基于事件画像和案例推理的社区工单处置

强海玲*,陈　剑,佘祥荣,陈健鹏,陈　钢
(长三角信息智能创新研究院,安徽 芜湖 241000)

摘　要:近年来,随着政府数字化转型的不断深入,越来越多的 12345 政务热线工单下发到社区进行处置。工单文本信息通常较为稀疏,主题序列涵盖城市治理方方面面。社区管理人员对工单进行处置往往花费较长时间,无法满足群众实时响应的需求。为了提升社区工单处置的质量和时效性,该文提出了一种基于事件画像和案例推理的工单处置决策方法。首先,基于统一标准地址库以三元组方式构建地名地址基因库用以获取地名中的谱特征,构建树集合以表征地址基因之间的层次关系,利用地址基因之间的关联关系对缺失地址元素进行补全和还原;其次,为了充分发掘社区工单文本的局部特征和全局特征,该方法通过基于 BiGRU、Self-Attention、CNN、CRF 的组合神经网络对社区工单事件进行有效提取;最后,在构建社区事件历史案例库的基础上使用关键词提取并计算事件之间的相似度。对比实验结果表明,该方法相较于其他方法能够取得更好的性能。

关键词:事件画像;案例推理;工单处置;地名地址基因;事件提取;组合神经网络
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2023)07-0012-08
doi:10.3969/j.issn.1673-629X.2023.07.002

Community Work-order Disposal Based on Event Portrait and Case-based Reasoning

QIANG Hai-ling*,CHEN Jian,SHE Xiang-rong,CHEN Jian-peng,CHEN Gang
(Yangtze River Delta Information Intelligence Innovation Research Institute,Wuhu 241000,China)

Abstract:In recent years,with the deepening of the government's digital transformation,more and more 12345 government hotline work-orders are issued to the communities for disposal. Work-order text information is usually sparse,and the topic sequences cover all aspects of urban governance. It often takes a long time for community managers to handle these work-orders,which cannot meet the needs of the people for real-time response. In order to improve the quality and timeliness of community governance,we propose a work-order disposal decision-making method based on event portrait and case-based reasoning. Firstly,the address gene database of geographical names is constructed in the form of triples to obtain the spectral features in geographical names based on the unified standard address database,and the tree set is constructed to represent the hierarchical relationship between address genes with the purpose of completing and restoring the missing address elements. Secondly,in order to fully explore the local and global features of the community work-order text,the community event contained in the work-order text is extracted by the combined neural network based on BiGRU,Self-Attention,CNN and CRF models. Finally,on the basis of building the historical cases for the community events,keyword extracting is used to calculate the similarity between events based on their keywords. Comparative experimental results show that the proposed method can achieve better performance than the baseline methods.

Key words:event portrait;case-based reasoning;work-order disposal;address gene;event extraction;combined neural network

0　引　言

近年来,各地政府尝试将信息技术嵌入服务流程中,逐步推动政务服务数字化转型。在此背景下,12345 便民服务热线受理的群众诉求工单也逐步下层到社区去处置。大多数 12345 便民服务热线工单文本都在 200 字以内,内容相对稀疏,且工单文本的主题序列涵盖城市生活的诸多领域[1]。对于社区管理人员来说,如何高效处置上级单位派发的工单,及时满足人民群众的诉求,对于提升政府治理能力具有重要意义。借助于用户画像理念,事件画像(Event Portrait)将事件视为一个对象,使用大数据画像技术对相关特征分析挖掘,留下对事件画像刻画有贡献的特征,实现对事

收稿日期:2022-09-13　　**修回日期**:2023-01-16
基金项目:国家自然科学基金(61976198);2021 年安徽省重点研究与开发计划(202104a05020071)
作者简介:强海玲(1987-),女,硕士,通信作者,研究方向为人工智能、社区治理。

件更为清晰、直观的勾勒描述。事件画像主要包括三类特征:事件主题内容(类型、触发词)、事件元数据信息(地理信息、时间信息、主客体等)、事件其他信息(任务、关键词等)。案例推理(Case-Based Reasoning, CBR)从历史案例库中获取经验和知识,针对新旧情况的差异做相应的调整,从而得到新问题的解决方案并形成新的案例[2]。

社区工单处置的目的是及时处置群众反映的事件,处置事件的措施能够为未来同类事件提供参考。事件处置旨在使用最佳手段解决问题,由于历史事件库中留存了大量已发生事件的解决方式,所以社区管理者可以从历史库中选择一个与待处置事件最为相似的事件加以参考,减少处置响应时间,提升决策效率和科学性。为了在社区海量事件库中快速、准确地检索工单所包含的相似事件及处置方案,该文提出一种基于事件画像和案例推理的社区治理工单处置决策方法。该方法通过地址标准化和事件提取构建事件画像,然后通过构建社区事件历史案例库并基于关键词提取来挖掘社区事件之间的相似性,进而找出与新事件相似的案例并给出参考解决方案,协助社区管理人员高质量、高效率地处置社区工单。

