DOI: 10.3979/j.issn.1673-825X.202104220124

类型感知的汉越跨语言事件检测方法

张　磊[1,2],高盛祥[1,2],余正涛[1,2],刘　畅[1,2],陈瑞清[1,2]
(1.昆明理工大学 信息工程与自动化学院,昆明 650500;2.昆明理工大学 云南省人工智能重点实验室,昆明 650500)

摘　要: 针对汉越跨语言事件检测缺少平行语料,越南语标注困难,需要统一跨语言语义空间,且触发词存在较大的歧义和局限性等问题,提出基于事件类型感知的汉越跨语言事件检测方法。构造类型感知的注意力机制突显事件特征,融入汉越的词位置、词性和命名实体信息,并通过梯度反转(gradient reversal layer,GRL),实现有标注汉语和无标注越南语之间的对抗训练,将从大量汉语新闻文本中学到的语言无关的事件类型特征融入到联合特征提取器中,进行汉越跨语言的无触发词事件检测,缓解越南语的数据稀缺和触发词的局限性。实验中提出的方法较最好的基线模型在准确率上提升了 4.32%。

关键词: 汉越跨语言事件检测; 无触发词; 事件类型感知; 梯度反转; 语言对抗
中图分类号: TN391.3　　**文献标志码:** A　　**文章编号:** 1673-825X(2022)05-0803-09

Chinese-Vietnamese cross-language event detection method based on event type awareness

ZHANG Lei[1,2], GAO Shengxiang[1,2], YU Zhengtao[1,2], LIU Chang[1,2], CHEN Ruiqing[1,2]
(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, P. R. China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, P. R. China)

Abstract: In view of the lack of parallel corpus for Chinese-Vietnamese cross-language event detection, the difficulty of Vietnamese labeling, the need to unify the cross-language semantic space, and the large ambiguity and limitations of trigger words, this paper proposes an event-type-aware-based Chinese-Vietnamese cross-language event detection method. First, an event-type-aware attention mechanism is constructed to highlight event features, integrate Chinese and Vietnamese language differences into position, part of speech, and named entity information, and achieve confrontation training between tagged Chinese and unlabeled Vietnamese through gradient reverse layer (GRL). The language independent event type features learned from a large number of Chinese news texts are integrated into the joint feature extractor to detect Chinese Vietnamese cross language non-trigger word events. It alleviates the data scarcity and the limitation of trigger words in Vietnamese. The accuracy of the proposed method is 4.32% higher than that of the best baseline model.

收稿日期: 2021-04-22　　**修订日期:** 2022-08-21　　**通讯作者:** 高盛祥　gaoshengxiang.yn@foxmail.com
基金项目: 国家自然科学基金(61972186,61761026,61762056);国家重点研发计划(2018YFC0830105,2018YFC0830101,2018YFC0830100);云南高科技人才项目(201606);云南省重大科技专项计划(202002AD080001-5);云南省基础研究计划(202001AS070014,2018FB104)
Foundation Items: The National Natural Science Foundation of China (61972186, 61761026, 61762056); The National Key Research and Development Program (2018YFC0830105, 2018YFC0830101, 2018YFC0830100); The Yunnan High-Tech Talent Project (201606); The Yunnan Major Science and Technology Special Project (202002AD080001-5); The Yunnan Basic Research Program (202001AS070014, 2018FB104)

Keywords: Chinese-Vietnamese cross-language event detection; non-triggers; event-type awareness; gradient reversal; language adversarial

0 引 言

事件检测是自然语言处理(natural language processing, NLP)的重要主题之一,目标是在纯文本中识别特定的事件类型。汉越跨语言事件检测就是在汉语和越南语上实现双语事件检测。

目前在汉越事件方面的跨语言工作还很有限,其涉及到跨语言语义表征问题。汉语语料丰富而越南语作为小语种语料稀缺、数据标注困难,且汉越同属孤立语系[1],既存在相似之处又存在明显的差异性,给汉越事件检测带来了挑战。跨语言事件检测方法目前还没有系统的分类,针对跨语言问题的解决大致有以下 3 类。

1) 基于多语言多任务的方法。Alexis 等[2]提出跨语言模型 XLM-R,刷新了多个跨语言任务的记录包括跨语言事件检测,但是需要大量的语料,其中,越南语同样面临数据稀缺的问题,且模型较大需要很长的训练时间。Andrew[3]在多种语言上进行训练,利用依赖于语言的特征和不依赖于语言的特征提高跨语言事件抽取性能。

2) 基于跨语言词向量映射的方法。Anders 等[4]提出一个共享的、跨语言的向量空间,但仅仅只是跨语言词向量映射导致双语之间的语义粒度过大,在具体的任务上不能准确反映语言本身的层级语义特征。Ananya 等[5]通过使用卷积神经网络(convolutional neural networks,CNN),将所有实体信息片段、事件触发词、事件背景放入一个复杂的、结构化的多语言公共空间,然后从源语言注释中训练一个事件抽取器,并将它应用于目标语言。

