文章编号:1006-2467(2021)S2-0086-06 DOI: 10.16183/j.cnki.jsjtu.2021.S2.014

基于语义特征抓取电网调度事件的检测技术

许 凌[1], 王兴志[1], 肖林朋[2,3]
(1. 国家电网有限公司华东分部,上海 200120; 2. 南瑞集团有限公司(国网电力科学研究院有限公司),南京 211106; 3. 北京科东电力控制系统有限责任公司,北京 100192)

摘 要: 电网运行管理文档中的领域用语专业且复杂,在信息化提取的过程中需要优秀且适用的事件检测方法抽取事件主体.而当前的中文事件检测方法大多采用词嵌入技术以捕获语义表示,但这些方法难以捕捉触发词与同一句子中的其他领域词汇间的依赖关系.基于上述现状,提出了一种混合表示的新架构,以表征两个字符和单词的语义和结构信息.该模型可以通过依赖性解析器生成的语义表示来捕获富语义特征.实验基于调度日志、调度检修票、调度预案语料使用不同模型进行评估,结果表明,该方法可以显著提高电网调度文本事件检测的性能.

关键词: 电网调度;事件检测;神经网络;语义特征
中图分类号: TP 311 **文献标志码**: A

Power Grid Dispatching Event Detection Technology Based on Semantic Feature Capture

XU Ling[1], *WANG Xingzhi*[1], *XIAO Linpeng*[2,3]
(1. East Branch of State Grid Corporation of China, Shangha 200120, China; 2. Nari Group Corporation (State Grid Electric Power Research Institute), Nanjing 211106, China; 3. Beijing Kedong Power Control System Co., Ltd., Beijing 100192, China)

Abstract: The domain terms in power grid operation management documents are professional and complex. In the process of information extraction, excellent and applicable event detection methods are needed to extract event subjects. However, most of the current Chinese event detection methods use the word embedding technology to capture semantic representation, but it is difficult for these methods to capture the dependency between trigger words and other domain words in the same sentence. Based on the above situation, this paper proposes a novel hybrid representation architecture to represent the semantic and structural information of two characters and words. The model can capture rich semantic features through the semantic representation generated by the dependency parser. The experiment is based on the corpus of dispatching log, dispatching maintenance ticket, and dispatching plan. The results show that this method can significantly improve the performance of power grid dispatching text event detection.

Key words: power grid dispatching; event detection; neural network; semantic feature

收稿日期:2021-09-01
基金项目:国家电网有限公司华东分部科技项目(SGHD0000DKJS2100225)
作者简介:许 凌(1981-),男,山东省平度市人,高级工程师,从事大电网调度运行及自动控制研究.
通信作者:肖林朋,男,中级工程师,电话(Tel.):15120045525; E-mail:1553990434@qq.com.

随着“双碳”、新型电力系统建设要求的逐步深入和研究，电网调度各类信息化、智能化系统的建设也突飞猛进. 但在其建设过程中，电网调度的各类留痕文档的体量和类型也逐步增多，需要更多的自动化、智能化手段帮助电网调度管理生产工作中抽取事件的主体，从而提升工作人员的工作效能. 其中，本文提到的事件是指在电网调度文档中提及的对电网运行具有一定影响的事情，事件检测的目标是通过机器学习的手段推断出文档中发生了何种事件. 在目前人工智能的发展历程中，事件抽取是自然语言处理的重要任务之一，而事件检测是事件抽取的关键步骤之一. 利用深度学习识别事件触发词，并进一步实现事件分类，旨在以纯文本中的特定类型识别特定类型的事件实例.

具体而言，给定句子后需要模型判断句子是否包含事件触发器，如果是，则需要识别特定事件类型. 例如，在句子“2021 年 10 月 30 日 08：30A 省 B 站 #3 主变由于二次回路误动作引发重瓦斯出口告警”中，事件检测系统将检测“二次回路误动作”触发的“重瓦斯出口告警”事件.

