

基于最大熵模型的语义角色标注

于江德¹,王希杰¹,余正涛²

(1 安阳师范学院 计算机与信息工程学院,安阳 455002; 2 昆明理工大学 信息工程与自动化学院,云南 昆明 650051)

摘 要: 提出了基于最大熵模型的语义角色标注方法,该方法以浅层句法分析为基础,把短语或命名实体作为标注的基本单元,将最大熵模型用于句子中谓词的语义角色标注.该方法的关键在于模型参数估计和特征选择.具体应用中采用 IIS 算法学习模型参数,并选择基于句法成分的、基于谓词的、句法成分-谓词关系、语义四类特征作为模型特征集.将该方法用于信息抽取中事件表述语句的语义角色标注,对“职务变动”和“会见”两类事件的表述语句进行事件要素的语义角色标注,在各自的测试集上分别获得了 76.3% 和 72.2% 的综合指标 F 值.

关键词: 语义角色标注;最大熵模型;参数估计;特征选择

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)08-0173-04

Semantic Role Labeling Based on Maximum Entropy Model

YU Jiang-de¹, WANG Xi-jie¹, YU Zheng-tao²

(1 School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China;

2 School of Information Engineering and Automation, Kunming University of Technology, Kunming 650051, China)

Abstract: A method based on maximum entropy model is proposed for Semantic Role Labeling (SRL). This method takes shallow syntactic parsing as base, and takes phrase or named entity as the labeled units, and maximum entropy model is trained to label the predicates' semantic roles in a sentence. The key of the method is parameter estimation and feature selection for maximum entropy model. In this paper, the IIS algorithm was employed for parameter estimation, and four categories features: features based on sentence constituents, features based on predicate, predicate-constituent features and semantic features as features set of the model were selected. The method is used to label semantic roles in an event mention sentence for information extraction. We got $F=76.3\%$ and $F=72.2\%$ results on different test set for "management succession" and "meeting".

Key words: semantic role labeling; maximum entropy model; parameter estimation; feature selection

1 引言

近些年来,国内外在语义角色标注方面已经进行了一些卓有成效的研究.目前人们大多采用统计学习的方法解决语义角色标注问题^[1-2].使用统计学习的方法进行语义角色标注,离不开标注好语义角色的语料资源,英语中较为知名的有 FrameNet^[3]和 PropBank^[4]两种.对于汉语的语义角色标注资源,Chinese PropBank 是 UPenn 基于 Chinese Penn TreeBank 标注的汉语浅层语义标注资源.文献[2]构建的汉语框架

语义知识库(Chinese FrameNet, CFN)是一个以框架语义学为理论基础、以真实语料为事实依据的语义词典,其资源将用语义 Web 标记语言描述.

文中提出了基于最大熵模型(Maximum Entropy Model, ME)的语义角色标注方法.实验结果表明,该方法有不错的标注性能.

2 最大熵模型

2.1 最大熵理论

最大熵理论反映了自然界的一条基本原则:事

物是约束和自由的统一体,并且在约束下事物总是争取最大自由度,即最大熵.因此,在已知条件下,熵最大的事物,最可能接近它的真实状态.可以应用最大熵理论来解决现实世界中的许多问题,这些问题都是在已经知道一些先验知识情况下,需要选择一个合适的模型对问题做出预测,无疑最大熵模型是解决此类问题的合适模型.具体来说,对于一个事件,往往只了解它的部分情况,对于其它情况则一无所知.那么对这个事件建立模型时,对于已知的部分要尽量地拟合,使模型符合已知情况.对于未知的情况,则保持均匀分布,即使该事件的熵最大.

2.2 最大熵模型的一般形式

对于语义角色标注等分类问题,给定一些训练样本 (x, y) ,其中 x 表示上下文, y 表示角色的类别,可根据这些已知的样本构建一个能够对实际问题进行准确描述的统计模型 $p(y|x)$ 用来预测未知的事件.该模型的概率分布与训练语料中的经验概率分布应该相符,训练样本中上下文信息与输出的经验概率分布 $\tilde{p}(x, y)$ 为:

$$\tilde{p}(x, y) = C(x, y) / \sum_{x, y} C(x, y) \quad (1)$$

其中, $C(x, y)$ 为 (x, y) 在训练样本中出现的次数.

