

基于最大熵分类器的语义角色标注^{*}

刘 挺, 车万翔⁺, 李 生

(哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

Semantic Role Labeling with Maximum Entropy Classifier

LIU Ting, CHE Wan-Xiang⁺, LI Sheng

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

+ Corresponding author: Phn: +86-451-86413683 ext 806, E-mail: car@ir.hit.edu.cn, http://ir.hit.edu.cn/~car/

Liu T, Che WX, Li S. Semantic role labeling with maximum entropy classifier. *Journal of Software*, 2007, 18(3):565-573. <http://www.jos.org.cn/1000-9825/18/565.htm>

Abstract: Semantic role labeling is a feasible proposal to shallow semantic parsing. A maximum entropy classifier is used in the semantic role labeling system, which takes syntactic constituents as the labeled units. The maximum entropy classifier is trained to identify and classify the predicates' semantic roles at the same time. Some useful features and their combinations are used in the classifier. In the post-processing step, only the roles with the highest probability among the embedding ones are kept. After predicting all the arguments, which have matched the constituents in full parsing trees, a simple rule-based post-processing is applied to correct the arguments that have not matched the constituents in these trees. $F1=75.49\%$ and $F1=75.60\%$ results are obtained on the development and test set respectively. So far as it is known, this is the best result based on single syntactic parser in literatures. Finally, some proposals for solving the difficulties in semantic role labeling and the future works are given.

Key words: semantic role labeling; shallow semantic parsing; maximum entropy classifier

摘 要: 语义角色标注是浅层语义分析的一种可行方案.描述了一个采用最大熵分类器的语义角色标注系统,该系统把句法成分作为语义标注的基本单元,用最大熵分类器对句子中谓词的语义角色同时进行识别和分类.最大熵分类器中使用了一些有用的特征及其组合.在后处理阶段,在具有嵌套关系的结果中,只有概率最高的语义角色被保留.在预测了全部能够在句法分析树中找到匹配成分的角色以后,采用简单的后处理规则去识别那些找不到匹配成分的角色.最终在开发集和测试集上分别获得了 75.49% 和 75.60% 的 $F1$ 值,此结果是已知的基于单一句法分析结果中最好的.最后提出了对该任务的一些难点问题的解决方案以及对语义角色标注发展的一个初步展望.

关键词: 语义角色标注;浅层语义分析;最大熵分类器

中图法分类号: TP18 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60575042, 60503072, 60675034 (国家自然科学基金)

Received 2005-09-27; Accepted 2006-03-09

(1) 研究背景

“语义角色标注(semantic role labeling)”是“浅层语义分析(shallow semantic parsing)”的一种实现方式.所谓语义分析,指的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这个句子意义(即句义)的某种形式化表示.例如,对于句子:“张三吃了苹果”和“苹果被张三吃了”,虽然它们的表述形式不同,但表示成语义的形式就统一为:“吃(张三,苹果)”.

对句子进行正确的语义分析,一直是从事自然语言理解研究的学者们追求的主要目标.然而,经过几十年的发展,目前还没有太多使用学习的方法来获取详细语义理解知识的研究.Gildea 等人^[1]使用统计的方法进行语义角色标注的研究可谓是对语义理解的一种新的探索.

语义角色标注只是标注句子中的一些成分为给定动词(谓词)的语义角色,这些成分作为谓词的参数被赋予了一定的语义含义.例如:“[委员会 Agent][明天 Tmp]将要[通过 V][此议案 Passive].”其中:“通过”为谓词;“委员会”、“此议案”和“明天”分别是其施事、受事和发生的时间.

进行语义角色标注的基础技术如词性标注、句法分析、统计学习方法等,目前已经比较成熟.同时,语义角色标注在问答系统、信息抽取、机器翻译等领域有着广泛的应用.

(2) 语料资源

与其他基于统计的自然语言处理技术一样,进行语义角色标注也需要好的语料资源.目前,英语较为知名的语义角色标注资源为 FrameNet^[2]和 PropBank^[3]两种.

Berkeley 开发的 FrameNet 以框架语义为标注的理论基础,对英国国家语料库进行部分标注.它试图描述一个词汇单元(动词和部分名词以及形容词)的框架,同时也试图描述这些框架之间的关系.从 2002 年 6 月发布开始到现在共标注了 49 000 句.其中,每个句子都标注了目标谓词(包括动词、部分名词和形容词)和其语义角色、该角色句法层面的短语类型(如 NP,VP 等)以及句法功能(如主语、宾语等).FrameNet 现包含 1 462 个谓词.图 1 是 FrameNet 中表示身体动作的一个语义框架以及其对一个句子的标注实例.FrameNet 更像是一部语义词典.