中文事件抽取最近取得了进展. 迄今为止，已经提出了许多方法[1-2]，并获得了较好的性能. 其中，传统媒体中基于文档的事件监测方法主要通过文本的相似性和聚类来检测事件，Yang 等[3-4]提出了基于文档方法事件检测的基本步骤，包括文本预处理、数据表示、数据组织或聚类，这些步骤至今也是许多事件检测方法基本组成部分. Salton[5]使用词频-逆文本频率(TF-IDF)根据语料库对文档中重要的词语进行加权，该方法也广泛地被后来的事件检测方法使用. 为了改善术语向量和词袋模型的不足，Kumaran 等[6]提出了一种融合了命名实体的向量空间模型(VSM)文本向量模型，用来加权重要的词特征，弥补 TF-IDF 的不足. Sankaranarayanan 等[7]提出了一种在 Twitter 中检测突发新闻的系统 TwitterStand，使用朴素贝叶斯分类器来过滤不相关推文，采用在线聚类方法，根据 TF-IDF 加权的推文词语向量的文本余弦相似度进行聚类，得到新闻类，检测突发新闻. Phuvipadawat 等[8]提出了一种在 Twitter 中收集、聚类、排名和追踪突发新闻的方法，通过“#breakingnew”等关键字获取推文，让 TF-IDF 结合斯坦福命名实体识别，增加命名实体、话题标签、用户名的权重用以计算推文的相似度，按相似度将推文分组，通过粉丝数、转推数和时间新鲜度对组打分排名.

但是，现有的事件抽取方法从纯文本中捕获足够的语义信息，发现它具有挑战性，因为单词可能在不同的句子中具有不同的含义. 例如，在表 1 的句子 1 中，单词“站”等同于电站类型，但在句子 2 中，相同的单词“站”表示身体姿态. 此外，因为触发器与单词完全匹配，单词触发器不匹配问题仍然存在. 在句子 1 中，该事件应由“装机容量”触发，这是一个交叉字触发器. 而通常因为词分割工具将“装机容量”划分为“装机”和“容量”，使得无法提取完整的触发器.

表 1　纯文本信息与触发词

Tab. 1　Plain text information and trigger words

编号	句子	触发词
1	A 省火电站总体装机容量为 x kW	电站类型
2	供电员工站立时的行为举止规范要求	身体姿态

为了提高事件检测质量，需要捕获诸如语法特征的其他信息；依赖性解析器是捕获语法特征的有效方法. 使用依赖性解析器，可以用包含依赖关系的结构化信息标记句子. 弧线箭头表示依赖关系，该关系将依赖词连接到句子中的字头. 例如，在图 1 中，弧线箭头表示将依赖词“引发”连接到实体的依赖关系(主谓和动宾). 依赖关系为模型提供了丰富的语义功能，并且合理地执行. 例如，在图 1 中，“引发”是触发字. 根据“引发”及其实体的动作和信号之间的依赖关系，可以利用其实体来协助触发器的分类和识别事件类型的“引发-warning”. 这些相关实体称为提示词，可以提供可用信息以帮助触发分类. 然而，通过传统的单词嵌入，很难充分利用这些提示词，因为它们分散在句子中. 因此，我们通过添加依赖关系来连接提示词来触发评论. 在大多数情况下，这些附加功能可以提供有用的信息.

通过此目标，本文提出了一种混合表示方法，用

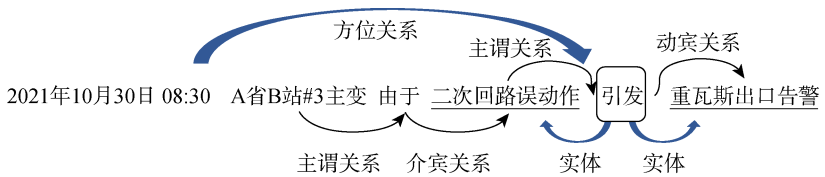

图 1　依赖关系

Fig. 1　Relationship of dependency

于从单词、字符和依赖关系中学习信息. 通过使用 Token-Level 神经网络学习两个单独的字符级和字级表示;并从依赖性解析器获取依赖性信息,通过独热编码生成其表示. 通过分析证明,从依赖性解析器生成的特征有利于事件检测. 最后,本文设计适当的混合范式以捕获混合表示,其模型分别实现了 81.43%和 73.85%的平均 $F1$ 值,可用于触发识别和触发分类.

