事件信息抽取中语义角色标注研究*)

于江德[1,2] 樊孝忠[1] 庞文博[1]

(北京理工大学计算机科学技术学院 北京 100081)[1] (安阳师范学院计算机科学系 河南安阳 455000)[2]

摘 要 文本信息抽取是处理海量文本数据的手段,事件信息抽取是其中最具挑战性的任务之一。提出了一种基于条件随机场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,将条件随机场用于句子中谓词的语义角色标注。应用该方法对"职务变动"和"会见"两类事件的事件要素及其语义角色进行标注,在各自的测试集上分别获得了77.3%和74.2%的综合指标F值。

关键词 语义角色标注,条件随机场,事件信息抽取,事件要素

Research on Semantic Role Labeling for Event Information Extraction

YU Jiang-De[1,2] FAN Xiao-Zhong[1] PANG Wen-Bo[1]

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081)[1]

(Department of Computer Science, Anyang Teachers' College, Anyang, Henan 455000)[2]

Abstract Text information extraction is an important means of processing large quantity of text. Event extraction is one of the most challenge tasks of the research on information extraction. A method based on conditional random fields (CRFs) 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 CRFs model is trained to label the predicates' semantic roles in a sentence. The method is used to label event argument and its roles on two test sets of management succession and meeting, and the F measure is 77.3% and 74.2% respectively.

Keywords Semantic role labeling, Conditional random fields, Event information extraction, Event argument

1 引言

对自然语言形式的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。在对句子进行语义分析时,一个重要的步骤是语义角色标注(Semantic Role Labeling,SRL)。所谓语义角色标注是根据一个句子中的动词(谓词)与相关的各类短语等句子成分之间的语义关系而赋予这些句子成分的语义角色信息。在自然语言理解的诸多领域,例如:问答系统、信息抽取、机器翻译、词汇语义排歧等,语义角色标注都有着广泛的应用。

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和实验。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。Gildea 等人[1]是第一个使用统计的方法进行语义角色标注研究的;Thompson 等[2]使用产生式模型(Generative model)进行语义角色标注;Liu 等[3]使用最大熵分类器来实现语义角色标注;文[4]将支持向量机(SVM)用于语义角色标注。

本文提出了一种基于条件随机场(Conditional Random Fields,CRFs)的语义角色标注方法。CRFs 是 Lafferty 等[5]于 2001 年提出的一种用于序列数据标注的条件概率模型。将本文提出的方法用于事件信息抽取(简称事件抽取,Event Extraction)中事件要素的语义角色判定。所谓事件抽取是指从自然语言形式的文本中自动地抽取用户感兴趣的事件以及卷入其中的特定类型的实体,并将这些信息转换为结构化的数据并存储的过程。例如,从包含职务变动的文本中抽取职务变动事件的详细信息(事件要素):人员、组织机构、职位、时间等。实验中对"职务变动"和"会见"两类事件抽取进行了语义角色标注,结果表明,该方法有较好的标注性能。

2 条件随机场

条件随机场是一种以给定的输入序列为条件来预测输出序列概率的无向图模型。用于模拟序列数据标注的 CRFs 是一个简单的链图,如图 1 所示。

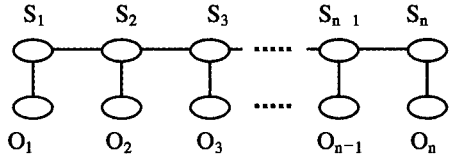

图 1 线链 CRFs 的图形结构

设 $O=\{o_1,o_2,\cdots,o_T\}$ 表示被观察的输入数据序列,例如一个事件表述语句中顺序出现短语或命名实体。$S=\{s_1,s_2,\cdots,s_T\}$ 表示被预测的状态序列,每一个状态均与一个语义角色(例如施事、受事)相关联。这样,在一个输入序列给定的情况下,参数为 $\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$ 的线链 CRFs,其状态序列的条件概率为:

$$P_\Delta(S|O)=\frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \quad (1)$$

其中,Z_O 是归一化因子,它确保所有可能的状态序列的条件概率和为 1。

$$Z_O=\sum_s\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \quad (2)$$

$f_k(s_{t-1},s_t,o,t)$ 是一个任意的特征函数,通常是一个二值表征函数。λ_k 是一个需要从训练数据中学习的参数,表示相应的特征函数 $f_k(s_{t-1},s_t,o,t)$ 的权重,取值范围可以是 $-\infty$ 到 $+\infty$。

