文章编号：1003-0077(2023)04-0045-07

# 基于中文触发词指导的越南语新闻事件检测

寇梦珂[1,2]，高盛祥[1,2]，余正涛[1,2]，毛存礼[1,2]，朱恩昌[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500)

**摘 要**：触发词的识别在事件检测任务中起着至关重要的作用。目前没有越南语触发词标记语料，而中文触发词标记语料较为丰富，根据表达相同观点但语言不同的句子通常有相同或相似的语义成分这一特征，该文提出一种基于中文触发词指导的越南语新闻事件检测方法。首先采用对抗学习的方法将两种语言映射到同一语义空间下，然后将映射后的中文触发词嵌入指导模型识别越南语新闻中的触发词信息，最后进行事件类型的分类。通过在越南语新闻事件检测的实验结果表明，在中文触发词指导下的越南语新闻事件检测取得了较好的效果。

**关键词**：越南语新闻；事件检测；触发词；对抗学习

**中图分类号**：TP391 **文献标识码**：A

## Vietnamese News Event Detection Based on the Guidance of Chinese Triggers

KOU Mengke[1,2], GAO Shengxiang[1,2], YU Zhengtao[1,2], MAO Cunli[1,2], ZHU Enchang[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunmong University of Science and Technology, Kunming, Yunnan 650500, China)

**Abstract**: The recognition of triggers plays an important role in detecting events. There is no Vietnamese triggers marking corpus at present, but the Chinese triggers marking corpus is relatively rich. Based on the observation that sentences conveying identical idea but in different languages usually have the same or similar semantic components, the paper puts forward a detection method of Vietnamese news event guided by Chinese triggers. Firstly, the two languages are mapped into the same semantic space by adversarial learning. Then the Chinese triggers embeddings are mapped into the guidance model to identify the triggers information in Vietnamese news. Finally, the events types in Vietnamese news are classified. The experimental result shows that the Vietnamese news event detection achieves a better performance on the guidance of Chinese triggers.

**Keywords**: vietnamese news; event detection; triggers; adversarial learning

## 0 引言

事件检测是事件抽取中的一个关键任务，可以从海量的新闻文本中快速、准确地获取事件信息，其研究成果应用于后续的信息检索和阅读理解以及问答任务。越南语新闻事件检测的目的是检测越南语新闻文本中的事件触发词，然后识别触发词所对应的事件类型，事件触发词通常是引发事件的单词。给定一个越南新闻句："Một chiếc xe tăng đã nổ súng vào một khách sạn.(译文：一辆坦克向旅馆开火。)"，需要先识别这个句子中是否包含事件触发词"nổ súng vào(译文：开火)"，并将其分类为特定的事件类型"Cuộc xung đột(译

**收稿日期**：2020-09-13 **定稿日期**：2020-10-20

**基金项目**：国家重点研发计划(2018YFC0830105，2018YFC0830100)；国家自然科学基金(61972186，U21B2027，61761026，61762056)；云南高新技术产业发展项目(201606)；云南省重大科技专项计划(202002AD080001，202103AA080015)；云南省基础研究计划(202001AS070014，2018FB104)；云南省科技人才与平台计划(202105AC160018)

文：冲突）”。

目前，Ji等人[1]将事件检测任务转换为分类任务，将文本语义信息直接嵌入低维空间，然后基于这些特征向量检测给定句子中的每个单词是否为触发词以及它触发的事件类型。由此可见，触发词是一个事件指称中最能代表事件发生的词，是决定事件类型的重要特征。Li等人[2]使用有监督方法识别触发词并进行事件检测。Tong等人[3]针对标记数据的长尾问题，利用开放域触发知识从大量未标记的数据中提取未标记触发词。目前，事件检测所用的语料大多为ACE数据集，在英文和中文新闻数据集中有大量的事件触发词标记数据，但是在越南语新闻数据集中没有触发词的标记数据。

训练一个事件检测模型需要标注大量的数据。数据标注的复杂性和高成本，使得现有的数据在不同语言之间是稀缺和不平衡的。根据表达相同观点但用不同语言表达的句子，通常有相同或相似的语义成分，于是Chen等人[4]使用英语在并行文档上标记事件，以获取用于增强中文事件检测的附加数据；Zhu等人[5]使用机器翻译来获取用于训练的附加标记数据；Hsi等人[6]将嵌入投影方法与双语事件抽取的多语种特征提取相结合。事件检测任务在英语、中文等大语种上的研究工作较多且取得了较好的效果，但是在越南语新闻事件检测方面的工作还相当有限。因此，为了解决越南语触发词标记语料缺失问题，本文根据多语言语义的一致性特征，针对中越相关新闻报道构建了中越双语新闻数据集。