1 方法

本文提出的模型通过两个阶段构成,并通过动态多池卷积神经网络(DMCNN)处理混合表示:第一阶段为触发识别,通过使用 Token-Level 神经网络利用触发器的字符组成结构以捕获包含有关角色的潜在触发语义;第二阶段用于确定事件的特定类型,其称为触发类型分类. 我们使用来自依赖性解析器提取依赖性信息,以在这两个阶段中通过 Token-Level 神经网络生成特征表示;最终将其与单词特征表示组合并获取混合表示.

电网调度语义事件检测方法的体系结构由 4 个部分组成,包括输入序列的表示、基于依赖解析器的特征表示、混合表示法和 DMCNN,其框架如图 2 所示.

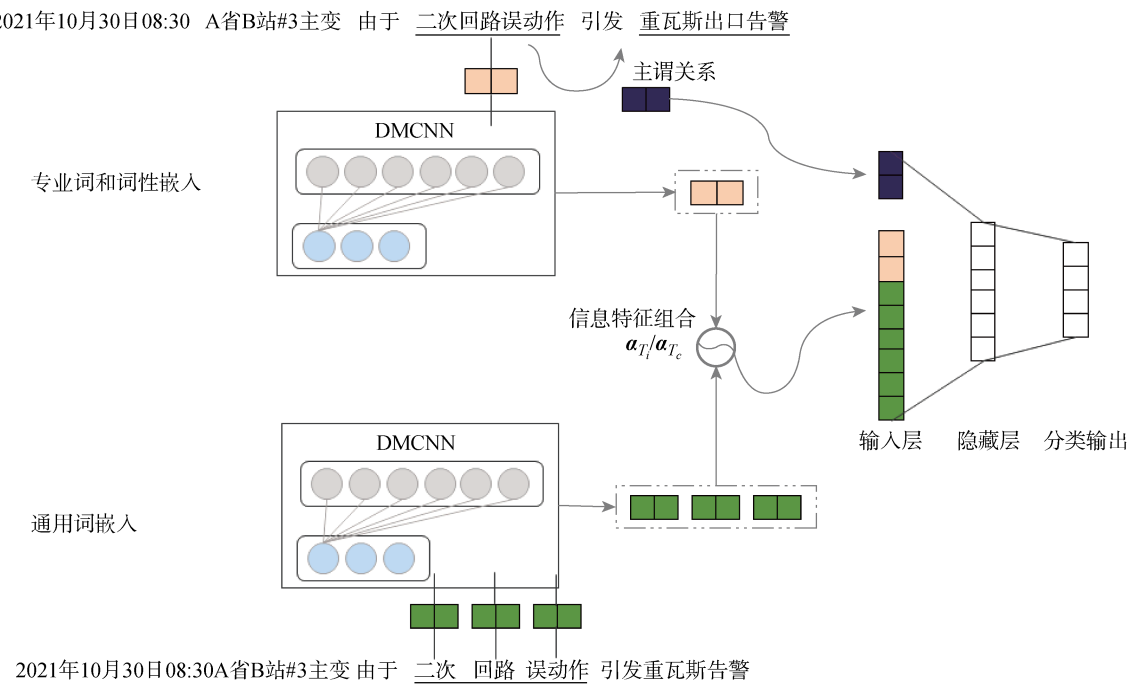

图 2 调度文本的语义事件检测框架

Fig. 2 Framework of scheduling text's semantic event detection

1.1 输入序列的表征

为了更好地捕获有关不同级别的信息,本文使用两级嵌入,即领域词汇嵌入和普通词汇级别嵌入. 为了进一步提高性能,本文使用预先训练的权重来初始化嵌入. 并根据两个 Token-Level 神经网络用于从领域词汇和普通词汇获取特征. 网络架构类似于 Nugget 区块提案网络(NPNs). 令 $T=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n\}$,其中 $\boldsymbol{t}_i(i=1,2,\cdots,n)$ 是句子中的 token 数,n 为句子能被分词产生的词汇数量. 并令 $\boldsymbol{x}_i$ 是 $\boldsymbol{t}_i$ 的嵌入(领域词汇或普通单词),本文引入窗口大小为 s 的卷积层,$\boldsymbol{t}_c$ 用来捕获文本输入的语义,c 表示卷积层,该卷积层表征句子的语义,其隐含层标识如下:

$$\boldsymbol{h}_{ij}=\tanh(w_i\boldsymbol{x}_{j:j+s-1}+b_i) \tag{1}$$

$$\boldsymbol{h}_i^{\text{left}}=\max_{j<c}\boldsymbol{h}_{ij},\quad \boldsymbol{h}_i^{\text{right}}=\max_{j\geqslant c}\boldsymbol{h}_{ij} \tag{2}$$

式(1)表示卷积过程,其中:w_i 表示第 i 层卷积层的滤波器;j 表示 token 中的索引;$\boldsymbol{x}_{j:j+s-1}$ 表示嵌入层 $\boldsymbol{x}_j$ 到 $\boldsymbol{x}_{j+s-1}$ 的串联;b_i 表示偏置.

