ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 4,Jul. 2017,pp. 838-843
DOI:10.16337/j.1004-9037.2017.04.023

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于最大熵的越南语新闻事件元素抽取方法

周 枫 庙介璞 潘清清 严 馨 余正涛

(昆明理工大学信息工程与自动化学院,昆明,650500)

摘 要:越南与中国一水相依,是重要的政治、军事和经济合作邻国,然而针对越南语新闻事件元素的提取研究非常匮乏。本文针对越南语特点,提出一种基于最大熵模型的越南语新闻事件元素抽取方法。该方法针对越语句子结构和词汇语义的特点,采用最大熵算法,选取上下文、邻近触发词以及邻近实体作为特征,定义特征模版,训练获得越南语新闻事件模型,实现新闻事件元素抽取。抽取实验结果表明本文提出的方法抽取新闻事件元素的准确率达到80%以上。

关键词:越南语;最大熵;机器学习;新闻事件元素抽取

中图分类号:TP391.1 **文献标志码**:A

Extractiond Method of Vietnamese News Event Elements Based on Maximum Entropy

Zhou Feng, Miao Jiepu, Pan Qingqing, Yan Xin, Yu Zhengtao

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: The study on extraction of Vietnamese news event elements is rare, while Vietnam is a significant neighboring country with political, military and economic cooperation, which is just at a distance of a river with us. According to the Vietnamese characteristics, this paper puts forward a method of Vietnamese news event element extraction based on maximum entropy model. This method selects the context, adjacent trigger words and neighboring entities as features, delimits feature templates, trains Vietnamese news events model and achieves the extraction of news event elements of Vietnamese on the basis of the characteristics of the Vietnamese sentence structure and lexical semantic using the maximum entropy algorithm. The experimental result of the extraction shows that the accuracy of the news event elements extracted by the method proposed in this paper reaches more than 80%.

Key words: Vietnamese; maximum entropy; machine learning; news event elements extraction

引 言

新闻是指报纸、电台、电视台和互联网等媒体所传播的信息的一种称谓。近年来,随着计算机技术的飞速发展,新闻的传播媒体也越来越向互联网倾斜。人们每天通过互联网新闻获取大量信息,“信息超载”成为人们天天面临的重要问题。从海量新闻文本中有效获取人们感兴趣的信息成为了国内外的一个研究热点,也就是新闻事件元素抽取[1]。新闻作为一种重要的传播媒介,在中越两国的发展和交流

* 基金项目:国家自然科学基金(61462055,61562049)资助项目。
收稿日期:2014-05-09;修订日期:2016-10-14

中发挥着重要的引导作用。中越两国悠久的历史渊源与现阶段特殊的地缘关系,使得汉越新闻信息的共享、交流尤为重要。面对海量的越语新闻信息,及时、精准地抽取越语新闻中的事件元素信息正是本文研究的主要问题。新闻事件元素抽取的任务,就是抽取出新闻事件信息中的时间、地点和参与者,即常用的 3 类事件元素:人名事件元素、地名事件元素和时间事件元素[2]。近年来,事件元素抽取吸引了大量的专家学者,并开展了大量的基础研究工作。事件抽取最初是由 MUC 会议[3](ACE 会议的前身)提出的,并作为一项重要的评测项目,吸引了众多的专家学者参与评测,ACE 会议也在 2005 年引入了事件抽取的评测任务。新闻事件元素抽取主要包括 3 个步骤:(1)新闻事件类型识别,即区分一个事件所属的类型;(2)事件类别的识别,关键任务是识别引起事件发生的词,是区分事件类别的关键因素;(3)识别出事件的相关元素,即事件发生的时间、地点和参与者。本文主要就越南语新闻事件元素抽取开展相关研究,抽取新闻中的 3 类事件元素,即地名、人名和时间 3 类事件元素,这 3 类事件元素都是新闻文本当中重要的实体信息,但事件元素和命名实体又有着本质的区别。针对每一篇新闻,其中都含有大量的实体信息,这些实体可能只有一部分是事件元素,而并不是所有实体都是事件元素,这就需要对命名实体进行分类,来确定哪些实体是事件元素。

