基于句法分析的中文语义角色标注实现*

冯娟娟　李晗静　李生

哈尔滨工业大学机器智能与翻译实验室 哈尔滨 150001

E-mail: jjfeng@mtlab.hit.edu.cn

摘　要：本文提出了一种面向三维场景生成的中文语义角色标注方法。首先构造了面向场景生成的中文语义信息语料库，在句法分析的基础上对该语料中出现频率较高的六种语义角色，应用最大熵模型，对语义角色进行了标注。在基础特征空间上整体 F 值达到 60.185%；在扩展特征空间上，整体 F 值达到 61.027%。使用了后处理规则，整体 F 值提高到 63.862%，表明通过使用句法分析的结果，结合扩展特征空间和后处理规则后的中文语义角色标注的方法取得了较好的效果。

关键词：语义角色标注，句法分析，最大熵模型

Chinese Semantic Role Labeling Based on Syntactic Analysis

Feng Juanjuan, Li Hanjing, Li Sheng

Machine Intelligence and Translation Laboratory, Harbin Institute of Technology, Harbin 15000

E-mail: jjfeng@mtlab.hit.edu.cn

Abstract: In this paper, we address a method for Chinese semantic role labeling based on the 3D scene generation. Firstly, construct the corpus orienting scene generation, which contains Chinese semantic information, and then we select the top 6 semantic roles as the targets to do experiments with syntactic analysis, using maximum entropy model. Under the basic feature space, the total value of F is 60.185%. After adding the extension feature space, this value goes up to 61.027%. Secondly, we propose some post process rules, and the total value of F is 63.862%. The results show that the performance of this method for Chinese semantic role labeling using syntactic analysis has a significant increase, after adding both extension features and post process rules.

Keywords: semantic role labeling, syntactic analysis, maximum entropy model.

1　引言

语义角色标注也称浅层语义分析，是语义分析的一种主要实现方式[1]。采用“谓语-角色”的结构形式，标记句子中的成分作为给定谓语的语义角色，包括施事、受事、时间、地点等。例如，“乌鸦站在大树上”，标注语义角色后的结果为“[乌鸦 ARG0][站 V][在大树上 ARGM-LOC]”。其中，“站”是目标动词，“乌鸦”是“站”的施事者，“在大树上”是动作发生的地点。

根据文献[2]，在语义角色的识别和分类中，主要有两类学习方法：基于特征的统计机器学习方法；基于 Kernel 的学习方法。基于特征的机器学习方法是指根据一定的语言学知识给出待标注单元的特征空间和所属的语义角色类型，从而用一个特征向量表示这个单元。2005 年，Xue 等人[3]采用句法成分作为标注单元，选取 CPB 中的 760 篇文档，包括训练集 661 篇，测试集 99 篇，使用位置，路径，谓语，短语类型等共 10 类特征，在测试集的 Precision 和 Recall 分别达到了 81.83%，82.91%。基于 Kernel 的学习方法是将低维线性不可分的问题映射到高维空间，变成线性可分问题。Moschitti[2]提出了利用句法分析树来计算 Kernel 函数来进行语义角色标注。

* NSFC60803094, HITQNJS.2008.050。

语料库是语义角色标注的基础，中文的浅层语义标注语料以 Chinese PropBank(CPB)最具代表性。Chinese PropBank 是 Upenn 基于 Chinese Penn 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。目前可使用的Chinese PropBank 语料是对 760 篇新闻语料进行了手工标注，包含 10384 个句子，4900 个动词，其中每个动词根据语义和用法不同定义为若干个框架(frame)，每个框架内包含的角色类型可能不同。CPB 的语义角色与 PropBank 类似，它包含 20 多个语义角色，其中核心的语义角色为 ARG0-5，其余的为附加语义角色，用 ARGM 加上附加标记来表示。

语义角色标注有着广泛的前景，在问答系统，信息抽取[4]等领域都有着广泛的应用。目前，我们主要研究如何从自然语言文本到相应的三维动态虚拟场景的转换，在该转换中，语义角色标注是基础工作之一。例如“乌鸦站在大树上”，通过寻找动词“站”对应的语义角色，可以为后端动画的生成提供必要的动作信息。因此，以这个任务为背景，我们提出了基于句法分析的语义角色标注，采用基于特征的机器学习方法。