PropBank 是 UPenn 在 Penn TreeBank 句法分析的基础上标注的浅层语义信息.与 FrameNet 不同的是,PropBank 只对动词(非系动词)进行标注,相应地被称作目标动词.而且只包含 50 多个语义角色,相同的语义角色由于目标动词不同会有不同的语义含义.其中,核心的语义角色为 Arg0~Arg5 共 6 种:Arg0 通常表示动作的施事;Arg1 通常表示动作的影响等等;其余的语义角色为附加语义角色,使用 ArgM 表示,如 ArgM-LOC 表示地点,ArgM-TMP 表示时间等等.图 2 是 PropBank 中对一个句子的标注实例.

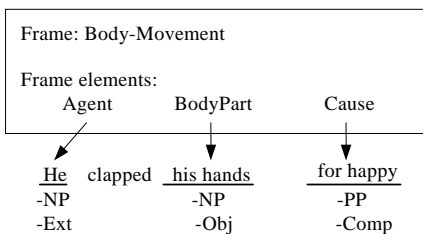


Fig.1 FrameNet and sentence annotation

图 1 FrameNet 框架以及句子标注

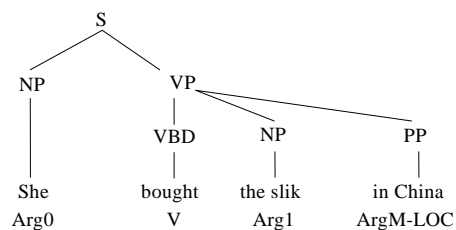


Fig.2 Propbank and sentence annotation

图 2 Propbank 及句子标注

与 FrameNet 不同,PropBank 几乎对 Penn TreeBank 中的每个动词及其语义角色都进行了标注,因此覆盖范围更广,可学习性更强.

除了英语以外,许多其他语言也建立了各自的浅层语义标注库.

例如:SALSA^[4]是基于 FrameNet 标注体系大量标注的德语语料库;Prague Dependency Treebank*项目进行大量的句法和语义标注(捷克语),甚至包括指代消解的标注等.Chinese PropBank**是 Upenn 基于 Chinese Penn

* <http://ufal.mff.cuni.cz/pdt/>

** <http://www.cis.upenn.edu/~chinese/cpb/>

TreeBank^{***}标注的汉语浅层语义标注资源.

(3) 统计学习方法

目前,人们大多采用统计学习的方法来解决语义角色标注问题.基本思想是以句子中一定的连续词语为标注的基本单元,然后根据一定的语言学知识列出该单元的各种特征,并与该单元的语义角色类型(可能不属于任何语义角色)组成学习的实例,最后使用某种学习算法对这些实例进行自动学习,以对新的实例进行预测.

Chen 等人^[5]使用决策树 C4.5 算法进行语义角色标注的实验.但是,决策树学习方法对于处理高维问题的效果并不理想.而近年出现的随机森林算法是对决策树算法的一种改进,Nielsen 等人将其应用于语义角色标注任务^[6].基于支持向量机(SVM)的语义角色标注系统获得了较好的结果^[7].然而,支持向量机的效率非常低.为加快学习速度,人们引入在线学习(错误驱动)算法.Winnow 和 Perceptron 及其各种变形算法等是常用的在线学习算法,也被成功地应用于语义角色标注中^[8,9].

(4) 标注单元

语义角色标注的基本单元可以是句法成分^[7]、短语^[10]、词^[7]或者依存关系^[11]等.在图 2 的句法分析树中,每个非终结节点如 S, NP 等都是句法成分.一般认为每个语义角色是与某一句法成分相对应的.如在图 2 的例子中,Arg0 对应一个 NP, ArgM-LOC 对应一个 PP 等.然而,我们很难自动获得这种深层句法分析的结果,尤其是除了英语以外的其他语言,而且,现有的句法分析系统在通用领域的表现也不尽如人意.

为此,有人试图将语义角色标注建立在浅层句法分析的基础上,毕竟浅层句法分析的鲁棒性要好于深层句法分析.虽然浅层句法分析不能获得全部的句法分析结果,只能获得非嵌套短语的信息,但是我们一般认为,一个非嵌套短语属于同一语义角色,因此产生了使用短语作为语义角色标注的基本单元的系统.词是比短语更细的语言单元,有些系统也使用词作为标注的基本单元,然而效果并不理想.

