软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(10):2612−2627 [doi: 10.3724/SP.J.1001.2012.04189]

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

基于 SVM 和扩展条件随机场的 Web 实体活动抽取*

张传岩, 洪晓光+, 彭朝晖, 李庆忠

(山东大学 计算机科学与技术学院,山东 济南 250101)

Extracting Web Entity Activities Based on SVM and Extended Conditional Random Fields

ZHANG Chuan-Yan, HONG Xiao-Guang+, PENG Zhao-Hui, LI Qing-Zhong

(School of Computer Science and Technology, Shandong University, Ji'nan 250101, China)

+ Corresponding author: E-mail: hxg@sdu.edu.cn

Zhang CY, Hong XG, Peng ZH, Li QZ. Extracting Web entity activities based on SVM and extended conditional random fields. *Journal of Software*, 2012,23(10):2612−2627 (in Chinese). http://www.jos.org.cn/1000-9825/4189.htm

Abstract: On the basis of the traditional methods extracting information, this paper defines the formal model of entity activity based on case grammar and presents a method based on supported vector machine and extended condition random fields to extract Web entity activities accurately. First, in order to automatically train the machine learning models, the study puts forward a heuristic method to transform the semantic role labeling training data into the training data of entity activity extraction. Next, the study trains a support vector machine classifier and extends condition random fields using the training data. Third, using the classifier, the study distinguishes the sentences that contain Web entity activities. The paper also proposes forward and extends condition random fields to model the frequency and relationship feature. The traditional conditional random fields cannot model this while the new model can label the entity activity information in natural language sentences more accurately. Finally, the experimental results show that the method is effective in multi-domains and can be applied to Web entity activity extraction.

Key words: information extraction; case grammar; entity activity; support vector machine; extended condition random fields

摘 要: 在传统信息抽取的基础上,研究 Web 实体活动抽取,基于格语法对实体活动进行了形式化定义,并提出一种基于 SVM(supported vector machine)和扩展条件随机场的 Web 实体活动抽取方法,能够从 Web 上准确地抽取实体的活动信息.首先,为了避免人工标注训练数据的繁重工作,提出一种基于启发式规则的训练数据生成算法,将语义角色标注的训练数据集转化为适合 Web 实体活动抽取的训练数据集,分别训练支持向量机分类器和扩展条件随机场.在抽取过程中,通过分类器获得包含实体活动的语句,然后利用扩展条件随机场对传统条件随机场中不能利用的标签频率特征和关系特征建模,标注自然语句中的待抽取信息,提高标注的准确率.通过多领域的实验,其结果表明,所提出的抽取方法能够较好地适用于 Web 实体活动抽取.

关键词: 信息抽取;格语法;实体活动;支持向量机;扩展条件随机场

中图法分类号: TP391 **文献标识码**: A

* 基金项目: 国家自然科学基金(61003051); 国家科技支撑计划(2009BAH44B02); 山东省自然科学基金(2009ZRB019RW); 山东省科技攻关计划(2010GGX10108)

收稿时间: 2011-08-15; 修改时间: 2011-11-02; 定稿时间: 2012-01-17

随着 WWW 的快速发展,Web 逐渐成为一个海量数据源,存在着大量有价值的信息.现有的信息抽取技术可以有效地抽取 Web 上的命名实体[1]、实体关系[2-4]、事件[5,6]以及 Web 数据对象[7,8],建立基于以上结构化数据的知识库,有效地为信息检索、信息集成、问答等系统服务.此外,在 Web 也存在大量关于实体行为活动的相关信息.例如,普遍存在于各大新闻网站上的有关世界各国政府要人的活动行程、各大公司的成长发展轨迹等等.但是,对 Web 上实体活动信息进行严格定义和抽取的研究,目前尚不多见.

实体活动信息在自然语言中体现为语句中的谓语动词和其所支配的其他语义成分,包括动作的发出者、动作的作用对象、动作发生的时间、地点等.格语法[9]描述了句子中核心谓词与周围体词的语义关系,所以格语法可以以多元关系的形式来刻画实体活动;但是格语法未能给出自然语言完整的语义格,所以不能直接适用于实体活动这一类特定自然语句信息的定义.

信息抽取是指从无结构或半结构化文本中抽取结构化信息的过程[10].近年来,以 Deep Web 为数据源的信息集成技术,有效地利用页面的结构特征进行信息抽取,取得了巨大进步;实体活动信息主要包含在 Web 页面无结构文本中,所以 Web 实体活动抽取要以自然语言为研究对象.目前,面向自然语言的信息抽取包括命名实体识别、实体关系抽取和事件抽取.本文将 Web 实体活动描述为一种多元关系,现有的关系抽取主要以二元关系[1]为研究对象,其中,OpenIE[4]抽取思想主张在进行关系抽取前无需预先定义关系类型,而是通过启发式和机器学习的方法对文本中实体间可能存在的所有关系进行一次性抽取.这种抽取思想符合我们对于 Web 实体活动抽取中无法预先定义活动类型以及大规模数据处理的要求.Web 实体活动抽取也可以视为以实体为中心的事件抽取,在现有的事件抽取中,识别事件参数和属性[5]的任务与实体活动抽取有相似之处,但是,由于复杂的自然语言处理和事先对待抽取事件进行定义的需要,使得事件抽取并不直接适用于 Web 实体活动抽取.

本文从信息抽取角度出发,研究 Web 实体活动抽取,力求从无结构的海量 Web 网页中获得结构化的 Web 实体活动信息,拓宽了信息抽取的研究范围,对现有的主要以二元关系抽取为研究任务的关系抽取技术提出了挑战.本文的贡献体现在以下几个方面:

1) 依据统计和需求,对实体活动进行了严格的定义,并基于语义格建立了实体活动的形式模型,拓宽了信息抽取的研究范围,提高了 Web 数据的利用率;
2) 提出一个完整的 Web 实体活动抽取框架,首先通过 SVM 发现有效语句,然后利用扩展条件随机场抽取实体活动信息;
3) 提出一种基于启发式的训练数据生成算法,将语义角色标注的训练数据转化为实体活动抽取的训练数据,避免手工标注训练数据;
4) 提出扩展条件随机场,对序列数据标注过程中普遍存在的语义标签之间的频率特征和关系特征建模,提高了传统链式条件随机场标注结果的准确性.

1 问题定义

1.1 实体活动

定义 1(实体活动). 反映了一个确定的实体,在某个确定的时间进行的一项确定的活动,而实体活动的集合则在一定程度上反映了实体的产生和实体的发展轨迹.

1.2 实体活动的形式模型

由于自然语言形式多变,为自然语言建立统一的、全面的形式模型是自然语言中一项复杂和极具挑战性的任务.在本文中,依据定义的限制,实体活动在自然语言中体现为一些包含与实体相关的特定信息的语句.所以,我们依据这些能够反映实体活动的相关信息,力求用最简洁的形式为实体活动建立形式模型.

Fillmore 在 20 世纪 60 年代中期提出格语法[9],是一套着重探讨句法结构与语义关系的语法理论和语义学理论.格语法中的"格"是"深层格",描述了句子中核心谓词(动词、形容词等)与周围体词(名词、代词等)的关系,这种关系是语义关系,是一切语言中普遍存在的现象,而实体活动主要体现为描述活动的动词和周围名词、代

词和短语之间的关系,所以本文基于语义格为实体活动建立形式模型.通过大量观察和统计分析,我们将最能体现活动的信息归纳为 8 类,并通过对传统语义格的扩展和规约,提出了实体活动的 8 元关系模型.

定义 2(实体活动的形式模型). 在本文中定义为 8 元关系模型

$$T\{agent,activity,\{object\},time,\{location\},\{cause\},\{purpose\},\{manner\}\}.$$

- *Agent*:主体,表示动作活动的发起、状态的主体或者是非自发动作的当事者,可以是有生命的人或者动物,也可以是无生命的实体,如公司、书籍等.根据实体活动的定义,只有当某一感兴趣实体处于语义格中的主体位置时,这个语句包含的其他格信息才被视为这个实体的活动信息;
- *Activity*:活动,传统的格语法中并没有活动格,我们将传统格理论中的核心动词视作实体活动的活动维度.同时,在原有核心动词的基础上,我们规定活动是能够体现整个语句核心语义的动词或者动词短语,它能够完整而简洁地代表一类动作活动;
- *Object*:客体,表示动作活动所涉及或者影响到的事物,可以是原先存在的,也可以是动作产生的;
- *Time*,*location*,*cause*,*purpose*,*manner* 分别表示实体活动的时间、地点、原因、目的和方式,其中,*manner* 是对原来语义格中工具格(instrument)和方式格(manner)的合取,其余维度的表意与原语义格相同.在实体活动模型 8 个维度中,我们规定,一个实体活动所必需的维度是 *agent*,*activity* 和 *time*,这 3 个维度唯一确定一个实体活动,其他维度在一个语句中很少同时出现,大多数情况下会有缺失现象.