最大熵原理表明, x, y 的正确分布应该是在满足已知条件(约束)的情况下,熵最大的分布.分布 $p(x, y)$ 的不确定性的数学度量是熵:

$$H(p) = - \sum_{x, y} p(x, y) \lg p(x, y) \quad (2)$$

条件分布 $p(y|x)$ 的不确定性的数学度量是条件熵:

$$H_t(p) = - \sum_{x, y} \tilde{p}(x) p(y|x) \lg p(y|x), \quad (3)$$

其中, $\tilde{p}$ 表示训练样本的概率分布.

研究者引入了特征函数的概念来表示已知的约束条件,特征函数一般情况下是一个二值函数 $f(x, y) \rightarrow \{0, 1\}$,形式如下:

$$f(x, y) = \begin{cases} 1 & \text{如果}(x, y) \text{ 满足某种约束} \\ 0 & \text{否则} \end{cases} \quad (4)$$

对于特征函数 f_i ,它相对于经验概率分布 $\tilde{p}(x, y)$ 的期望值为

$$E_{\tilde{p}} f_i = \sum_{x, y} \tilde{p}(x, y) f_i(x, y) \quad (5)$$

特征函数 f_i 相对于模型 $p(y|x)$ 的期望值为

$$E_p f_i = \sum_{x, y} \tilde{p}(x) p(y|x) f_i(x, y) \quad (6)$$

这样,模型的约束可以表示为特征的期望等于训练数据特征的期望,即

$$E_p(f_i) = E_{\tilde{p}}(f_i) \quad (7)$$

那么,对于 k 个特征,求解的模型集合为:

$$C = \{p \mid E_p f_i = E_{\tilde{p}} f_i, i = \{1, \dots, k\}\} \quad (8)$$

而最好的模型是使其熵最大的模型 p^* :

$$p^* = \arg \max_{p \in C} H(p) \quad (9)$$

求解这个最优解的经典方法是拉格朗日乘子算法,可以证明 p^* 满足下面的形式:

$$p(y|x) = \frac{1}{Z(x)} \exp \left[\sum_{i=1}^k \lambda_i f_i(x, y) \right] \quad (10)$$

其中

$$Z(x) = \sum_y \exp \left[\sum_{i=1}^k \lambda_i f_i(x, y) \right] \quad (11)$$

k 为特征的数目, $Z(x)$ 为归一化因子,保证对所有 x , $\sum_y p(y|x) = 1$.参数 λ_i 指示了特征 f_i 对于模型的重要程度.通过在训练集上进行学习,求出 λ_i 的值,就得到了概率分布函数,完成了最大熵模型的构造,式(10)即为最大熵模型的一般形式.

2.3 参数估计和特征选择

使用最大熵模型进行语义角色标注首先要建立模型,其中的两个关键问题是:参数估计和特征选择.参数估计是从训练集学习每一个特征的权重参数,即求解向量 $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$ 的过程.而特征选择是筛选出对最大熵模型有表征意义的特征.

2.3.1 参数估计

参数估计就是从训练数据集学习权重参数向量 Λ 的过程.传统的办法是Della Pietra等人在GIS (Generalized Iterative Scaling Algorithm)算法^[5]的基础上提出的IIS (Improved Iterative Scaling Algorithm)算法.文中采用IIS算法求解模型参数,其过程为:

输入:一个标注过的事件表述语句集和一组特征函数 $f_1, f_2, \dots, f_n$.

(1) 对所有 $i \in \{1, 2, \dots, k\}$,初始化 $\lambda_i = 0$;

(2) 对所有 $i \in \{1, 2, \dots, k\}$,重复以下过程直到所有 λ_i 收敛:

① 根据式(10)求解每个事件表述语句的 $p_{\Delta}(y|x)$;

② 对每个参数 λ_i

(a) 求出下式的解 $\Delta \lambda_i$:

$$\sum_{x, y} \tilde{p}(x) p(y|x) f_i(x, y) \exp(\Delta \lambda_i f_i^{\#}(x, y)) = \tilde{p}(f_i) \quad (12)$$

其中 $f_i^{\#}(x, y) \equiv \sum_{i=1}^n f_i(x, y)$;

(b) 根据 $\lambda_i = \lambda_i + \Delta \lambda_i$ 更新 λ_i .

