计算机科学
COMPUTER SCIENCE

基于深度对比孪生网络的事件辨重方法

李子琛, 易修文, 陈顺, 张钧波, 李天瑞

引用本文

李子琛, 易修文, 陈顺, 张钧波, 李天瑞. 基于深度对比孪生网络的事件辨重方法[J]. 计算机科学, 2024, 51(12): 30-36.

LI Zichen, YI Xiuwen, CHEN Shun, ZHANG Junbo, LI Tianrui. Deep Contrastive Siamese Network Based Repeated Event Identification [J]. Computer Science, 2024, 51(12): 30-36.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300025

基于深度对比孪生网络的事件辨重方法

李子琛[1]　易修文[2,3]　陈　顺[1,2,3]　张钧波[1,2,3]　李天瑞[1]
1 西南交通大学计算机与人工智能学院　成都 611756
2 北京京东智能城市大数据研究院　北京 100176
3 京东城市(北京)数字科技有限公司　北京 100176
(zichen_li@126.com)

摘　要　在中国,市民可以通过拨打12345市民热线,向政府报告生活中遇到的问题并寻求帮助。然而,有许多重复的事件被多次上报,这给负责事件分派的工作人员带来了很大的压力,也会导致事件的处置效率变低,浪费社会公共资源。对重复事件的判断需要精确分析文本语义和上下文关系,为了解决这个问题,文中提出了一种基于深度对比孪生网络的事件辨重方法,通过评估两个事件的描述文本之间的相似性,辨别出具有相同诉求的事件。首先通过召回和过滤的方法来减少候选事件的数量;然后通过对比学习构造任务,微调预训练的BERT模型,学习易于辨识的事件描述语义表征;最后引入事件标题作为上下文信息,并通过带有分类器的孪生网络来识别重复事件。在南通市12345事件数据集上进行了实验,结果表明,该方法在各项评估指标上均优于基线方法,特别是在与辨重任务场景相关的F0.5分数上,能够有效地辨别重复事件,提高事件处置的效率。

关键词:12345热线;重复事件识别;对比学习;孪生网络;城市计算
中图分类号　TP399

Deep Contrastive Siamese Network Based Repeated Event Identification

LI Zichen[1],YI Xiuwen[2,3],CHEN Shun[1,2,3],ZHANG Junbo[1,2,3] and LI Tianrui[1]
1 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 JD Intelligent Cities Research,Beijing 100176,China
3 JD Intelligent Cities Technology Co.,Ltd.,Beijing 100176,China

Abstract　In China,citizens can report issues they encounter in daily life to the government and seek assistance by calling the 12345 citizen hotline. However,many events are reported multiple times,which places significant pressure on the staffs responsible for event allocation,resulting in low efficiency of event disposal and waste of public resources. Identifying repeated events requires precise analysis of textual semantics and contextual relationships. To address this problem,this paper proposes an event repetition identification method based on a deep contrastive siamese network. By evaluating the similarity between the descriptions of events,the method identifies events with the same demands. First,it reduces the number of events through retrieval and filtering. Then,it fine-tunes a pre-trained BERT model through contrastive learning to learn distinct semantic representations of event descriptions. Finally,the event title is introduced as contextual information,and a siamese network with a classifier is used to identify repeated events. Experimental results on the 12345 event dataset of Nantong demonstrate that the proposed method outperforms baseline methods across various evaluation metrics,particularly in the F0.5 score,which is relevant to the repetition task scenario. The proposed method can effectively identify repeated events and improve the efficiency of event handling.

Keywords　12345 hotline,Repeated event dispatch,Contrastive learning,Siamese network,Urban computing

1　引言

城市中的各种大小事项频发不断,例如空气污染、入学难题和停车困难等,为了及时解决这些问题,中国各地的地方政府建立了12345市民热线。当市民遇到难题时,可以通过12345市民热线将事件上报至政府,寻求帮助。一旦热线分派人员收到市民上报的事件,就会分析事件所对应的诉求,并将其分配给职责匹配的相关部门。2021年1月6日,国务院办公厅印发了《关于进一步优化地方政务服务便民热线的指导意见》。该指导意见旨在通过优化流程和资源配置,实现热线受理与后台办理服务紧密衔接,确保企业和群众反映的问题和合理诉求及时得到处置和办理。这一措施意在使政务服务

到稿日期:2024-03-04　返修日期:2024-07-17
基金项目:国家重点研发计划(2023YFC2308703);北京市科技新星(Z211100002121119)
This work was supported by the National Key R&D Program of China(2023YFC2308703) and Beijing Nova Program(Z211100002121119).
通信作者:易修文(xiuwenyi@foxmail.com)