1.3 Web实体活动抽取

Web 实体活动抽取,面向 Web 网页,在给定感兴趣命名实体集合的前提下,从网页文本中抽取实体的活动信息.基于此,我们给出了 Web 实体活动的形式化定义.

定义 3(Web 实体活动抽取). 这是指给定命名实体集合 E 和 Web 页面集合 W,从 W 中抽取 T,使得 T 中 $agent\in E$,其中,$E=\{e_1,e_2,\ldots,e_i,\ldots,e_n\}$,$e_i$ 是给定的命名实体,T 为实体活动.

2 Web 实体活动抽取

由第 1.3 节中 Web 实体活动抽取的任务定义,我们提出一种自动、准确的 Web 实体活动抽取框架,适用于面向 Web 的信息抽取.抽取框架如图 1 所示.

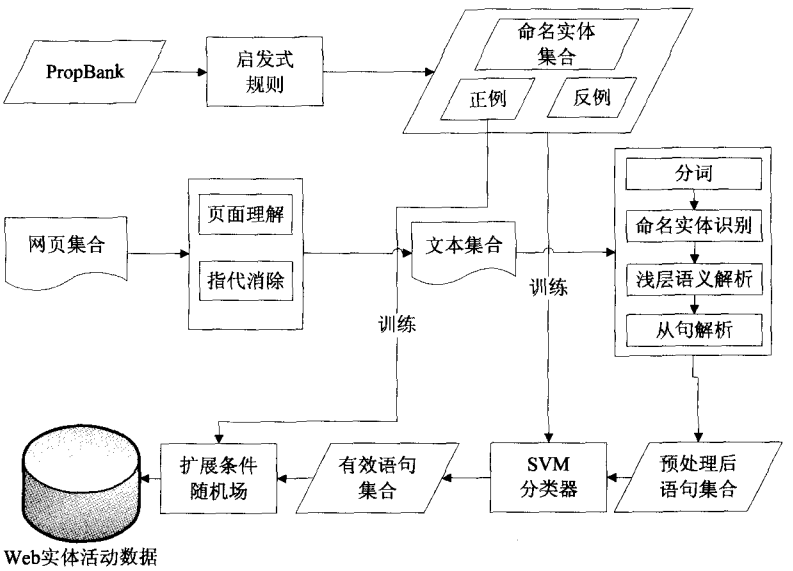

Fig.1 Web entity activity extraction framework

图 1 Web 实体活动抽取框架

如图 1 所示,整个抽取框架分模型训练、数据预处理和数据抽取 3 部分:

- 在模型训练部分,将 PropBank 的数据集通过启发式规则转换成我们需要的训练数据,分别训练支持向量机(support vector machine,简称 SVM)和扩展条件随机场(extended condition random fields,简称 ECRFs);
- 数据预处理部分,对从 Web 上爬取到本地的网页进行预处理,处理的内容包括页面理解和指代消除:页面理解的作用是识别网页中的有效文本信息,过滤掉广告、导航链接等噪音;指代消除是将句子中的代词替换为相应的命名实体.接下来,对文本集合中每一条自然语句进行预处理,包括分词(POS)、命名实体识别(NER)、浅层句法分析(shallow parsing)和从句处理,在本文中采用的是 CoNLL2005[11]大会提供的处理工具;
- 最后是数据抽取阶段,对预处理后的语句通过 SVM 分类器识别有效语句,即包含实体活动的语句,然后,利用 ECRFs 对包含实体活动的语句进行标注,抽取实体活动信息.

2.1 训练数据的自动生成

语义角色标注(semantic role labeling,简称 SRL)是对自然语句中各个语义角色进行分析标注的过程,这些语义角色即为"语义格",然后通过语义框架网识别语义角色的具体含义[11,12].SRL 是在格理论的基础上进行的,与 Web 实体活动的抽取任务有相似之处,所以,我们可以通过一些启发式规则,将现有 SRL 语料资源转化为 Web 实体活动抽取的语料资源.目前,SRL 较为著名的语料资源有 FrameNet 和 PropBank[13],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 相比信息更全面.其核心语义角色为 *Arg*0~*Arg*5 共 6 种,核心动词表示为 REL,其余的语义角色为附加语义角色,使用 *ArgM* 表示.例如,*ArgM*-TMP 表示时间,*ArgM*-LOC 表示地点等等.

本文以 PropBank 为训练数据的数据源,通过研究各个语义角色的含义,设计了基于启发式规则的训练数据自动生成算法.在实体活动模型 T 中,我们将 PropBank 中 *Arg*0~*Arg*5 共 6 种核心语义角色精简归纳为 *agent* 和 *object*,并将部分核心语义信息融合到 *activity* 中(很多情况下,单个动词不能完整地代表一类活动,例如"make","take"等等),同时,在 PropBank 大量的附加语义角色中选择了 5 种所有活动中通用且较常见的信息.综上所述,我们根据 T 的定义和 PropBank 的特点,构建了训练数据的生成算法,见表 1.

Table 1 Training dataset automatic generation algorithm

表 1 训练数据自动生成算法

输入:PropBank 标注实例,若语句 $S_i \in PropBank$,则存在 REL$\in S_i$,REL 是 S_i 的核心动词,存在 S_i 的非空核心语义角色集合 $ArgI=\{Argi|0\leqslant i\leqslant 5\}$,可能存在附加语义角色:*ArgM*-LOC,*ArgM*-TMP,*ArgM*-MOD,*ArgM*-MAN,...;
输出:集合 C 和集合 E,其中,C 为包含实体活动的语句集合,E 为语句中 *agent* 集合.

1. 如果?$\in S_i$||!$\in S_i$,则 $S_i \in \overline{C}$,结束;
2. 如果 *Arg*-MOD$\in S_i$||*Arg*-NEG$\in S_i$,则 $S_i \in \overline{C}$,结束;
3. 如果 $ArgI=\{Arg0\}$,则 *agent*=*Arg*0;如果 $ArgI=\{Arg1\}$&S_i 不是被动句,则 *agent*=*Arg*1;转步骤 8;
4. 如果 $ArgI=\{Arg0,Argi|1\leqslant i\leqslant 5\}$,则 *agent*=*Arg*0,*object*=*Argi*;如果 $ArgI=\{Arg1,Argi|2\leqslant i\leqslant 5\}$,则 *agent*=*Arg*1,*object*=*Argi*;
5. *Activity*=REL,转步骤 8;
6. 如果 $ArgI=\{Argi,Argj,Argk|1\leqslant i<j<k\leqslant 5$ 且 $i=0||1\}$,则 *agent*=*Argi*,*object*=*Argk*,*activity*=REL+*Argj*,转步骤 8;
7. 如果 $ArgI=\{Argi,Argj,Argk,Argl|1\leqslant i<j<k\leqslant 5$ 且 $i=0||1\}$并且 S_i 中存在 from *Argk* to *Argl*,则 *agent*=*Argi*,*object*=from *Argk* to *Argl*,*activity*=REL+*Argj*;
8. 如果 *agent* 是代词,则 $S_i \in \overline{C}$,结束;否则,S_i 中其他附加语义角色按以下方式转化:*location*=*ArgM*-LOC,*time*=*ArgM*-TMP,*manner*=*ArgM*-MAN,*purpose*=*ArgM*-PNC,*cause*=*ArgM*-CAU,*otherInfo*=*others*;
9. $S_i \in C$,*agent*$\in E$,否则 $S_i \in \overline{C}$;

根据我们对实体活动的定义,S_i 如果包含实体活动信息,从句型上必须是陈述句,从时态上必须是过去式或现在时.所以算法中,步骤 1 首先将疑问句、祈使句排除,步骤 2 将时态不合理的语句排除.在剩余的语句中,按照 S_i 中包含的完成核心语义标签数目分类进行语义标签的转换.由核心语义和实体活动的定义得知,只有 *Arg*0 和 *Arg*1 有成为 *agent* 的可能,所以将不包含 *Arg*0 和 *Arg*1 的 S_i 排除.其中,步骤 7 考虑到同时包含 4 个核心语义角

色的情况,通常以"起点~终点"的形式出现在语句中;对于其他特殊情况,我们不予考虑.一个语句中同时出现 5 个或者更多语义角色的情况极为少见,所以我们视其为 $\bar{C}$.