给定一个由公式(1)定义的条件随机场模型,在已知输入

*)基金项目:由教育部博士点基金项目(20050007023)支持。于江德 博士生,讲师,研究方向为自然语言处理、信息处理、文本数据挖掘等。

数据序列 O 的情况下，最可能的标记序列可以由下式求出：

$$S^* = \arg\max_S P_\Delta(S|O) \quad (3)$$

最可能的标记序列可以由上式通过类似于隐马尔可夫模型(Hidden Markov Model，HMM)中的韦特比算法动态规划求出。

2.1 参数估计

要建立 CRFs 模型还有两个关键的问题：参数估计和特征选择。参数估计就是从训练数据集学习权重参数向量 Λ 的过程，这个过程通过最大对数似然估计实现。将训练数据集表示为：$D=\{O^{(i)},S^{(i)}\}_{i=1}^N$，其中，每个 $O(i)=\{o_1^{(i)},o_2^{(i)},\cdots,o_T^{(i)}\}$ 是一个输入数据序列；$S^{(i)}=\{s_1^{(i)},s_2^{(i)},\cdots,s_T^{(i)}\}$ 是相应的输出数据序列。在训练数据集 D 下对数似然为：

$$L_\Delta=\sum_{i=1}^{N}\log P(S^{(i)}|O^{(i)}) \quad (4)$$

将 CRFs 模型中的条件概率公式(1)代入(4)式，可得：

$$L_\Delta=\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\log Z_{O^{(i)}}\right) \quad (5)$$

为了避免对于大量参数估计时出现过拟合(over-fitting)，对数似然经常需要将参数作先验分布的调整，采用高斯先验调整后，(5)式变为：

$$L_\Delta=\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\sum_{i=1}^{N}\log Z_{O^{(i)}}-\sum_k\frac{\lambda_k^2}{2\sigma^2} \quad (6)$$

其中最后一项是用于进行调整的特征参数的高斯先验值，σ^2 表示先验方差。优化(6)式需要一个迭代的过程，传统的办法是 Della Pietra 等人在 GIS 算法的基础上提出的 IIS 算法。文[5]在进行 CRFs 参数估计时采用了基于 IIS 的算法。但 IIS 算法或基于 IIS 的算法计算量大，求解速度慢，所以本文采用 L-BFGS 算法[6]对 CRFs 的参数进行估计。L-BFGS 是一种充分利用以前的梯度和修改值来近似曲率值的二阶方法。使用 L-BFGS 算法进行 CRFs 训练只要求提供似然函数的一阶导数，训练数据集的对数似然函数的一阶导数为：

$$\frac{\partial L_\Delta}{\partial\lambda_k}=\sum_{i=1}^{N}\sum_{t=1}^{T}f_k(s_{t-1}^{(i)},s_t^{(i)},o^{(i)},t)-\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s,s'}f_k(s,s',o^{(i)},t)p(s,s'|o^{(i)})-\frac{\lambda_k}{\sigma^2} \quad (7)$$

其中，第一项为特征 f_k 在经验分布下的期望值；第二项为特征 f_k 在模型 Λ 下的期望值。对它们的计算，可采用动态规划高效实现。

2.2 特征选择

针对语义角色标注这一任务，参照文[4,7]对中文文本中语义角色的特征进行了深入分析，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。

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。本文用到的部分基于句法成分的特征如表 1 所示。

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。本文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。

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 2 示意了谓词到 She 所在 NP 的路径为 VBD↑VP↑S↓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

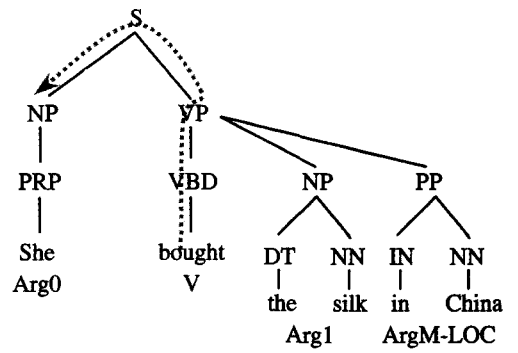

图 2 语义角色标注中路径示意

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 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 CRFs 的语义角色标注