针对越南语新闻触发词标记缺失的问题，在中越相关新闻的数据集下，将两种语言的新闻文本映射到同一个语义空间下，使用中文触发词指导越南语新闻事件检测。首先在编码过程中融入实体信息，采用对抗学习的方法将两种语言映射到同一语义空间下，然后通过注意力机制，引入映射后的中文触发词嵌入指导模型关注越南语新闻中的触发词信息，最后利用得到的注意力上下文向量进行事件类型的多分类，进而实现越南语新闻事件检测。

## 1 相关工作

事件检测（Event Detecting，ED）是近年来自然语言处理领域的一个研究热点。事件检测旨在检测非结构化文本中的事件触发词。事件检测任务所用的方法主要分为四类：基于特征的方法、基于表示的方法、基于外部资源的方法、基于神经网络的方法。

（1）在基于特征的方法中，由于自然语言的灵活性和多样性，同一触发词可以在不同的上下文中触发不同的事件，事件触发词可能非常模糊，Ahn等人[7]使用词汇特征、句法特征和外部知识来提取事件；Li等人[8]提出联合模型捕捉触发词和参数的组合特性，这些传统方法严重依赖于手工特征，限制了模型的可伸缩性和健壮性。

（2）在基于表示的方法中，Chen等人[9]使用CNN进行自动特征提取。Nguyen等人[10]使用更复杂的结构来构建事件之间的依赖关系。Nguyen等人[11]对非连续 $n$-grams进行建模以获得更高的性能，此类方法避免了复杂的特征工程，能够自动学习数据中的显著特征，从而获得更好的性能。

（3）在基于外部资源的方法中，Chen等人[12]使用FreeBase通过远程监控标注新的训练数据、Liu等人[13]利用FrameNet中的事件来引导训练。Liu等人[14]利用了额外的参数信息和FrameNet，基于外部资源的方法则将特征学习与来自外部资源的大量数据结合起来，以实现进一步的改进。

（4）随着深度学习的发展，神经网络可以自动学习输入序列的特征，进行词级分类。这些模型只能在固定大小的窗口中捕获本地上下文特性。Chen等人[15]设计综合模型来探索触发词之间的相互依赖性，但它们的性能受到特定语言中标记数据量的限制。近年来，神经网络作为一种从数据中自动学习文本表示的有效方法，在英语语言环境下取得了不错的效果。

不同的语言通常有不同的特点和习语。多语言事件检测尝试在不同语言之间传递知识以提高性能。一种语言的歧义表达在另一种语言中可能有非歧义对应，这可以为消除歧义提供补充信息。Feng等人[16]提出了多语言事件检测方法，使用混合神经网络训练了一个不需要任何编码特征的多语言事件检测模型，Huang等人[17]为了有效地转移资源，开发了一种利用多级对抗性迁移的新的神经结构，在词级对抗训练中，在不使用任何平行语料库或双语词典的情况下，将源语言单词投影到目标语言相同的语义空间中。本文首次构建了中越双语新闻语料库，并使用对抗学习的方法将两种语言映射到同一语义空间下。

由于在一个给定的句子中标记触发词是非常耗时的，Liu等人[18]提出根据目标事件类型计算句子的表示来解决在未标记触发词基础上的事件检测。

为了解决越南语新闻触发词标记缺失的问题，本文同样采用未标记触发词的方法进行越南语新闻事件检测，不同的是本文首次使用与越南语语义相似的中文新闻中可标记的事件触发词来辅助越南语新闻事件检测。这在一定程度上解决了越南语触发词标记缺失的问题。

## 2　方法

本文提出了一种基于中文触发词指导的越南语事件检测方法，其核心是在中越相关新闻数据集下，利用对抗学习的方法将两种语言映射到同一语义空间下，使用中文触发词指导越南语新闻事件检测。整体框架如图 1 所示：左边这部分采用对抗学习的方法将两种语言映射到同一个语义空间下，右边这部分首先融入了实体向量来捕捉词级别的语义信息，使用 BiLSTM 捕获句子的语义信息，然后通过注意力机制将映射后的中文触发词指导模型找到越南语句子中的触发词信息，最后利用得到的注意力上下文向量进行事件类型的多分类，进而实现越南语新闻事件检测。