2.2 基于SVM的有效语句识别

语句进行预处理后,要识别有效语句,这是一个二分类问题.由于我们训练数据有限,选取特征丰富,所以选择 SVM 分类器.这是因为 SVM 采用了结构风险最小化原则和核函数思想,把非线性空间的问题转化到线性空间,在解决有限样本、非线性以及高维模式识别问题中表现出许多优势,较好地适用于我们的分类问题.

2.2.1 SVM 原理

SVM 的基本思想是,在特征空间构造出最优超平面,使得距离超平面最近的不同样本集之间的距离最大,从而达到最大的泛化能力[14].下面介绍 SVM 在本问题中的应用.

设二元分类问题的最优超平面为

$$w\times x+b=0 \tag{1}$$

其中,x 为多维向量,$w\times x$ 表示向量 w 与向量 x 的内积.最优平面要求:如果训练样本没有被平面错误分开,并且距平面最近的训练样本与平面的间距最大,即最小化 $\frac{1}{2}\|w\|^2$ 约束条件为各数据点(x_i,y_i)到分类平面的距离大于 1,保证训练样本被正确划分.另外,考虑到可能存在一些样本不能被正确分类,因此引入非松弛变量$\xi_i,i=1,2,\ldots,n$,则最优解问题为

$$\left.\begin{array}{c}\min\limits_{w,b}\left(\frac{1}{2}\|w\|^2+c\sum\limits_i\xi_i\right)\\ y_i(w\times x_i+b)\geqslant 1-\xi_i\\ i=1,2,\ldots,n\\ \xi_i\geqslant 0\end{array}\right\} \tag{2}$$

构造公式(2)的 Lagrange 函数,设α_i为 Lagrange 乘子,根据 Kuhn-Tucker 条件,可将公式(2)的最优化问题转化为其 Wolfe 对偶问题,是不等式约束下二次函数寻优问题,存在唯一解,解此问题后得到的最优分类函数为

$$f(x)=sign\{\sum_1^n\alpha_i^*y_i(x_i\cdot x)+b^*\} \tag{3}$$

2.2.2 特征选择

首先,本文给出有效语句的明确定义,根据定义确定 SVM 的分类特征.

定义 4(有效语句). 即包含实体活动的自然语句,是指满足以下 3 个条件的语句:

(1) 语句的核心语义是活动,该活动能够对现有实体产生影响;

(2) 活动的主体是我们感兴趣的命名实体,即:$agent\in E$,E 是我们感兴趣的命名实体集合;

(3) 活动必须有明确的时间.

- 条件 1 主要是排除自然语句中状态说明性的句型,例如"Google is an American multinational public corporation";
- 条件 2 保证我们抽取的实体活动都是我们感兴趣实体的活动信息;
- 条件 3 保证能够抽取到时间信息维度.

凡是不满足以上 3 个条件中任意一条的语句,都是不包含实体活动的语句.

为了提高 SVM 效率,我们利用 SVM 的特点对特征进行了选择.由公式(3)计算得到 $w^*=\sum_1^n\alpha_i^*x_iy_i$,最后求得权重向量 w^*.向量的维度是最初选择的特征数量,设为 m,w^*中每一个分向量的代表对应特征的权重.设 $w^*=\{w_1,w_2,\ldots,w_m\}$,如果存在 $w_i=0,1\leqslant i\leqslant n$,则删除 w_i对应特征向量,对分类结果没有影响.然后,对 w^*中剩余非零分量进行排序,从权值最低的分量开始删除,直到删除 w_j分类器的准确性出现明显下降,保留 w_j,停止删除.

成为包含实体活动的语句的必要条件是语句中各个名词(时间名词、地点名词等各类实体名称)与代表活动的动词之间存在关系.OpenIE[4]利用贝叶斯分类器判断两个名词之间二元关系存在与否,本文对 OpenIE 中的

规则和特征进行了扩展,得到了 17 种 SVM 特征向量,见表 2.

Table 2 SVM feature vectors and introduction

表 2 SVM 特征向量及说明

特征	特征说明
包含时间	是/否
包含地点	是/否
是否包含命名实体集合 E',$E'\cap E\neq\varnothing$,E 为任务给定集合	是/否
包含系动词,如"is","are"等	是/否
包含表示领属关系的动词,例如"have","possess"等	是/否
动词词性	及物动词或者不及物动词
动词与 e_i 的相对位置,$e_i\in E'$	前/后
动词与 e_i 的间隔	间隔单词的数量
动词与 e_i 之间是否有句中标点符号,",",";	是/否
任何 $e_i,e_j\in E'$,$i\neq j$,e_i,e_j 的相对位置	前/后
e_i,e_j 间隔单词数量	单词数量
e_i,e_j 间隔是否包含标点,",",";	是/否
e_i,e_j 间隔单词词性	单词词性
是否现在式	是/否
是否过去式	是/否
是否陈述句	是/否
语句是否是被动句	是/否

以上分析都是基于一个语句包含一个动词这种理想条件下进行的.为了保证 Web 实体活动抽取能够处理语句包含多个动词的情况,我们提出一个简单而有效的启发式规则,将一个包含多动词的语句变为包含单个动词的语句.这样,可以保证我们每次只研究一个活动,提取一个活动的相关信息.

启发式规则. 设语句 S 包含动词 $\{v_1,v_2,\ldots,v_n\}$,$n\geqslant 1$,将句子 S 复制 n 份,第 i 份以 $v_i(1\leqslant i\leqslant n)$ 为 S 的核心动词. SVM 依据定义 4 判断第 i 份是否为有效语句.

2.3 Web实体活动抽取

Web 实体活动抽取旨在从自然语句中抽取相关实体的活动信息.自然语句具有很强的序列性,通过浅层句法分析(shallow parsing)可以将语句划分成很多语义块(chunk),并划分出从句(clause),而这些 chunk 和从句正是我们要抽取的 T 的各个维度信息.通过大量观察和分析,我们规定:在对主句进行实体活动抽取时,从句作为一个整体进行抽取;在抽取从句信息时,将从句视为 chunk 的集合.因此,通过一定的方法将语句中的 chunk 和从句进行正确的标注,就能够实现 Web 实体活动抽取任务.在标注时,有些 chunk 能够单独成为某一维度信息,有些则是连续的 2 到多个成为一个维度信息.而 clause 往往单独成为一个维度信息,如 manner,cause,purpose 等.所以,Web 实体活动抽取问题最终演变成序列数据的语义切割标注问题.条件随机场(conditional random fields,简称 CRFs)是目前处理序列分割和标注的最好的统计机器学习模型[15],它利用事先定义好的标签集,通过统计训练数据的各种特征信息进行标注.

在利用 CRFs 进行序列数据的切割标注时,T 的每个维度标签都有-B,-M,-E 这 3 种状态,分别表示某一维度标签的开始、中间和结束.例如"On April 13, 2007, Google reached an agreement to acquire DoubleClick for $3.1 billion.",Shallow parsing 处理后为"[PP On] [NP April 13, 2007], [NP Google] [VP reached] [NP an agreement] [VP to acquire] [NP DoubleClick] [PP for] [NP $3.1 billion].".我们的标注结果是

[PP On] [***time*** NP April 13, 2007], [***agent*** NP Google] [***activity-B*** VP reached] [***activity-M*** NP an agreement] [***activity-E*** VP to acquire] [***object*** NP DoubleClick] [PP for] [***manner*** NP $3.1 billion].

目前,利用 CRFs 进行序列语义标注时,只对训练数据中蕴含的状态特征和状态转移信息进行挖掘利用[15],而忽略了标注过程中状态的本身特征和状态之间的关系特征,例如某一标注状态在序列中的出现次数、某两个状态之间的前后关系.在本文的标注工作中,这两种特征体现得非常明显,例如:*agent*,*activity* 和 *time* 出现且必须出现 1 次;*location* 如果出现,则仅能出现 1 次;*activity* 与 *object* 存在明显的但并不是强制性的前后关系,以及其

他我们不能够直接通过观察获得的特征.因此,本文在传统 CRFs 的基础上提出扩展的条件随机场,用统计的方法挖掘训练数据集中蕴含的各种特征信息,并以带权重的扩展边的形式添加到传统 CRFs 中,提高了语义标注的准确度.