式(2)通过使用动态多池来提供句子不同部分的重要特征,其中:$\boldsymbol{h}_i^{\text{left}}$ 关联卷积层 $\boldsymbol{t}_c$ 左侧的汇集结果;$\boldsymbol{h}_i^{\text{right}}$ 关联卷积层 $\boldsymbol{t}_c$ 右侧的汇集结果;通过连接 $\boldsymbol{h}_i^{\text{left}}$ 和 $\boldsymbol{h}_i^{\text{right}}$ 获取卷积层 $\boldsymbol{t}_c$ 的领域词汇表示 $\boldsymbol{f}_{\text{bword}}$;通过在普通词汇级序列上使用相同的过程,还可以获得普通词汇级表示 $\boldsymbol{f}_{\text{nword}}$.

1.2 基于依赖分析的特征表征

依赖性解析器是基于依赖关系的依存句法分析的重要部分. 句法依赖关系可用于获得深度语义信

息. 通过直接将它们结合到嵌入中，在神经网络模型中使用句法依赖项. 在这项工作中，使用三个不同的特征抽象层来表示三个功能：POS 表示词性标注；DR 表示依赖关系分析；DIS 表示与头的距离. POS 是单词的词性，它在自然语言中发挥着重要作用，例如命名实体的识别、语法分析和事件抽取. 名词或代名词可以作为句子中的主题，但不能相互转化. 这是因为语法组件对语音组件有限制. 因此，POS 适用于抽象特征，以表达文本语义信息的不同特征. 本文将 POS 作为一种特征来加强基于单词的特性. 中文中的常用词性大概有 52 个，本文将使用 52 维的独热向量来代表句子 POS 的部分. 这意味着句子中的每个单词的 POS 可以表示为 52 维特征向量. 52 维矢量的每个维度代表了 POS 的一部分. 例如，表示一个单词的位置，如果其中一个值为 1，并且剩余的 51 维值为 0.

DR 依赖关系表达了句子组件之间的语义关系. 对于事件检测任务，触发词通常是谓语（即动词）. 一般在电网调度语料库中，触发词为动词对象的角色占 19%. 因此，本文认为可以使用依赖关系来改善触发器检测. 在依赖关系的特征层上，矢量维度为 23（22 个关系类型和 1 个“其他”类型）. 22 种依赖关系经常使用句法依赖项，并且为了降低特征表示的复杂性，将其他依赖关系分类为“其他”类型.

DIS 是依赖路径的长度. 具体地说，如果一个单词和头部直接相关，则定义 DIS=1. 如果路径包括一个中间依赖性，则 DIS=2. 例如，在图 1 的句子中，二次回路误动作引发重瓦斯告警的依赖路径如下：

二次回路误动作→引发→重瓦斯告警

其中“二次回路误动作”是这句话的头，“引发”在头部和“重瓦斯告警”之间创造了中间依赖. 因此，对于“重瓦斯告警”这个领域词汇，其 DIS=2. 头部是句法依赖项中的根节点，通常表示句子中的核心字，用来触发代表事件中的重要内容. 基于对电力调度语料库中触发器的分析，本文发现头和触发词是相似的. 这意味着距离头部的距离可用于测量单词是否在某种意义中被触发. 在 DIS 的特征层上，使用 7 维向量表示 0 到 6 的距离. 另外，如果其大于 6，仍然将其值识别为 6.

1.3　混合表征学习

对于中文事件检测，只有使用领域词汇表示或普通词汇级表示无法获得足够的信息. 例如，如果理解字符级表示中的“引发”，是一个由“引”和“发”组成的触发器. 在单词级表示中，单词级序列既可以提供更明确的信息，又将“引”的语义区分开.