1　相关研究

1.1　事件提取

文献[3]提出一种事件要素注意力与编码层融合的事件触发词提取模型,该模型能够有效地利用事件要素信息,提高触发词提取性能。该模型采用的预训练语言模型还需改进,以针对跨文档的时间聚合。文献[4]利用图卷积神经网络和自注意机制来学习不同句法距离的相关词之间的依赖关系,通过 mask 算法计算词间的句法距离矩阵完成事件提取任务。该模型可以同时预测文本中所有提及对之间的关系,然而该模型在捕捉远距离关系时性能不足。文献[5]利用 TAC-KBP 时隙填充来表示事件的模糊时间跨度,基于共享参数和时间关系构建文档级事件图,并使用图形注意网络传播时间信息避免了在传统模型中的错误传播。该模型没有在事件和关系间构建更鲁棒的结构化约束。文献[6]提出了全方位事件提取方法,将两类任务视作同一个任务,避免了上游任务对下游任务的影响,使用神经网络学习特征,引入注意力机制突出重点信息,但该模型中针对事件触发词的检测未能完全作用于下游任务中。文献[7]提出了具有共享表示学习和结构化预测的联合事件和时间关系提取模型,利用结构化推理和学习方法来分配事件标签和时间关系标签,提高了事件和时间关系提取的性能。该模型采用结构化推理和学习方法的结合方式,使得计算时间相较于其他模型较高。文献[8]针对滑坡灾害指标在数值上的分布特点进行区间划分,应用剪枝的方法对所有指标进行筛选,然后计算综合相似度,进而得到供参考的相似案例,在模型性能上优于其他模型。该模型针对指标所涉及不同领域的问题考虑不足,且数据集的选取较为单一。

1.2　案例推理

为了改善案例推理检索算法的预测结果质量,文献[9]提出一种改进的 KNN 案例推理检索算法。相较于传统 KNN 案例推理检索算法,改进的 KNN 案例推理检索算法预测结果的精度显著提高。但案例间的相似度计算没有考虑到特征权重对结果的影响,缺乏对特征权重的有效分配,因此预测精度仍有待继续提升。为了让决策者快速借鉴历史案例做出满意的应急决策,文献[10]基于案例推理提出针对滑坡灾害的应急相似案例智能生成方法,并以湖北省秭归县滑坡为例验证该方法的有效性。该方法兼顾考虑了指标在整体和局部的情况,有效地找出与目标案例最为相似的历史案例,但大多数自然灾害都伴发次生灾害,应急方案的生成应该联系次生灾害动态方案的生成。为了探索智能化跨域立案,对案件进行适用法律条文自动推荐,文献[11]基于民间借贷案件提出了基于案例推理的推荐方法,并取得了较好的准确率和召回率,但收集的验证集数据量较少,无法准确地反映模型的各项指标,有必要在更多的数据集上进行验证。文献[12]基于案例推理原理,利用不同类型特征属性将案件量化来构建案例库,并使用 K 近邻算法计算历史案例与目标案例的相似度,匹配出相似度最高的历史案例。该算法在特征属性方面有待进一步完善,比如添加社区纠纷特征属性,将会优化社区纠纷调解方案的选择。为了提高突发事件发生时公安指挥部门处置决策方案的及时性和科学性,文献[13]提出基于案例推理和规则推理的公安突发事件辅助决策算法,使用 CBR 和 RBR 相结合的方法构建该辅助决策算法,通过引入分级检索算法提升 CBR 检索案例的速度,并改进 KNN 算法提升相似度计算的有效性。

2　事件画像

在社区事件上报过程中,由于表达习惯的不同,同一地理实体可能对应多种不同的地名描述,这些地名指代往往存在模糊性、随机性、多样性等特点。此外,事件文本在空间序列上呈现出楼宇、小区、社区、街道的逐层扩散特征,在主题序列上通常覆盖城市生活各个方面的连续扩展特征[14]。如果仅基于词级语义而忽视句级别语义特征来处理工单文本,只能关注到浅层文本信息,在事件主题的挖掘上会有较大的偏差。

因此,在面向社区治理的事件画像构建过程中要解决地址标准化和事件提取问题。事件画像使用地名地址基因对事件文本中包含的地址信息进行提取,而后对相邻的地理元素进行完整性判断,并将不完整的地址基因扩充为完整基因集,进一步合并后将每一个地址基因扩充成标准化地址,基于该地址将事件分拨给对应网格下的社区管理员。

2.1 地址标准化

地名地址在形式上可以分解为若干地名地址要素,引起相互之间的关联与派生关系,单个地名地址要素或若干个地名地址要素的组合形成地名地址基因。尽管地名指代的描述存在不确定性,但针对同一事件的地名地址使用一般存在描述相似性,即地名描述中所包含的地名地址基因往往是相似的。该文基于统一的标准地址库构建地名地址基因库,并构建树集合以表征地址基因之间的层次关系,如图 1 所示。