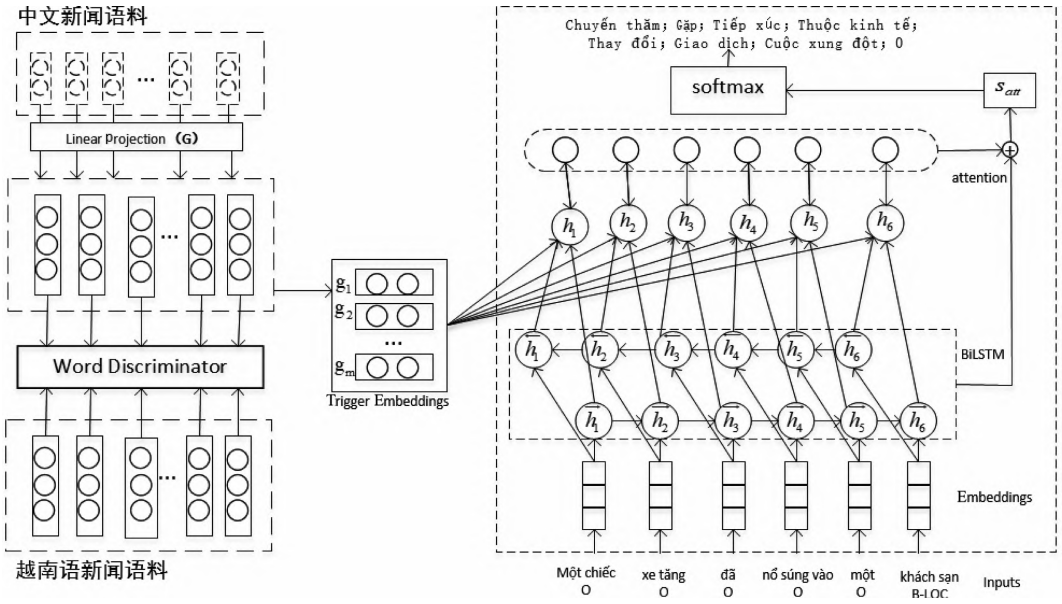


图 1　模型整体框架

### 2.1　输入层

从大量未标记数据中学习的词向量被证明能够捕获单词的有意义的语义规则。首先，在越南语语料上预训练了越南语词向量，得到词向量词表 $\boldsymbol{W}^{N\times d_z}$。然后根据 underthesea 工具[①]中的实体标记类型，为每种实体标记都随机初始化一个实体向量，得到实体向量词表 $\boldsymbol{E}^{M\times d_o}$。通过查找词向量词表和实体向量词表，所有的输入单词和实体标记都转化为低维向量。

给定一个包含 $n$ 个词的越南语新闻句 $S=\{w_1,w_2,\cdots,w_n\}$，对于 $S$ 中的每一个词 $w_i$，都被 underthesea 标记出实体类型 $e_i$；然后通过词向量词表 $\boldsymbol{W}^{N\times d_z}$ 查询到 $w_i$ 对应的词向量 $\boldsymbol{V}_{w_i}$，并且通过实体向量词表 $\boldsymbol{E}^{M\times d_o}$ 查询到 $e_i$ 对应的实体向量 $\boldsymbol{V}_{e_i}$；最后，将词向量与实体向量拼接起来作为 $w_i$ 最终的向量表示 $\boldsymbol{V}_i$，如式(1)所示。

$$\boldsymbol{V}_i=[\boldsymbol{V}_{w_i};\boldsymbol{V}_{e_i}] \tag{1}$$

将 $S$ 中的每个词 $w_i$ 都按照上述方式表示成向量 $\boldsymbol{V}_i$，用⊕运算符进行向量方向的拼接操作，则句子 $S$ 的语义表示矩阵 $\boldsymbol{M}_s$，如式(2)所示。

$$\boldsymbol{M}_s=\boldsymbol{V}_1\oplus\boldsymbol{V}_2\oplus\cdots\oplus\boldsymbol{V}_n \tag{2}$$