3) 一般跨语言任务中还有翻译和对抗的方法。Guillaume[6]引入翻译系统桥接 2 种语言,但机器翻译模型需要大量的平行句对来训练,汉越上准确度还有待提升,且引入了噪声,如翻译错误和对齐错误等。Mei 等[7]基于对抗学习自适应(adversarial based domain adaptation)通过判别器和生成器将源域和目标域数据在数据空间或者特征空间对齐,在目标域数据未标记的情况下学习域之间的映射,即学习语言不变性的特征,实现跨语言的域适应(domain adaptation)。而 Yaroslav 等[8]令对抗学习(adversarial learning)与领域自适应相结合,并提出了梯度反转概念[9],使得模型的训练不需要如同生成对抗网络(generative adversarial net,GAN)的复杂训练过程。

主流方法依赖人工标注数据和平行语料,对于越南语来说,只有少量汉-越平行语料,有标注数据稀缺,且人工标注代价昂贵。事件检测的许多先进模型严重依赖监督方法中的巨量有标签数据,由于没有足够的越南语数据作为统计学习模型的训练支撑,模型在越语上的性能往往不佳。

人工标注需要对每个事件句标注事件触发器,注重文本中的事件触发器提及,即触发词的标注,忽视文本中蕴含的大量事件语义信息。在汉越事件检测中还存在触发词歧义和局限性等问题,如: S1: Một triệu người tị nạn(难民) ở Việt Nam, chỉ có Trung Quốc đã đứng ra(出席、出面) cứu trợ(救济)。译文: 越南百万难民,只有中国出面救济。

句中的词“đứng ra”原意“出席”指向会议事件,但通过前文的“người tị nạn(难民)”和“đứng ra”后面的“cứu trợ(援助)”才能准确的理解“đứng ra”的意思为“出面(救助)”。相比于“đứng ra”,“cứu trợ(援助)”更适合作为该句的触发词,或者给予更多的权重使其影响该句的事件类型归属。

触发词被定义为词或短语,注释者需要从给定句子中严格标注“最清晰”的词,越南语由音节构词,音节可以独立使用,或组成多音节词,事件触发词的标注和识别存在歧义,使得越南语事件检测受限于多音节词歧义。

本文提出一种事件类型感知的汉越跨语言事件检测模型。采取语言对抗的方式训练大量有标注的汉语语料和无标注越南语语料,在汉语和越南语分布之间存在转移的情况下,训练语言鉴别器,迁移汉语中事件类型信息到共享的特征提取器中,因此,经汉语训练的分类器也可以用于越南语[10]。融合词位置、词性和命名实体信息,通过基于事件类型的注意力机制凸显事件相关词的语义贡献,探索在没有清晰定义触发词的情况下检测事件,模糊触发词对于事件类型的影响。

1 基于事件类型感知的特征提取网络 F

每个事件类型都与某些词组相关联,准确捕获

这些词的语义信息[11]就能更好地判断其到底更偏向哪种事件类型。本文任务中没有标注触发词，为了捕获这些词的语义信息，提出基于事件类型感知的特征提取网络，将词位置、词性和命名实体等信息与双语词向量拼接，即汉语和越南语的差异信息送入模型。通过多个二进制分类对该任务进行建模，给定一个句子，它将被放入所有候选事件类型的二进制分类器中，输入是<sentence，事件类型>。网络在特征提取阶段依据候选事件类型等信息，使得注意力机制将更多的权重分配给与本事件类型相关的词上，如图 1 所示。

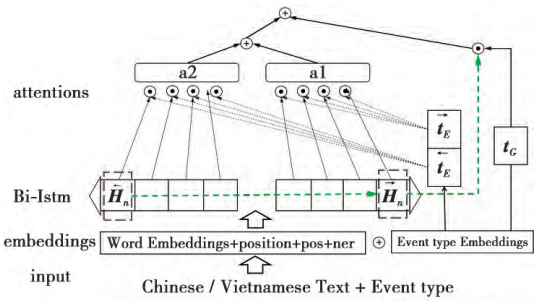

图 1　事件类型感知的联合特征提取网络

Fig.1　Joint feature extraction network for event type awareness

1.1　汉越双语词向量预训练

从大量未标记数据中学习的单词嵌入已经被证明能够捕获单词语义规则，本文中输入文档为单词序列 $X=w_1\cdots w_n$，其中每个 w_i 由其词嵌入 v_i 表示。由于汉越各自训练出的单语词嵌入向量矩阵 C 和 V 分布不同，需要找到最佳映射矩阵 W_C、W_V，使 CW_C、VW_V 在同一语义空间下。使用 Artetxe 等[12]的无监督方法，结合自学习算法来逐渐优化映射矩阵 W，将 W 约束为正交矩阵，即

$$WW^{\mathrm{T}}=W^{\mathrm{T}}W=I \tag{1}$$

在语义不变性情况下，汉越语义相同的词嵌入在公共语义空间中的距离更近，缓解汉语和越南语之间的差异，让跨语言模型更加健壮。并通过 Stanford NLP 得到汉越词位置、词性和名实体等信息扩充双语词向量。

1.2　基于 Bi-LSTM 的语义编码

长短时记忆网络（long short term memory，LSTM）[13]对依次输入的文本词嵌入，计算其上下文表示，使信息在其中持续保存，考虑词之间的依赖性，保留了诸如词顺序等重要信息，缓解了长期困扰循环神经网络的梯度消失或梯度爆炸问题[14-16]。而双向长短时记忆网络（Bi-LSTM）由一个前向 LSTM 与一个后向 LSTM 组成，某一时刻的输出由这两个方向上的状态共同决定，在具有记忆特性和顺序语义的基础上增加了捕捉逆序语义的能力，如图 2 所示。