本文第二部分主要介绍实验所需的语料库及其构建；第三部分重点描述用于语义角色标注的特征空间，包括基础特征空间和扩展特征空间，以及标注步骤；第四部分分析了系统的实验结果；第五部分对本文进行总结，并对后期工作进行展望。

2 语料库的构建

我们使用的语料是《伊索寓言》，由于国际上没有针对本研究领域的标注语料，在参考了 CPB 的标注规范后，仿照其标注形式，构建了《伊索寓言》语料库。标注的步骤是：选择待标注单元；确定要标注的语义角色类型；确定待标注的典型动词；仿照 CPB 的标注形式完成标注。

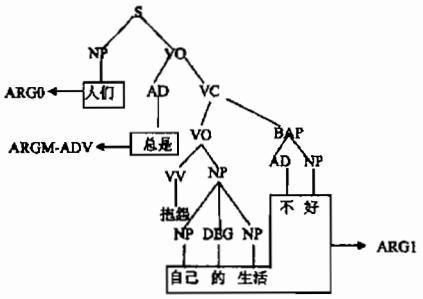

图 1 《伊索寓言》中一个句子的标注实例

我们将句法成分作为待标注单元，利用实验室开发的 Parser 工具[5]，加工《伊索寓言》，如图 1。由于语料中动词的语义角色大部分是施事、受事、时间、地点等，因此选取 ARG0，ARG1，ARG2，ARGM-LOC，ARGM-TMP，ARGM-ADV 作为待标注的语义角色类型。将《伊索寓言》中频率出现最高的前 56 个动词作为标注对象，见附表 1。图 1 中“抱怨”，标注后的形式为：399 139 抱怨 137--ARG0 140:144--ARG1 138--ARGM-ADV 139—rel，以上各项的意义依次为：文章号 动词编号 动词本身 ARG0 词语的编号--ARG0 ARG1 词语的开始编号:结束编号--ARG1 ARGM-ADV 词语的编号--ARGM-ADV 动词的编号--rel；以上所有编号都是词语在一篇文章中的位置编号，从 1 开始。

3 语义角色标注系统

3.1 基础特征空间

在 Xue 等人[3,6]的工作中，提出了很多有效的特征，具体说明如下：

(1) 句法成分与谓语动词的位置关系：这是一个三值特征，0 表示在谓语动词前面，1 表示在谓语动词后面，2 表示不能确定与谓语动词的位置关系。

(2) 句法成分到谓语动词的路径：如图 1，ARG0 对应的节点 NP，其路径为 NP[up]S[down]VO[down]VC[down]VO[down]VV。

(3) 中心词特征：根据文献[7]，统计语义角色中频率排名前 20 的中心词及其词性，见附录 2。

(4) 谓语：按照文献[3]对于动词分类原则，将 56 个动词进行分类。

(5) 子类框架：谓语动词父节点及其子节点。如图 1，“抱怨”的子类框架是 VO-VV-NP。

(6) 短语类型：加工后的《伊索寓言》共有短语类型 31 个。但是其中包含了很多含义相近的类型，如图 1 中的 VO，VC，参照 CPB 中给出的短语类型，将这些动词短语类型合并，统一标记成 VP，其他的短语类型也作类似处理。经过合并，共计包含短语类型 16 个。

(7) 句法成分的左右兄弟节点的类型。

(8) 句法框架：句法框架特征包含围绕动词和围绕谓语动词的名词短语，如图 1 中的句法成分 ARG0 对应的节点 NP 的句法成分是 NP-VV-NP-NP。

3.2 扩展特征空间

根据文献[8]，列出了一些新增特征，定义为扩展特征空间。共包括 4 类，即路径上所包含的子句数；路径上的名词短语数；句法成分和谓语动词的最近祖先；句法成分和谓语动词的位置关系：是否满足兄弟关系；是否属于同一动词短语的儿子节点；是否属于同一子句的儿子节点。

