marcu D, Roukos S, eds. Proc. of the HLT-NAACL. Stroudsburg: ACL, 2004. 249−256.

[18] Xue N, Palmer M. Automatic semantic role labeling for Chinese verbs. In: Kaelbling LP, Saffiotti A, eds. In: Kaelbling LP, Saffiotti, eds. Proc. of the IJCAI. San Francisco: Morgan Kaufmann Publishers, 2005. 1160−1165.

[19] Liu T, Che WX, Li S. Semantic role labeling with maximum entropy classifier. Journal of Software, 2007,18(3):565−573 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/565.htm [doi: 10.1360/jos180565]

[20] Wang HL. Research on feature-based semantic role labeling for English and Chinese [Ph.D. Thesis]. Suzhou: Soochow University, 2008 (in Chinese with English abstract).

[21] Che WX. Kernel-Based semantic role labeling [Ph.D. Thesis]. Harbin: Harbin Institute of Technology, 2008 (in Chinese with English abstract).

[22] Liu HJ, Che WX, Liu T. Feature engineering for Chinese semantic role labeling. Journal of Chinese Information Processing, 2007, 21(2):75−80 (in Chinese with English abstract).

[23] Ding JT, Wang HL, Zhou GD, Zhu QM, Qian PD. On optimized combination of features in semantic role labeling. Computer Applications and Software, 2009,26(5):17−21 (in Chinese with English abstract).

[24] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[25] Joachims T. Making Large-Scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A, eds. Advances in Kernel Methods—Support Vector Learning. Cambridge: MIT Press, 1999.

附中文参考文献:

[19] 刘挺,车万翔,李生.基于最大熵分类器的语义角色标注.软件学报,2007,18(3):565−573. http://www.jos.org.cn/1000-9825/18/565.htm [doi: 10.1360/jos180565]

[20] 王红玲.基于特征向量的中英文语义角色标注研究[博士学位论文].苏州:苏州大学,2009.

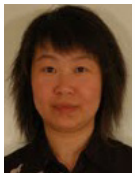
[21] 车万翔.基于核方法的语义角色标注研究[博士学位论文].哈尔滨:哈尔滨工业大学,2008.

[22] 刘怀军,车万翔,刘挺.中文语义角色标注的特征工程.中文信息学报,2007,21(1):75−80.

[23] 丁金涛,王红玲,周国栋,朱巧明,钱培德.语义角色标注中特征优化组合研究.计算机应用与软件,2009,26(5):17−21.

李世奇(1984－),男,黑龙江哈尔滨人,博士生,主要研究领域为自然语义处理,语义角色标注.

赵铁军(1962－),男,博士,教授,博士生导师,主要研究领域为机器翻译,自然语言处理,人工智能.



李晗静(1974－),女,博士,副教授,主要研究领域为自然语言处理,文景转换.

刘鹏远(1974－),男,博士,讲师,主要研究领域为词义消歧.

刘水(1981－),男,博士生,主要研究领域为句法分析.