2.3.1 条件随机场

在以下CRFs的模型介绍中,随机变量X表示需要标记的观察序列集,随机变量Y表示相应的标记序列,$Y_i \in Y$被限定在一个大小为 N 的有限字符集内.随机变量 X 和 Y 是联合分布,但在判别式模型(CRFs 是判别式概率模型)中需要构造一个关于观察序列和标记序列的条件概率模型 $P(Y|X)$,下面给出 CRFs 的定义.

定义 5. 设 $G=(V,E)$是一个无向图,$Y=\{y_v|y_v \in V\}$,Y 中元素与无向图 G 中的顶点一一对应.如果每个随机变量 y_v 相对于图 G 服从马尔可夫属性,即 $p\{y_v|\{y_w\}_{w\neq v},X\}=p\{y_v|y_u,X,(u,v)\in E\}$,则称$(X,Y)$是一个条件随机场(conditional random fields,简称 CRFs)[15].

根据 CRFs 的基础理论,当 $G=(V,E)$为一条一阶链时,在给定观测序列 X 的条件下,标注序列 Y 的概率分布 $P(Y|X)$为

$$P(Y|X)=\frac{1}{Z(x)}\exp\left(\sum_k \lambda_k f_k(y_i,x,i)+\sum_k \mu_k g_k(y_{i-1},y_i,x,i)\right) \tag{4}$$

其中,$f(y_i,x,i)$是观察序列上的状态特征函数,$g(y_{i-1},y_i,x,i)$是状态转移特征函数.给定训练样本$\{(x^{(k)},y^{(k)})\}$预定义的特征函数,可以从样本中学习一个 CRFs 模型.样本参数$\Lambda=\{\lambda_k,\mu_k\}$可以使用极大似然、极大后验或 Quasi-Newton 等方法估计.$Z(x)$是归一化因子,表示为

$$Z(x)=\sum_y \exp\left(\sum_y \lambda_k f_k(y_i,x,i)+\sum_k \mu_k g_k(y_{i-1},y_i,x,i)\right) \tag{5}$$

2.3.2 扩展条件随机场

本文提出扩展条件随机场,建模传统链式条件随机场(linear-chain CRFs)中无法利用的状态变量 Y 的特征,包括状态的频率特征和状态之间的关系特征,提高 CRFs 的标注准确率.扩展条件随机场的定义如下:

定义 6. 设 $G=(V,E)$是一个链式条件随机场,X 是序列观测数据随机变量,Y_1 是状态标注序列随机变量,$Y_i \in Y$在每次标注过程中出现的频率为状态的频率特征(frequency feature),$y_u,y_v(u\neq v),y_u$ 和 y_v 间存在的逻辑关系称为状态的关系特征(relationship feature).我们称在 G 中引入状态频率特征和状态的关系特征的条件随机场为扩展的条件随机场(extended condition random fields,简称 ECRFs).

传统的 Linear-Chain CRFs 与 ECRFs 的原理图如图 2、图 3 所示.

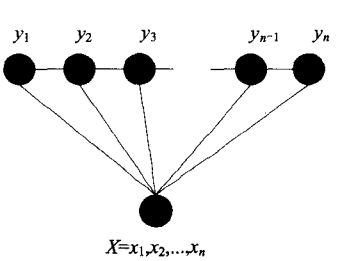

Fig.2 Linear-Chain CRFs graph model

图 2 链式 CRFs 图模型

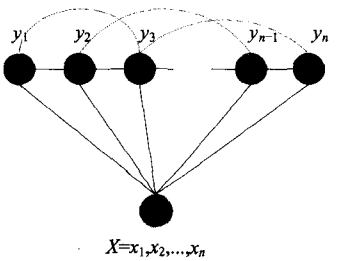

Fig.3 Linear-Chain ECRFs graph model

图 3 链式 ECRFs 图模型

在图 2 所示的 Linear-Chain CRFs 中,整个观察序列上只能定义两类特征函数 $f_k(y_i,x,i)$和 $g_k(y_{i-1},y_i,x,i)$;而对于其他一些对于提高标注结果准确率十分有利的特征,例如状态频率特征、状态的共现特征、顺序特征等都没有进行有效的利用.所以,我们在ECRFs中充分挖掘这些特征信息,建立了区别于原有转移边的第2类边,用这些边表示状态的其他特征信息.我们有如下特征定义:

定义 7. ECRFs 中的扩展特征:

(1) 状态频率特征:表示在每次状态标注过程中,某一状态 $y_i \in Y$ 出现的概率;

(2) 共现(occurrence feature):状态的关系特征之一,揭示了两个不同状态在一次标注过程中共现的概率;

(3) 顺序(sequence feature):顺序关系表示这两个不同状态在一次标注中同时出现时前后次序的概率,状态之间的顺序关系以共现为前提条件;

(4) 在 $G(X,Y)$中,$y_i \in Y, y_j \in Y, i \neq j$. y_i 和 y_j 之间存在共现关系,则称(y_i,y_j)是一条共现边(occurrence edge,简称 OE);如果(y_i,y_j)是 OE,且 y_i 和 y_j 之间存在顺序关系,则(y_i,y_j)是顺序边(sequence edge,简称 SE);如果 $y_i=y_j$,则称(y_i,y_j)是一条等值边(equal value edge,简称 EVE),主要用于约束状态频率特征,设 E'为 3 种边的集合,统称为扩展边.

在给定观测序列 X 的条件下,标注序列 Y 的概率分布 $P(Y|X)$为

$$P(Y \mid X)=\frac{1}{Z(x)}\exp\left(\sum_k \lambda_k f_k(y_i,x,i)+\sum_k \mu_k g_k(y_{i-1},y_i,x,i)+\sum_{e\in E',k}\chi_k h_k(y_{i-1},y_i,x,i)\right) \tag{6}$$

其中,$h_k(y_{i-1},y_i,x,i)$是扩展边特征函数;λ_k,μ_k,χ_k 分别是特征函数的权重;$Z(x)$是归一化因子,表示为

$$Z(x)=\sum_y \exp\left(\sum_k \lambda_k f_k(y_i,x,i)+\sum_k \mu_k g_k(y_{i-1},y_i,x,i)+\sum_{e\in E',k}\chi_k h_k(y_{i-1},y_i,x,i)\right) \tag{7}$$

2.3.3 基于扩展条件随机场的 Web 实体活动抽取

利用 ECRFs 进行自然语句的语义标注工作,最终实现 Web 实体活动抽取需要完成以下工作:

(1) 建立扩展边;

(2) 参数估计;

(3) 推理.

2.3.3.1 扩展边的建立

扩展边要在推理之前建立,边建立的关键在于寻找边的始末节点.扩展边的创建是以图 G 中相关顶点的语义标签已知为前提的,所以我们首先估算图 G 中各个节点属于某一语义标签的置信度,当置信度高于某一阈值时,创建相应的扩展边.目前,存在很多种方法估算图 G 顶点参数值的置信度[16].

在 ECRFs 中,主要利用训练数据中的文本特征来判断某一数据元素属于不同语义标签的置信度,即将数据元素对于某一标签的发射概率作为置信度.

构建扩展边的算法见表 3.

Table 3 Extended edge generation algorithm

表 3 扩展边建立算法

输入:一条链式数据 LD,$LD=\{v_1,v_2,\ldots,v_n\}$,语义标签集合 $L=\{y_1,y_2,\ldots,y_m\}$;
输出:E'_{OE},E'_{SE},E'_{EVE}.

1. 对于$\forall y_a \in L$,当$\forall v_i \in LD, p(v_i=y_a)>\alpha$,令 $v_i \in tempSet_a, 1\leqslant a\leqslant m$;
2. $\forall |tempSet_a| \geqslant 2$,连接 $tempSet_i$ 中两两顶点,将所建立的边添加至 E'_{EVE};
3. 对于$\forall v_i,v_j, i\neq j, p(v_i=y_a)>\alpha, p(v_i=y_b)>\alpha$,且 $p(y_a,y_b)>\beta$;
4. 边(v_i,v_j)添加到边集合 E'_{OE};
5. 对$\forall v_i,v_j, i\neq j, p(v_i=y_a)>\alpha, p(v_i=y_b)>\alpha$,且 $p(y_a,y_b)>\beta$,且 $p'(y_a,y_b)>\delta$或者 $p'(y_a,y_b)\leqslant 1-\delta$;
6. 边(v_i,v_j)添加到边集合 E'_{SE};
7. 返回 E'_{OE},E'_{SE},E'_{EVE}.