3.3 标注步骤

语义角色标注系统共包含 4 个步骤：剪枝、识别、分类和后处理。剪枝是指根据启发式规则，删除大部分不可能成为语义角色的句法成分。根据文献[6]，得到启发式规则，见公式(a)。

$$\begin{aligned} &\text{if} \quad \exists N_1, \text{满足} N_1 \in \{\text{Neighbor(N)}\} \text{且} \text{Node\_Type}(N_1)==\text{PP} \\ &\text{then} \quad \text{Arg\_Set(P)}=\text{Arg\_Set(P)} \cup \{\text{Children}(N_1)\} \end{aligned} \tag{a}$$

并作如下定义：P 表示目标谓语动词，N 表示当前句法成分所在的节点，Node_Type(N)表示节点 N 的类型，Children(N)表示节点 N 的所有孩子节点，Neighbor(N)表示节点 N 的邻居节点，Arg_Set(P)表示谓语动词 P 的所有待标注单元的集合，Parent(N)表示节点 N 的父节点，Root(P)表示谓语动词 P 所在的句法分析树的根节点。基于规则(a)，剪枝算法步骤如下：

(1) 在句法分析树中选取目标谓语动词作为当前节点 N

(2) 对于$\forall N_2 \in \{\text{Neighbor(N)}\}$，$\text{Arg_Set(P)} = \text{Arg_Set(P)} \cup \{N_2\}$

(3) 运用启发式规则(a)

(4) if $\exists N_3$满足$\text{Parent(N)} == N_3$ then $N=N_3$, goto(2); Utill $N==\text{Root(P)}$

识别是指在具体分析每个语义角色之前，首先判断待标注单元是否为语义角色，用二元分类器把候选角色分为语义角色和非语义角色。分类是指将上一阶段识别出的语义角色分到对应的类别中，本文我们使用最大熵分类器[9]实现对语义角色的分类。

根据《伊索寓言》中动词的特点，我们制定了如下后处理规则：

1. 限定动词作用域。把手工标注(见附表1)的56个动词分为两类：说，说明，回答，告诉，有，属于第一类，作用域是整个句法分析树；其余动词属于第二类，作用域是它所在的子句。

2. 确定语义角色对应的节点。语料中很多语义角色是由若干个非叶节点的叶节点组成，如图1，“抱怨”的ARG0是NP，但是ARG1对应的内容是“自己的生活不好”，但却不存在一个节点，使得以它为根节点的子树的所有叶节点中当且仅当包括“自己”，“的”，“生活”，“不”，“好”。为了使识别出的句法成分尽可能与语义角色对应的实际内容相匹配，采用了如下策略：

(1) 寻找语义角色的候选节点，满足以该节点为根的子树的所有叶子节点中尽可能多地包含“自己”，“的”，“生活”，“不”，“好”。这里选出的候选节点为：NP，BAP，VC，VO；

(2) 从上述的候选节点中删除异常节点，删除的原则是：以该节点为根的子树的所有叶子节点中包含了“自己”，“的”，“生活”，“不”，“好”以外的叶子节点，如，候选节点VC和VO中后包含了“抱怨”这个叶子节点，删除异常节点后，候选节点剩下NP和BAP；

(3) 在(2)的基础上，确定最终语义角色对应的句法成分节点，满足的条件是：以该节点为根的子树中所包含的叶子节点数目最多，则抱怨的ARG1对应的句法成分便是NP。

3. 角色合并。统计发现上述六种角色的数目相差很大，各个语义角色的数量见表1，为了优化系统的性能，我们将ARGM-ADV ARGM-LOC ARGM-TMP这三种角色都统一成ARGM。

表1 在56个动词中出现上述6种语义角色的数量

	ARG0	ARG1	ARG2	ARGM-ADV	ARGM-LOC	ARGM-TMP
训练语料	1708	1742	239	343	33	59
测试语料	595	606	65	125	10	20
总计	2303	2348	304	468	43	79