在嵌入层之后，可以获得领域级特征表示 $\boldsymbol{f}_{\text{bword}}$，一个普通级特征表示 $\boldsymbol{f}_{\text{nword}}$ 和一个总体特征表示 $\boldsymbol{f}'_{\text{F}}$. 本文通过学习两个门来模拟触发识别和事件类型分类器的信息流，公式如下：

$$\boldsymbol{\alpha}_{T_i} = \text{SIGM}(\boldsymbol{W}T_i\boldsymbol{f}_{\text{bword}} + \boldsymbol{U}T_i\boldsymbol{f}_{\text{nword}} + \boldsymbol{V}T_i\boldsymbol{f}'_{\text{F}} + \boldsymbol{b}T_i) \quad (3)$$

$$\boldsymbol{\alpha}_{T_c} = \text{SIGM}(\boldsymbol{W}T_c\boldsymbol{f}_{\text{bword}} + \boldsymbol{U}T_c\boldsymbol{f}_{\text{nword}} + \boldsymbol{V}T_c\boldsymbol{f}'_{\text{F}} + \boldsymbol{b}T_c) \quad (4)$$

式中：SIGM 为 sigmoid 函数；$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{V}$ 均为权重矩阵；$\boldsymbol{b}$ 表示偏置.

基于图 2 中的三个特征图层，可获得三个特征表示. 通过连接这三个特征表示来构建一个 82 维矢量 $\boldsymbol{f}'_{\text{F}}$. 最后，可获得特征和单词作为新的表示的串联：$\boldsymbol{f}'_{\text{h}} = [\boldsymbol{f}'_{\text{F}} \quad \boldsymbol{f}_{\text{bword}}]$.

根据引入触发扣核生成器和事件类型分类器的栅格，可以获得作为输入的最终矢量：

$$\boldsymbol{f}_{T_i} = \boldsymbol{\alpha}_{T_i}T_i\boldsymbol{f}_{\text{bword}} + (1-\boldsymbol{\alpha}_{T_i})\,\boldsymbol{f}'_{\text{h}} \quad (5)$$

$$\boldsymbol{f}_{T_c} = \boldsymbol{\alpha}_{T_c}T_i\boldsymbol{f}_{\text{nword}} + (1-\boldsymbol{\alpha}_{T_c})\boldsymbol{f}'_{\text{h}} \quad (6)$$

式中：$\boldsymbol{f}_{T_i}$ 是触发识别的混合特征；$\boldsymbol{f}_{T_c}$ 是事件类型分类器的混合功能；$\boldsymbol{\alpha}_{T_i}$ 和 $(1-\boldsymbol{\alpha}_{T_i})$ 分别代表触发识别中 $\boldsymbol{f}_{\text{nword}}$ 和 $\boldsymbol{f}'_{\text{h}}$ 的重要性；$\boldsymbol{\alpha}_{T_c}$ 在事件类型分类器中扮演类似的作用.

1.4　动态多池化卷积神经网络

传统的卷积神经网络仅使用一个池层实现最大操作. 这意味着传统的卷积神经网络仅捕获句子的表示中最重要的信息. 在事件检测中，一个句子可能包含两个或多个事件，并且参数可以使用不同的触发来获取不同的角色. 但是，传统的卷积神经网络只会捕获整个句子的最有用信息，并将同一个句子中的其他信息丢失. 为了解决上述问题，通过 DMCNN 在不丢失最大池值的情况下获得更有价值的信息. 本文将使用类似的神经网络，图 3 描述了其触发检测的架构.

首先，获得嵌入和相对位置 $\boldsymbol{x}_i$ 的串联，将 $\boldsymbol{x}_i$ 串联到上文中定义的特征向量中；随后将词汇级别特征作为卷积层的输入，以捕获组成语义并获得特征图. 具体地，卷积操作通过利用卷积核来扫描 W 个单词的窗口来产生新的特征，此处 W 表示进行卷积的词汇大小，一般为 3 或 5. 一个卷积核产生一个位置的一个特征 $\boldsymbol{c}_{ij} = \sigma(w_j\boldsymbol{x}_{i:i+j} + b)$，其中：$\sigma$ 是非线性函数（一般情况使用 tanh）；j 的范围为 1 到 m；m 是卷积核的数量，i 为词汇的数量.