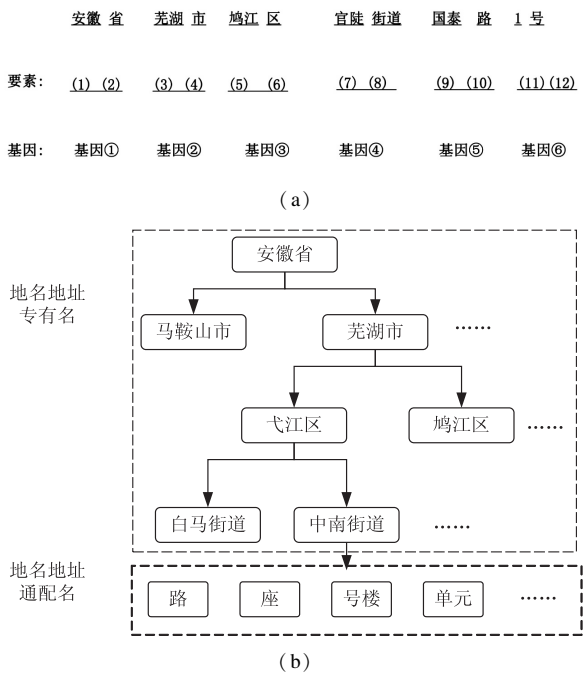

图 1 地名地址基因示例

2.1.1 地名地址基因库

地址是一种将若干地址元素按照一定规则排列组织起来的短文本,因而如何寻找各个地址元素之间的“位置点”是地址分词的关键问题。该文利用统计学特征确定标准地址库中地址的落差点,通过递增切分的方法,对递增的短语在整个地址库中的数量进行统计,根据地址元素的使用频次会随地址描述逐渐精确而逐渐降低的规律,当待判断短语后缀超过落差点后,对应的短语在整个地址库中出现的数量将发生明显下降,据此可划分出落差点集合 M 。

由于存在落差点之间的元素长度过短、错误或并非完整元素等情况,落差点并不完全等于后缀点,但落差点中包含划分地址中专有名词的后缀点,且后缀点之间的内容构成地址要素。为了对 M 中的元素是否为正确的后缀点做出判断,按照地址构成方式的规则设计决策树,然后根据决策树对每一个落差点 $m_i \in M$ 是否构成后缀词或后缀点做出判断,依据判定成功的后缀点进行分词,并对两个后缀点之间的地址要素加以记录,如图 2 所示。

经过分词后,标准化的地址描述所包含的地址要素被划分为专有地址部分与通配地址部分,同时获得一个包含专有地址名词基因的词表 WordList。针对专有地址部分,基于标准地址自身的前后文关系,结合地址信息本身所包含的层次,为提取后的专有地名元素赋予先后序关系标记,构成三元组:

$$(\mathrm{id}, \mathrm{ele}_i, \mathrm{seqmark}_i)$$

其中,id 表示对地址元素的唯一标识, ele_i 表示专有地名元素, $\mathrm{seqmark}_i$ 是以地址元素所属行政区划层级表示的先后序标记。针对每一条地址,将对应的三元组元素按前后序关系构建一棵子树 Tree_i ,将每一棵子树完全相同三元组的节点进行合并,合并后的若干棵树构成的集合 TreeSet 构成了一个基本的地名地址基因库 AddressDB,其中包含以地理要素为基础构建的地名地址基因及其对应的层次关系。

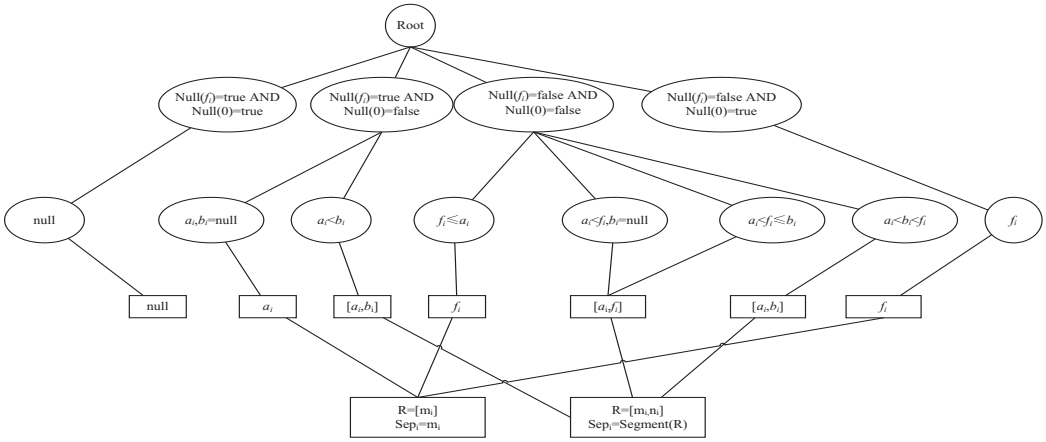

图 2 基于后缀点的地址要素提取

2.1.2　地址基因匹配

利用构建好的地名地址基因库对事件文本进行要素抽取,使用获得的词表 WordList 与全文进行匹配,提取其中的专有地址基因集合 D_1。针对“号”“号楼”“栋”“幢”等地址通配名进行逆向增字匹配,匹配到通配名后向前判断通配名之前的字符是否为阿拉伯数字、以汉字表达的数字或英文字母,若符合则将其加入匹配结果并继续判断,直到判断为否为止,构建通配地址基因集 D_2。对集合 D_1 中的两个相邻元素 d_1、d_2,两个元素在事件文本中对应起始位置 loc 若满足:

$$\mathrm{loc}_{\mathrm{ad}_j} - \mathrm{loc}_{\mathrm{ad}_i} = \mathrm{length}(\mathrm{ad}_i)$$