4 实验结果

实验所用的语料是《伊索寓言》，共10卷，432篇，2477句。其中训练语料324篇，1849句，测试语料108篇，628句。采用准确率(Precision)，召回率(Recall)和F-Score对系统的性能进行评价。我们共进行了三组实验，用于测试不同的特征空间和后处理规则对系统的贡献。其中BFS表示基础特征空间，EFS表示扩展特征空间，Post_Processing表示后处理规则。

表2 BFS、BFS+EFS、BFS+EFS+Post_Processing的整体性能

	BFS	BFS+EFS	BFS++EFS+Post_Processing
Precision(%)	59.633	60.467	63.25
Recall(%)	60.748	61.597	64.486
F-Score(%)	60.185	61.027	63.862

表3 BFS+EFS+Post_Processing中系统的性能

	ARG0	ARG1	ARG2	ARGM	Total
Precision(%)	65.060	65.667	51.351	33.846	63.25
Recall(%)	63.780	79.435	30.645	19.820	64.486
F-Score(%)	64.414	71.898	38.384	25	63.862

表2列出了系统在BFS、BFS+EFS、BFS++EFS+Post_Processing这三组实验中的整体性能。

实验一应用了基础特征空间，整体的F-Score是60.185%。实验二使用了基础特征空间和扩展特征空间的组合，整体的F-Score是61.027%。实验三在实验二的基础上增加了后处理规则，整体的F-Score达到63.862%。可以看出使用后处理规则能够显著提高系统的整体性能。表3给出了实验三的结果，可以看到对ARG0和ARG1识别效果较好，而对于数据非常稀疏的ARGM，通过后处理中的合并原则，也能达到一定的识别效果。

图2，3，4分别列出了在上述三组实验中，ARG0，ARG1，ARG2的Precision、Recall和F-Score。可以看到实验三对于ARG0的识别效果最好。在实验二中，尽管系统的整体性能的增加幅度不是很大，但是对于数据同样稀疏的Arg2，它的识别效果得到了较大的提升。