之后，根据句子中的卷积核的数量，特征映射输出 $\boldsymbol{c}_j$ 分为 i 部分. 例如，如果一个句子有一个卷积

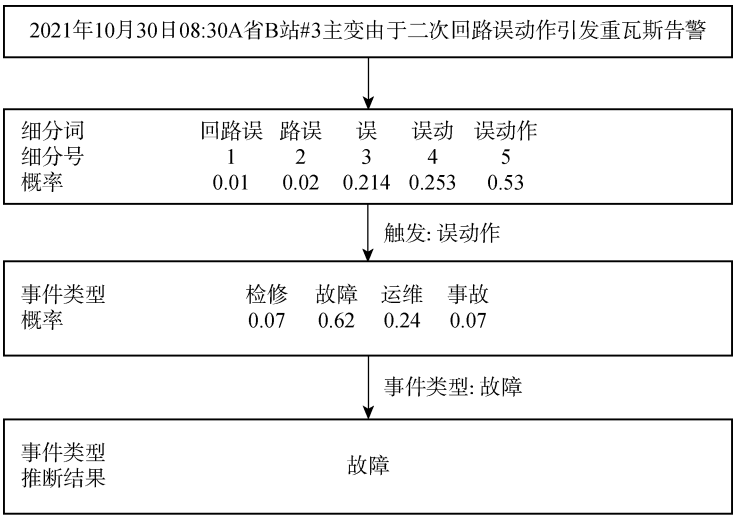

图 3 触发检测框架

Fig. 3 Framework of trigger detection

核,那么句子将被分成两个部分,当这句话有两个卷积核时,这两个触发器将句子分为三个部分. 通过动态多池,一个过滤器的最终输出由 $\boldsymbol{p}_{ji}=\max(\boldsymbol{c}_{ji})$ 给出,以获得每个特征映射的 $\boldsymbol{p}_{ji}$,并且所有 $\boldsymbol{p}_{ji}$ 连接到形成的最终结果.

最后将上面的特征向量和词汇特征连接到一个矢量 $\boldsymbol{f}_{\text{word}}$ 中. 采用类似的方法来获取字符级向量 $\boldsymbol{f}_{\text{char}}$,为触发识别和触发类型分类层产生两个混合表示.

1.5 训练和分类

在训练过程中,本文将事件检测视为多级分类问题. 将 $\boldsymbol{f}_{\text{N}}$ 作为卷积神经网络的输入. 输入是分词后的句子. 并加入 Dropout 防止过度拟合.

另外,卷积层用于捕获语义并卷积核生成特征映射. 传统的卷积神经网络通常通过 PoolMax 层获得一个最大值. 本文使用动态多池通过将每个特征映射分成多个部分来获得多个最大值. 通过连接句子级和词汇特征,得到一个矢量 $\boldsymbol{f}_{\text{F}}$,并将其馈入分类器. 获取最终输出的过程可以表示如下:

$$\boldsymbol{Z}=\boldsymbol{W}_{\text{T}}\boldsymbol{f}_{\text{F}}+\boldsymbol{b} \tag{7}$$

式中:$\boldsymbol{W}_{\text{T}}$ 是转换矩阵;$\boldsymbol{b}$ 是偏置项;$\boldsymbol{Z}$ 是网络的最终输出.

最后,使用 SoftMax 函数的激活功能来预测每个类型的概率分数,选择最高概率作为最终结果. 使用两个不同的分类器来预测触发器检测中的跨度类型和触发类型分类中的事件类型. 在图 3 中,举例说明分类器如何获取触发单词并预测事件类型. 当分类器采用潜在触发或潜在触发器时,一个分类器包含此字符的不同可能为依据,触发抽取取最佳匹配单词. 然后,另一个分类器将预测此触发的事件类型.

2 实验准备数据集和评价指标

本文使用电网调度中的调度日志、调度检修票、调度预案语料来评估研究的网络. 在评估语料库中,有 23 967 份文件. 表 2 显示了语料库中的每个事件类型的数量. 并使用了 20 967 个文档作为训练集,2 000 个文档作为开发集,其余 1 000 个文档用于测试. 本文遵循标准评估程序:如果其事件子类型和偏移符合参考触发器,则触发器是正确的. 最后本文使用精度(P),召回(R)和 $F1$ 值来评估结果.