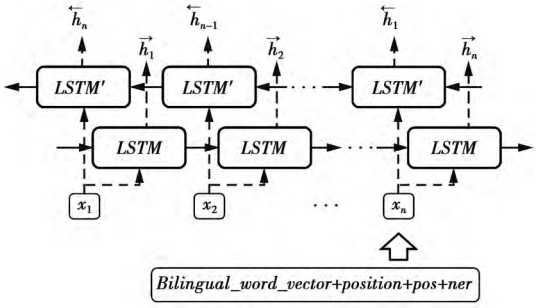

图 2　Bi-LSTM 结构图

Fig.2　Bi-LSTM structure diagram

汉越和越南语都没有严格意义的形态变化，语义主要靠语序和词序来表达，句子成分中主语、谓语和宾语相似，如：tôi ăn cháo，我吃饭（主语-谓语-宾语）。但在句法结构上存在语序差异[17]，如越南语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，特别是定语要放在所修饰的中心词之后[18]，这类越南语修饰语后置的语序与汉语相反，存在明显的镜像差异。如：新学期开始了，Ky hoc mơ i đã bắt đâu，后置修饰定语“mơi（新）”位于“Ky hoc（学期）”之后；一只可爱的狗，Một con chó dễthương，其中，修饰语“dễ thương（可爱的）”位于名词“chó（狗）”之后。

使用 Bi-LSTM 对输入的汉越词嵌入向量提取特征，输出 2 个方向的隐藏状态向量序列，越南语和汉语都存在与之相反的序列，不同于主流的正逆序列拼接再计算，注意力机制将分别依据 2 个事件类型嵌入矩阵对 2 个方向的隐藏状态序列计算注意力权重，减小 2 种语言的差异。

图 2 中，根据当前词嵌入向量 v_i，前一个正序隐藏层状态 $\overrightarrow{h}_{i-1}$ 和逆序隐藏层状态 $\overleftarrow{h}_{i-1}$，得到当前正序隐藏层状态和逆序隐藏层状态，即

$$\overrightarrow{h}_i=lstm(\overrightarrow{h}_{i-1},v_i) \tag{2}$$

$$\overleftarrow{h}_i=lstm(\overleftarrow{h}_{i-1},v_i) \tag{3}$$

1.3　事件类型感知的注意力机制

模型中的注意力机制借助事件类型来计算句子的向量表示，根据输入的事件类型 T 查表得到随机

初始化的 3 个事件类型嵌入: 正序$\overrightarrow{t_E}$、逆序$\overleftarrow{t_E}$指导注意力机制关注事件类型信息(事件相关词的局部语义信息),t_G 拟合句子的全局语义信息。句子事件类型的偏向性依赖事件相关词的局部语义和句子全局语义信息,句子总的表示由局部和全局信息加权得到。实验证明,通过调整局部信息和全局信息的加权比,模型能够更好的分类句子所属的事件类型。

给定句子隐藏状态向量输出$\overrightarrow{H}$、$\overleftarrow{H}$的第 k 个隐藏状态$\overrightarrow{h_k}$、$\overleftarrow{h_k}$,第 k 个词嵌入向量的注意力分数$\overrightarrow{a_k}$、$\overleftarrow{a_k}$由以下方程计算(逆序计算同理)。

$$\overrightarrow{a_k} = \frac{\exp(\overrightarrow{h_k} \cdot \overrightarrow{t_E})}{\sum_i \exp(\overrightarrow{h_i} \cdot \overrightarrow{t_E})} \tag{4}$$

$$\overleftarrow{a_k} = \frac{\exp(\overleftarrow{h_k} \cdot \overleftarrow{t_E})}{\sum_i \exp(\overleftarrow{h_i} \cdot \overleftarrow{t_E})} \tag{5}$$

模型中,目标事件类型的相关词预计获得比其他词更高的注意力权重。句子的表示 S_{att} 为

$$S_{\text{att}} = a_1^{\text{T}}\overrightarrow{H} + a_2^{\text{T}}\overleftarrow{H} \tag{6}$$

(6) 式中: a_1,a_2 是正逆序计算出的注意力权重向量;$\overrightarrow{H}$、$\overleftarrow{H}$为正序逆序的隐藏状态。Bi-LSTM 的最后输出$\overrightarrow{H_n}$和$\overleftarrow{H_n}$含有句子的全局信息,2 个拼接得到 H_n,通过目标事件类型 T 查表得到全局 t_G,S_{global} 期望捕获整个句子语义。

$$S_{\text{global}} = H_n \cdot t_G^{\text{T}} \tag{7}$$

$\mu \in [0,1]$ 是 S_{att}和 S_{global}之间权衡的超参数,输出定义为 S_{att} 和 S_{global} 的加权和: $\mu \cdot S_{\text{att}} + (1-\mu) \cdot S_{\text{global}}$。