为了能够更好地理解和描述越南语新闻事件,以便于对其中的事件元素进行抽取,需要分析越南语新文本的特点以及对越南语新闻事件当中包含的重要事件元素进行抽取。越南语新闻文本中的事件元素可以认为是新闻事件的关键信息,主要包括事件发生的时间、地点以及参与者等。通过对越南语新闻文本的分析可以发现,新闻的重要信息主要集中在新闻的标题、各段首句与新闻关键词对应的关键句中。本文提取这些句子用于实体识别,并对实体识别的结果进行分析,找出读者感兴趣的信息。

1　新闻事件元素抽取

新闻事件元素抽取是信息抽取领域的一个重要研究方向,其任务主要是从无结构或半结构化的新闻文本中抽取出有用的信息,如新闻事件发生的时间、地点、事件的主要参与者和事件的前因后果等。这对正确快速地从大量新闻文本中获取读者感兴趣的信息非常重要。现阶段对于事件元素抽取主要有基于模式匹配的方法和基于机器学习分类器的方法。模式匹配的方法是早期信息抽取的主要方法之一。Chklovski[4]利用语句匹配模板(Lexcial-Syntactic pattern,LSP)抽取具有事件关系的资源,并将抽取的结果整理成一个称为“VerbOcean”的知识库。文献[5]利用将模式泛化为特征性量的方法通过多特征层级模式来识别事件论元角色。模式匹配依赖人为的预先定义的事件模板,需要消耗大量的人力资源并且可移植性不强,召回率低,但是拥有不错的准确率。近年来对于机器学习算法在事件抽取领域的应用越来越多,对于支持向量机、最大熵算法的应用越来越广泛。文献[5]于 2002 年首次将最大熵分类器融入到事件元素抽取中。文献[6]结合 MegaM 和 Timbl 两种机器学习方法分别实现了事件抽取中事件类别识别和事件元素识别,在 ACE 英文语料上均取得了不错的效果。Ritter 在大规模 Twitter 上提取潜在事件[7],提高了 14%的 F 值。国内对于中文事件元素的相关研究起步较晚,Li Qi 通过全局特征联合提取事件元素与触发词[8],在 ACE 语料上取得了不错的效果。丁效[9]将最大熵分类器应用于音乐领域的事件抽取上,并采用触发词聚类的方法自动发现事件类型,完成了在特定领域事件抽取的一次尝试。而对于越南语的事件元素抽取,国内外还未见相关研究资料报道。本文针对越南语新闻的特点,主要抽取新闻事件发生的时间、地点和新闻事件的主要参与者,也就是事件抽取中最主要的时间、地点和人物 3 类事件元素。选取最大熵分类器作为事件元素的分类器。原因在于使用最大熵方法建模时,实验者只需集中精力选择特征,而不需要花费精力去考虑如何使用这些特征,所以可以更加灵活地选择特征,结合更加丰富的信息,且不需要额外的独立假设或内在约束。

越南语是一种孤立语[10],动词无变化,词既没有文法上的性别跟数的形式,也没有文法上格的变化,形容词也不需要跟被修饰的名词保持文法上的性、数和格上的一致。它以不变的根词的词序和虚词来表示语法关系。句子的主语在谓语之前,宾语和补语在动词之后,名词修饰语一般在名词之后,但数

词、量词修饰语在名词之前。词序或虚词改变后，语义也随之而变。越南语的文句机构是：主语-谓语-宾语(SVO)。

跟多数东南亚语言(泰语、老挝语和马来语等)一样，越南语也是形容词后置的语言。所以“越南语”就不是“ Việt(越)Nam(南)tiếng(语)”，而是“tiếng Việt Nam”；“京族的正式语言”就应该写成“ngôn ngữ(言语)chính thức(正式)của[属于、的] Quốc gia(民族)Kinh(京)”。