我们用到 3 个阈值:α,β,δ,其中,α是某一数据元素为某一确定语义标签的阈值,β和δ是语义标签之间是否存在共现和顺序特征的阈值,阈值大小的设置见实验部分.算法中涉及到的公式:$p(v_i=y_a)$表示数据元素 v_i 属于标签 y_a 的置信度,在本文中等价于发射概率;$p(y_a,y_b)$表示标签 y_a 和 y_b 的共现概率:

$$p(y_a,y_b)=\frac{N(y_a,y_b)}{N(y_a \| y_b)} \tag{8}$$

其中,$N(y_a||y_b)$是数据集中包含 y_a 或者 y_b 的训练数据数目,$N(y_a,y_b)$是训练数据集中标签 y_a 和 y_b 同时出现的次数.$p'(y_a,y_b)$是标签 y_a 和 y_b 共现时,y_a 出现在 y_b 之前的概率.如下所示:

$$p'(y_a,y_b)=\frac{N'(y_a,y_b)}{N(y_a,y_b)} \tag{9}$$

其中,$N'(y_a,y_b)$是 y_a 出现在 y_b 之前的次数.

2.3.3.2 参数估计

在 ECRFs 模型中,参数估计首先是针对不同特征函数分别估计参数:一部分是针对普通边和节点上特征函数的参数估计,另一部分是针对扩展边上特征函数的参数估计;然后,由于扩展边的引入导致多重边的产生,多重边的存在使得模型的推导变得困难,所以有必要对多重边上的不同参数进行参数融合,消除多重边.

与传统的链式 CRFs 的参数估计相似,ECRFs 模型中使用最大似然估计来估计针对普通边和节点上的特征函数的权重参数.即,在给定一个具有概率分布为 $\tilde{p}(X,Y)$ 的训练集 $D=\{(x^i,y^i)\},1\leqslant i\leqslant n$ 上估计参数 $\Lambda=\{\lambda_1,\lambda_2,\ldots,;\mu_1,\mu_2,\ldots\}$的值,使得该训练集数据的对数似然函数 $L(\Lambda)$达到最大.似然函数表示为

$$L(\Lambda)=\sum_i \tilde{p}(x^i,y^i)\log p(y^i\mid x^i,\Lambda) \tag{10}$$

由于 $L(\Lambda)$为凹函数,导数为 0 时为最值点.故对Λ求导,则偏导数公式为

$$\frac{\partial L(\Lambda)}{\partial \lambda_k}=E_{\tilde{p}(x,y)}[f_k]-E_{p(y|x,\Lambda)}[f_k] \tag{11}$$

$$\frac{\partial L(\Lambda)}{\partial \mu_k}=E_{\tilde{p}(x,y)}[g_k]-E_{p(y|x,\Lambda)}[g_k] \tag{12}$$

令公式(11)、公式(12)等于 0,函数 $L(\Lambda)$取得最大值.为减少计算量,$E_{p(y|x,\Lambda)}[f_k]$和 $E_{p(y|x,\Lambda)}[g_k]$使用向前-向后(forward-backward)算法求解.为了避免对大量参数估计时出现的过拟合问题,需要将参数作先验分布调整.采用高斯先验调整后,公式(10)转化为

$$L(\Lambda)=\sum_i \tilde{p}(x^i,y^i)\log p(y^i\mid x^i,\Lambda)-\sum\frac{\Lambda^2}{2\delta^2} \tag{13}$$

其导数变为

$$\frac{\partial L(\Lambda)}{\partial \lambda_k}=E_{\tilde{p}(x,y)}[f_k]-E_{p(y|x,\Lambda)}[f_k]-\frac{\lambda_k}{\sigma^2} \tag{14}$$

$$\frac{\partial L(\Lambda)}{\partial \mu_k}=E_{\tilde{p}(x,y)}[g_k]-E_{p(y|x,\Lambda)}[g_k]-\frac{\mu_k}{\sigma^2} \tag{15}$$

其中,σ^2 表示先验方差,Λ的参数估计问题可以用 GIS,IIS 等最优化方法来解决.

扩展边的参数估计,*OE* 扩展边(y_a,y_b)特征函数的权重初始值等于 $p(y_a,y_b)$,$p(y_a,y_b)$的计算见公式(8).*SE* 扩展边(y_a,y_b)初始权重为 $p'(y_a,y_b)$,$p'(y_a,y_b)$的计算见公式(9).

对等值边(EVE)的参数估计,我们用等值顶点出现的概率计算其初始权重:

$$W(y_a=y_b)=\begin{cases}\dfrac{N(x)}{N}\\ (p(y_a))^x\end{cases} \tag{16}$$

其中,$x=|tempSet_a|$;$N(x)$表示训练数据中,标签 y_a 在 1 次标注过程中出现 x 次的次数,N 为训练数据集大小;$p(y_a)=\dfrac{N(y_a)}{N(Lable)}$,$N(y_a)$表示训练数据中标签 y_a 出现的次数,$N(Lable)$是训练数据中总的标签数量.

多重边是指,在 $G=(V,E)$中,v_i 和 v_j 由于不同的特征产生,如状态转移特征、共现、顺序等,连接两个顶点而具有不同权重的边.多重边的存在会导致 ECRFs 的推导变得极具困难,因此我们将具有不同权重的多重边加以融合,使得边(v_i,v_j)唯一标识一个边,并具有一个融合后的权重.多重边权重融合公式为

$$\chi'_{i,j}=\frac{\sum_1^n \varphi_k\chi_{i,j}}{n} \tag{17}$$

其中,$\chi'_{i,j}$ 是融合后的权重,$\chi_{i,j}$是相对于不同特征产生的多重边的权重.公式是对多重边上的权重进行加权平均:

$$\sum_1^n \varphi_k = 1.$$

2.3.3.3 推 理

推理算法的时间复杂度直接影响模型的性能.在 ECRFs 模型中,由于扩展边的引入使得模型中包含了环,并且环的距离可能比较长或发生重叠,导致精确推理算法的时间复杂度呈指数级增长,所以精确的推理算法不再适合.本文使用 Loopy Belief Propagation 算法[17]进行近似推理,Loopy Belief Propagation 算法是对向前-向后算法的归纳.向前-向后算法的时间复杂度为 $O(n^2T)$,其中,n 代表状态集合的大小,T 代表观察序列的长度.在 ECRFs 模型中,对单条扩展边进行处理的代价等同于向前-向后算法中对单条邻接边的处理代价.因此,在 ECRFs 模型中,Loopy Belief Propagation 算法的时间复杂度变为 $O(|L|^2(T+M))$,其中,$|L|$代表语义标签集合的大小,T 代表一条自然语句预处理后数据元素的个数,M 代表扩展边的条数,M 与 T 的平方成正比.Loopy Belief Propagation 是一种不保证收敛的迭代算法,但是相关研究[17]和本文的实验都表明,其在实际应用中具有较好的近似推理效果,可以有效地推理语义标注序列.

3 实 验

为了对本文提出的 Web 实体活动抽取方法进行评估,本文从如下 4 个方面进行了测试:

(1) SVM 分类器性能;

(2) 传统线性链 CRFs 与 ECRFs 标注结果的比较;

(3) ECRFs 阈值对抽取结果的影响;

(4) SVM+ECRFs 与单独使用 CRFs/ECRFs 的性能比较.

3.1 数据集

以下是对所提出方法进行评估的真实数据集:

(1) 人物实体活动数据集(person activity dataset,简称 P)

该数据集以人物的活动为研究对象,这些人都是来自不同领域的名人,如科学家、政治家、体育名人、娱乐明星、商界名人等 200 位.我们从维基百科(http://en.wikipedia.org/)的人物简介页面以及 Reference 链接的新闻页面中选择 100 句包含该人物活动的语句,共 20 000 条,这部分数据是人物活动正例数据集(positive person activity dataset,简称 PP).然后,针对每个人物随机选择 400 条不包含该人物活动信息的句子构成反例数据集(negative person activity dataset,简称 NP),共 80 000 条.P 包含 PP 和 NP 共 100 000 条语句.

(2) 公司实体活动数据集(company activity dataset,简称 C)

该数据集以公司的活动为研究对象,这些公司是包括 Microsoft,Google,Apple,IBM 等在内的 100 家公司.方法同上,我们分别得到公司活动正例数据集(positive company activity dataset,简称 PC)和反例数据集(negative company activity dataset,简称 NC),分别含有 10 000 条和 40 000 条,C 共含数据 50 000 条.

通过分析统计,在简介性页面,实体活动正例和反例比值接近 1:4;而在新闻页面,远小于 1:4.本文中,正例数据是我们研究的主要对象.为保证主要数据的数量,正例和反例的比值选择 1:4.