2 汉越跨语言事件检测模型

模型由 3 部分构成: ①前文构造的基于事件类型感知并融入词位置; ②词性和命名实体等信息的汉越特征提取器 F; ③基于标准多层前馈网络的事件检测器 P 和语言鉴别器 Q,如图 3 所示。

F 旨在学习有助于预测事件分类器 P 的特征,并抑制语言鉴别器 Q,而训练有素的 Q 无法分辨出 F 提取特征的语种,这个特征可以看作是 2 种语言共有的,即语言无关而事件类型相关的。在 F 和 Q 之间有一个梯度反转层,使 F 的参数在 Q 和 P 中都参与梯度更新,一个最小化 P 分类误差; 一个是最大化 Q 分类误差。Q 试图为汉语输出更高的分数,为越南语输出更低的分数,因此,Q 是对抗性的。

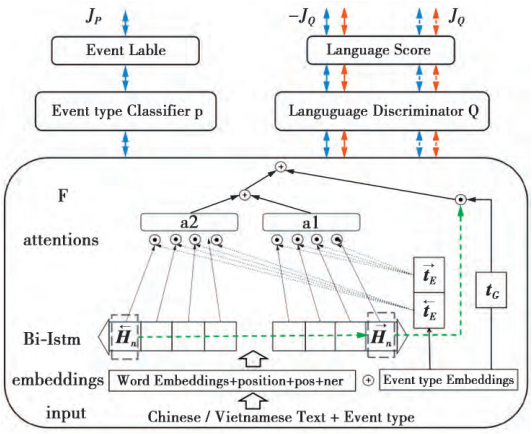

图 3 越南语为目标语言的跨语言事件类型感知模型

Fig.3 Cross-language event type awareness model with Vietnamese as the target language

根据 F 提取的隐藏特征 $f(x)$,P 末尾使用 softmax 层分类事件类型,Q 是一个二进制分类器,末尾有一个 sigmoid 层为语言打分,始终在 [0,1],表示输入文本 x 为汉语或越南语,训练过后,打分应趋向于 0.5。考虑汉语和越南语的联合隐藏特征 F 的分布 P_{F}^{ch}、P_{F}^{vi} 为

$$P_{\text{F}}^{\text{ch}} = P(F(x) \mid x \in chinese) \tag{8}$$

$$P_{\text{F}}^{\text{vi}} = P(F(x) \mid x \in vietnamese) \tag{9}$$

在训练时,未标记的汉语(蓝线)和越南语(黄线)通过语言鉴别器,而有标签的汉语文本通过事件检测器。J_p 和 J_q 是培训目标 P 和 Q。F、P 的参数一起更新(实线)。-JQ 和 JQ 的意思是期望最大化语言鉴别器 Q 的分类损失,为了学习语言不变的特征,对抗训练将使这两个分布尽可能接近以获得更好的跨语言泛化。根据 Kantorovich Rubinstein 对偶性[19] 最小化 P_{F}^{ch} 和 P_{F}^{vi} 之间的 Wasserstein[20] 距离 W,该距离有连续性,训练时提供更好的梯度。

$$W(P_{\text{F}}^{\text{ch}}, P_{\text{F}}^{\text{vi}}) = \sup_{\|g\|_L \le 1} \underset{f(x) \sim P_{\text{F}}^{\text{ch}}}{E}[g(f(x))] - \underset{f(x') \sim P_{\text{F}}^{\text{vi}}}{E}[g(f(x'))] \tag{10}$$

(10) 式中对于所有的 x 和 y,函数 g 应当满足利普希茨(Lipschitz)连续条件[21]。为了近似地计算 $W(P_{\text{F}}^{\text{vi}}, P_{\text{F}}^{\text{vi}})$,使用语言鉴别器 Q 作为公式中的函数 g,这需要 Q 的参数被剪辑到 $[-c,c]$。让 Q 参数化为 θ_q,然后 Q 的目标 J_q 变为

$$J_q(\theta_f) = \max_{\theta_q} \underset{F(x) \sim P_{\text{F}}^{\text{ch}}}{E}[Q(F(x))] - \underset{F(x') \sim P_{\text{F}}^{\text{vi}}}{E}[Q(F(x'))] \tag{11}$$

事件检测器 P 由 θ_p 参数化,使用 $L_p(\hat{y}, y)$ 二分

类交叉熵损失，l_p 是 P 预测正确标签的对数似然函数。在本文的方法中，每一个训练样本是 <句子，事件类型> 对，预测标签是 0 或 1 取决于该句子是否属于这个事件类型，因此，负样本会比正样本多很多，所以，本文给事件检测器增加了一个偏向正样本的损失。给定所有的训练样本数量为 M，x，y，θ 是模型的参数，δ 是 L2 normalization 权重。$1+y^{(i)}\cdot\beta$ 是偏置对于负样本为 1，对于正样本为 $1+\beta$，β 大于 0，本文为 Q 寻求以下损失函数的最小值为

$$J_p(\theta_f)=\frac{1}{M}\sum_{i=1}^{M}L_P(P(F(x)),y)(1+y^{(i)}\cdot\beta)+\delta\|\theta\| \tag{12}$$