便民热线接得更快、分得更准、办得更实，从而进一步畅通政府与群众的互动渠道，提高政务服务水平，推进国家治理体系和治理能力现代化，不断增强人民群众的幸福感。

随着城市化的发展，公共事务的服务质量对城市的发展和社会稳定具有重要意义。因此，政府部门需要尽快解决市民反馈的问题，并确保信息的准确性和完整性。然而，由于缺乏有效的重复事件识别机制，同一事件经常会被多人或同一人多次上报，导致了重复事件的出现。重复上报的事件会导致重复指派，浪费社会公共资源，延长事件的处理时间，降低老百姓对政府部门的信任度。为解决这个问题，需要在事件指派之前对反馈的问题进行去重处理，识别和去除重复事件。目前，重复事件的辨别主要依靠人工，这种方法存在速度慢、成本高、容易漏判或误判等问题，而且重复事件的数量非常庞大，加重了业务人员的工作负担，因此需要寻求一种更加高效的解决方案。在这种情况下，构建一种由数据驱动的重复事件识别方法是非常必要的。

在辨别重复事件的过程中，最主要的问题在于判断两个描述文本是否是同一事件。描述文本中包含了事件背景、参与者、发生时间及地点等多个要素。正确的相似性判断不仅需要精确分析文本中的这些要素，还需深入理解文本隐含的上下文关系。这个问题可以被归类为文本语义相似度(STS)任务。近年来，通过预训练模型解决文本语义相似度任务已经成为主要方法之一。然而，对于具有高度相似文本格式的文本语义相似度任务来说，通过预训练模型产生的表征是难以区分的。除此之外，辨别重复事件还存在两个主要的挑战：1)事件描述涵盖多个维度，包括但不限于时间和空间，因此不能将其简单视作纯文本。此外，事件描述中含有大量专有名词、方言以及俚语，这些语言特征使得已有的预训练模型难以直接用于处理事件描述；2)在上报的事件数据中，重复事件仅占少数，且在这些重复事件中，只有极少数被热线服务人员进行了标记，标记样本的稀缺性限制了传统有监督学习方法在此类问题上的效果。

深度学习作为目前主流的人工智能方法之一，也在文本语义相似度任务中得到了广泛的应用。深度学习具有强大的拟合和泛化能力，可以有效捕捉文本序列中的语义信息，为许多自然语言处理任务提供了支持，例如机器翻译、文本摘要和问答系统等。此外，随着计算机硬件能力的提升，近年来还出现了一些基于预训练的深度学习模型，如 BERT[1]，ELMo[2]，XLNET[3]等，这些模型通过在大规模的文本数据集中预训练神经网络，能够捕捉更丰富的上下文信息，并能适应更加复杂的任务场景，在文本语义相似度任务中表现优异。

为了解决上述问题，本文提出了一种基于深度对比孪生网络的重复事件识别方法。该方法有两个主要贡献：1)通过对政务事件语料进行对比学习，并微调预训练 BERT 模型的参数，解决了 BERT 生成的语义向量相近的问题，从而生成更易于区分的事件描述表征，以符合事件辨重任务的需求；2)将事件标题作为上下文信息引入模型，通过交叉注意力机制融合事件描述表征和事件标题表征，以加强事件描述中的核心语义，并利用带有分类器的孪生网络来识别事件是否重复。本文在南通市 12345 热线事件数据集上进行了验证，实验结果表明了本文方法的有效性，能够有效辨别重复的政务事件。

2 相关工作

2.1 12345 热线

目前，政务服务 12345 热线事件处理主要依赖于人工处理。近年来，一些学者试图通过信息化、智能化的方式来提高管理城市事件的效率[4-5]。作为人工智能重要分支之一的深度神经网络[6]也被用于城市事件分类的研究。还有学者在对事件进行分类之前使用了短文本聚类来提高准确率[7]。Peng 等[8]则尝试构建了城市事件管理系统。Luo 等[9]在海口市 12345 热线投诉文本数据集上对一些自然语言处理中的深度学习方法如 FastText，TextCNN，TextRNN 等进行了比较。

2.2 文本语义相似度

文本语义相似度(STS)指在给定两个文本片段的情况下，评价它们之间语义等价性的度量[10]。传统方法通常基于人工特征和浅层模型，如向量空间模型、word2vec[11]、Glove[12]等，但这些方法在处理大规模文本数据时的性能有限，准确率不高。深度学习技术的出现为解决这一问题提供了新的途径。常用模型包括基于卷积神经网络[13-14](CNN)、循环神经网络(RNN)[15]和变换器[16](Transformer)等。这些模型能够从大规模的数据中学习到丰富的特征，从而提高模型的准确率和泛化能力。在实际应用中，文本语义相似度任务还需要考虑多个方面的因素，如上下文信息、命名实体识别、词义消歧等。因此，一些特定领域的语料库和知识库也被广泛应用于文本语义相似度任务中，以提高模型的性能。Katherine 等[17]提出了一种结合 SimHash 和文档向量相似度完成对文本数据集去重的方法；Bikash 等[18]提出了一种结合局部敏感哈希和词嵌入，对多学科学术文献进行去重的方法。在处理热线事件的辨重任务时，由于事件描述文本中通常包含较多的专有名词，因此有必要设计一种能够充分考虑场景特性的事件辨重方法，以更准确地识别和区分具有相似性但不完全相同的事件。