3.2 评估标准

本文采用检验 Web 信息抽取结果的常用标准:查全率、查准率、$F1$ 测度和实例抽取准确率这 4 项指标,对实验结果进行综合评价.另外,本文还将实例准确率(instance accuracy)作为评估标注.实例准确率为每个数据元素均被正确抽取的 T 占总的测试数据的比例.

3.3 实验结果与分析

3.3.1 SVM 分类器性能评估

SVM 分类器是经典的分类器模型,在本文中,SVM 分类器是在给定命名实体集合和待抽取语句的前提下,将自然语句集合划分为包含实体活动的语句集合和不包含实体活动的语句集合.以数据集合 P 和 C 作为测试数

据,测试 SVM 的性能,结果见表 4.

Table 4 Result of SVM classifier

表 4 SVM 分类结果

数据集	有效语句集合		
	Precision (%)	Recall (%)	*F*1 (%)
P	93.42	87.65	90.44
C	92.13	86.41	89.17

实验结果表明,SVM 分类器在识别包含实体活动的语句时,查准率达到了较高水平.我们分析了错误数据出现的原因,主要集中在复杂语句上,特别是动词多于 2 个时,语句中出现长距离依赖,SVM 的性能会有所降低.同时,查准率的提高也牺牲了一定的召回率.这是因为,我们在设置 SVM 特征时,许多长距离依赖下的实体活动会被认为依赖不存在,而将该语句放入反例集合.实验 *F*1 值在 90%左右,说明利用算法 1 自动生成的训练数据在对 SVM 进行训练时具有良好的效果.

3.3.2 ECRFs 与传统条件随机场标注结果的比较

本节在两个数据集上通过实验比较了 ECRFs 与 Linear-Chain CRFs 模型在 Web 实体活动抽取上的性能.在整个 Web 实体活动抽取流程中,ECRFs 是对 SVM 分类后的语句集合进行标注.为了单独测试 ECRFs 的性能,为保证测试的开放性,我们分别从 PP 和 PC 数据集中随机抽取各 5 000 条语句手工标注,作为测试数据.表 5 和表 6 显示了在每个语义格上的查全率、查准率、*F*1 值以及 *F*1 平均值.

Table 5 Performance contrast on PP dataset

表 5 PP 数据集上性能比较

标签	Linear-Chain CRFs			ECRFs		
	Recall (%)	Precision (%)	*F*1 (%)	Recall (%)	Precision (%)	*F*1 (%)
agent	77.61	71.47	74.41	89.29	79.45	84.08
activity	80.46	71.15	75.51	78.52	91.78	84.63
object	64.83	72.78	68.57	73.72	89.47	80.83
Time	78.61	62.59	69.69	93.28	67.23	78.14
location	80.12	64.71	71.59	87.48	70.45	78.04
cause	67.67	55.91	61.23	70.64	64.63	67.50
purpose	70.48	55.65	62.19	75.17	60.23	66.87
manner	64.12	65.15	64.63	64.67	65.44	65.05
Average *F*1	–	–	68.48	–	–	75.64

Table 6 Performance contrast on PC dataset

表 6 PC 数据集上性能比较

标签	Linear-Chain CRFs			ECRFs		
	Recall (%)	Precision (%)	*F*1 (%)	Recall (%)	Precision (%)	*F*1 (%)
agent	79.47	71.57	75.31	91.78	80.81	85.94
activity	76.51	64.83	70.18	72.71	89.64	80.29
object	67.48	75.78	71.38	75.14	92.59	82.95
Time	80.48	56.51	66.39	90.42	61.89	73.48
location	79.43	54.40	64.57	89.69	67.58	77.08
cause	64.71	56.31	60.21	74.19	65.72	69.69
purpose	62.19	60.45	61.30	83.72	57.31	68.04
manner	52.67	77.12	62.59	91.72	52.78	67.00
Average *F*1	–	–	66.49	–	–	75.56

从表中可以看出,基于 ECRFs 的实体活动抽取方法在总体性能上要优于基于 Linear-Chain CRFs 的方法.*F*1 的平均值在两个数据集上分别提高了 7.16%和 9.07%,并且每个字段的 *F*1 值均有所提高.实验结果表明,通过增加扩展边,可以充分利用自然语句中 chunk 的语义标签之间潜在的频率特征和共现特征,进一步降低 chunk 语义标注的错误率.另外,对于一些出现频率比较特殊的语义标签,例如 *agent*,*activity*,*time*,在每次标注过程中出现且仅出现 1 次,语义标注准确性的增长尤为明显,高于平均 *F*1 增长值.例如,*activity* 和 *object* 语义标签之前存在显著的顺序关系,通常情况下,*object* 位于 *activity* 之后,*object* 语义标签标注的准确度增幅也明显高于平均 *F*1 增长

值.在数据集 PP 和 PC 上,ECRFs 的平均 $F1$ 值分别是 75.64%和 75.56%,相差 0.08%,这说明我们的抽取方法在不同领域的抽取任务中性能稳定.实例标注准确率如图 4 所示,可以看出,在两个数据集上,与 Linear-Chain CRFs 相比,ECRFs 的准确率均有不同程度的提高,平均值分别提高了 10.97%和 9.25%.这说明,ECRFs 还可以提高整个 Web 实体活动的标注性能.

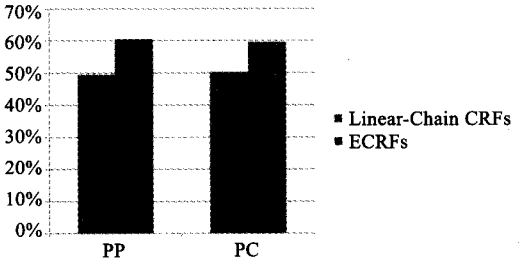

Fig.4 Instance tagging accuracy of ECRFs and Linear-Chain CRFs

图 4 ECRFs 和 Linear-Chain CRFs 实例标注准确率

3.3.3 ECRFs 阈值对标注结果的影响

在算法 2 中,我们用到 3 个阈值:α,β,δ.本节分别对以上 3 个阈值进行单独测试.

在数据集 PC 和 CC 上,分别在区间[0.8,0.99]中采用步长 0.01,设置不同的可信度阈值进行实验.在对一个阈值进行测试时,其他两个阈值不变.本实验选取 5 个可信度阈值(α,β,δ:0.8,0.85,0.9,0.95,0.99),分析不同阈值对 Web 实体活动抽取性能的影响.

图 5~图 7 是 ECRFs 模型分别在数据集 PC 和 CC 上改变不同可信度阈值得到的 $F1$ 平均值的变化.实验结果表明:在两个数据集上,随着 α 阈值的增大,$F1$ 的平均值也越大;但到达某一数值后,就会下降.这主要因为阈值越大,说明数据元素标注某一标签的可信度就越高,标注准确度也随之提高;但是,如果阈值过大,能够建立的扩展边减少,条件随机场不在受扩展边的约束,此时,语义标注的性能就会下降.最坏的情况是,所有的语义标签都不受扩展边影响.在实验中,当可信度阈值取到 0.99 时,ECRFs 和 CRFs 语义标注的 $F1$ 平均值相同.同理,在图 6 和图 7 中,随着 β 和 δ 的变化,$F1$ 平均值的变化趋势与 α 相似.另外,在图 6 中,α 由 0.95 增长到 0.99,$F1$ 平均值并没有发生变化.这主要与本文标记任务的特殊性有关,α 是控制共现特性的阈值,当两个标签共现概率高于 0.95 的同时也会很大几率高于 0.99.例如,(*agent*,*activity*)在每次标记过程中都会出现.