### 2.2　中越双语词向量映射

本文采用对抗学习的方法将越南语和中文映射到同一语义空间下。首先采用 skip-gram 扩展模型[19]的方法预训练出中文词向量 $\boldsymbol{V}_c=[a_1,a_2,\cdots,a_{d_s}]\in\boldsymbol{C}^{K\times d_s}$ 和越南语词向量 $\boldsymbol{V}_w=[b_1,b_2,\cdots,b_{d_z}]\in\boldsymbol{W}^{N\times d_z}$。其中，$K$ 和 $N$ 分别是词表大小，$d_s$ 和 $d_z$ 分别表示的是中文词向量维度和越南语词向量维度。

① https://github.com/undertheseariplp/underthesea

然后使用映射函数 $f$ 将中文投影到与越南语相同的语义空间中，如式(3)所示。

$$\boldsymbol{V}_{\tilde{c}} = f(\boldsymbol{V}_c) = \boldsymbol{V}_c\boldsymbol{U} \tag{3}$$

其中，$\boldsymbol{U}^{d_s \times d_z}$ 是映射矩阵，$\boldsymbol{V}_{\tilde{c}} \in \widetilde{\boldsymbol{C}}^{K \times d_z}$ 是投影后的中文词向量，$\widetilde{C}^{K \times d_z}$ 是投影后的中文词向量词表。

约束转换矩阵 $\boldsymbol{U}$ 与奇异值分解(SVD)正交，以减少参数搜寻空间，如式(4)所示。

$$\boldsymbol{U} = \boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}, \quad \boldsymbol{A}\sum\boldsymbol{B}^{\mathrm{T}} = \mathrm{SVD}(\boldsymbol{V}_c\boldsymbol{V}_c^{\mathrm{T}}) \tag{4}$$

为了优化映射函数 $f$，本文引入一个多层感知机作为词鉴别器，使用越南语词向量和映射后的中文词向量作为输入，输出单个标量。$D(w_i^*)$ 表示 $w_i^*$ 来自于越南语词表的概率。单词鉴别器使用二元交叉熵损失，如式(5)、式(6)所示。

$$L_{\mathrm{dis}}^{w} = -\frac{1}{I_{s;z}} \cdot \sum_{i=0}^{I_{s;z}} (y_i \cdot \log(D(w_i^*)) + (1-y_i) \cdot \log(1-D(w_i^*))) \tag{5}$$

$$y_i = \delta_i(1-2\varepsilon) + \varepsilon \tag{6}$$

其中，$\delta_i = 1$ 表示单词是来自于 $z$，$\delta_i = 0$ 表示单词是来自于 $s$。$I_{s;z}$ 表示从 $z$ 和 $s$ 的词汇表中一起采样出的单词的个数。$\varepsilon$ 是添加到正负标签的平滑值。

映射函数 $f$ 和单词鉴别器 $D$ 是两个对抗层，翻转单词标签，通过最小化损失来优化，如式(7)、式(8)所示。

$$L_f^{w} = -\frac{1}{I_{s;z}} \cdot \sum_{i=0}^{I_{s;z}} ((1^- y_i) \cdot \log(D(w_i^*)) + y_i \cdot \log(1-D(w_i^*))) \tag{7}$$

$$y_i = \delta_i(1-2\varepsilon) + \varepsilon \tag{8}$$

本文采用对抗学习的方法来将两种语言映射到同一语义空间下，使用随机梯度下降(stochastk gradient descent，SGD)依次训练单词鉴别器和映射函数，以最小化 $L_{\mathrm{dis}}^{w}$ 和 $L_f^{w}$。

触发词向量：给定中文新闻文本并标记句子中的触发词。通过映射矩阵将中文触发词映射到与越南语同一语义空间下，得到映射后的中文词向量词表 $\widetilde{\boldsymbol{C}}^{K \times d_z}$。然后从向量词表 $\widetilde{\boldsymbol{C}}^{K \times d_z}$ 中取出映射后的中文触发词向量 $\boldsymbol{V}_g = [r_1, r_2, \cdots, r_{d_z}] \in G^{m \times d_z}$。所有映射后的中文触发词转化为一组映射向量 $\boldsymbol{G}^{m \times d_z} = \{\boldsymbol{V}_{g_1}, \boldsymbol{V}_{g_1}, \cdots, \boldsymbol{V}_{g_m}\}$ 用于捕获越南语句子中隐藏的触发词。

### 2.3 BiLSTM 层

使用 BiLSTM 对输入的越南语新闻句子进行建模，运行在单词和实体嵌入的连接序列上。BiLSTM 可以看作两个单向的 LSTM，包含正向的 LSTM 和逆向 LSTM，使当前时刻的输出能与前一时刻的状态和后一时刻的状态都产生联系。