2 基于最大熵的越南语新闻事件元素抽取

2.1 特征模板定义

越南语新闻事件元素的类型选择可以看成是一个二元分类问题，为每类事件元素定义二元分类器，这里只需要定义人物、时间和地点这 3 个二元分类器。特征定义如下：

(1)上下文特征 F_a，在考虑越南语新闻文本上下文特征时，要考虑候选实体是不是事件元素，就需要考虑该实体的左右两侧越南语词的特征以及对应词性的特征，因此一个候选实体是不是事件元素与它左右的词特征和词性特征密切相关。因此上下文特征对于新闻事件元素抽取有至关重要的作用。本文选择越南语新闻文本中实体两侧的词和词性作为基本特征：(a)：实体左侧 n 的词语，(b)：实体右侧 n 的词语，(c)：实体左侧 n 个词性，(d)：实体右侧 n 个词性，其中 n 为 1～4 之间的整数。

(2)邻近触发词特征 F_b，在识别越南语新闻事件元素时，实体左右两侧的触发词对确定该实体是不是事件元素有着非常重要的作用，比如李克强在北京会见阮善仁，“会见”是触发词，它决定了事件的类型，同时也标志着该句子里的人名和地名都是事件元素，所以实体邻近的触发词对事件元素的提取起到至关重要的作用。又因为越南语动词没有变化，既没有时态也没有格上的不同，所以选取实体左右侧最邻近的一个触发词作为特征：(a)实体左侧最近的一个触发词，(b)实体右侧最近的一个触发词。

(3)邻近实体类型特征 F_c，新闻事件元素左右两侧 1～2 个实体很有可能也是事件元素，这个跟该实体的类型有很大关系。比如在会见事件中，一个人名事件元素后若紧跟一个地名和一个时间实体的话，那么该地名实体和时间实体很有可能也是事件元素。越南语句子的主语在谓语之前，宾语和补语在动词之后，名词修饰语一般在名词之后，但数词、量词修饰语在名词之前。词序或虚词改变后，语义也随之而变。越南语的文句机构是：主语-谓语-宾语(SVO)。例如在访问事件中，人名实体后紧跟一个地名实体，那么这个人名实体就很可能是访问事件的人名事件元素。所以选取实体左右两侧实体的类型作为特征融入：(a)同一句子中实体左侧的 m 个实体的类型，若没有则以 null 代替。(b)同一句子中实体右侧的 m 个实体的类型，若没有则以 null 代替，其中 m 是 1～2 之间的整数。综上所述取的越南语新闻事件元素特征如表 1 所示。

针对一条越南语新闻片段来详细描述各个特征，例如：Chiều 12/1/2014, Thủ tướng Nguyễn Tấn Dũng dẫn đầu đoàn Việt Nam đã đến thủ đô Phnom Penh, bắt đầu chuyến thăm làm việc tại Campuchia.(2014 年 1 月 12 日下午，越南总理阮晋勇率领代表团到柬埔寨首都金边开始国事访问。)

首先对该则新闻进行分词、词性标注和命名实体识别，结果如下：Chiều/N 12/1/M ,/, Thủ tướng/N Nguyễn Tấn Dũng/Np đến/V đầu/V đoàn/N Việt Nam/Np đã/R đến/V thủ đô/N Phnom Penh/Np,/, bắt đầu/V chuyến/V thăm/V làm việc/V tại/E Campuchia/Np ./. 可以发现，结果中有 6 个动词，但是这些动词不全是事件触发词，所以进一步

表 1 越南语新闻事件元素特征定义

Tab. 1 Features of Vietnam news event elements

特征	特征意义
WL_2	实体左侧向前两个词及词性
WL_1	实体左侧向前 1 个词及词性
WR_2	实体右侧向前 2 个词及词性
WR_1	实体右侧向前 1 个词及词性
VL	实体左侧的触发词
VR	实体右侧的触发词
EL_2	实体左侧向前 2 个实体的类型
EL_1	实体左侧向前 1 个实体的类型
ER_2	实体右侧向前 2 个实体的类型
ER_1	实体右侧向前 1 个实体的类型