输出:最优参数值 Δ ,最优模型.

2.3.2 特征选择

最大熵模型的一个主要优点是能够在同一个模型中集成不同的特征.所以,建立最大熵模型的一个关键是如何针对特定的任务为模型选择合适的特征集,用简单的特征集表示复杂的语言现象.针对语义角色标注这一任务,在详细分析输入文本的上下文信息的基础上,参照文献[6]对语义角色标注及汉语语义角色的特征分析,将特征分成四类:基于句法成分的特征、基于谓词的特征、句法成分-谓词关系特征和语义特征.

基于句法成分的特征是指句子中一个句法成分中所包含的短语、命名实体、词语所具有的类别、词性等特征以及该句法成分前后的标注单元所具有的特征,这些特征是最基本的特征.文中用到的部分基于句法成分的特征如表 1 所示.

表 1 基于句法成分的特征列表

特征	特征含义描述
PhraseType	句法成分的短语类别
NEType	句法成分包含的命名实体及类别
HeadWord	句法成分包含的中心词
POS	句法成分的词性标注
FirstWord	句法成分中的第一个词
LastWord	句法成分中的最后一个词
Prepositions	句法成分的前置词
PreviousUnit	句子中该句法成分的前一个标注单元
NextUnit	句子中该句法成分的后一个标注单元
WordNumber	句法成分中词的数目

基于谓词的特征主要指待处理的句子中谓词(目标动词)的类型、位置、词义等所具有的特征.文中用到的基于谓词的特征如表 2 所示.

表 2 基于谓词的特征列表

特征	特征含义描述
PredicatePOS	谓词词性标注
PredicatePostion	谓词的位置
PredicateVoice	谓词的语态
PredicateSense	谓词的词义

句法成分和谓词之间有各种各样的外在关系,如句法成分相对于谓词的位置等.这类特征如表 3 所示.

语义特征是该句法成分及其上下文所具有的一些语义特征,这些特征对确定语义角色有重要作用.这类特征如表 4 所示.

表 3 句法成分-谓词关系特征列表

特征	特征含义描述
Path	句法树中,从谓词到句法成分的句法路径.图 1 中谓词到 She 所在 NP 的路径为 VBD $\uparrow$ VP $\uparrow$ S $\downarrow$ NP.
PathLength	句法成分到谓词的路径长度
Position	句法成分和谓词的位置关系

表 4 语义特征列表

特征	特征含义描述
Independent	句法成分先于谓词所表示的事件独立存在
Causation	该句法成分所指的事物施行某个动作、或造成某个事件或状态
StateChange	使事物或事件状态发生变化
Affected	所指事物承受某一动词的动作、行为的影响

3 基于最大熵模型的语义角色标注

3.1 语义角色标注的一般步骤

依据上述的方法建立好最大熵模型之后,接下来就可以采用最大熵模型来进行语义角色标注了.进行语义角色标注的基本单元可以是句法成分、词、短语、命名实体等.在图 1 的句法分析树中,每个非终结点,如 NP、PP 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 1 中,Arg0 对应一个 NP,ArgM-LOC 对应一个 PP 等.然而,这种深层句法分析的结果很难自动得到的,尤其对汉语而言,深层次的句法分析还不成熟.所以,文中的语义角色标注是建立在浅层句法分析基础之上,待标注的基本单元是短语或命名实体等.

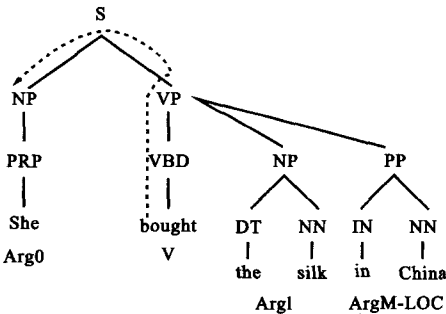


图 1 语义角色标注中路径示意图

语义角色标注一般分为四个阶段.第一个阶段首先确定目标动词;第二个阶段是过滤掉不可能成为语义角色的成分;第三个阶段是识别目标动词可支配的短语或命名实体(动词论元)的边界;第四个阶段是确定每个论元的语义角色.