以上的方法都是建立在短语结构句法分析方法的基础上的.Hacioglu^[11]使用依存句法分析结果进行语义角色标注,也可以取得与基于短语结构句法分析相似的效果.

(5) 特征选择

Pradhan 等人^[7]总结了目前常用的基本特征,包括谓词、路径、短语类型、位置、语态、中心词、动词子类框架等.这些特征都从不同的侧面反映了待标注单元的语义角色信息.但是,这些仅仅反映了待标注单元的局部信息.一些全局特征,如标注单元语义角色的序列、全句含有目标动词的数目等,也可以作为特征加入.

(6) 标注步骤

语义角色标注一般分为 4 个阶段,即首先过滤掉不可能成为语义角色的句法成分,然后识别出哪些标注单元是某一目标动词的语义角色;接着将确定为语义角色的单元进行具体语义角色的分类;最后,对标注的语义角色进行推理,找到合理的角色组合.也有些系统将某些步骤进行合并,如合并识别和分类步骤,直接对每个单元进行语义角色的分类,即非语义角色也作为一个类别.

(7) 评价方法

在角色识别阶段,通常采用信息检索中的准确率(precision)、召回率(recall)和 F-Score 来评价系统的性能.而在角色分类阶段,由于此时输入的都是正确的语义角色,我们需要做的仅是将其划分成不同的类别,仅使用分类的精确率(accuracy)就可以衡量分类的效果.最终整体考虑识别和分类的结果时使用整体的准确率、召回率和 F-Score.

对于语义角色标注,国际上在 2004~2005 年举行过 3 次评测,分别为 Senseval-3^{****}以及 CoNLL(Conference on Computational Linguistics Learning)会议主办的 SRL(semantic role labeling) Shared Task 2004^[12], 2005^[13].其中:Senseval-3 是以 FrameNet 为训练和测试语料;而 CoNLL SRL Shared Task 则是以 PropBank 为语料.

*** <http://www.cis.upenn.edu/~chinese/>

**** <http://www.cs.unt.edu/~rada/senseval/senseval3/workshop.html>

1 系统描述

1.1 数据资源

我们使用 CoNLL2005 评测提供的数据作为训练、开发和测试集.主要来自以华尔街日报(WSJ)为标注来源的 PropBank.其中:Section 02-21 为训练集;Section 24 为开发集;测试集除了 Section 23 以外,还包括一部分来自 Brown 语料的标注结果.另外,此次评测还提供了两个完全句法分析的结果,分别来自 Collins^[14]和 Charniak^[15].此外,Chieu 和 Ng^[16]还提供了命名实体自动识别的结果.

1.2 标注单元的选择

我们采用句法成分作为标注的单元,主要是因为英文中,完全句法分析技术相对成熟,结果较为可信.然而,毕竟句法分析效果并不完美,所以,不可能每一个角色都能够在句法分析树中找到与之匹配的句法成分.据统计,在训练集中,大约 10%的语义角色找不到与之相匹配的句法成分.表 1 中列出了前 5 个最不匹配成分的角色类型.

Table 1 Top-5 arguments with no matched constituents
表 1 前 5 个最不匹配成分的角色类型

Args	Cha parser	Col parser	Both
AM-MOD	9 179	9 205	9153
A1	5 496	7 273	3822
AM-NEG	3 200	3 217	3185
AM-DIS	1 451	1 482	1404
A0	1 416	2 811	925

这里,Charniak 句法分析器有 10.08%的角色不匹配,而 Collins 句法分析器为 11.89%.因此我们看到,Charniak 句法分析器在语义角色标注问题上性能优于 Collins 句法分析器.所以,我们采用 Charniak 分析器产生的句法分析结果作为语义标注的输入.

经过对训练数据的分析我们看到:对于某些角色(如 AM-MOD 和 AM-NEG),这两个分析器都得到了大量不匹配的结果.同时发现:使用一些简单的后处理规则就能够提高这些角色的准确率,而其他角色的不匹配现象是由于句法分析器造成的.在第 2.3 节,我们对比了使用后处理规则和不使用后处理规则的结果.