则判定两元素为相邻,反之判定为不相邻。对相邻的基因元素,利用地名地址基因库中专有地址基因三元组中包含的先后续标记关系 seqmark 对相邻元素的完整程度进行判断:若两个相邻元素的标记之间存在缺失值,则说明两个地理元素之间存在要素缺失。

根据构建出的地址基因库中的三元组树从上到下搜索,对不符合条件的相邻地址基因进行补充,生成新的完整地名地址基因作为事件中提取得到的地名地址信息。基于标准地址库进行搜索匹配,将关键地址基因扩充成完整的标准化地址,逐层解析并对应分拨到具体网格,进而完成对事件文本中地理要素的提取。

2.2　事件提取

卷积神经网络(Convolution Neural Network,CNN)可以有效提取文本局部特征,但在卷积和池化操作时会丢失工单文本序列中词汇的位置和顺序信息,因此不能很好地捕捉工单文本的全局信息[15]。门控循环单元(Gated Recurrent Unit,GRU)可以有效获取序列化句子的层级特征,然而单从一个方向提取特征不能完整地表示整个句子的上下文特征。双向 GRU 网络(Bidirectional GRU,BiGRU)可以有效获取文本上下文依赖关系和全局信息[16]。Attention 机制能够凸显文本的重要特征以便更好地提取关键信息[17]。为此,在 BiGRU 网络中引入 Attention 机制,使其在计算语义信息时根据其重要程度赋予不同的权重。条件随机场(Conditional Random Field,CRF)可以有效解决序列标注问题[18],可以为每个词语分配标记并计算整个序列得分,有效提取社区事件文本中的事件特征。基于上述模型优势,该文提出一种混合神经网络用于提取社区事件文本中的事件信息,如图 3 所示。

首先,为了方便对较长的社区事件描述文本进行处理,使用 Jieba 等中文分词工具对社区事件描述文本进行分词,获得分词列表,并配合使用自定义停用词表提升对未能正确识别的专有名词的识别效果;接着,使用 word2vec、glove 等词向量工具对分词后的结果进行编码,转换为高维空间中的向量 e,以方便下一步分别从全局特征和局部特征角度进行语义信息提取;然后,将词向量输入到卷积神经网络中,提取事件描述文本中的局部特征,同时,通过双向 GRU 网络,基于上下文信息提取事件描述文本的全局信息,作为对卷积神经网络无法捕捉到的远距离语义关联信息的补充,提高对各事件要素识别的准确度。完成全局特征与局部特征的提取后,对两类特征进行融合,并输入到自注意力机制模型中,从融合后的信息中提取出描述文本中的关键信息;最后,将自注意力模型得到的结果与卷积神经网络得到的结果进行融合,并输入 CRF 模型,输出最高概率的要素识别序列,作为对事件中各个关键事件要素的提取结果。

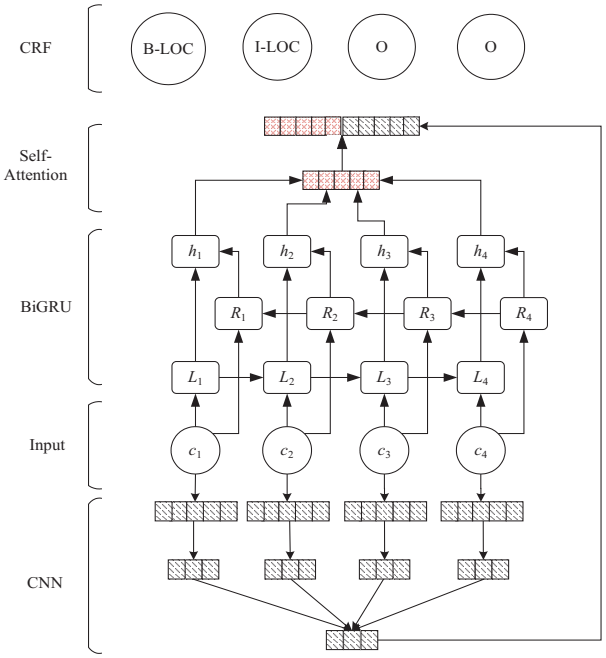

图 3　社区工单事件提取模型

2.2.1　BiGRU 模块

在社区工单文本事件提取中,当前时间步长的隐藏状态与前一时刻和下一时刻相关联。采用单向 GRU 网络对文本序列建模时,状态总是由前向后传递,因此仅能获取文本前文信息,难以获取整个文本的上下文信息。BiGRU 由前向 GRU 单元和后向 GRU 单元组成。GRU 由更新门和重置门组成。门结构可以选择保存上下文信息来解决 RNN 梯度消失或爆炸的问题。在性能与 LSTM 相当的情况下,GRU 的结构比 LSTM 更简单,训练速度更快。前向 GRU 单元的隐藏层表示为 $\overrightarrow{h_t}$,后向 GRU 单元的隐藏层表示为 $\overleftarrow{h_t}$。单向 GRU 在 t 时刻的隐藏层输出计算公式如下:

$$\overrightarrow{h_t} = \mathrm{GRU}(x_t, \overrightarrow{h_{t-1}}) \tag{1}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(x_t, \overleftarrow{h_{t-1}}) \tag{2}$$