2.3 对比学习

对比学习旨在通过学习两个或多个样本之间的相似度或差异性来进行分类、聚类或检索等任务。相对于传统的监督学习方法，对比学习可以利用较少的标注数据进行训练，并且在处理高维度数据时具有更好的性能，尤其是在无监督或半监督学习任务中表现出更好的效果。对比学习的应用领域非常广泛[19]，可以用于解决许多复杂问题，例如人脸识别、物体检测和图像检索[20]、文本分类、情感分析和机器翻译等任务[21]。

3 问题定义

事件：一个事件可被定义为 $e=\{t,l,h,d\}$，其中 t 为上报时间，l 为事件发生地点，h 为事件标题，d 为事件描述。表 1 列出了热线事件的示例。

事件辨重：给定一个新上报的事件 e_{new} 和候选事件集合 $C=\{e_1,e_2,\cdots,e_n\}$，目标是辨别新事件 e_{new} 与候选事件集合中的某一事件 e_i 是否匹配。

表 1 12345 政务热线事件示例

Table 1 Example of 12345 hotline events

时间	地点	标题	事件描述
2019 年 10 月 19 日 11:17:23	湾子头新寓	关于住户承重墙装修的问题	2021 年 1 月 18 日，湾子头新寓 7 幢 xxx 室装修时将承重墙敲了个门洞，存在非常大的安全隐患。
2021 年 1 月 18 日 23:14:35	小海街道海棠花园	关于工地施工噪音扰民的问题	2019 年 10 月 19 日，小海街道海棠花园二期 29 幢南边有工地正在施工，机器轰鸣，噪音扰民。

4 整体框架

图 1 给出了本文方法的框架，由 3 部分组成。1）召回和过滤。使用空间、时间和语义相似度等条件对候选事件进行初步筛选，大幅度减少候选事件的数量，提高后续模型的训练和辨识速度。2）对比表征学习。通过数据增强构造三元数据集，并使用对比学习的方式微调预训练的 BERT 模型，使其学习政务事件相关的领域知识，并能够生成易于区分的事件描述表征。3）上下文孪生网络。为提高重复事件辨别的准确性，我们引入了事件标题作为上下文信息，并通过交叉注意力机制融合事件描述与标题的表征向量，以增强事件描述的核心语义，再通过带有分类器的孪生网络结构识别重复事件。上述步骤可以更好地刻画事件的语义特征，进而更准确地发现事件之间的异同点，高效地完成事件辨重任务。

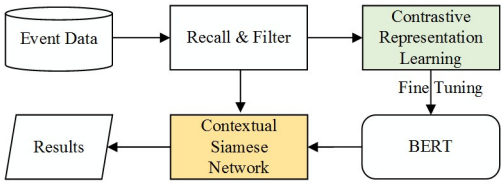


图 1 事件辨重方法方法的整体框架

Fig. 1 Framework of repeated event indentification approach

5 方法

5.1 事件召回与过滤

据统计，12345 热线每天接收到的事件常常数以万计，如果直接使用新上报的事件与所有历史事件构建事件对，这上万的事件对将会使得识别的时间成本十分高昂，很难在线部署。为解决该问题，本文通过以下两个步骤来减少候选事件的数量：事件召回和过滤。该方法的流程如图 2 所示。

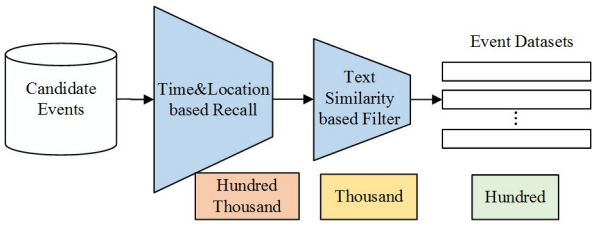


图 2 召回和过滤方法

Fig. 2 Recall and filter methods

在召回步骤中，选择具有相同街道地址且上报时间间隔在 7 天内的事件，对候选事件进行初步召回，过滤掉大量的候选事件。之后，在过滤步骤中，使用 TF-IDF 的方法得到事件描述的词向量集合，并通过计算余弦距离的方法，过滤掉事件描述相似度低于 50% 的事件。经过召回步骤对候选事件进行初步筛选后，事件的数量级能够从数十万降低到以千为单位；而通过过滤步骤进一步筛选后，候选事件的数量级可以从千减少到百，这极大地了加快了后续模型辨别重复事件时的处理速度以及训练速度。