最后，由 θ_f 参数化的联合特征提取器 F 最小化事件检测器损失 J_p 和语言鉴别器损失 J_q。

$$J_f=\min_{\theta_f}J_p(\theta_f)+J_q(\theta_f) \tag{13}$$

梯度反转层在正向传播时不会产生任何影响，在反向传播时，会将参数乘上 $-\lambda$，即分类器损失 $\frac{\partial L_d}{\partial\theta_f}$ 会被替换为 $-\lambda\frac{\partial L_d}{\partial\theta_f}$，这里，$\lambda$ 是一个平衡 P 和 Q 对 F 影响的协调训练超参数。通过这种方式，可以使用标准反向传播训练整个网络，模型随着梯度的更新既能学习到基于标签分类的歧视性特征又能学习到可供转移的语言不变特征。

3　实验及结果分析

3.1　数据和模型参数

中国和越南新闻网站在 2008—2020 年，基于相似主题和板块爬取大量的汉越可比语料，汉语数据 21 万条，其中 20 万条为训练集，1 万条为测试集。越南语数据 143 061 条为训练集，8 236 条为测试集，对所有汉语语料和越南语测试集进行事件类型标注<sentence，事件类型>，而越南语训练集不做标注，如图 4 所示。

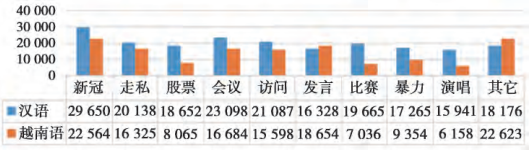

图 4　汉越可比语料规模

Fig.4　Chinese-Vietnamese comparable corpus scale

实验配置为 window10、Python3.7、Pytorch0.4.0。本文汉语和越南语采用 Pennington 等[22] 提出的 Glove 词向量初始化新闻文本，词向量维度 L 为 100，剔除词频小于 5 的词。为缓解过拟合现象，引入 dropout 用于事件检测器的全连接层。采用自适应矩估计 Adam 训练模型进行优化，它是一个基于随机梯度的优化器，具有自适应估计。

因为 F 和 Q 的训练可能不完全同步，在实践中观察到 F 比 Q 训练得更快，对模型的拟合速度和效果会产生一定影响[23]。所以每 k 轮训练将 λ 置为非零值一次。当 $\lambda=0$ 时，Q 的梯度不会反向传播到 F，这允许 Q 在 F 进行一次更新之前进行更多的迭代次数以适应 F。训练时设置了 2 个学习率：①Q 的多次自迭代 lr_1；②Q 和 F 的共同训练 lr_2。

3.2　评价指标

使用准确率 P、召回率 R 和 $F1$ 值衡量模型是否可以正确分类汉越双语新闻文本所属事件类型。TP 是所有含事件的文本中被正确检测出数量，FP 是检测出事件的文本中错误的和实际不含事件的数量，FN 是检测出不含事件而实际含事件的数量。P 是所有的文本中正确检测出事件类型的比例，R 是正确检测出事件类型的占实际含有事件类型的文本的比例，$F1$ 是 P 值和 R 值的一个综合度量值。

$$P=\frac{TP}{TP+FP} \tag{14}$$

$$R=\frac{TP}{TP+FN} \tag{15}$$

$$F_1=\frac{2PR}{P+R} \tag{16}$$

3.3　对比模型实验

实验评估有 7 个模型，如表 1 所示。

表 1　不同联合特征提取网络对比实验结果

Tab.1　Comparison of experimental results of different joint feature extraction networks

模型	Acc/%	Recall/%	F_1
CNN	44.64	45.81	0.452 2
RNN	37.41	35.75	0.365 6
Bi-LSTM-ATT	45.23	42.69	0.439 2
CNN-LSTM-ATT	47.66	44.64	0.461 0
平均网络模型	32.27	30.56	0.313 9
本文模型(Ⅰ)	51.98	49.87	0.509 0
本文模型(Ⅱ)	49.73	47.26	0.484 6

1) 本文模型(Ⅰ)：本文提出的基于类型感知的汉越跨语言事件检测模型。

2) 本文模型(Ⅱ)：本文提出的基于类型感知的

汉越跨语言事件检测模型,未扩展融合位置、词性和命名实体信息。

3) 平均网络模型: 将事件类型感知的联合特征提取网络替换为平均网络。

4) Bi-LSTM-ATT: Zhou 等[24]根据位置特征扩充字向量特征,通过多层注意力机制,提高 LSTM 模型输入与输出之间的相关性。

5) CNN: Zeng[25]首次提出用卷积神经网络从多个 level 提取词语和语句级别的特征,并进行了一层简单的卷积处理,来精简参数量并保留关键信息。

6) RNN: Wang[26]使用双向 RNN 网络,但两个方向的信息不是拼接,而是相加。

7) CNN-LSTM-ATT: 吴汉瑜等[27]结合 CNN 与 LSTM 的优点,能够捕获不同层次的关键模式信息和全局结构信息并对它们进行融合。

由表 1 可知,平均网络就是对输入文本的向量序列各取平均,这是最基本的特征提取方式,由于不区分各词向量之间的重要程度,事件检测的准确率只有 32.27%。RNN 取得除平均网络外最低的准确率和 F_1,而 CNN 能捕捉局部相关的关键信息,同时不存在 RNN 的梯度消失和梯度爆炸问题,相比于 RNN 取得了 6.23% 的提升。Bi-LSTM-ATT 解决了 RNN 存在的问题,加上使用了注意力机制为不同的信息分配不同权重,较 CNN 获得了 1.59% 的提升。CNN-LSTM-ATT 同时具有循环神经网络可以提取文本的全局结构信息,卷积神经网络局部特征提取和注意力机制的优点,准确率达到了 47.66%。