1.3 标注步骤

由于最大熵模型的速度很快,并且分类时类别数量对最大熵的分类效率影响不大,因此,我们采用一步到位的方法对与谓词相关的全部句法成分进行预测,看其属于哪类语义角色,不属于任何角色的成分被赋予空类别.之后,再使用规则进行后处理.

1.4 最大熵分类器

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^[17].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,而不用顾及它们之间的相互影响.另外,相对 SVM 等基于空间距离的分类方法,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,便于后续推理步骤使用.同时,最大熵的训练效率相对较高.上述优点使其成功应用于信息抽取、句法分析等多个自然语言处理领域.

在预测一个句法成分是否为某一语义角色过程中会涉及各种因素,假设 X 就是一个由这些因素构成的向量,变量 y 的值为语义角色类型, $p(y|X)$ 是指系统对某个句法成分预测为某一语义角色的概率.这个概率可以用上述思想来估计.最大熵模型要求 $p(y|X)$ 在满足一定约束的条件下,必须使下面定义的熵取得最大值:

$$H(p) = -\sum_{X,y} p(y|X) \log p(y|X).$$

这里的约束条件实际上就是指所有已知的事实,一般可以用以下的方式来表述:

$$f_i(X, y) = \begin{cases} 1, & \text{if } (X, y) \text{ satisfies certain condition} \\ 0, & \text{else} \end{cases}, i=1,2,3,\dots,n.$$

称 $f_i(X, y)$ 为最大熵模型的特征, n 为所有特征的总数. 可以看到, 这些特征描述了向量 X 与语义角色 y 之间的联系. 最终概率输出为

$$p \times (y | X) = \frac{1}{Z(X)} \exp \left(\sum_i \lambda_i f_i(X, y) \right),$$

$$Z(X) = \sum_y \exp \left(\sum_i \lambda_i f_i(X, y) \right),$$

其中, λ_i 是每个特征的权重.

我们使用张乐的最大熵模型工具包****和带有高斯先验平滑的 L-BFGS 参数估计算法^[18].

1.5 特征选择

我们选取了表 2 所列特征用于语义角色的分类, 这些特征主要参考了 Pradhan 等人的工作^[7]并加以改进.

Table 2 Features list

表 2 特征列表

句法成分相关特征	
短语类型	
中心词词干	Collins 在其学位论文 ^[14] 中描述了识别一个成分中的中心词的规则. 而对于介词短语, 我们可以使用短语中的最后一个名词来代替传统的中心词. 最后采用基于规则的方法 ^[19] 对一个成分中的中心词提取词干.
最后一个词的词干	一个成分中最后一个词的词干
最后一个词的词性	
命名实体	如果一个成分以命名实体结尾, 则把命名实体的类型作为特征, 类型包括 3 种: 地点 (LOC)、组织 (ORG)、人物 (PER)
中心词的词性	
句法成分前一个词	
句法成分后一个词	
句法成分前面第 2 个词	
句法成分后面第 2 个词	
谓词相关特征	
谓词原形	
语态	我们用规则识别谓词是主动语态还是被动语态, 即谓词的词性为 VBN, 并且前面有 AUX 则为被动语态.
子类框架	谓词所在 VP 的子类框架, 如在图 2 中, 子类框架为 VP->VBD, NP, PP
谓词的词性	
谓词的后缀	我们使用谓词的最后 3 个字母
谓词-句法成分关系特征	
路径	句法树中, 从句法成分到谓词之间的句法路径. 如在图 2 中, She 所在 NP 到谓词的路径为 NP↑S↓VP↓VBD.
位置	句法成分和谓词之间的位置关系, 有两个取值: “前”和“后”. 对于“覆盖”的情况, 我们采用启发式规则直接将其忽略, 因为这种情况下, 句法成分不可能成为谓词的角色.
路径长度	一个成分和它的谓词之间的长度
部分路径	从一个成分到这个成分与它的谓词的最近共同祖先之间的路径
从句层级	在一个成分和它的谓词之间的路径上的从句的数量
谓词到句法成分的路径上 VP 个数	

由于最大熵分类器不能自动地对特征进行组合, 因此, 我们使用上述一些特征的组合来构造组合特征, 对语义角色标注有潜在帮助的组合特征被逐一叠加起来. 同时, 我们去掉那些在实际实验中造成性能下降的组合特征. 最后保留了如下的组合特征:

**** http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

- 位置+语态;
- 路径长度+从句层级;
- 谓词+路径;
- 路径+位置+语态;
- 路径+位置+语态+谓词;
- 中心词词干+谓词;
- 中心词词干+谓词+路径;
- 中心词词干+短语类型;
- 从句层级+位置+谓词;
- 位置+最后一个词的词性;
- 句法成分前一个词+第 2 个词;
- 句法成分后一个词+第 2 个词.