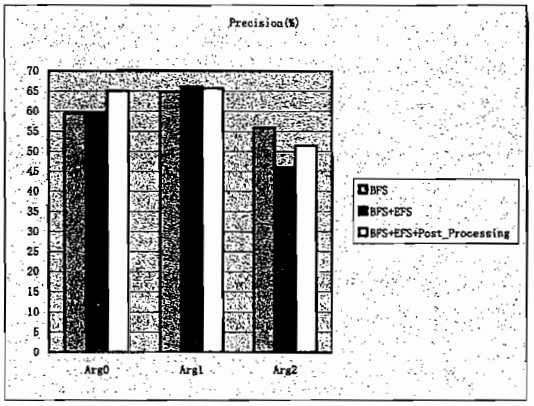

图2 ARG0~ ARG 2分别在BFS、BFS+EFS、BFS+EFS+Post_Processing中的Precision

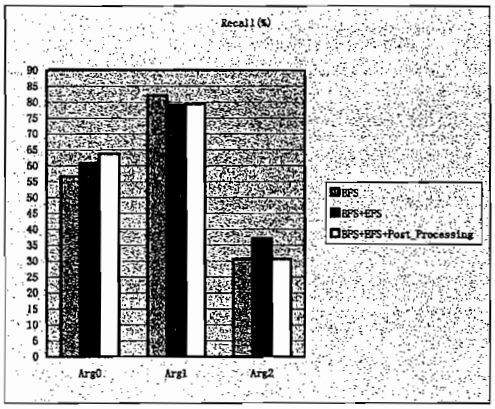

图3 ARG0~ ARG 2分别在BFS、BFS+EFS、BFS+EFS+Post_Processing中的Recall

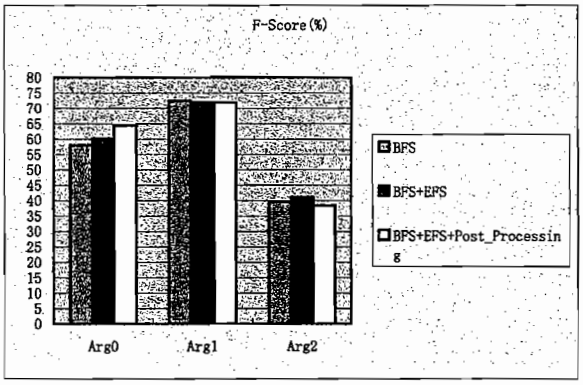

图4 ARG0~ARG 2 分别在 BFS、BFS+EFS、BFS+EFS+Post_Processing 中的 F-Score

5 结论

实验三的整体性能得到了显著的提高。这是由于在后处理规则中，我们将 56 个动词分成了两类，对于第一类动词，它们的宾语通常是一句话，而主语往往隐含的给出，因此通过限定动词的作用域以及语义角色的层次，可以有效地找到真正的语义角色。对于数据稀疏的 ARGM，通过合并，能更好的突出 ARGM 的共同特征。但是对 ARGM 和 ARG2 的识别效果较差，原因有如下几方面：在汉语里，ARG2 的用法非常灵活，对于不同的动词就有不同的含义，增加了识别

的难度；在《伊索寓言》中，ARGM-TMP，ARGM-LOC 并没有明显的语义特征，因此仅通过语言学特征，很难对上述这两类语义角色做出正确的识别。

目前我们的工作是在句法分析的结果上，利用基于特征的机器学习方法，来处理语义角色标注问题，因此下一步我们将使用基于 Kernel 的学习方法进一步改善这一领域内的中文语义角色标注系统的性能。

参 考 文 献

[1] 车万翔,刘挺,李生.自动浅层语义分析.中国中文信息学会二十五周年学术会议. 2006.

[2] Alessandro Moschitti. A study on convolution kernels for shallow statistic parsing. In proceedings of ACL-2004, 2004, pages 335–342.

[3] N. Xue and M. Palmer. Automatic Semantic Role Labeling for Chinese Verbs. In proceedings of IJCAI-2005, Edinburgh, Scotland, 2005.

[4] M. Surdeanu, S. Harabagiu, J. Williams and P. Aarseth. Using Predicate-Argument Structures for Information Extraction. In proceedings of ACL-2003, 2003.

[5] 曹海龙.基于词汇化统计模型的汉语句法分析.哈尔滨工业大学博士学位论文. 2006.

[6] N. Xue and M. Palmer. Calibrating Features for Semantic Role Labeling. In proceedings of EMNLP-2004, 2004.

[7] Daniel Collins and Daniel Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245-288.

[8] 刘怀军,车万翔,刘挺. 中文语义角色标注的特征工程. 第三届学生计算语言学研讨会. 2006, 75~80.

[9] Ting Liu, Wanxiang Che, Sheng Li, Yuxuan Hu, and Huaijun Liu. 2005. Semantic role labeling system using maximum entropy classifier. In Proceedings of CoNLL-2005, 2005, pages 189–192.

附表 1 已标注动词列表

动词	数量	动词	数量	动词	数量	动词	数量	动词	数量	动词	数量	动词	数量
说	776	说明	131	回答	80	叫	41	告诉	28	指责	17	发出	15
许诺	12	道	11	答应	11	讲	10	发誓	9	笑	9	嘲笑	10
抱怨	7	表示	6	接受	6	谈	5	显示	5	争	5	招待	5
是	542	成	15	成为	13	变成	11	作	9	作为	7	行	6
发	5	显得	5	有	343	带	34	得到	34	受	20	带来	18
忍受	18	受到	16	享受	9	获得	6	拥有	5	过	24	出去	20
离开	14	回到	11	回家	11	外出	8	经过	7	进去	6	到	220
来	184	住	62	起来	60	站	52	出	46	出来	46	拿	43

附表 2 语义角色中出现的频率最高的前 20 个中心词及其词性

词/词性	词频	词/词性	词频	词/词性	词频	词/词性	词频	词/词性	词频
这/代词	666	一/数词	119	人/名词	77	人们/名词	49	那/代词	44
他/代词	326	在/介词	84	狐狸/名词	75	狮子/名词	48	个/量词	42
我/代词	194	对/介词	78	那些/代词	69	她/代词	48	他们/代词	41
你/代词	186	不/副词	77	有些/代词	50	狼/名词	46	狗/名词	41