3.1 语义角色标注的一般步骤

依据上述的方法构建好 CRFs 模型之后，接下来就可以采用条件随机场来进行语义角色标注了。由于对汉语而言，深层次的句法分析还不成熟，因此，本文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。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,也是本文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。

本文将基于CRFs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。"职务变动"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

4.1 实验数据集

实验采用的训练数据集是借助辅助工具依据触发词从人民日报1995年全年的生语料中抽取出来的职务变动、会见两类事件表述语句,共得到18264条语句,这些语句经过简单的人工筛选后剩下5100条语句。这些语句经过分词、词性标注、命名实体识别后、浅层句法分析、语义角色标注后作为训练数据集,用于构建CRFs模型。测试数据集是依据是否包含触发词从人民日报1998年1月的熟语料中抽取出来的上述两类事件的表述语句,这些标注过的语句经过命名实体识别、浅层句法分析、语义角色标注后用于验证基于CRFs的语义角色标注方法。

4.2 性能评估

在对语义角色标注性能进行评估时,采用了三个评测指标:查准率(P)、查全率(R)、综合指标F值(F)。计算公式如下:

$$P=\frac{\text{准确标注的语义角色数}}{\text{标记的所有语义角色数}} \tag{8}$$

$$R=\frac{\text{准确标注的语义角色数}}{\text{应该标注的语义角色数}} \tag{9}$$

$$F=\frac{(\beta^2+1)PR}{(\beta^2P+R)} \tag{10}$$

其中,β决定对P侧重还是对R侧重,通常设定为1、2或1/2。本文β取值为1,即对二者一样重视。

4.3 实验结果及分析

4.3.1 两类事件语句的语义角色标注实验

我们首先对"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CRFs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。然后就不同特征集对CRFs抽取性能的影响进行比较实验,相应的实验结果和分析在下一小节介绍。

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	0.847	0.826	0.836	A0-PER	0.835	0.809	0.821
A0-ORG	0.819	0.802	0.810	A1-PER	0.816	0.782	0.798
Start-PER	0.823	0.768	0.794	ArgM-LOC	0.796	0.758	0.776
Left-PER	0.725	0.769	0.746	ArgM-TMP	0.750	0.696	0.722
Position	0.836	0.758	0.795	ArgM-CAU	0.645	0.622	0.633
ArgM-LOC	0.760	0.735	0.747	ArgM-MNR	0.708	0.693	0.700
ArgM-TMP	0.815	0.706	0.757				
ArgM-CAU	0.702	0.695	0.698				

4.3.2 特征集的影响

为了比较不同的特征集对语义角色标注性能的影响,在进行基于条件随机场语义角色标注时,起先只使用基于句法成分的特征集,然后再加入谓词特征、句法成分-谓词关系特征和语义特征,并观察不同特征空间对标注性能的影响,结果如表6所示。从表6可以看出,加入谓词特征后对标注性能影响并不大,平均F值从72.5%提升到了73.1%;加入句法成分-谓词关系特征后对标注性能影响较大,平均F值从72.5%提升到了74.8%;加入语义特征后标注性能也有不小的提升,平均F值从72.5%提升到了74.5%。

表6 不同特征集的语义角色标注结果

不同特征集	P平均值	R平均值	F平均值
句法成分特征集	0.734	0.716	0.725
+谓词特征	0.739	0.724	0.731
+句法成分-谓词关系特征	0.757	0.740	0.748
+语义特征	0.756	0.736	0.745
全部特征集	0.792	0.755	0.773

小结 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。本文将条件随机场模型用于语义角色标注,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。

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]. In: Proceedings of ECML-2003, LNAI 2837, Springer Berlin Heidelberg, 2003. 397~408

3 Liu T, Che W X, Li S, et al. Semantic role llabeling system using maximum entropy classifier [C]. In:Proceedings of CoNLL-2005. Ann Arbor, Michigan, 2005. 189~192

4 Pradhan S, Hacioglu k, Krugler V, et al. s Support vector learning for semantic argument classification [J]. Machine Learning, 2005

5 Lafferty J, Pereira F, McCallum A. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]. In:Proceedings of 18th International Conference on Machine Learning, 2001. 282~289

6 Byrd R H, Nocedal J, Schnabel R B. Representations of quasi-Newton matrices and their use in limited memory methods [J]. Mathematical Programming, 1994, 63(2):129~156

7 袁毓林. 论元角色的层级关系和语义特征[J]. 世界汉语教学, 2002(2):10~22