本文模型准确率相较于 CNN-LSTM-ATT 和 Bi-LSTM-ATT 得到了 2.07% 和 4.5% 的提升。分析原因,虽然 CNN-LSTM-ATT 和 Bi-LSTM-ATT 都使用了注意力机制,但都是无外部信息的自注意力,没有利用事件相关信息,输入文本的词向量之间的权重分配依据文本自身。而本文中基于事件类型信息的注意力机制根据重要的事件类型等外部信息,指导句子中的词获得占比权重,即结合事件任务特性学习句子向量特征。在扩展词位置、词性和命名实体信息后,准确率达到了 51.98%,表明模型不仅可以获取句子内部特定事件和词之间的依赖关系,也可以捕获更多有利于检测事件的相关的特征信息。

为了验证在没有标注触发词下模型仍然能有效利用事件相关词的语义信息来检测事件类型,设置 1 个正类、3 个负类事件类型标签(暴力、走私、会议、演唱),让模型输出如下 2 个例句的注意力权重热力图,如图 5 所示。

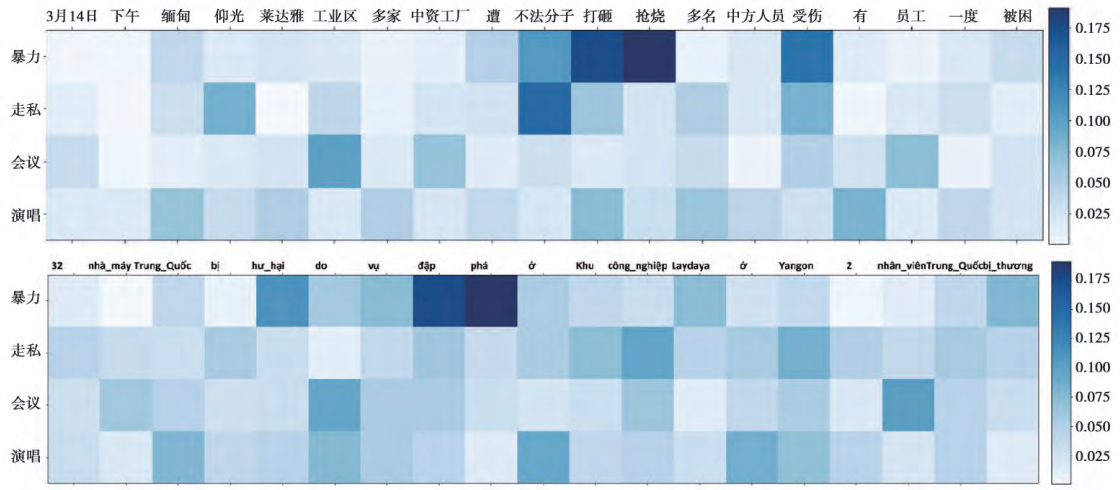

图 5 例句中词的注意力权重

Fig.5 Attention weight of words in example sentences

S2: 3 月 14 日下午,缅甸仰光莱达雅工业区多家中资工厂遭不法分子打砸抢烧,多名中方人员受伤,有员工一度被困。

S3: 32 nhà máy Trung Quốc bị hư hại do vụ đập phá ở Khu công nghiệp Laydaya ở Yangon, 2 nhân viên Trung Quốc bị thương(仰光莱达雅工业区打砸事件造成 32 家中资工厂受损,2 名中国员工受伤)。

汉语和越南语句子中,“打砸抢烧”、“hư_hại(损害)”和“đập_phá(破坏)”是暴力事件重要的词,模型有效捕捉到了这一点,为其分配了较高注意力权重。在走私、会议和演唱等负样本事件中没有相关的词,模型为每个词分配了几乎相等的注意力

权重。

3.4　语言对抗训练的有效性

语言鉴别器的对抗训练实现了汉越的跨语言事件检测,为进一步验证模型跨语言的有效性,验证模型通过大量有标签汉语与无标签越南语对抗训练。

提升越南语事件检测准确率的效果,与去掉语言鉴别器的实验作对比,添加汉语验证集,从第 5 轮到第 30 轮的迭代次数中对比汉语事件检测和越南语事件检测准确度提升情况,结果如表 2 所示。

由表 2 可知,在没有语言对抗的实验中,应为训练集中有事件类型标注,汉语事件检测的准确率随着多轮迭代得到大幅提升,而越南语没有标注,事件检测准确率提升几乎没有。在语言对抗的实验中,通过汉越语言的对抗训练,越南语准确率得到明显提升,在 30 轮后较无语言鉴别器的情况提升 0.319,证明语言鉴别器的对抗训练的确有助于无标签越南语事件检测的准确率提升。

同时在汉语语料规模不变的情况下,探究无标签越南语料规模对模型性能的影响,使越南语数据从 5 万到 14 万,每增加 1 万验证模型在越南语事件检测的准确率,如表 3 和图 6 所示。