将一个包含 $n$ 个词的越南语新闻句子中的每个词的词向量按顺序输入 BiLSTM 单元构成的神经网络中，得到句子的一组隐状态 $H = \{h_1, h_2, \cdots, h_n\}$，$h_i$ 是句子中第 $i$ 个词的隐状态。在这一阶段的每个步骤中，前向 BiLSTM 在 $t$ 时刻的输入 $w_t$ 和先前隐状态 $h_{t-1}$ 计算当前的隐状态 $\overrightarrow{h}_t$，接着反向运行 BiLSTM 来生成后向隐状态 $\overleftarrow{h}_t$。

前向的 LSTM 与后向的 LSTM 结合组成 BiLSTM。与 LSTM 不同的是，输入层的数据会经过向前和向后两个方向进行推算，最后输出的隐状态再进行拼接，再作为下一层的输入，如式(9)所示。

$$\boldsymbol{h}_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{9}$$

### 2.4 触发词注意力机制层

每种类型的事件通常由一组特定的词触发，这些词称为事件触发词。例如，“Cuộc xung đột(冲突)”事件通常是由“对抗”、“袭击”等词触发的。因此，事件触发词是完成事件检测任务的重要线索。根据一组中文触发词向量 $\boldsymbol{G}^{m \times d_z} = \{\boldsymbol{V}_{g_1}, \boldsymbol{V}_{g_1}, \cdots, \boldsymbol{V}_{g_m}\}$ 和经 BiLSTM 编码的越南语新闻句子的一组隐状态 $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$，计算每个触发词向量 $\boldsymbol{V}_g = [r_1, r_2, \cdots, r_{d_z}] \in G^{m \times d_z}$ 与各个隐状态 $\boldsymbol{h}_i \in \boldsymbol{H}$ 之间的注意力权重，得到一组注意力权重向量 $\boldsymbol{\Lambda}^{m \times n} = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$。具体来说，给定 $G$ 中第 $k$ 个中文触发词向量 $\boldsymbol{V}_{g_k}$ 与 $t$ 时刻的隐藏状态 $\boldsymbol{h}_t$ 之间的注意力权重由式(10)计算，在这个模型中，越南语新闻目标事件类型的触发词期望获得比其他词更高的权重。

$$\alpha_k = \frac{\exp(\boldsymbol{h}_t \cdot \boldsymbol{V}_{g_k}^{\mathrm{T}})}{\sum_{i=1}^{n} \exp(\boldsymbol{h}_i \cdot \boldsymbol{V}_{g_k}^{\mathrm{T}})} \tag{10}$$

计算完 $\boldsymbol{V}_{g_k}$ 与所有时刻的隐藏状态 $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$ 之间的注意力权重后，得到一个注意力权重向量 $\boldsymbol{\alpha}_k = [\alpha^1, \alpha^2, \cdots, \alpha^n]$。遍历完 $\boldsymbol{G}^{m \times d_z} = \{\boldsymbol{V}_{g_1}, \boldsymbol{V}_{g_1}, \cdots, \boldsymbol{V}_{g_m}\}$，则得到一组注意力权重向量 $\boldsymbol{\Lambda}^{m \times n} = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$。然后，求取这一组权重向量中权重最大的元素，把该元素所在的那一个向量作为当前输入句子的最终注意力权重向量，记为 $\boldsymbol{\alpha}_{max} = [\alpha^1, \alpha^2, \cdots, \alpha^n]$。因为 $G$ 中与当前输入句子最相关的中文触发词向量将求得更大的注意力权重。

最后，将 $\boldsymbol{\alpha}_{\max}$ 与 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$ 进行加权求和，即可得到当前输入句子的向量表示 $\boldsymbol{s}_{att}$，如式(11)所示。

$$\boldsymbol{s}_{\text{att}}=\sum_{i=1}^{n}\alpha^i\boldsymbol{h}_i \tag{11}$$

其中，$i=1,2,\cdots,n$。

### 2.5　输出层

将当前输入句子的向量表示 $\boldsymbol{s}_{\text{att}}$ 输入 softmax 层，得到预测事件类型的概率分布 $p$，如式(12)所示。

$$p=\text{softmax}(\boldsymbol{W}\cdot\boldsymbol{s}_{\text{att}}+\boldsymbol{b}) \tag{12}$$