进行去停用词处理,只留下 dân dâu 和 thâm,这两个动词就是该则新闻的触发词。新闻中的实体有 Chiêu 12/1(12 月 1 日下午)、Nguyên Tên Dêng(阮晋勇)、Viât Nam9 越南 0,Phnom Penh(金边)和 Campuchia(柬埔寨)。分别是时间、人名、地名、地名和地名这 5 个实体。考虑人名事件元素 Nguyên Tân Dũng(阮晋勇),根据上述定义的特征,这里设 $n=2$,$m=2$,Nguyên Tân Dũng 为人名事件元素。F_a:Nguyên Tân Dũng(阮晋勇)前两个词分别为"Thù tuóng(总理)"和",",词性分别为"N"和",",后两个词分别为"dân fâu(领导)"和"doàn(代表团)",词性分别为"V"和"N"。F_b:Nguyên Tân Dũng(阮晋勇)左侧并没用触发词,设为 null,右侧触发词是 dàn dâu。F_c:Nguyên Tán Dũng(阮晋勇)左侧仅有一个时间命名实体(TIM),右侧的两个实体类型分别是地点命名实体 LOC 和 LOC。

可以用特征向量来表述该人名事件元素,其中在描述特征时,若多个单子词表示一个词语,则将这多个单子词用下划线进行连接,以方便构建特征向量,如图 1 所示。其中第 1 列为标记列,其后依次为左侧第 1 个词以及该词的词性,左侧最近的 1 个触发词,右侧最近的 1 个触发词,左侧的第 1 个实体类型,右侧的第 1 个实体类型,其他标签以此类推,这里的实体类型采用简洁定义为 PER,LOC 和 TIM。

```
people Thủ_tướng N , , dẫn_đầu V đoàn N null dẫn_đầu TIM null LOC LOC
```

图 1　事件元素特征集

Fig. 1　Feature set of event elements

针对收集的越南语新闻语料,一共标注了 300 篇越南语新闻,并对人名、地名和时间 3 类事件元素分别构建了一个最大熵的分类器,以地名事件元素为例,图 2 为地名事件元素标注的特征集合。

```
place phòng N Quốc Np đã R tích_cực A ơn trợ null null LOC LOC
place vận_tải V Giao_thông N để E trao_đổi V vận_tải trao_đổi null null LOC null
nplace Vương_quốc Np Thái_từ N và CC Công_nương N tiếp có PER null LOC null
place Thượng_viện N của E thăm V Việt_Nam Np biểu thăm null null LOC null
place thăm V Campuchia Np . . null null thăm null LOC null null null
place null null null null mong V WHO Np null mong null null null null
place Công_nương N và CC sắp V thăm V từ sắp null null loc null
place thăm V sắp V . . thăm null loc null null null
```

图 2　地名事件元素标注特征集合

Fig. 2　Feature set of place event elements

2.2　模型构建

最大熵模型[11]的基本思想就是为所有已知的因素构建模型,同时把未知的因素排除在外。在预测一个候选实体是否属于时间、地点或是人名事件元素的过程中,会涉及各种各样的因素,这里假设 x 就是一个由这些因素构成的向量,变量 y 即为当前特征影响下对应的事件元素类型。$p(y|x)$ 指模型将候选实体在特征 x 下预测为事件类型 y 的概率。最大熵模型要求 $p(y|x)$ 在满足一定约束条件的情况下,必须使得下面定义的熵取得最大值,则

$$H(p)=-\sum_{x,y}p(y\mid x)\log p(y\mid x) \tag{1}$$

这里的约束条件实际上就是指所有已知的事实,可以表示为

$$f_i(x,y)=\begin{cases}1 & (x,y)\text{ 满足一定条件}\\ 0 & i=1,2,3,\cdots,n\end{cases} \tag{2}$$

式中:$f_i(x,y)$ 为最大熵模型的特征。最终概率输出为

$$p^*(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_i\lambda_i f_i(x,y)\right) \tag{3}$$