5.2 对比表征学习

近年来，由于 BERT 模型具有出色的性能和良好的泛化能力，因此基于预训练模型 BERT 的方法被广泛用作文本语义编码器，用于将事件描述转化为句子表征。然而，现有的预训练模型不能直接用在政务事件的辨重上。这不仅是因为政务事件描述中含有许多专有名词、俚语和方言，也由于 BERT 会将所有的句子都映射到一个较小的语义空间，这使得大多数句子在语义空间中都很接近，难以区分。此外，在上报的事件中只有小部分是重复的，且只有更加少量的重复事件被业务人员标记出来，因此传统的有监督学习方法很难训练出一个有效的重复事件辨别模型。

因此，本文提出一种基于对比表征学习的方法解决了上述问题，可以较好地抽取事件描述的语义特征，具体过程如图 3 所示。我们在真实事件数据集上进行数据增强，得到涵盖样本、正例、负例的三元数据集，并使用该数据集进行对比学习，微调预训练的 BERT 模型。经微调后的 BERT 能够产生区分度大的事件描述表征，使得该表征能更加适应辨重任务的需要。该方法分为以下两个部分。

1）构建三元数据集。该部分的目的是生成用于后续对比学习的三元数据集。通过构造三元数据集，可以避免设置不同的阈值对来判断两个事件之间是否重复所造成的影响。首先，对事件描述进行文本挖掘，以获得事件描述的词向量序列，步骤如下：分词、去除停用词以及词嵌入。之后，使用 Paragram Phrase Embedding[22] 方法，将一个句子中的所有词向量进行平均计算，生成一个事件描述的句表征。在生成三元数据集的过程中，选择与锚描述最为相似的事件描述作为三元组中的正样本，同时选择相似度排名第四和第五的句子作为负样本。该策略旨在放大与锚描述具有部分相似性的样本在表征空间中的距离，从而提升模型的区分能力。通过这种方法，本文构建了一套专门用于后续模型微调阶段的数据集。生成句表征及余弦距离的计算式如下：

$$g_{\text{phrase}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_{x_i} \tag{1}$$

$$Sim(\boldsymbol{A},\boldsymbol{B})=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{2}$$

式(1)中，$\boldsymbol{W}$ 为事件描述中的词向量，n 为词的数量；式(2)中，$\boldsymbol{A},\boldsymbol{B}$ 为两个向量，n 为向量的维度。

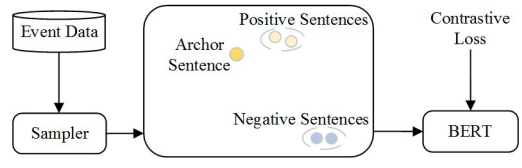


图 3 对比表征学习

Fig. 3 Contrastive representation learning

2）对比学习。以 BERT 为基础的预训练模型在许多任务中都有着很好的表现。然而，直接使用 BERT 的输出作为句子的特征向量，在文本语义相似度任务上的效果却不是很好。由于 BERT 的词表示空间具有各向异性，因此句子对之间相似分数普遍偏高，这被称为表征坍缩[23]。对比学习的目标是学习一种编码器，这个编码器可以使相似的句子在编码后的表征空间中更加相近，不同的句子则相距尽可能远。因此，可以用对比学习的方法来解决 BERT 的坍缩问题，帮助模型学习到更好的文本表示，从而提高下游任务的性能。

本文使用上文中构造的三元数据集，通过对比学习微调加载的预训练 BERT。对于一个锚定事件 s，使用相似的事件描述对(s,s^+)作为正样本对，使用不同的事件描述对(s,s^-)作为负样本对。对比学习的损失函数如下：

$$\mathscr{L}=-\log\frac{e^{sim(s_i,s_i^+)/\tau}}{\sum_{j=1}^{N}(e^{sim(s_i,s_j^+)/\tau}+e^{sim(s_i\cdot s_j^-)/\tau})} \tag{3}$$

其中，s_i 表示三元组中的锚定事件描述，sim 表示余弦相似度，N 表示训练时每个 batch 的大小，τ 表示温度参数。

该训练目标函数鼓励相似的事件描述在嵌入空间中离得更近，不同的事件描述在嵌入空间中离得更远。经过对比学习的微调进而引入政务事件相关的领域知识，也能使其更多地关注抽象语义信息，以生成更具差异化的事件表征。

5.3 上下文孪生网络

事件描述的文本中包含许多与事件核心语义无关的噪声信息，这不利于判断事件之间的语义相似度，而事件的标题往往是一个事件的高度概括。因此，我们引入事件标题作为上下文信息，对事件语义进行增强，并借助孪生网络识别重复事件。图 4 给出了上下文孪生网络的框架。

为了生成融合表征，将新事件和候选事件的描述和标题分别输入参数共享的 BERT 编码器与上下文编码器，然后使用注意力机制融合事件描述与标题表征，最后通过一个分类器来辨别重复事件。方法的具体描述如下。