表 2 触发识别与触发分类实验结果

Tab. 2 Experimental results of trigger recognition and trigger classification %

模型	触发识别			触发分类		
	P	R	$F1$	P	R	$F1$
FBRNN	62.23	37.8	45.98	51.87	32.78	40.82
DMCNN	56.27	23.82	46.92	67.21	32.98	45.98
CLIZH	63.21	23.45	33.72	60.76	35.21	36.87
NPN	65.87	66.78	33.87	56.98	45.89	40.30
JMCEE	67.28	67.89	36.89	49.90	35.89	40.37
本文方法	69.34	59.34	49.38	68.21	46.98	49.90

3 实验设置

对于分类器,本文划分为正例和负例. 包含在触发单词中的字符被视为一个正实例,否则该字符被视为否定实例. 将正例与负例的比率设置为 1∶20.

为了有效地训练神经网络，将 500 作为句子长度的最大限制，并且在字级表示中的最大句子长度为 130. 嵌入尺寸和 Char 嵌入维度的单词为 100. 对于卷积神经网络的激活功能，使用 sigmoid 函数. 因为标记的语料库中的大多数触发器不超过 3，为了评估字符是否被视为正实例，限制了 3 的触发长度，将批量尺寸设置为 128，Dropout 设置为 0.5，并将矩阵的初始化设置为在 −0.1 至 0.1 之间的均匀分布.

4　实验结果

使用五类模型进行实验对比.

FBRNN：使用前后循环神经网络（RNN）来检测单词或短语的事件. 该方法是使用 RNN 进行事件检测的第一次尝试.

DMCNN：使用动态多池层根据事件触发器和参数进行实验. 此事件抽取方法在不使用复杂的自然语言处理工具的情况下自动提取词法级和句子级别特征.

CLIZH：在 LSTM 编码器中加入了许多启发式功能，在 TAC KBP 2017 评估中取得了最佳性能.

NPN：该网络使用混合表示捕获特征，并设计了三种混合范例（CONCAT、General 和 Casst-Complete）以获得混合表示.

JMCEE：采用预训练 BERT 编码器，并利用多组二进制分类器来确定事件检测.

正如表 2 所示，本文方法在触发识别的任务中具有最好的精确率和 $F1$ 值. 但是召回率与最佳结果相比，本文的模型大概低了 8%. 在触发分类的任务中，本文方法可以实现全指标的最佳，证明了本文方法的有效性.

5　结语

基于语义特征抓取电网调度文本预计事件检测技术方法的四个部分，分别为：输入序列的表示、基于依赖解析器的特征表示、混合表示法和动态多轮询卷积神经网络，通过以上四个部分可以有效地将电网调度文本进行检测，并实现基于语义特征成功抓取电网调度事件的工作.

此外，可以通过实验进一步研究并利用深度学习识别事件触发词，并进一步实现事件分类，旨在以纯文本中的特定类型识别特定类型的事件实例. 更深层次地通过自动化、智能化手段帮助电网调度管理生产工作中抽取事件的主体，从而提升工作人员的工作效能.

参考文献：

[1] ZHANG T T, JI H, SIL A. Joint entity and event extraction with generative adversarial imitation learning [J]. **Data Intelligence**, 2019, 1(2): 99-120.

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

[3] YANG Y M, PIERCE T, CARBONELL J. A study of retrospective and on-line event detection [C]// **21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. Melbourne: ACM, 1998: 28-36.

[4] YANG Y M, ZHANG J, CARBONELL J, *et al*. Topic-conditioned novelty detection [C] // **Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. Edmonton: ACM, 2002: 688-693.

[5] SALTON G. Automatic text processing: The transformation analysis and retrieval of information by computer [M]. Reading: Addison-Wesley, 1989.

[6] KUMARAN G, ALLAN J. Text classification and named entities for new event detection [C]// **27th Annual International Conference on Research and Development in Information Retrieval**. Sheffield: ACM, 2004: 297-304.

[7] SANKARANARAYANAN J, SAMET H, TEITLER B E, *et al*. TwitterStand: News in tweets [C]// **17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. Seattle: ACM, 2009: 42-51.

[8] PHUVIPADAWAT S, MURATA T. Breaking news detection and tracking in twitter [C]// **2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology**. Toronto: IEEE, 2010: 120-123.

（本文编辑：黄伟）