式中 λ_i 为每个向量的权重,且

$$z(x)=\sum_y\exp\left(\sum_i\lambda_i f_i(x,y)\right) \tag{4}$$

3 实验结果及分析

实验所采用的越南语新闻语料来源于互联网爬取的越南语新闻文本，该语料集主要包括政治、经济、社会、文化、生活、技术和科学 7 大类型。一共标注了 3 000 篇越南语新闻文本。数据集中的 5/6 作为训练越南语新闻事件元素最大熵模型。数据集中的 1/6 作为测试集，用于验证抽取的效果。数据集基本情况如表 2 所示。

表 2 越南语新闻事件元素识别训练测试数据集

Tab. 2 Vietnam news event element recognition training test data sets

训练语料集/篇	测试语料集/篇
3 000	500

为了验证本文抽取的越南语新闻事件元素的效果，采用基于模板和最大熵模型相结合的事件元素抽取方法进行实验，若模板匹配成功，则抽取出对应的事件元素，若模板匹配不成功，则采用最大熵模型进行事件元素抽取，本文选取传统的准确率，召回率及 F 值作为评测标准。实验评测结果如表 3 所示。

表 3 不同模板＋特征的实验结果

Tab. 3 Experimental results of different template＋feature

模板＋特征	人名事件元素			地名事件元素			时间事件元素		
	P	R	F	P	R	F	P	R	F
模板＋F_a	51.23	37.46	43.28	55.32	40.12	46.51	58.43	43.21	49.68
模板＋F_a＋F_b	55.67	40.98	47.21	60.12	48.43	53.65	66.56	50.32	57.31
模板＋F_a＋F_c	78.45	60.89	68.56	79.67	68.95	73.88	80.14	72.39	76.07
模板＋F_a＋F_b＋F_c	81.25	70.12	75.28	83.67	75.34	79.29	84.13	77.68	80.78

$$F = 2 \times P \times R/(P+R) \times 100\% \tag{5}$$

式中：P 为准确率，R 为召回率，其中

$$P = \text{正确识别的总数}/(\text{正确识别的总数} + \text{不正确识别的总数}) \times 100\% \tag{6}$$

$$R = \text{正确识别的总数} / \text{正确的总数} \times 100\% \tag{7}$$

经过实验验证，在采用模板和最大熵模型相结合的事件元素抽取方法中，最终确定当 $n=2$，$m=2$ 并在最大熵模型训练迭代次数为 150 时，在开发集上达到效果最优。从表 3 中的数据可以看出，采用不同的特征和模板相结合的方法识别效果不同，在仅使用模板与上下文基础特征 F_a 时效果较差，F 值不足 50%。在模板与上下文特征的基础上分别加入临近触发词特征 F_a 与相邻实体特征 F_c，效果均较模板＋F_a 有所提高，其中模板＋F_a＋F_c 的准确率、召回率均高于其他两种组合，可见临近实体特征对事件元素的抽取效果影响最大。当使用模板＋F_a＋F_b＋F_c 时，识别效果最好，两者相结合的方法在越南语新闻事件元素的抽取上取得了较高的准确率和召回率。由于越南语新闻事件元素抽取方面的研究尚处于起步阶段，未见有相关文献研究可用于对比，故本文参考文献[9]中的中文研究结果作为对比以说明本文特征选取的有效性。

如表 4 所示，与文献[9]相比，在人名、地名以及时间元素上的准确率，召回率与 F 值虽有差距，但大多属于正常误差范围，仅时间元素的抽取效果低于中文抽取效果 5.18%。分析越南语新闻文本发现，越南语在人物和地点上多为专有名词，且在表达方式上与中文相同或类似。而在时间上越南语分词工具将 năm(年)与 2014 切分成两个词，然而 năm 在越南语中除了年以外还有很常用的“男的”意思。诸如此类的多义词造成了一定程度上的识别偏差。