根据前向 GRU 单元和后向 GRU 单元的隐藏层输

出,得到 BiGRU 在 t 时刻的隐藏层输出:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{3}$$

2.2.2 Self-Attention 模块

自注意力机制被认为是查询到一系列键值对的映射,在此基础上可以更好地为重要信息分配权重,更准确地理解序列语义。BiGRU 难以从句子序列中捕获重要信息,在社区工单文本信息中,对于事件类型判定价值较高的信息往往集中在部分关键词。例如,工单文本“来电人反映融创招商星河万里楼盘经常夜间施工,噪音太大了,影响周边居民休息”属于噪音污染,文本中“夜间施工”“噪音太大了”等信息对于事件提取较为重要,而“融创招商星河万里”“楼盘”等字词对事件提取的帮助较小,可能会削弱事件提取效果。自注意力机制善于获取文本特征内部的相关性,在 BiGRU 网络捕捉社区工单文本的上下文特征后,该文采用自注意力机制来提取社区工单文本句子中的重要信息。融合自注意力机制后的 BiGRU 网络能够得到注意力的概率分布,降低无效信息的干扰,从而提升事件提取性能。多头注意力机制的计算公式如下:

$$m(h_t) = \text{concat}(\text{score}_1(h_t), \text{score}_2(h_t), \cdots, \text{score}_h(h_t))W^O \tag{4}$$

其中,h_t 为 BiGRU 的隐藏层输出,score_i 为第 i 个自注意力机制的输出,h 为重复次数。

利用 Attention 函数对 score_i 进行计算:

$$\text{score}_i(\boldsymbol{h}_t) = \text{attention}(h_t\boldsymbol{W}_i^Q, h_t\boldsymbol{W}_i^K, h_t\boldsymbol{W}_i^V) \tag{5}$$

其中,$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$ 和 $\boldsymbol{W}^O$ 为参数矩阵,用于将输入 $\boldsymbol{h}_t$ 映射到不同的向量空间,分别为 $\boldsymbol{W}_i^Q \in R^{d\times d_Q}$,$\boldsymbol{W}_i^K \in R^{d\times d_Q}$,$\boldsymbol{W}_i^V \in R^{d\times d_i}$,$\boldsymbol{W}^O \in R^{hd_i\times d}$。$d$ 是 BiGRU 网络隐藏层的输出向量维度,d_Q 和 d_V 是向量空间维度。

函数 Attention 是扩张点积的自注意力机制操作,计算公示如下所示:

$$\text{attention}(Q,K,V) = \text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V \tag{6}$$

其中,$\sqrt{d}$ 起到缩放调整的作用,使内积不会太大。Self-Attention 网络在时间 t 的隐藏层输出定义为 $H_t = m(h_t)$,该层输出序列为 $H = (H_1, H_2, \cdots, H_n)$。

2.2.3 CNN 模块

该文利用 CNN 的特性提取社区工单文本局部特征形成语义特征向量,相当于采用卷积核在输入矩阵上滑动进行乘积求和的过程。首先,将词向量序列通过 $M\times N$(M 为卷积长度,N 为词向量长度)卷积核在输入层从上到下滑动进行等长卷积操作,得到特征向量 c,$M\times N$ 的卷积核根据不同 M 大小提取不同长度相邻单词的特征。卷积操作所生成的每一个特征向量 c 送到池化层用以生成潜在的局部特征。采用最大池化策略对输出结果进行池化来获取最重要特征:

$$\hat{c} = \max\{c\} \tag{7}$$

2.2.4 CRF 模块

CRF 是解决序列标注问题的主流方法,它无需设置规则就可以捕捉序列元素中相邻元素的影响,且不局限于任意时刻的观察值。为此,该文利用 CRF 为每个词语分配标记并计算整个序列的得分。BIO(B-begin,I-inside,O-outside)标签体系在精度和训练复杂度上具备优势,该文选择该标签体系对序列进行标注。在工单文本的标签定义中,数据集内凡是带有事件地点(LOC)和事件触发词(TRG)标记的词,词中第一个字重新标记为 B-LOC 和 B-TRG,词中剩余的字重新标记为 I-LOC 和 I-TRG,对两位以上连续数字做合并标记,其他词一律标记为 O(含标点符号)。CRF 为每个词语分配标记,并计算整个序列得分。首先,将 Self-Attention 和 CNN 的结果进行拼接以将二者特征相结合;然后,通过全连接层降维后使用 softmax 得到置信度分布 conf:

$$\text{conf} = \text{softmax}(\text{linear}[\boldsymbol{V}_{\text{Self-Attention}}, \boldsymbol{V}_{\text{CNN}}]) \tag{8}$$

最后,采用 CRF 网络计算序列标签得分,最终结果为最高得分的标注序列。序列得分由词语标记得分和标记转移得分共同构成。

假设 label 是标签序列$(\text{label}_1, \text{label}_2, \cdots, \text{label}_k)$,则序列 seq 的 CRF 得分计算如下:

$$\text{score}(\text{seq},\text{label}) = \sum_{i=1}^{n} \text{conf}_{i,\text{label}_j} + \sum_{i=2}^{n} T_{\text{label}_{i-1},\text{label}_i} \tag{9}$$