1.6 后处理阶段

对于一个成分嵌套在另一个成分中的情况,上面描述的系统有可能对这两个成分都进行了标注.但是,根据语义角色标注的规范,这是不允许的.因此,我们的做法是在嵌套关系发生时,只保留一个角色.因为最大熵分类器能够很容易地预测每一个输出的概率,于是,我们只把嵌套关系中概率最高的成分标注出来.

在对能够与成分完全匹配的角色进行标注之后,我们需要解决第 2.2 节中描述的那些很少能够找到与之匹配的角色的成分,例如:AM-MOD,AM-NEG 和 AM-DIS 等.为此,我们给出了如下一些简单的后处理规则:

- 用 V 标注目标动词及其后面的小品词;
- 将目标动词组块中的 not 和 n't 标注为 AM-NEG;
- 将目标动词组块中的情态动词,例如带有 MD 词性的词,标注为 AM-MOD;
- 将处于一个包含目标动词的从句之首且带有 CC 和 RB 词性的词标注为 AM-DIS.

2 实验及讨论

2.1 数据和评价方法

如第 2.1 节介绍,我们是用 CoNLL2005 评测提供的实验数据,数据来自 PropBank 以及 Brown 语料.其中:PropBank 的 Section 02-21 为训练集,共 39 832 句;Section 24 为开发集,共 1 346 句;测试集包括 Section 23 的 2 416 句以及 Brown 语料 3 个单元共 426 句.测试结果用主办方提供的 srl-eval.pl^{*****}脚本进行评测,评测指标为准确率、召回率和 F1 值.

2.2 实验结果

上面描述的全部特征及其组合均已被应用到我们的系统中.表 3 给出了我们的系统在开发集和测试集上取得的最佳性能.从测试结果中可以看出,该系统在 Brown 语料库上的性能比在华尔街日报语料库上的性能差很多,其中的原因是,句法分析器是在华尔街日报语料库上构建的,应用到新的语料库上性能会下降.

最终设定高斯先验值为 2,并在最大熵工具包中使用 1 000 次迭代.在 P4 2.8G 的 CPU 和 2G 内存的计算环境下,对所有训练集训练时间为 20 小时,训练集中包含了 39 832 个句子,150 多万样本和 250 多万特征.在开发集上测试时间为 160 秒.

CoNLL 2005 Shared Task 最佳系统^[9]的性能见表 4.我们系统的最终 F1 性能与之相差不到 3 个百分点.究其原因,主要是因为此次参赛成绩较高的系统均采用多句法分析结果相结合的策略,这在提高系统性能的同时,势必会给训练以及预测过程增加较多的工作,使其系统变得繁琐且不易扩展.而在采用单一句法分析结果的系统

***** <http://www.lsi.upc.es/~Esrlconll/srl-eval.pl>

中,我们目前报告的性能已经超过了 Surdeanu 和 Turmo^[20]报告的结果(见表 5),它是此次评测中使用单一句法分析的最佳系统.

Table 3 Overall results

表 3 整体实验结果

	Precision (%)	Recall (%)	F1
Development	79.65	71.34	75.49
Test WSJ	81.30	73.37	77.13
Test Brown	71.65	60.36	65.52
Test WSJ+Brown	80.02	71.65	75.60

Table 4 The best result on CoNLL 2005 Shared Task

表 4 CoNLL 2005 Shared Task 最佳系统性能

	Precision (%)	Recall (%)	F1
Development	80.05	74.83	77.35
Test WSJ	82.28	76.78	79.44
Test Brown	73.38	62.93	67.75
Test WSJ+Brown	81.18	74.92	77.92

Table 5 The best result on CoNLL 2005 Shared Task with single parser

表 5 CoNLL 2005 Shared Task 中使用单一句法分析结果的最佳系统性能

	Precision (%)	Recall (%)	F1
Development	79.14	71.57	75.17
Test WSJ	80.32	72.95	76.46
Test Brown	72.41	59.67	65.42
Test WSJ+Brown	79.35	71.17	75.04