3.2 事件抽取中的语义角色标注

事件抽取的过程一般包括两个阶段:第一个阶段是通过触发词探测发现候选事件所在的语句.触发词(trigger)是能够很好地概述出该事件的中心意义的词.例如,职务变动事件中的“任命”、“辞去”等词语.也是文中对事件表述语句进行语义角色标注时的谓词.第二阶段是从候选事件语句中抽取事件的事件要素,在此阶段就需要对与触发词相关联的短语或命名实体进行语义角色标注的确定.

文中将基于最大熵模型的语义角色标注方法应用到“职务变动”类事件抽取和“会见”类事件抽取中.所谓“职务变动”类事件抽取就是从相关的文本中抽取“职务变动”事件信息.也就是从自然语言形式的文本中发现“职务变动”类事件并抽取该事件的过程.该类事件的事件要素包括:事件发生的时间、人物、组织机构、职位等.“会见”类事件抽取就是从相关的文本中抽取“会见”事件信息.也就是从自然语言形式的文本中发现“会见”类事件并抽取该事件的过程.该类事件的事件要素包括:会见的时间、参与者、会见地点、持续时间等.深入分

析这两类事件的大量表述语句之后,确定对“职务变动”类事件标注的语义角色包括八个:下达任命的人(Arg0 - PER)、下达任命的组织(Arg0 - ORG)、被任命者(Start - PER)、离任者(Left - PER)、职位(Position)、事件发生时间(ArgM - TMP)、事件发生的地点(ArgM - LOC)、原因(ArgM - CAU).确定对“会见”类事件标注的语义角色包括六个:施行会见的人(Arg0 - PER)、被会见者(Arg1 - PER)、事件发生的地点(ArgM - LOC)、事件发生时间(ArgM - TMP)、原因(ArgM - CAU)、方式(ArgM - MNR).

4 实验结果及其分析

为验证文中提出的基于最大熵模型的语义角色标注方法,实验时分别对“职务变动”和“会见”两类事件在各自的训练集和测试集上进行.另外,在实验中还探讨了不同特征集对实验性能的影响.

对“职务变动”和“会见”两类事件表述语句进行语义角色标注的实验,其中最大熵模型中使用了前面提到的全部四类特征集.表5显示了语义角色标注结果.

表5 “职务变动”和“会见”类事件语义角色标注结果

语义角色	P	R	F	语义角色	P	R	F
A0 - PER	0.824	0.816	0.820	A0 - PER	0.825	0.791	0.808
A0 - ORG	0.813	0.810	0.811	A1 - PER	0.802	0.729	0.764
Start - PER	0.825	0.782	0.803	ArgM - LOC	0.767	0.785	0.776
Left - PER	0.705	0.728	0.716	ArgM - TMP	0.745	0.660	0.700
Position	0.826	0.782	0.803	ArgM - CAU	0.635	0.621	0.628
ArgM - LOC	0.738	0.725	0.731	ArgM - MNR	0.680	0.635	0.657
ArgM - TMP	0.801	0.726	0.762				
ArgM - CAU	0.620	0.705	0.660				

由于文本事件信息抽取中事件要素及其语义角色的研究相对较少,而事件要素及其角色识别问题的实质是对事件表述语句中的触发词的论元进行语义角色标注.所以可以从另一个侧面来验证本组实验的结果,即从语义角色标注来对实验结果做出评判.近几年,语义角色标注成为国内外 NLP 领域研究的热点问题,国内哈尔滨工业大学的刘挺教授、车万翔等对该问题进行了深入的研究,他们开发的语义角色标注系统在 CoNLL2005 评测提供的测试数据上评测结果的综合指标 $F(\beta$ 取值为 1) 值为 75.6%.从表5可以看到两类事件的事件要素及其语义角色识别的综合指标 F 平均值分别达到了 76.3%

和 72.2%.这个结果表明,文中的方法有较好的标注性能.

5 结束语

语义角色标注的研究正越来越受到重视,其应用范围也非常广泛.文中将最大熵模型模型用于语义角色标注,并对模型建立的两个关键问题:参数估计和特征选择进行了详细论述.并将该方法应用于事件信息抽取中“职务变动”和“会见”两类事件表述语句的语义角色标注,实验结果表明该方法有较好的标注效果.