其中,$\text{conf}_{i,\text{label}_j}$ 表示 seq_i 的标签是 label_j 的置信度,$T_{\text{label}_{i-1},\text{label}_i}$ 表示从标签 label_{i-1} 过渡到标签 label_i 的概率。损失函数定义如下:

$$\text{LOSS}_{\text{CRF}} = -\text{score}(\text{seq},\text{label}) + \log\sum_{Y} e^{\text{score}(\text{seq},\text{label})} \tag{10}$$

其中,Y 为所有可能的标签序列集合。

2.2.5 模损失函数

采用联合损失函数进行模型训练:

$$\text{LOSS}_{\text{CRF}} = -S(x,y) + \log\sum_{Y} e^{S(x,y)} \tag{11}$$

$$\text{LOSS} = -\sum_{E}\log p_{E_j} + \text{LOSS}_{\text{CRF}} \tag{12}$$

其中,j 是社区工单 E 所属的事件类型。

3 案例推理

基于案例推理的社区工单处置主要包含两部分:(1)构建历史社区事件案例库;(2)基于事件关键词的检索、重用、修正、保存。

3.1 历史案例库

面向海量社区工单事件,构建一个可以快速检索

的历史事件案例库。历史事件案例主要是文本类型数据,包括对事件整体情况的描述,还包括对事件的解决方案的描述和对事件求解效果的描述。单个历史事件案例可以表示为:

<事件描述,事件解决方案描述,效果描述>

为了快速检索历史事件案例库,需要对事件描述生成标签来完成快速检索。该文使用 XLNet 预训练语言模型[19]对历史事件描述进行关键词提取,将关键词作为事件案例的标签,并对标签进行编码处理。某个历史事件案例可以表示为:

<标签编码集合,事件描述,事件解决方案描述,效果描述>

利用标签编码可以实现对历史事件案例的快速检索。事件描述关键词提取算法如下:

Step1:对事件描述进行句子分割并使用 jieba 分词进行分词处理得到 document tokens 和 sentence tokens,并将分词后的 token 进行词性标注得到带有词性标签的 label token 序列。

Step2:使用 NP-chunker 根据词性标签从 label token 序列中提取名词 token(NP),得到的 NP 作为候选关键词,如图 4 所示。

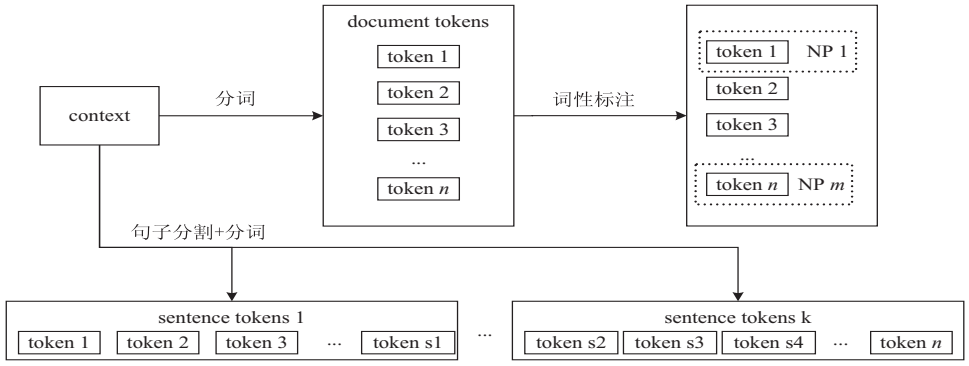

图 4　候选词生成

Step3:将所有 document tokens 使用 XLNet 生成词向量,再使用 SIF 权重将词向量组成 word level 的文本向量。

Step4:将所有 sentence tokens 使用 XLNet 生成词向量,使用 SIF 权重将词向量组成多个句向量。根据文本的内容层次分布(如文章的中心内容主要集中在第一句或最后一句,句子长度等),使用加权平均的方法将多个句向量组合成 sentence level 的文本向量。最后,将 word level 和 sentence level 的文本向量加权组成 document vector,如图 5 所示。

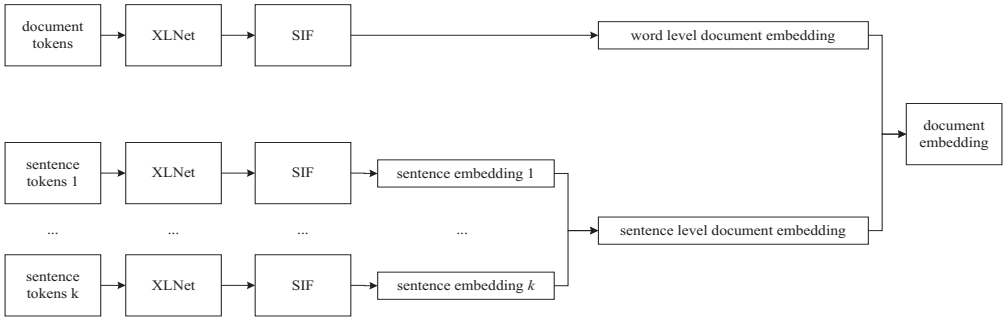

图 5　文档级语义向量构建

Step5:将每个 label token 使用 XLNet 生成 word vector,计算与 document vector 之间的距离。将此距离视为候选关键字与文档主题之间的相似度,选择最相似的候选关键词的前 N 个作为最终关键词,如图 6 所示。