4 结束语

针对越南语新闻文本进行事件元素抽取任务，本文提出了一种基于模板和最大熵模型相结合的越南语新闻事件元素抽取方法。首先介绍了越南语新闻事件元素抽取的主要困难，介绍了越南语新闻的特点，根据该特点，并结合模板与最大熵模型进行越南语新闻事件元素的抽取。实验结果表明，上述方

表 4　不同评测对象的实验结果

Tab. 4　Experimental results of different evaluation object

评测对象	人名事件元素			地名事件元素			时间事件元素		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
音乐领域事件元素抽取	82.14	69.70	75.41	82.35	76.36	79.25	88.29	83.76	85.96
越南语新闻事件元素抽取	81.25	70.12	75.28	83.67	75.34	79.29	84.13	77.68	80.78
差距	−0.89	+0.42	−0.13	+1.02	−1.02	+0.04	−4.16	−6.08	−5.18

法在越南语新闻事件元素的抽取上取得了很好的效果。由实验结果可以看出临近实体特征对抽取效果显著。由于中国与越南的毗邻关系,且中文事件抽取技术已经相当成熟,下一步准备借助于中越双语对齐平行语料库,并挖掘更多实体特征与句法信息以提高越南语新闻事件元素抽取的性能。

参考文献:

[1] Wang Wei. Chinese news event 5W1H semantic elements extraction for event ontology population[C]// Proceedings of the 21st International Conference Companion on World Wide Web. JAPAN:ACM, 2012:197-202.

[2] Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ACE) program-tasks, data, and evaluation[C]//LREC. Lisbon:ELRA/ELDA,2004:837-840.

[3] Walker C S, Strassel S Medero J, et al. ACE 2005 multilingual training corpus[J]. Linguistic Data Consortium, Philadelphia, 2006:57.

[4] Chklovski T , Pantel P. VerbOcean: Mining the web for fine-grained semantic verb relations[C]//Conference on empirical methods in natural language processing. Spain:ACL,2004:33-40.

[5] Chieu H L, Ng H T. Named entity recognition: A maximum entropy approach using global information[C]//Proceedings of the 19th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics. San Francisco :Morgan Kaufmann Publishers,2002: 1-7.

[6] Ahn D. The stages of event extraction[J]. Proceedings of the Workshop on Annotating and Reasoning about Time and Events. Sydney, Australia:ACL, 2006: 1-8.

[7] Ritter A, Oren E, Sam C. Open domain event extraction from twitter[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Japan: ACM, 2012:1104-1112.

[8] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features[C]//Association for Computational Linguistics. Bulgaria:ACL,2013: 73-82.

[9] 丁效,宋凡,秦兵,等. 音乐领域典型事件抽取方法研究[J]. 中文信息学报,2011,25(2):15-20.
Ding Xiao, Song Fan, Qin Bing, et al. Research on typical event extraction method in the field of music[J]// Journal of Chinese Information Processing, 2011,25(2):15-20.

[10] 刘华明,毕学慧,王维兰,等. 基于最大熵和局部优先度的裂痕唐卡分割[J]. 数据采集与处理,2015,30(2):433-434.
Liu Huaming, Bi Xuehui, Wang Wailan, et al. Segmentation of RIP Tangka based on maximum entropy and local prlority [J]. Journal of Data Acquisition and Processing, 2015,30(2):434-433.

[11] Berger A L, Pietra V J D, Pietra S A D. A maximum entopy approach to natural language processing[J]. Computational liguistics, 1996,22(1):39-71.

作者简介:

周枫(1958-),男,副教授,研究方向:自然语言处理、软件工程应用技术。

庙介璞(1991-),男,硕士研究生,研究方向:自然语言处理,E-mail:iambeichen@163.com。

潘清清(1998-),男,硕士研究生,研究方向:自然语言处理,事件抽取等。

严馨(1969-),女,副教授、硕士生导师,研究方向:自然语言处理、数据挖掘理论及应用等。

余正涛(1970-),男,教授、博士生导师,研究方向:自然语言处理、机器翻译和机器学习等。