(下转第 180 页)

的极大浪费.因此,可以利用图形化界面实现多个测试程序并行运行,充分发挥多核系统的性能优势.

在自动生成 Python 脚本时,可通过图形化界面,让 M5 用户根据需要进行多个测试程序,根据系统中可用的处理器数量,同时生成多个进行 Python 脚本,并启动多个 M5 运行实例实现测试程序的并行模拟,既充分利用本地系统资源,又提高了工作效率.此外,将多个测试程序的输出信息同时显示在图形化界面,提供了直观的对比,使得 M5 用户可以更好的分析所设计系统的性能.

文中使用双核系统对多个测试程序的并行模拟进行了简单测试,实验中 M5 采用用户级模拟模式,设定的模拟平台为 ALPHA.从韩国 SNU(Seoul National University)^[5]实时研究组的 SNU-RT 测试程序集中挑选了 3 个典型的程序,对并行测试环境进行了分析.这 3 个程序分别为 bs、fir 和 crc.bs 是一个实现 15 个整型元素数组二分检索算法的程序;fir 是实现了有限冲激响应和高斯函数的程序;crc 是一个循环冗余校验示例程序.测试结果如表 2 所示.

表 2 串行模拟与并行模拟比较

测试组	基准测试程序	测试程序耗时(s)	串行耗时(s)	并行耗时(s)	加速比
1	bs	0.61	1.64	1.02	1.61
	fir	1.03			
2	crc	1.79	2.82	1.84	1.53
	fir	1.03			
3	bs	0.61	2.40	1.81	1.32
	arc	1.79			

从表 2 中可以看出,在多核处理器上采用并行模拟的方法可以显著缩短模拟所需的时间,从而提

高了工作效率.

5 结束语

文中针对 M5 模拟器使用中出现的两个主要问题,提出一种对系统结构进行图形化配置与验证的测试方法,该方法能够根据功能部件之间的互连关系及属性配置自动生成 M5 运行所需的 Python 脚本.并可根据用户的配置和系统中可用的处理器数量,同时生成多个 Python 脚本,实现对多个测试程序的并行模拟.通过上述两种方法减少了系统结构设计者的工作量,提高了 M5 的使用效率.下一步将对该图形化辅助工具进行完善,使其能够生成针对不同模拟器的测试脚本,例如实现对 Simics 的支持.

参考文献:

- [1] Nathan L Binkert, Ronald G Dreslinski, Lisa R Hsu, et al. TheM5 simulator: modeling networked systems[J]. IEEE Micro, 2006, 26(4): 52-60.
- [2] Peter S Magnusson, Magnus Christensson, Jesper Eskilsson. Simics: a full system simulation platform[J]. Computer, 2002: 50-58.
- [3] 许建卫, 杨伟, 潘晓雷, 等. 体系结构模拟器的技术和发展[J]. 信息技术快报, 2008, 6(1): 1-25.
- [4] 时磊, 逢珺, 杨磊, 等. M5 模拟器的内核分析及应用[J]. 微计算机应用, 2009, 30(4): 30-34.
- [5] SNU. SNU real-time benchmarks[EB/OL]. [2009-12-17]. <http://archi.snu.ac.kr/realtime/benchmark/>.

作者简介:

计卫星 男, (1980-), 博士, 讲师. 研究方向为计算机系统结构及并行计算.

(上接第 176 页)

参考文献:

- [1] Gildea D, Jurafsky D. Automatic labeling of semantic roles[J]. Computational Linguistics, 2002, 28(3): 245-288.
- [2] Thompson CA, Levy R, Manning CD. A generative model for semantic role labeling[C]// Proceedings of ECML-2003, LNAI 2837. Berlin, Heidelberg Springer, 2003: 397-408.
- [3] Che W, Zhang M, Liu T, et al. A hybrid convolution tree kernel for semantic role labeling[C]// Proceedings of ACL-2006. Sydney, 2006.
- [4] Baker C F, Fillmore C J, Lowe J B. The berkeley

frameNet project[C]// Proceedings of ACL & Coling-1998. Canada: Quebec, 1998: 86-90.

- [5] Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480.
- [6] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

作者简介:

于江德 男, (1971-), 博士, 副教授. 研究方向为自然语言处理、信息抽取、文本数据挖掘等.