表 2　语言对抗对准确率的影响

Tab.2　Impact of language discriminator on accuracy

Epoch	Chinese (with Q)	Vietnamese (with Q)	Chinese (without Q)	Vietnamese (without Q)
5	0.493	0.293	0.482	0.093
7	0.587	0.371	0.575	0.125
10	0.669	0.405	0.689	0.086
13	0.684	0.426	0.782	0.167
16	0.712	0.459	0.803	0.142
19	0.731	0.462	0.837	0.098
21	0.773	0.482	0.859	0.257
24	0.791	0.509	0.874	0.182
27	0.805	0.516	0.883	0.195
30	0.816	0.529	0.895	0.210

表 3　无标注语料规模对准确率的影响

Tab.3　Impact of unlabeled corpus size on accuracy

语料(万)	5	6	7	8	9	10	11	12	13	14
VI 准确率/%	43.29	43.89	44.84	45.26	46.83	47.91	48.24	49.56	51.06	51.98

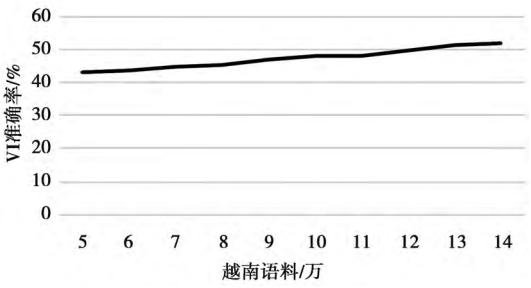

图 6　越南可比语料规模与准确率

Fig.6　Comparable corpus scale and accuracy rate in Vietnam

由表 3 可知,14 万的越南语无标签语料比 5 万在准确率上提升了 0.086 9,且由图 6 可知,在目前的 10 万级语料下,越南语料的规模与模型准确率呈正相关,而无标签的汉越可比语料获取相对容易,可以通过构建更多的汉越可比语料来显著提升跨语言事件检测模型的性能。

3.5　模型训练关键参数对性能的影响

本文模型中协调训练的参数 k 对实验结果影响较大。遂通过评估不同参数数值对越南语事件检测准确率的影响,分析参数特性以实现模型性能最好。实验中模型迭代至第 30 轮,其余参数相同。验证 k 与越南语事件检测准确率的关系,如图 7 所示。

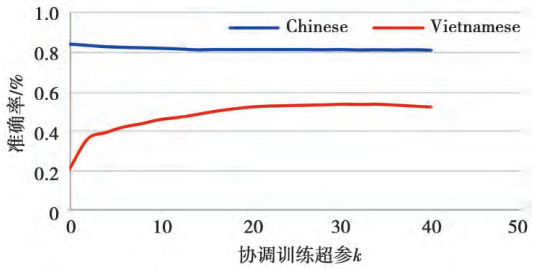

图 7　参数 k 对汉越事件检测准确率的影响

Fig.7　Influence of k on the accuracy of C-V incident detection

由图 7 分析可知,随着 k 的增大汉语的准确率几乎无变化,而越南语的准确率则提升明显。协调训练的参数 k 表示每 k 轮次训练中只有一次 Q 的梯度能反向传播到 F,即模型中一条汉语有标签语料训练一次,相应就有 k 对汉越可比语料输入了语言鉴别器 Q 进行对抗训练。分析 k 变大对于每条汉

语的有标签学习无影响,所以汉语准确率不受 k 影响,但同时让更多汉越可比语料进行了语言的对抗训练,加强了有标注汉语信息向越南语训练模型的参数迁移能力,使得无标注越南语的准确率提升。

但过大的 k 会让模型参数量翻倍,对模型收敛速度、内存和加速显卡的显存容量十分不友好,可以看到,当 k 值达到并超出 [20,30] 后出现了下降的趋势。所以本文将 k 置于一个既能使汉语标注信息充分向越南语训练模型传递,又能使模型收敛速度不太慢,同时目前硬件条件能满足的大小。模型中其余参数经过验证皆为最优,如表 4 所示。

表 4 模型的超参数

Tab.4 Hyper parameters of the model

参数	数值与范围
δ	0.000 1
β	0.9
μ	0.35
k	25
λ	0.01
lr_1	0.000 5
lr_2	0.000 5
dropout	0.2
Q 剪辑范围	[-0.01,0.01]

4 结束语

本文提出一种基于事件类型感知的汉语跨语言事件检测模型,利用丰富的汉语语言信息提高了越南语事件检测的准确性,缓解了越南语数据稀疏的问题,并通过基于事件类型感知的特征提取网络,模糊处理事件触发词,缓解了传统方法中单语歧义性和触发词局限性等问题。实验表明,将事件类型等语言无关而使事件相关的语义信息融入特征提取阶段会使跨语言事件检测性能有所提高。

参考文献:

[1] 武氏河.越南语与汉语的句法语序比较 [J].云南师范大学学报,2005,3(6):65-68.
WU S H. Comparison of sentence and French order between Vietnamese and Chinese [J]. Journal of Yunnan Normal University. 2005, 3(6): 65-68.