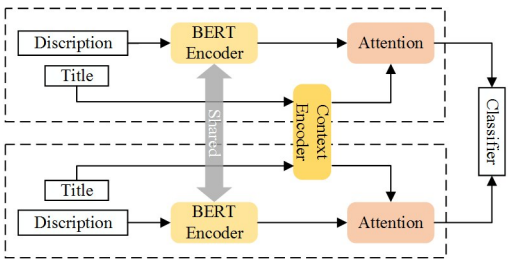


图 4 上下文孪生网络的结构

Fig. 4 Architecture of contextual siamese network

1）上下文特征融合。由于事件标题可以被视为一个事件的总结，因此使用与 3.2.1 节中相同的方法得到标题的句表征。之后，将标题的句表征送入具有多层全连接网络的上下文编码器以学习隐藏标题表征。同时，事件描述表征也被送入 BERT 编码器来得到隐藏描述表征。相比直接拼接，注意力机制能更好地利用标题中的信息，从多个角度加强事件描述的核心语义。我们对标题使用与上一节构建三元数据集时相同的方式得到事件标题的特征向量，并将标题特征向量输入一个多层感知机（MLP）进行特征抽取，再将事件描述送入 BERT 中得到事件描述的特征向量，之后将事件标题的特征向量作为注意力网络的 ***Query***，事件描述的特征向量作为注意力网络的 ***Key*** 和 ***Value***，使用交叉注意力机制融合标题特征与事件描述特征。融合表征能够使孪生网络中的分类器更加高效地捕捉事件的核心语义，提升分类器辨重的效果。交叉注意力机制的表达式如下：

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{4}$$

其中，d_k 为 Q,V 的维数。

2）孪生网络。BERT 结构通常使用交叉编码器的方法，对两个输入句子进行编码。这种方法每次都需要对新句和候选句进行交叉编码，耗时长，不适合用在大规模的语义文本相似任务上。孪生网络采用两个不同的输入，通过两个具有相同架构、参数和权重的子网络来计算差异。它可以在本地离线存储候选事件的表征，因此每次只需要对新上报的事件进行编码，该特点使得孪生网络适合作为服务在线部署。

基于上述分析，本文使用孪生网络来计算新事件与候选事件的融合特征之间的差异。将两个融合向量 $\boldsymbol{v}_1$，$\boldsymbol{v}_2$ 和它们的元素差分 $|\boldsymbol{v}_1-\boldsymbol{v}_2|$ 连接起来，并输入到一个带有多层感知器的分类器中，分类器的输出为标量。该过程的表达式如下：

$$z=\boldsymbol{W}(\boldsymbol{v}_1,\boldsymbol{v}_2,|\boldsymbol{v}_1-\boldsymbol{v}_2|) \tag{5}$$

$$\hat{y}=\mathrm{Sigmoid}(z) \tag{6}$$

其中，$\boldsymbol{W}$ 为可学习的参数矩阵，$\hat{y}$ 是事件辨重结果，v_i 为融合向量。

本文使用均方误差函数作为孪生网络训练时的损失函数，计算式如下：

$$L=MSE(y,y')=\frac{\sum_{i=1}^{n}(y_i-y_i')^2}{n} \tag{7}$$

其中，y_i 和 y_i' 分别代表事件对真实的重复情况和预测的重复情况。

本文方法的整体流程如算法 1 所示。

算法 1 基于深度对比孪生网络的事件辨重

输入：历史事件集 E_h，新上报的事件 E_{new}

输出：重复事件集合 E_r

1. 通过时间、地点以及文本相似度对 E_h 进行召回和过滤，得到过滤后的事件集合 E_f
2. 利用 E_f 构造三元政务事件数据集 E_t
3. 通过无监督对比学习对预训练的 BERT 模型进行微调，数据集使用 E_t
4. BERT 模型抽取 E_{new} 和 E_h 中的事件描述特征 F_{nd} 和 F_{hd} 以及事件标题特征 F_{nt} 和 F_{ht}
5. 将事件描述特征和事件标题特征输入注意力网络得到融合特征 F_n 和 F_h
6. 将融合特征 F_n 和 F_h 拼接成 $(F_n,F_h,|F_n-F_h|)$ 后输入分类器中，得到重复事件集合 E_r

6 实验设计

6.1 数据集

在实验中，本文使用了来自南通市 2019－2020 年的 12345 市民热线的共 213717 条无标记事件数据，以及 14869 条人工标注好的重复事件对。通过文本相似度的方法从 2019 年以及 2020 年的无标记事件中生成不重复事件对，并

将它们与人工标记的重复事件对结合，组成后续实验中使用的数据集。由于在真实上报事件中，重复事件数量要远少于不重复事件数量，因此将数据集中的正负样本比例设为1:10。实验时，训练集与测试集的比例为 8:2。