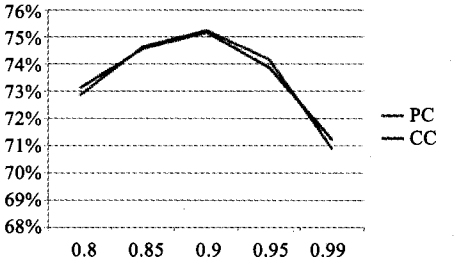

Fig.5 Change of $F1$ average value with α (β=0.9, δ=0.9)

图 5 $F1$ 平均值随阈值 α 的变化(β=0.9,δ=0.9)

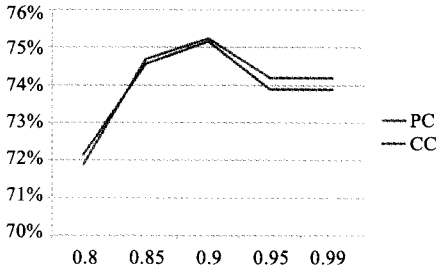

Fig.6 Change of $F1$ average value with β (α=0.9, δ=0.9)

图 6 $F1$ 平均值随阈值β的变化(α=0.9,δ=0.9)

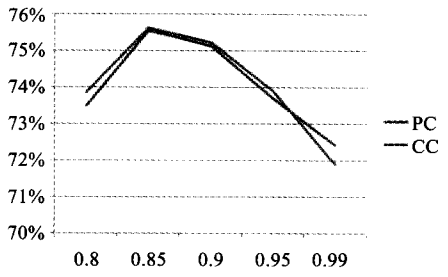

Fig.7 Change of $F1$ average value with δ (α=0.9, β=0.9)

图 7 $F1$ 平均值随阈值δ的变化(α=0.9,β=0.9)

3.3.4 SVM+ECRFs 与单独使用 CRFs/ECRFs 的性能比较

本节在两个数据集上通过实验比较了 SVM+ECRFs,Linear-Chain CRFs+Rules 和 ECRFs+Rules 的整体抽取性能.本文先通过 SVM 识别有效语句,然后利用 ECRFs 标注自然语句.同样,也可以对自然语句直接进行标注;然后,通过简单的启发式规则判定标注结果是否是感兴趣的实体活动.

根据实体活动定义,这些启发式规则包括:

(1) *agent*∈*E*,即标注结果的 *agent* 属于感兴趣的 Web 命名实体集合;

(2) 标注序列中包含 *time* 语义标签;

(3) *activity* 是非系动词或者状态动词;

(4) 语句是陈述语句.

表 7 和表 8 显示了在每个语义格上的查全率、查准率、$F1$ 值以及 $F1$ 平均值.

Table 7 Performance contrast on P dataset

表 7 P 数据集上性能比较

标签	SVM+ECRFs			Linear-Chain CRFs+Rules			ECRFs+Rules		
	Precision (%)	Recall (%)	$F1$ (%)	Precision (%)	Recall (%)	$F1$ (%)	Precision (%)	Recall (%)	$F1$ (%)
agent	80.78	64.18	71.52	46.43	40.39	43.19	50.82	65.91	57.38
activity	68.27	78.20	72.89	48.04	51.96	49.92	54.31	59.10	56.60
object	70.38	76.19	73.16	32.78	46.91	38.59	50.14	53.67	51.84
time	74.65	65.16	69.58	49.38	42.63	45.75	56.41	43.14	48.89
location	72.55	60.45	65.94	52.82	45.29	48.76	55.81	44.23	49.34
cause	62.12	49.47	55.07	41.53	30.52	35.18	48.72	35.64	41.16
purpose	61.87	49.81	55.18	35.21	38.43	36.74	41.92	42.45	42.18
manner	64.88	57.21	60.80	45.23	39.42	42.12	44.77	38.62	41.46
Average $F1$	–	–	65.52	–	–	42.53	–	–	48.61

Table 8 Performance contrast on C dataset

表 8 C 数据集上性能比较

标签	SVM+ECRFs			Linear-Chain CRFs+Rules			ECRFs+Rules		
	Precision (%)	Recall (%)	$F1$ (%)	Precision (%)	Recall (%)	$F1$ (%)	Precision (%)	Recall (%)	$F1$ (%)
agent	78.79	67.85	72.91	45.17	41.34	43.17	48.71	65.32	55.80
activity	69.44	78.90	73.86	48.51	57.23	52.51	51.38	67.45	58.32
object	68.31	74.15	71.11	38.19	40.31	39.22	47.94	57.49	52.28
time	73.81	58.87	65.49	51.81	38.13	43.92	56.29	46.84	51.13
location	69.31	57.78	63.02	49.92	43.36	46.40	58.28	45.11	50.85
cause	64.62	50.57	56.73	40.81	28.15	33.31	42.12	45.30	43.65
purpose	59.42	51.19	54.99	37.52	37.45	37.48	43.27	38.15	40.54
manner	57.56	54.56	56.01	43.97	40.47	42.14	49.58	35.24	41.19
Average $F1$	–	–	64.27	–	–	42.29	–	–	48.65

从表 7 和表 8 可以看出,基于 SVM+ECRFs 的 Web 实体抽取方法在总体性能要优于基于 CRFs+Rules 和

ECRFs+Rules 的方法.与 CRFs+Rules 和 ECRFs+Rules 相比,*F*1 的平均值在 P 数据集上分别提高了 22.99%和 16.91%,在 C 数据集上提高了 21.98%和 15.62%,并且每个字段的 *F*1 值均有所提高.实验结果表明,对于自然语句,先经过 SVM 分类处理,选择包含实体活动的语句进行标注抽取的方法,能够明显提高 Web 实体活动抽取的准确率.对导致 CRFs+Rules 和 ECRFs+Rules 算法效果下降的错误数据进行归类,分析原因:首先,SVM 可以比较准确地分类自然语句,对于不包含实体活动的语句的识别能力高于基于规则的方法,提高抽取结果的召回率;其次,对于包含多命名实体或者多动词的自然语句中,SVM 能够判断感兴趣的命名实体在语句中是否为核心实体,命名实体的对应动词是否能成为活动,从而在整体上提高了抽取算法的准确性.

图 8 显示了 SVM+ECRFs,CRFs+Rules 和 ECRFs+Rules 在实例标注准确率这项指标上的性能比较.从图 8 可以看出,在两个数据集上,与 CRFs+Rules 和 ECRFs+Rules 相比,SVM+ECRFs 的实例标注准确率均有不同程度的提高,实例标注准确率的平均值分别在 P 数据集上提高了 13.89%和 10.57%,在 C 数据集上提高了 11.48%和 10.81%.这说明,SVM+ECRFs 不仅能够提高单个数据元素的标注性能,而且还可以提高整个 Web 实体活动的标注性能.

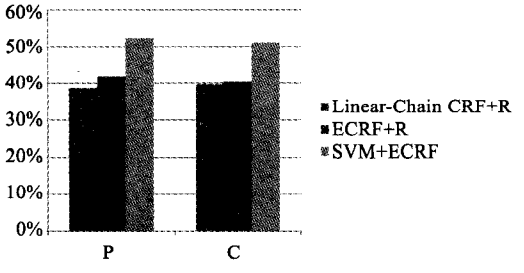

Fig.8 Instance tagging accuracy

图 8 实例准确率对比图

4 相关工作

近年来,随着互联网技术的发展,Web 逐渐成为一个海量信息源,越来越多的信息抽取(information extraction,简称 IE)技术以 Web 上的数据为研究对象.基于不同的 Web 数据源特征,如抽取命名实体[1]、实体关系[2,3,5]、事件[5,6]、数据对象[7,8,17]等等,建立结构化数据的知识库,服务于数据集成、信息检索、问答等系统.如果将这些数据以实体为中心分类,分别是实体名称(命名实体识别)、实体关系(实体关系抽取)、实体属性(数据对象、事件).但是对于实体的另一类重要信息——实体活动,相关研究并不多见.

实体活动描述了实体的行为活动,而同一个实体一系列行为活动的集合构成一个实体的踪迹,是 Web 上极具研究价值的信息.要抽取实体活动,必须首先建立实体活动的形式模型,Fillmore 的格语法[9]为模型的建立提供了很好的借鉴.但是格语法认为无生命实体是不能发出活动的,无法成为主格,而且并没有给出统一完成的格系统.因此,我们对语义格进行了规约和扩展,使得模型中用到的语义格更加符合对实体活动的描述.

我们将实体活动抽取定义为多元关系,目前,关系抽取主要以二元关系抽取为主,典型的二元关系抽取方法有基于 Bootstrapping[18]的 relation-specific extraction[2]和 OpenIE[3,4],其中,OpenIE[3,4]主张不事先定义抽取关系类型,而是训练一个分类器,判断一个句子中是否有关系值得抽取,这种一次性抽取待抽取文档中所有关系的思想符合实体活动抽取的任务需求.多元关系抽取多年来一直是关系抽取中的难点,文献[19]把多元关系分解为二元关系,利用二元关系抽取中的 Bootstrapping 思想进行多元关系抽取,但是抽取方法的准确性主要依赖于结构良好的种子实例,在实体活动类型未知的前提下并不适用.文献[20]提出实体踪迹(entity activity)的概念,在实体踪迹模型中,从核心实体、时间、地点和活动等信息维度对踪迹进行描述,并提出一种统计概率模型来判别一个句子是否为某个实体的踪迹.文献[20]以搜索引擎为背景,侧重算法的识别效率,模型对实体活动的刻画不够丰富;文献[20]提出的贝叶斯概率模型并没有明确指出句子中代表活动的单词或短语,与本文细粒度的实体活动抽取不同.事件抽取是在深层 NLP 的基础上,识别句子中动词所代表的事件类型,然后抽取事件的参数及属

性信息,最后进行事件融合[5,6].但是,事件抽取以事件为研究中心,不符合我们以实体为中心的研究需求;并且,事件抽取预先定义了事件类型,无法处理 Web 上大量未知的事件类型.