其中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为 softmax 层的权重和偏置。

### 2.6　损失和优化

将优化目标函数定义为交叉熵损失如式(13)所示。

$$J(\Theta)=-\ln p(y\mid s,\Theta)+\lambda_1 L_{\text{dis}}^{w}+\lambda_2 L_f^{w} \tag{13}$$

其中，$s$ 表示训练数据集中的所有样本数，$y$ 表示真实事件类型的类别，$L_{\text{dis}}^{w}$ 为词鉴别器，$L_f^{w}$ 为映射函数，$\lambda_1$ 和 $\lambda_2$ 是超参数，$\Theta$ 表示所有参数。在优化方法上，采用随机梯度下降(SGD)来最小化目标函数。

## 3　实验

### 3.1　数据集

到目前为止，还没有公开的中越双语新闻数据集，因此，根据事件检测任务中通用的 ACE2005 数据集的格式构建了中越双语新闻数据集。首先在越南网站爬取了 813 篇越南语新闻文本，根据越南新闻文本在百度咨询上爬取了相对应的 4 065 篇中文新闻。参考 ACE 的事件标注体系标记出中文新闻文本中的触发词和事件类型、越南语新闻文本中的事件类型。其中越南语共 25 089 条新闻句，并将划分为训练集、验证集、测试集。本文构建的语料中划分了 7 种事件类型和 1 种非事件类型，如表 1 所示。

**表 1　事件类型及触发词**

| 事件类型 | 事件触发词 |
|---|---|
| chuyến thăm(访问) | 拜访，出访，考察，探访…… |
| Gặp(会见) | 接待，见面，接见，会谈…… |
| Tiếp xúc(合作) | 联合，交流，合作，合同…… |
| Thuộc kinh tế(经济) | 衰退，下降，上升…… |
| Thay đổi(换届) | 推举，选举，推选，投票竞选…… |
| Giao dịch(贸易) | 出口，进口，转让…… |
| Cuộc xung đột(冲突) | 争端，对抗，侵犯，冲突，袭击…… |

### 3.2　实验设置

本文的模型采用 Tensorflow 深度学习框架实现。在所有的实验中，使用最大长度为 40 的句子作为输入，也就是说当句子的长度大于 40 时，截取前 40 个词。当句子的长度不足 40 时，使用特殊字符填充。使用 300 维的预训练词向量，随机初始化实体向量维度为 50、批量大小为 100。为了防止过拟合，在分类时加入 L2 正则项。

### 3.3　评价标准

使用准确率($P$)、召回率($R$)和 $F_1$-measure($F_1$)来评估效果。

$P$：正确预测的事件在总预测事件中所占的比例。

$R$：正确预测的事件在真实事件中所占的比例。

$$F_1=\frac{2\times P\times R}{P+R} \tag{14}$$

### 3.4　实验结果分析

为验证本文方法是否能提升事件检测的效果，设置了第一组实验。在越南语新闻数据集上将本文方法与 CNN、GCN 进行比较，在不标记触发词的基础上与基线方法(TBNNAM)作对比。实验结果如表 2 所示。

**表 2　不同方法性能对比**　　(单位：%)

| 方法 | $P$ | $R$ | $F$ |
|---|---|---|---|
| CNN | 73.25 | 66.15 | 69.27 |
| GCN | 75.60 | 64.90 | 70.12 |
| TBNNAM | 76.18 | 64.82 | 70.73 |
| ours | 77.34 | 66.95 | 71.58 |

通过对比实验可知，GCN 的效果优于 CNN，主要是由于引入依存句法信息可以捕获到 CNN 未捕获到的信息。TBNNAM 方法中 LSTM 最后一个状态所捕获的全局信息对该任务也很重要，这种由注意力机制捕获的全局信息和局部信息是互补的。但是本文使用的 BiLSTM 比 LSTM 能够捕获句子中更多的语义信息。实验结果表明，本文方法效果

更好。

针对词嵌入层融入的编码特征进行探究，为了验证将实体信息融入词向量是否能提升事件检测的效果，设置了第二组实验。该组实验对比了加入实体前后对模型的影响。实验结果如表 3 所示。

**表 3 编码特征对模型性能的影响**

（单位：%）

| 编码特征 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 无实体向量 | 70.41 | 66.72 | 68.91 |
| 有实体向量 | 77.34 | 66.95 | 71.58 |