6.2 实验设置

在实验中，使用经过预训练的 Roberta-tiny 模型作为对比表征学习部分中的默认参数，输出维度为 312，采用 First-last Average 池化方法。学习率设置为 0.0001，使用 AdamW 作为优化器，mini-batch 的大小设置为 64。在对比学习部分中，损失函数的温度参数 τ 设置为 0.1。在孪生网络部分，在上下文编码器和分类器中使用 ReLU 作为非线性的激活函数，分类器的全连接层数设置为 3 层。本文使用腾讯 AI Lab 提供的 Embedding Corpus for Chinese Words and Phrases[24] 作为词向量表。

6.3 评估指标

事件辨重是一个二分类任务，因此采用了准确率(Accuracy)和 AUC 作为评估指标。对于二分类分类器来说，输出结果标签往往取决于输出的概率以及预定的概率阈值，例如 0.5，其阈值选取在一定程度上也反映了分类器的分类能力。AUC 能够直观地反映分类器具有的分类能力。AUC 的标准如下：

$$\begin{cases} \text{完美分类器}, & AUC=1 \\ \text{优于随机分类器}, & 0.5<AUC<1 \\ \text{差于随机分类器}, & 0<AUC<0.5 \end{cases} \tag{8}$$

在现实场景中进行重复事件辨别时，一个事件如果被视为重复事件，便不会被分派到任何部门。因此如果出现误判，则会导致上报的问题无人解决。在本任务场景中，辨别结果的精确率(Precision)的重要性要大于召回率(Recall)。因此，我们不仅使用 AUC 和准确率作为评估指标，还使用了对精确率更敏感的 F0.5(F0.5-Measure)分数作为评判指标。F0.5 分数的计算方法如式(9)－式(11)所示：

$$Precision=\frac{TP}{TP+FP} \tag{9}$$

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

$$F_{0.5}=(1+0.5^2)\cdot\frac{Precsion\cdot Recall}{0.5^2\cdot Precison+Recall} \tag{11}$$

其中，TP 为被正确分类为正例的数量，FP 是被错误分类为正例的数量，FN 是被错误分类为负例的数量。

6.4 实验环境

本文使用基于 CUDA11.6 的深度学习框架 Pytorch 1.13.1搭建网络模型。实验环境如下：CPU 为 Intel(R) Xeon(R) CPU @ 2.20GHz，GPU 为 NVIDIA TESLA T4 16GB，内存大小为 12GB，操作系统为 Ubuntu 20.04.5 LTS。

6.5 实验结果

6.5.1 对比实验

为了验证文本提出的基于深度对比孪生网络的事件辨重方法的有效性，本文以多种文本语义相似度方法为基线，具体如下。

TF-IDF：用于评估一个词对语料库中文档的重要性。实验时与余弦距离结合使用。

DAN[25]：一种通过求词嵌入平均值，再通过多层全连接网络构建前馈神经网络的模型。

TextCNN[13]：一种用卷积核对词嵌入进行卷积运算和最大池化的方法。

BERT[1]：一种基于双向 Transformer 编码器实现的预训练模型，使用 BERT-Base 版本参数。

E5[26]：一种利用对比学习和大规模文本对数据集进行弱监督训练的文本嵌入模型。

RoBERTa[27]：一种基于 BERT 的改进模型，其可以通过 BERT 加载。在实验中，使用 RoBERTa-tiny 版本参数。

SBert[28]：一种使用 BERT 作为编码器的双塔分类模型，其子网络都是用 BERT，且共享参数，使用 RoBERTa-tiny 版本作为 BERT 的参数。

SimCSE[21]：一种基于句嵌入的对比学习方法。

评测结果如表 2 所列。首先，由表 2 可知，本文方法在各项指标上都优于 Baseline 的模型。TF-IDF 是一种简单的非深度学习统计方法。显然，以深度学习为基础的其他模型的各项分数都远高于 TF-IDF。而以 BERT 为基础的方法又普遍优于基于深度学习的 DAN，TextCNN 方法。在以 BERT 为基础的方法中，虽然 Bert-Base 的参数量远大于 RoBERTa-tiny，但可以看到在各项指标的评估中，BERT 与其他方法的分数相差不大。而相比 RoBERTa，SBERT，SimeCSE，E5，本文方法的 AUC 分别提高了 2.4%，3.6%，2.5%，1.5%，Accuracy 分别提高了 6.5%，5.1%，3.2%，2.3%，这表明使用三元事件数据集对比学习微调后的预训练模型能够更好地生成在语义空间中易于区分的事件语义表征，提高了事件辨重的效果。在针对辨重任务场景特化的 F0.5 分数上，本文方法相比其他基线方法均有大幅度的提高，这表明本文方法可以很好地适应辨别重复事件的任务场景，能够满足事件辨重的需要。

表 2 对比实验结果

Table 2 Comparison experimental results

Model	*Accuracy*	*F0.5*	*AUC*
TF-IDF	0.324	0.139	0.664
DAN	0.732	0.314	0.823
TextCNN	0.764	0.331	0.874
BERT	0.798	0.369	0.915
RoBERTa	0.781	0.364	0.914
SBERT	0.795	0.353	0.902
SimCSE	0.814	0.387	0.913
E5	0.823	0.398	0.921
Ours	**0.846**	**0.537**	**0.938**