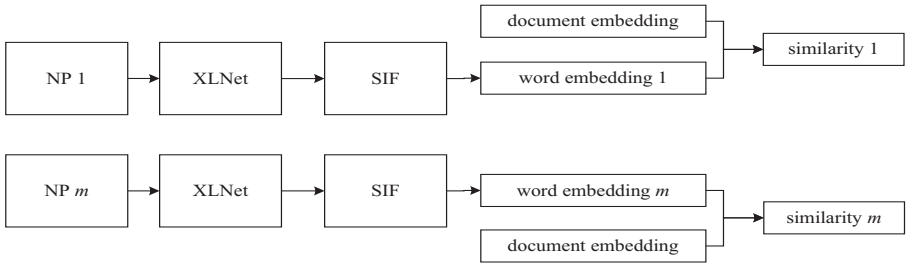

图 6　最终关键词筛选

3.2 案例检索

当社区新发生事件上报后,对新事件描述进行关键词提取,将关键词作为事件标签,并对标签进行编码处理,得到新事件表示:

<标签编码集合,事件描述>

用标签编码集合中的每个标签编码在历史事件案例库中进行查询,查找包含新事件标签的所有历史事件案例作为候选集。对候选集中的所有事件描述与新事件描述进行事件画像相似度计算,运用 Glove 模型、word2vec 模型训练生成词向量,计算标签词向量的相似度,设定阈值 0.8 作为两个事件是否相似的判定标准。具体算法流程如下:

Step1:收集历史案例库中的所有事件文本描述构建数据集 A,以维基百科语料库构建数据集 B;

Step2:对数据集 A 和数据集 B 中的文本进行清洗、去除停用词、分词操作,并进行合并构成数据集 C;

Step3:基于数据集 C 训练 Glove 或 word2vec 模型,得到数据集 C 中所有词对应的词向量,并建立对应的词表 T;

Step4:根据词表 T 查询事件标签词集合中的所有标签词的词向量表示,并进行加权融合得到事件描述的词向量表示结果 V;

Step5:通过余弦公式计算候选集中的所有事件描述与新事件描述的相似度,输出相似度大于 0.8 的所有事件描述;

Step6:对输出的所有事件描述进行相似度排序,输出 Top-N 结果。

将相似度排序前 3 个($N=3$)事件案例推送给社区管理人员。社区管理人员根据系统提供的案例处置当前新事件并生成新的解决方案。此时,当前新事件可以表示成:

<标签编码集合,事件描述,事件解决方案描述>

最后,将该新事件存入历史事件案例库。

4 实验分析

4.1 实验环境

该文使用基于 CUDA 9.0 的深度学习框架 PyTorch 1.0.1 搭建网络模型,实验操作系统为 CentOS 7.3,内存 64 GB,CPU 为 Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz,GPU 为 NVIDIA Tesla A100。

4.2 数据集

从 2019 年 1 月 1 日-2022 年 6 月 30 日期间安徽省芜湖市社区事件中挑选了 20 000 个事件构建了实验数据集,数据集描述如表 1 所示。同时,对 60 000 条工单文本进行了统计分析,事件文本长度的均值为 173 个字,且 95% 的事件文本长度在 200 个字以内。

表 1 数据集描述

数据总量	训练集	验证集	测试集	事件类型
20 000	13 000	4 000	3 000	61

4.3 实验设置

在整体网络训练过程中,所用模型的超参数如表 2 所示。模型使用学习率为 1e-5 的 Adam 优化器。

表 2 超参数设置

参数名	参数值
Epoches	20
BatchSize	128
SequenceLength	200
CNN-KernelSize	2,3,4
BiGRU-HiddenSize	128

4.4 基线模型对比

采用精确率(precision)、召回率(recall)和加权 F1 值作为评价指标。为了验证文中事件案例检索方法的性能,与以下基线模型进行了对比:(1)使用 TF-IDF 结合 word2vec 提取事件文本特征,并用 XGBoost[20]、LightGBM[21]和文献[12]所采用的 KNN 方法预测社区事件类型;(2)TextCNN 模型[22]和 HAN[23]模型;(3)RoBERTa 模型[24]和 ELECTRA 模型[25]。对比实验结果如表 3 所示。

表 3 对比实验结果

模型	precision	recall	加权 F1
XGBoost	0.63	0.70	0.66
LightGBM	0.64	0.70	0.60
文献[11]	0.67	0.72	0.67
文献[12]	0.65	0.69	0.66
TextCNN	0.67	0.70	0.72
HAN	0.73	0.78	0.74
RoBERTa	0.73	0.75	0.75
ELECTRA	0.71	0.74	0.76
文中方法	0.79	0.79	0.78