2.3 后处理

另外,我们还对后处理规则进行了测试,在开发集上使用后处理规则后,F1 值从 71.36% 提高到 75.49%。可见,我们的后处理规则是非常有效的。

2.4 性能曲线

因为训练语料足够大,所以能够获得学习曲线,以便估计用多少训练数据才是充分的。我们把训练数据集中的 20 个 Section 划分为 5 部分,每部分 4 个 Section,每个 Section 大约 8 000 个句子。图 3 显示出性能随着训练集合大小的变化而变化的曲线。可以看到,当全部训练数据都被用上的时候,系统的性能最好。

随着训练集的增长,系统的性能不断提高,提高的比率有逐渐放缓的趋势。因此,在当前状态下,训练数据的增长对系统的性能提高仍有帮助。

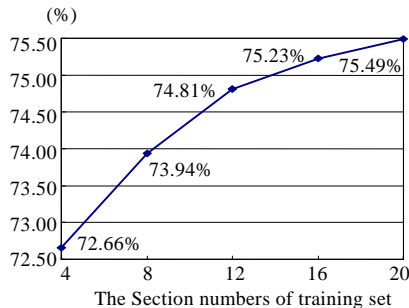


Fig.3 Performance curve effecting of training set

图 3 随测试集大小变化的性能曲线

3 结论及展望

我们描述了一个采用最大熵分类器的语义角色标注系统。该系统以句法成分为标注单元。最大熵分类器的

训练速度很快,这使得我们能够将识别和分类两个任务在一个阶段完成.本文使用一些有用的特征及其组合,对于具有嵌套关系的成分,只保留概率最大的.在预测了全部能够在句法分析树中找到匹配成分的角色之后,我们采用简单的后处理规则去识别那些找不到匹配成分的角色.基于句法成分的方法依赖于句法分析的性能,对 WSJ 语料和 Brown 语料的对比测试也说明了这一点.

我们今后的工作拟在以下方面继续深入.

3.1 克服数据稀疏问题

目前,语义角色标注的研究刚刚起步,可利用的资源还很有限,尤其是汉语等语言的语义角色标注,可供学习的语料库非常有限.因此,与其他自然语言处理问题相比,语义角色标注遇到的数据稀疏问题更加严重.另外,虽然可以利用自动句法分析的结果进行语义角色标注,但是,基于自动句法分析的效果一定不如人工加工的标准分析效果好.同时,目前的句法分析器在有些语料库,如 Penn TreeBank 上表现得较为理想,而在解决新领域中的问题时,就表现得差强人意了,因此,基于句法分析的语义角色标注的性能在新领域必然也会下降.

通过分析现有系统以及对语义角色标注任务的深入认识,我们认为:要想解决数据稀疏问题,必须对词语进行适当的聚类.例如,“我爱吃苹果”与“我爱吃香蕉”中的“苹果”和“香蕉”都是“吃”的受事,也就是它们的语义角色相同.如果能将类似的词语进行聚类,就可以在一定程度上克服数据稀疏问题,提高系统的泛化能力.在此,可以使用词义消歧技术.基于此思想,我们也可以试图寻找泛化能力更强的特征,以提高系统的召回率,同时提高系统的性能.

3.2 新的分析模型

语义角色标注规定对于一个目标动词,不允许出现重复的语义角色.同时,各种语义角色之间也存在着一定的约束关系,如受事通常出现在主动形式的动词之后等.目前,大多数系统并没有考虑这一点,它们只是对每个句法成分进行分类,除了不允许出现词语覆盖以外,没有考虑其他的约束条件.为了解决此问题,通常有两种做法:首先是将语义角色的约束关系直接加入到模型之中,使用序列模型解决语义角色标注问题,此方法多被用于词性标注、信息抽取等领域.另外,也可以采用重排序技术,即首先使用无约束的方法列出前 n 个最好的分析结果,然后利用约束条件对这 n 个结果进行重新排序,此方法在句法分析领域取得了较好的效果^[21].

3.3 汉语语义角色标注

不同于英语,汉语的词尤其是动词,有着灵活多样的用法.因此,汉语的浅层句法分析遇到的困难和问题也比英语多.尤其是目标动词不容易确定,因为汉语的动词用法灵活而且没有形态的变化,例如:“驳回/v 了/u 他/r 的/u 上诉/v 请求/v ./w”,其中有 3 个动词,而只有第 1 个动词“驳回”是目标动词.另外,由于汉语自动分析的基础模块,如分词、词性标注、句法分析等尚不成熟,因此,对于完全自动的汉语语义角色标注准确率目前还很低^[22].可以说,对汉语语义角色标注的研究还任重而道远.