6.5.2 消融实验

为了验证本文方法中各模块的有效性，本文进行了如下消融实验。

1)上下文信息模块：分别进行了没有额外模块的原生孪生网络(SN)，通过潜在狄利克雷分布模型[29](LDA)提取上下文特征，通过多层感知器(MLP)提取上下文特征的实验。

2)上下文特征融合方法：分别进行了直接拼接(Concatenate)，通过交叉注意力机制(Attention)融合上下文特征的实验。

实验结果如表 3 所列。相比没有任何额外组件的孪生网络(SN)，添加了上下文信息的网络在所有指标上都有所提升。在提取上下文信息的方法上，本文测试了两种方法：潜在狄利克雷分布(LDA)或多层感知器(MLP)。从实验结果可以看出，提取事件标题中的上下文信息后，辨重效果有了明显提升，特别是 F0.5 指标有了大幅提高。还可以看出，利用 MLP 的方法相比 LDA 对模型的提升效果更为显著，3 项

评估指标均有所提升。而在融合事件描述与标题的特征向量的方法上，本文测试了两种方法：直接拼接描述和标题或使用自注意力机制。从实验结果可以看出，采用交叉注意力机制网络的融合方法可以进一步提高模型的性能，这说明注意力机制可以更有效地融合两个特征，增强事件特征的核心语义，并提高后续事件辨重的准确性。

表 3 消融实验结果

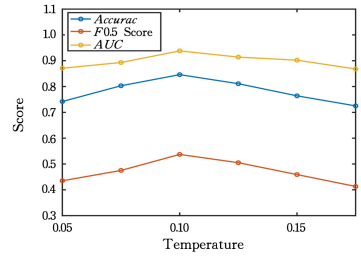
Table 3 Ablation experiments results

Module	*Accuracy*	*F*0.5	*F*1
SN	0.821	0.473	0.906
SN+LDA+Concatenate	0.832	0.504	0.921
SN+MLP+Concatenate	0.837	0.515	0.926
SN+MLP+Attention	**0.846**	**0.537**	**0.938**

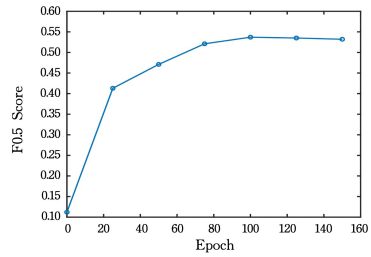
6.5.3 参数设置影响分析

本节探究对比表征学习中的损失函数温度 τ 的设置对后续模型分类效果的影响、epoch 次数对模型表现的影响。由图 5(a)可知，随着损失函数中的温度逐渐升高，其训练得到的预训练模型在之后的辨重任务中的效果逐渐变好，并在温度为 0.1 时达到最佳效果。随着温度进一步升高，辨重效果逐渐下降，这表明过高的温度会使得对比损失函数对负样本的敏感性下降，导致模型学习时没有轻重；而过低的温度则会使模型特别注重困难的负样本，而这些负样本可能是潜在的正样本，导致模型的泛化能力变差。这在一定程度上反映出应根据任务场景设置损失函数的温度。

在实际的任务场景中，除预测精度外，模型能否快速训练部署和更新迭代也是重要的考量因素。从图 5(b)可以看出，本文方法在较少次数迭代后已经有不错的表现，能够满足任务场景所需的上线要求。由图 5(b)可知，随着分类器的全连接层数增加，分类效果逐渐提高，并在层数为 3 层时达到最佳效果。随着层数的进一步增加，分类效果降低，这表明过多的全连接层容易出现过拟合，使模型泛化能力变差。这也在一定程度上反映出应根据实际情况对模型的深度进行设置。



(a)不同温度设置下的实验结果



(b)不同迭代次数下的实验结果

图 5 不同参数设置下的实验结果

Fig. 5 Experiments results with different parameter settings

6.6 样例分析

为直观地展示本文方法的性能，从南通市的 12345 上报事件数据集中选取了 3 组有代表性的事件对进行样例分析。3 组案例具有难易度的区别，案例分析围绕最简单的基础方法 TF-IDF、BERT+分类器、本文方法及人工标注的标准结果展开，结果如表 4 所列。其中，0 代表两个事件不重复，1 代表两个事件重复。