通过对比实验可知，实体的标注可以捕捉词语的语义信息。加入实体向量之后，模型的准确率 $P$、召回率 $R$、$F_1$ 值与本文模型相比均有所上升，也就证明了加入实体向量后可以提高事件检测的性能。

由于缺失越南语触发词标记语料，本文引入丰富的中文触发词语料来解决越南语触发词语料缺失问题，为了验证融入中文触发词是否能提升事件检测的效果，设置了第三组实验。该组实验对有无中文触发词对事件检测的效果做了对比，实验结果如表 4 所示。

**表 4 中文触发词指导的模型性能对比**

（单位：%）

| 融入特征 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 无触发词 | 60.67 | 56.34 | 56.28 |
| 有触发词 | 77.34 | 66.95 | 71.58 |

通过对比实验可知，有触发词标记的效果明显优于无触发词标记的效果。根据不同的语言对于同一新闻事件句下的表达具有一致性的特征，使用中文触发词可以找到对应的越南语句子中的触发词，进而完成越南语新闻的事件检测。

## 4 总结

本文提出了基于中文触发词指导的越南语新闻事件检测方法，表达相同想法但语言不同的句子通常有相同或相似的语义成分，使用丰富的中文标记语料可以解决越南语标记语料缺失的问题。通过对比实验，发现本文的效果最佳，证明了基于中文触发词指导的方法在越南语新闻事件检测任务上达到了很好的效果。在未来的工作中，将考虑一个句子有多个事件的情况，利用句子中的实体信息表示提升越南语新闻事件检测任务的性能。

## 参考文献


[1] JI H, GRISHMAN R. Refining event extraction through cross-document inference[C]//Proceedings of ACL-08: HLT. Columbus, Ohio: ACL, 2008: 254-262.

[2] LI Q, JI H, HUANG L. Joint event extraction via structured prediction with global features [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: ACL, 2013:73-82.

[3] TONG M, XU B, WANG S, et al. Improving event detection via open-domain trigger knowledge[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 5887-5897.

[4] CHEN Z, JI H. Can one language bootstrap the other: A case study on event extraction[C]//Proceedings of the NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing, Boulder, Colorado: ACL, 2009: 66-74.

[5] ZHU Z, LI S H, Zhou G D. Bilingual event extraction: A case study on trigger type determination[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: ACL, 2014: 842-847

[6] HSI A, YANG M, Carbonell J. Leveraging multilingual training for limited resource event extraction [C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING Organizing Committee, 2016: 1201-1210.

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

[8] LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: ACL, 2014: 402-412.

[9] CHEN Y B, XU L H, LIU K. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China: ACL, 2015: 167-176.

[10] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: NAACL, 2016: 300-309.

[11] NGUYEN T H, GRISHMAN R. Modeling skip-grams for event detection with convolutional neural networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, Texas: ACL, 2016: 886-891.

[12] CHEN Y B, LIU S L, ZHANG X, et al. Automatically labeled data generation for large scale event extraction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017: 409-419.

[13] LIU H L, LIU K, HE H Z. A probabilistic soft logic based approach to exploiting latent and global information in event classification[C]//Proceedings of the 30th AAAI Conference on Artificial Intelgence. Phoenix, Arizna, USA, 2016: 2993-2999.

[14] LIU S L, CHEN Y B, LIU K, et al. Exploiting argument information to improve event detection via supervised attention mechanisms[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017: 1789-1798.

[15] CHEN Y B, YANG H, LIU K, et al. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: EMNLP, 2018: 1267-1276.

[16] FENG X H, HUANG L F, TANG D Y, et al. A language-independent neural network for event detection[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany : ACL, 2016: 66-71.

[17] HUANG L F, JI H, MAY J. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguitics: Human Language Technologies, Volume 1. Minneapolis, Minnesota: NAACL, 2019: 735-744.

[18] LI S L, LI Y, ZHOU X P, et al. Event detection without triggers[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguitics: Human Language Technologies. Minneapolis, Minnesota: NAACL, 2019: 735-744.

[19] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed represertations of words and phraqses and their compositiontity[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 3111-2119.

寇梦珂(1995—)，硕士，主要研究领域为信息检索、自然语言处理。
E-mail: kmk0829@163.com

余正涛(1970—)，教授，博士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail: ztyu@hotmail.com

高盛祥(1977—)，通信作者，副教授，硕士生导师，主要研究领域为自然语言处理、信息检索、机器翻译。
E-mail: gaoshengxiang.yn@foxmail.com