如本文开始所述,人们对于自然语言理解的最终目标是真正深层次的语义分析,以期进行自动的知识获取、推理等.因此,在语义角色标注的基础上进行深层次的语义分析,将成为未来研究的重点.为达到此目标,我们必须建立性能更加卓越的语义角色标注系统.

References:

- [1] Gildea D, Jurafsky D. Automatic labeling of semantic roles. *Computational Linguistics*, 2002,28(3):245-288.
- [2] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project. In: Boitet C, Whitelock P, eds. *Proc. of the ACL&Coling'98*. Montreal: ACL, 1998. 86-90.
- [3] Palmer M, Gildea D, Kingsbury P. The Proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 2005, 31(1):71-106.
- [4] Erk K, Kowalski A, Pado S, Pinkal M. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In: Hinrichs EW, Roth D, eds. *Proc. of the ACL 2003*. Sapporo: ACL, 2003. 537-544.

- [5] Chen J, Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments. In: Hinrichs EW, Roth D, eds. Proc. of the EMNLP 2003. Sapporo: ACL, 2003. 41–48.
- [6] Nielsen RD, Pradhan S. Mixing weak learners in semantic parsing. In: Lin D, Wu D, eds. Proc. of the EMNLP 2004. Barcelona: ACL, 2004. 80–87.
- [7] 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.
- [8] Carreras X, Màrques L, Chrupala G. Hierarchical recognition of propositional arguments with perceptrons. In: Ng HT, Riloff E, eds. Proc. of the CoNLL 2004. Boston: ACL, 2004.106–109.
- [9] Punyakanok V, Koomen P, Roth D, Yih W. Generalized inference with multiple semantic role labeling systems. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL, 2005. 181–184.
- [10] Hacioglu K, Pradhan S, Ward W, Martin JH, Jurafsky D. Semantic role labeling by tagging syntactic chunks. In: Ng HT, Riloff E, eds. Proc. of the CoNLL 2004. Boston: ACL, 2004. 110–113.
- [11] Hacioglu K. Semantic role labeling using dependency trees. In: Nirenburg S, ed. Proc. of the Coling 2004. Geneva: COLING. 2004. 1273–1276.
- [12] Carreras X, Màrques L. Introduction to the CoNLL-2004 Shared Task: Semantic role labeling. In: Ng HT, Riloff E, eds. Proc. of the CoNLL 2004. Boston: ACL, 2004. 89–97.
- [13] Carreras X, Màrques L. Introduction to the CoNLL-2005 Shared Task: Semantic role labeling. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL, 2005. 152–164.
- [14] Collins M. Head-Driven statistical models for natural language parsing [Ph.D. Thesis]. Pennsylvania University, 1999.
- [15] Charniak E. A maximum-entropy-inspired parser. In: Nirenburg S, ed. Proc. of the NAACL 2000. Washington: ACL, 2000. 132–139.
- [16] Chieu HL, Ng WT. Named entity recognition with a maximum entropy approach. In: Daelemans W, Osborne M, eds. Proc. of the CoNLL 2003. Edmonton: ACL, 2003. 160–163.
- [17] Berger AL, Della Pietra SA, Della Pietra VJ. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39–71.
- [18] Chen SF, Rosenfeld R. A Gaussian prior for smoothing maximum entropy models. Technical Report, CMU-CS-99-108, 1999.
- [19] Porter M. An algorithm for suffix stripping. Program, 1980,14(3):130–137.
- [20] Surdeanu M, Turmo J. Semantic role labeling using complete syntactic analysis. In: Knight K, Ng HT, Oflazer K, eds. Proc. of the CoNLL 2005. Ann Arbor: ACL, 2005. 221–224.
- [21] Collins M, Duffy N. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In: Proc. of the ACL 2002. Philadelphia: ACL, 2002. 263–270.
- [22] Xue N, Palmer M. Automatic semantic role labeling for Chinese verbs. In: Charniak E, Lin DK, Kaelbling LP, Saffiotti A, eds. Proc. of the IJCAI 2005. Edinburgh: Professional Book Center, 2005. 1160–1165.



刘挺(1972 -),男,黑龙江哈尔滨人,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息检索.



李生(1943 -),男,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.



车万翔(1980 -),男,博士生,讲师,主要研究领域为自然语言处理.