表 4 3 组样例的对比分析结果

Table 4 Comparative analysis results of three groups of cases

事件对内容	TF-IDF	BERT	本文方法	人工标注
1)5 月 6 日 22:00 有施工人员在天都花苑南侧沿河路施工，噪音扰民 2)5 月 6 日 22:00 有施工人员在天都花苑南侧沿河路施工。某某，13**********)	1	1	1	1
1)服务对象来电反映鑫湖国贸 2 号楼 27 层至 30 层停水，联系物业无人处理 2)鑫湖国贸 2 号楼 27—30 层从 6 月 7 日下午开始水压很小，到晚上就停水了，联系了物业但没有人处理	0	1	1	1
1)服务对象是通州区农村户口，此前在启东市工作缴纳公积金，现已离职，希望咨询是否可以提取公积金，如何提取 2)服务对象是启东农村户口，公积金缴纳在通州区，现已离职，咨询如何提取公积金	1	1	0	0

结束语 本文提出了一种由数据驱动的基于深度对比孪生网络的重复事件识别方法。本文方法由 3 部分组成：召回和过滤、对比表征学习和上下文孪生网络。在南通市 12345 热线数据集上的实验结果表明，本文方法在多项指标上都超过基线方法，达到了较好的效果。而在 F0.5 分数上的大幅提高，表明本文方法能够较好地满足政务事件辨重任务的需要，更加精确地辨别重复事件，降低误判带来的影响，提高事件处置的效率。在后续工作中，我们将考虑在孪生网络中加入更多的事件相关信息，如事件的时间、归口等，这些也许能够进一步提高模型对重复事件辨别的效果。

参 考 文 献

[1] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies,2019:4171-4186.

[2] SARZYNSKA-WAWER J,WAWER A,PAWLAK A,et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research,2021,304:114135.

[3] YANG Z,DAI Z,YANG Y,et al. Xlnet:Generalized autoregressive pretraining for language understanding[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:5753-5763.

[4] ZHENG Y P,MA X L. Using Government Hotline Data to Promote Smart Governance—The Case of Guangzhou Government Hotline[J]. E-Government,2018(12):18-26.

[5] MA X L, ZHENG Y P, ZHANG C W. The Big Data Empowering Effect of Government Hotlines on City Governance Innovation: Value, Status and Issues[J]. Documentation, Informaiton & Knowledge, 2021, 38(2): 4-12.

[6] CHENG X M, CHEN G, CHEN J P, et al. RAVA: An Reinforced-Association-Based Method for 12345 Hotline Events Allocation[J]. Journal of Chinese Information Processing, 2022, 36(10): 155-166, 172.

[7] PU X, LONG K, CHEN K, et al. A semantic-based short-text fast clustering method on hotline records in Chengdu[C]// 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress. 2019: 516-521.

[8] PENG X, LI Y, SI Y, et al. A social sensing approach for everyday urban problem-handling with the 12345-complaint hotline data[J]. Computers, Environment and Urban Systems, 2022, 94: 101790.

[9] LUO J Y, QIU Z, XIE G Q, et al. Research on civic hotline complaint text classification model based on word2vec[C]// 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE, 2018: 180-1803.

[10] CHANDRASEKARAN D, MAGO V. Evolution of semantic similarity— a survey[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-37.

[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of the 1th International Conference on Learning Representations. 2013.

[12] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[13] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1746-1751.

[14] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 562-570.

[15] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2873-2879.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[17] KATHERINE L, DAPHNE L, NYSTROM A, et al. Deduplicating Training Data Makes Language Models Better[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 8424-8445.

[18] BIKASH G, LUCAS A, PETR K. Deduplication of Scholarly Documents using Locality Sensitive Hashing and Word Embeddings[C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 901-910.

[19] JAISWAL A, BABU A R, ZADEH M Z, et a l. A survey on contrastive self-supervised learning[J]. Technologies, 2020, 9(1): 2.

[20] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. PMLR, 2020: 1597-1607.

[21] GAO T, YAO X, CHEN D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 6894-6910.

[22] WIETING J, BANSAL M, GIMPEL K, et al. Towards Universal Paraphrastic Sentence Embeddings[C]// Proceedings of the 4th International Conference on Learning Representations. 2016.

[23] CHEN X, HE K. Exploring Simple Siamese Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15750-15758.

[24] SONG Y, SHI S, LI J, et al. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 175-180.

[25] IYYER M, MANJUNATHA V, BOYD-GRABER J, et al. Deep Unordered Composition Rivals Syntactic Methods for Text Classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1681-1691.

[26] WANG L, YANG N, HUANG X, et al. Text Embeddings by Weakly-supervised Contrastive Pre-training[J]. arXiv: 2212.03533, 2022.

[27] LIU Y, OTT M, GOYAL N, et al. Roberta: A Robustly Optimized BERT Pretraining Approach[J]. arXiv: 1907.11692, 2019.

[28] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3982-3992.

[29] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(Jan): 993-1022.

LI Zichen, born in 1997, postgraduate. His main research interests include urban computing and deep learning.

YI Xiuwen, born in 1991, Ph. D, data scientist, researcher, is a senior member of CCF (No. 45025M). His main research interests include spatio-temporal data mining and deep learning.

(责任编辑:喻藜)