CRFs 是目前最好的用于解决序列数据分割标注问题的统计模型[15],但是由于 Linear-Chain CRFs 自身的限制,很多有用的特征不能完全利用,所以近年来许多研究者在 Linear-Chain CRFs 的基础上尝试集成更多的标签间的复杂关系.Sutton 等人[21]提出一种动态条件随机场(dynamic CRFs),作为一种特殊情况,利用阶乘 CRFs,在相同的观察序列下同时解决两种 NLP 任务,通过建模两个任务之间的相互联系提高准确率.黄健斌等人[22]针对 Web 记录集成中的模式匹配问题,提出了混合链条件随机场模型(MSCRFs),通过在内容相似的 Web 数据元素之间建立跳边,加强了对长距离依赖关系的处理.文献[23]提出一种 Skip-Chain CRFs,在已有的 Linear-Chain CRFs 的基础上添加成对的相似单词之间的关联边,通过对这些跳边的建模提高模型准确度.以上模型都是在基于内容的相似性方面建立跳边,不应用于单条自然语句.文献[17]在 2DCRFs 的基础上提出了二维关联边条件随机场(2DCC-CRFs),对二维环境下的长距离依赖建模,提高 Web 语义标注的准确率.文献[24]提出约束条件随机场(CDRFs),在 CRFs 的基础上引入可信约束和逻辑约束,其中,可信约束是指数据元素取某个语义标签的可信度,而逻辑约束是指语义标签之间存在的约束关系;最后,利用动态规划的方法求解 CCRFs.但在文献[24]中,标签之间的逻辑约束是人工的硬性规定,并未通过统计的方式对其建模.

5 结束语

本文从信息抽取的角度出发,研究 Web 实体活动抽取的方法.我们首先基于格语法定义了实体活动的多元关系模型,然后提出了基于 SVM 和扩展条件随机场的 Web 实体活动抽取方法.实验结果表明,该方法能够较好地适用于 Web 实体活动抽取.但是,由于 Web 信息的海量性,有必要在保证抽取效果的同时提高抽取效率,这是我们下一步研究的方向.

References:

[1] Weikum G, Theobald M. From information to knowledge: Harvesting entities and relationships from Web source. In: Paredaens J, Van Gucht D, eds. Proc. of the 29th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2010). ACM Press, 2010. 65−76. [doi: 10.1145/1807085.1807097]

[2] Hoffmann R, Zhang C, Weld D. Learning 5000 relational extractor. In: Hajic J, Carberry S, Clark S, eds. Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). ACL Press, 2010. 286−295.

[3] Zhu J, Nie ZQ, Liu XJ, Zhang B, Wen JR. StatSnowball: A statistical approach to extracting entity relationships. In: Quemada J, León G, Maarek YS, Nejdl W, eds. Proc. of the 18th Int'l Conf. on World Wide Web (WWW 2009). ACM Press, 2009. 101−110. [doi: 10.1145/1526709.1526724]

[4] Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O. Open information extraction from the Web. In: Veloso MM, ed. Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007). San Francisco: Morgan Kaufmann Publishers, 2007. 2670−2676. [doi: 10.1145/1409360.1409378]

[5] Ahn D. The stages of event extraction. In: Proc. of the Workshop on Annotating and Reasoning about Time and Events. Sydney: ACL Press, 2006. 1−8.

[6] Li SQ, Liu PY, Zhao TJ, Lu Q, Li HJ. PKU_HIT: An event detection system based on instances expansion and rich syntactic features. In: Proc. of the 5th Int'l Workshop on Semantic Evaluation (ACL 2010). ACL Press, 2010. 304−307.

[7] Zhai YH, Liu B. Web data extraction based on partial tree alignment. In: Ellis A, Hagino T, eds. Proc. of the 14th Int'l Conf. on World Wide Web (WWW 2005). Chiba: ACM Press, 2005. 76−85. [doi: 10.1145/1060745.1060761]

[8] Cortez E, da Silva AS, Goncalves MA. ONDUX: On-Demand unsupervised learning for information extraction. In: Elmagarmid AK, Agrawal D, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Indianapolis: ACM Press, 2010. 807−818. [doi: 10.1145/1807167.1807254]

[9] Fillmore CJ. Frames and the semantics of understanding. Quademi di Semantica, 1986,6(2):222−253.

[10] Sarawagi S. Information extraction. Foundations and Trends in Databases, 2008,1(3):261−377. [doi: 10.1561/1500000003]

[11] Carreras X, Marquez L. Introduction to CoNLL-2005 shared task: Semantic role labeling. In: Proc. of the 9th Conf. on Computational Natural Language Learning. Ann Arbor, 2005. 152−164.

[12] 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]

[13] Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 2005, 31(1):71−106. [doi: 10.1162/0891201053630264]

[14] Cortes C, Vapnik V. Support vector networks. Machine Learning, 1995,20(3):273−297. [doi: 10.1007/BF00994018]

[15] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). Williamstown: Morgan Kaufmann Publishers, 2001. 282−289.

[16] Nie ZQ, Wu F, Wen JR, Ma WY. Extracting objects from the Web. Technical Report, MSR-TR-2004-128, Microsoft Research, 2004..

[17] Ding YH, Li QZ, Dong YQ, Peng ZH. Semantic annotation of Web data based on ensemble learning and 2d correlative-chain conditional random fields. Chinese Journal of Computers, 2010,33(2):267−278 (in Chinese with English abstract).

[18] Agichtein E, Gravano L, Snowball: Extracting relations from large plain-text collections. In: Proc. of the 15th ACM Conf. on Digital Libraries. New York: ACM Press, 2000. 85−94. [doi: 10.1145/336597.336644]

[19] Xu FY, Hans U, Li H. Automatic event and relation detection with seeds of varying complexity. In: Proc. of the 21st National Conf. on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conf. (AAAI 2006). Boston: AAAI Press, 2006. http://www.aaai.org/Library/Workshops/ws06-07.php

[20] Yao CL. Research on the extraction of Web entities and discovery of entity activities [Ph.D. Thesis]. Beijing: Peking University, 2008 (in Chinese with English abstract).

[21] Sutton F, Rohanimanesh K, McCallum A. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. (ICML 2004). Alberta: ACM Press, 2004. 693−723.

[22] Huang JB, Ji HB, Sun HL. Integration of heterogeneous Web records using mixed skip-chain conditional random fields. Journal of Software, 2008,19(8):2149−2158 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2149.htm [doi: 10.3724/SP.J.1001.2008.02149]

[23] Charles S, Andrew M. Collective segmentation and labeling of distant entities in information extraction. Technical Report, TR 04-49, University of Massachusetts, 2004.

[24] Dong YQ, Li QZ, Zheng YQ, Xu XY, Zhang YX. Semantic annotation of Web object using constrained conditional random fields. In: Chen L, Tang CJ, Yang J, Gao YJ, eds. Proc. of the 11th Int'l Conf. on Web-Age Information Management (WAIM 2010). LNCS 6184, Berlin, Heidelberg: Springer-Verlag, 2010. 28−39. [doi: 10.1007/978-3-642-14246-8_6]

附中文参考文献:

[12] 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]

[17] 丁艳辉,李庆忠,董永权,彭朝晖.基于集成学习和二维关联边条件随机场的Web数据语义标注方法.计算机学报,2010,33(2):267−278.

[20] 姚从磊.Web实体抽取与实体踪迹发现研究[博士学位论文].北京:北京大学,2008.

[22] 黄健斌,姬红兵,孙鹤立.基于混合跳链条件随机场的异构Web记录集成方法.软件学报,2008,19(8):2149−2158. http://www.jos.org.cn/1000-9825/19/2149.htm [doi: 10.3724/SP.J.1001.2008.02149]

张传岩(1985−),男,山东滨州人,硕士生,主要研究领域为信息抽取,数据集成.

彭朝晖(1978−),男,博士,副教授,CCF高级会员,主要研究领域为数据库,信息检索.

洪晓光(1964−),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为Web数据管理,数据库优化.

李庆忠(1965−),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为大规模网络数据管理,数据集成.