从表 3 可以看出,文中方法取得了比其他基线方法更好的实验效果。基于 TextCNN 和 HAN 的方法在事件分类效果上好于基于 XGBoost、LightGBM 和 KNN 的方法,原因在于后者仅简单的对事件文本中的词向量进行加权平均,无法挖掘事件文本中更深层次的语义信息。使用 Attention 机制的方法拥有更高的分类准确率,因为 Attention 机制可以让模型更加关注那些对预测贡献较大的特征。此外,由于预训练语言模型包含了大量的先验知识,基于 RoBERTa 和 ELECTRA 的方法也取得了较好的实验效果。该文引

入了基于 CNN 的局部特征提取和基于 BiGRU - SelfAttention 的全局特征提取,兼顾到了事件文本局部信息和上下文语义信息,充分发挥了各网络的优势,因而取得了更好的性能。

4.5　案例检索性能

表 4 展示了在不同事件规模下文中案例检索方法的性能。可以看出,在事件规模达到 20 000 件时,检索仅需 4.61 s。相较于文献[11]和文献[12],文中方法更能够满足社区工单处置的实时性要求。

表 4　案例检索性能

事件规模	检索时间/s		
	文献[11]	文献[12]	文中方法
5 000	8.58	5.17	1.92
10 000	15.09	11.03	3.15
15 000	18.64	15.60	3.80
20 000	27.11	23.24	4.61

5　结束语

针对社区管理人员处置工单耗时长、效率低的问题,提出了一种基于事件画像和案例推理的智能社区工单处置决策方法。通过统计学方法和决策树将地名地址描述拆分为地名地址基因,并以三元组形式构建了地名地址统一基因库,以此获取地名中的谱特征。考虑到工单文本局部特征和全局特征对于事件提取效果均存在影响,设计了一种基于 BiGRU、Self - Attention、CNN、CRF 的组合神经网络对工单文本进行事件提取。在工单事件处置方面,该方法利用历史事件案例构建案例库,对事件描述关键词进行编码,以关键词编码进行查询实现快速的事件检索,使用相似度实现精准的事件匹配并给出工单处置的参考解决方案。下一步工作是引入在线学习(Online Learning)来进一步提升社区工单处置的效率。

参考文献:

[1] 陈　钢. 融合 RoBERTa 和特征提取的政务热线工单分类[J]. 计算机与现代化,2022(6):21-26.

[2] 孔　钦,叶长青. 基于案例推理的故障诊断算法[J]. 计算机系统应用,2016,25(1):181-186.

[3] PAN Z,HUANG D G. Research on trigger word extraction based on the fusion of event argument attention and encoder layer[J]. Journal of Chinese Computer Systems,2021(4):673-677.

[4] AHMAD W U, PENG Nanyun, CHANG Kaiwei. GATE: graph attention transformer encoder for cross-lingual relation and event extraction[C]//Proceedings of the 35th conference on artificial intelligence. [s. l.]:AAAI,2021:12462-12470.

[5] WEN Haoyang, QU Yanru, JI Heng, et al. Event time extraction and propagation via graph attention networks[C]//Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Minneapolis: Association for Computational Linguistics, 2021:62-73.

[6] SHAFIEIBAVANI E, YEPES A J, ZHONG X, et al. Global locality in biomedical relation and event extraction[C]//Proceedings of the 19th SIGBioMed workshop on biomedical language processing. [s. l.]:[s. n.],2020:195-204.

[7] HAN Rujun, NING Qiang, PENG Nanyun. Joint event and temporal relation extraction with shared representations and structured prediction[C]//Proceedings of the 2019 conference on empirical methods in natural language processing. Beijing: Association for Computational Linguistics,2019:434-444.

[8] 盛煜堃,彭艳兵. 基于注意力机制 BiLSTM 的事件抽取方法[J]. 电子设计工程,2020,28(8):170-173.

[9] 孙宝贵,车文刚,廖江福. 一种改进的 KNN 案例推理检索算法[J]. 计算机工程与科学,2021,43(12):2263-2271.

[10] 姚　鑫,郭海湘,顾明赟,等. 基于案例推理的滑坡灾害应急相似案例智能生成研究[J]. 系统工程理论与实践,2021,41(6):1570-1584.

[11] 陈志奎,刘　杰,丁　锋,等. 基于案例推理的民间借贷案件适用法律推荐[J]. 计算机技术与发展,2021,31(5):198-203.

[12] 陈国清. 基于案例推理的社区纠纷调解方案检索算法[J]. 电子设计工程,2019,27(18):10-15.

[13] 蔡胜胜,凡　亮. 基于案例推理和规则推理的公安突发事件辅助决策算法[J]. 计算机与现代化,2019(9):7-11.

[14] 马　亮,郑跃平,张采薇. 政务热线大数据赋能城市治理创新:价值、现状与问题[J]. 图书情报知识,2021(2):4-12.

[15] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha: Association for Computational Linguistics,2014:1746-1751.

[16] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning phrase representations using RNN encoder - decoder for statistical machine translation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha: Association for Computational Linguistics, 2014: 1724-1734.

[17] 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. Long Beach: [s. n.], 2017:6000-6010.

[18] SUTTON C, MCCALLUM A. An introduction to conditional random fields[J]. Foundations and Trends in Machine Learning, 2010,4(4):267-373.

[19] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceedings of the 31th conference on advances in neural information processing systems. Long Beach: [s. n.],2019:5754-5764.

[20] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco: Association for Computing Machinery,2016:785-794.

(下转第 46 页)