[2] ALEXIS C, KARTIKAY K, NAMAN G, et al. Unsupervised crosslingual representation learning at scale [C]// In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Seattle, Washington, United States: [s.n.], 2020: 8440-8451.

[3] ANDREW H, YANG Y M, JAIME C, et al. Leveraging Multilingual Training for Limited Resource Event Extraction [C]//The 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: [s.n.], 2016: 1201-1210.

[4] ANDERS S, IVAN V, SEBASTIAN R, et al. Cross Lingual Word Embeddings [J]. Morgan & Claypool Synthesis Lectures on Human Language Technologies, 2020, 46(1): 245-248.

[5] SUBBURATHINAM A, LU D, JI H, et al. Cross-lingual Structure Transfer for Relation and Event Extraction [C]//In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). HongKong, China: [s.n.], 2019, 313-325.

[6] GUILLAUME L, ALEXIS C. Cross-lingual language model pretraining [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: [s.n.], 2019: 7057-7067.

[7] WANG M, DENG W H. Deep visual domain adaptation: A survey [J]. Computer Vision and Pattern Recognition. 2018, 312(3): 135-153.

[8] YAROSLAV G, EVGENIYA U, HANA A J. Domain-adversarial training of neural networks [J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.

[9] YAROSLAV G, VICTOR L. Unsupervised domain adaptation by back propagation [J]. Proceedings of the 32nd International Conference on Machine Learning, 2015, 37(15): 1180-1189.

[10] SHAI B D, JOHN B, KOBY C, et al. A theory of learning from different domains [J]. Machine Learning, 2010 79(1-2): 151-175.

[11] IGNACIO I, MOHAMMAD T P, ROBERTO N. Embeddings for Word Sense Disambiguation: An Evaluation Study [EB/OL].(2016-08-01) [2021-04-06]. http://sidenoter.nii.ac.jp/acl_anthology/P16-1085/.

[12] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings [EB/OL]. (2018-05-16) [2021-04-06]. https://arxiv.org/abs/1805.06297.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[14] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities [J]. Proceedings of the National Academy of Sciences, 1982, 79

(8): 2554-2558.

[15] JORDAN M I. Serial order: Aparallel distributed process sing approach [M]. San Diego: University of California, Institute for Cognitive Science, 1986.

[16] ELMAN J L. Finding structure in time [J]. Cognitive Science, 1990, 14(2): 179-211.

[17] 王振晗,何建雅琳,余正涛,等.融合句法解析树的汉越卷积神经机器翻译 [J].软件学报, 2020, 31(12): 3797-3807.
WANG Z H, HE J Y L, YU Z T, et al. Chinese Vietnamese Convolutional Neural Machine Translation with Syntactic Parsing Tree [J]. Journa of Software, 2020, 31(12): 3797-3807.

[18] 黄敏中.实用越南语语法 [M].北京: 北京大学出版社, 1997.
HUANG M Z. Practical Vietnamese Grammar [M]. Beijing: Peking University Press, 1997.

[19] CÉDRIC V. Optimal transport: old and new [M]. Berlin: Springer Science & Business Media, 2008.

[20] MARTIN A, SOUMITH C, LÉON B. Wasserstein generative adversarial networks. [EB/OL]. (2017-08-06) [2021-04-06]. https://proceedings.mlr.press/v70/arjovsky17a.html.

[21] SONI K, SONI R P. Lipschitz Behavior and Characteristic Functions [J]. SIAM Journal on Mathematical Analysis, 2012,4(2): 302-308.

[22] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha, Qatar: [s.n.], 2014: 1532-1543.

[23] CHEN X, SUN Y, BEN A, et al. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification [J]. Transactions of the Association for Computational Linguistics, 2018(6): 557-570.

[24] ZHOU P, SHI W, TIAN J, et al. Attention Based Bidirectional Long Short Term Memory Networks for Relation Classification [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: [s.n.], 2016: 207-212.

[25] ZENG D J, LIU K, LAI S W, et al. Relation classification via convolutional deep neural network [C]//Proceedings of COLING. Dublin, Ireland: [s.n.], 2014: 2335-2344.

[26] ZHANG D X, WANG D. Relation classification via recurrent neural network [J]. Big Data Mining and Analytics, 2015, 1(3): 234-244.

[27] 吴汉瑜,严江,黄少滨,等.用于文本分类的 CNN_BiLSTM_Attention 混合模型 [J]. 计算机科学, 2020, 47(11A): 24-27.
WU H Y, YAN J, HUANG S B, et al. CNN_BiLSTM_Attention Hybrid Model for Text Classification [J]. Computer Science, 2020, 47(11A): 24-27.

作者简介:

张　磊(1996—),重庆人,男,硕士研究生,主要研究方向为跨东南亚语言新闻事件关系识别与演化分析研究。E-mail: 1655011716@qq.com。

高盛祥(1977—),女,云南大理人,副教授,博士,硕士生导师,主要研究方向为自然语言处理、机器翻译及信息检索。E-mail: gaoshengxiang.yn@foxmail.com。

余正涛(1970—),男,云南宣威人,教授,博导,博士,研究方向为自然语言处理、信息检索、机器翻译、机器学习、智能系统与决策分析等。E-mail: ztyu@hotmai.com。

(编辑: 刘　勇)