软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(12):3829−3838 [doi: 10.13328/j.cnki.jos.006110]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

案件要素句子关联图卷积的案件舆情摘要方法*

韩鹏宇[1,2], 余正涛[1,2], 高盛祥[1,2], 黄于欣[1,2], 郭军军[1,2]

[1](昆明理工大学 信息工程与自动化学院,云南 昆明 650504)

[2](云南省人工智能重点实验室(昆明理工大学),云南 昆明 650504)

通讯作者: 余正涛, E-mail: ztyu@hotmail.com

摘 要: 案件舆情摘要是从涉及特定案件的新闻文本簇中,抽取能够概括其主题信息的几个句子作为摘要.案件舆情摘要可以看作特定领域的多文档摘要,与一般的摘要任务相比,可以通过一些贯穿于整个文本簇的案件要素来表征其主题信息.在文本簇中,由于句子与句子之间存在关联关系,案件要素与句子亦存在着不同程度的关联关系,这些关联关系对摘要句的抽取有着重要的作用.提出了基于案件要素句子关联图卷积的案件文本摘要方法,采用图的结构来对多文本簇进行建模,句子作为主节点,词和案件要素作为辅助节点来增强句子之间的关联关系,利用多种特征计算不同节点间的关联关系.然后,使用图卷积神经网络学习句子关联图,并对句子进行分类得到候选摘要句.最后,通过去重和排序得到案件舆情摘要.在收集到的案件舆情摘要数据集上进行实验,结果表明:提出的方法相比基准模型取得了更好的效果,引入要素及句子关联图对案件多文档摘要有很好的效果.

关键词: 案件舆情摘要;图卷积;案件要素;句子关联图

中图法分类号: TP18

中文引用格式: 韩鹏宇,余正涛,高盛祥,黄于欣,郭军军.案件要素句子关联图卷积的案件舆情摘要方法.软件学报,2021,32(12):3829−3838. http://www.jos.org.cn/1000-9825/6110.htm

英文引用格式: Han PY, Yu ZT, Gao SX, Huang YX, Guo JJ. Case-related public opinion summarization method based on graph convolution of sentence association graph with case elements. Ruan Jian Xue Bao/Journal of Software, 2021,32(12):3829−3838 (in Chinese). http://www.jos.org.cn/1000-9825/6110.htm

Case-related Public Opinion Summarization Method Based on Graph Convolution of Sentence Association Graph with Case Elements

HAN Peng-Yu[1,2], YU Zheng-Tao[1,2], GAO Sheng-Xiang[1,2], HUANG Yu-Xin[1,2], GUO Jun-Jun[1,2]

[1](Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China)

[2](Yunnan Key Laboratory of Artificial Intelligence (Kunming University of Science and Technology), Kunming 650504, China)

Abstract: The case-related public opinion summarization is the task of extracting a few sentences that can summarize the subject information from some case-related news documents. The case-related public opinion summarization can be regarded as a multi-document summarization in a specific field. Compared with the general multi-document summarization, the topic information can be characterized by some case elements that run through the entire text cluster. In text clusters, sentences and sentences are associated with each other, case elements also have associations of varying degree with sentences. These associations play an important role in extracting abstract sentences. A case-related public opinion summarization method based on graph convolution of sentence association graph with case elements is proposed, which uses graph structure to model all text clusters, with sentences as the main node, words and case elements as

* 基金项目: 国家重点研发计划(2018YFC0830105, 2018YFC0830101, 2018YFC0830100); 国家自然科学基金(61761026, 61972186, 61762056); 云南省自然科学基金(2018FB104)

Foundation item: National Key Research and Development Project (2018YFC0830105, 2018YFC0830101, 2018YFC0830100); National Natural Science Foundation of China (61761026, 61972186, 61762056); Natural Science Foundation of Yunnan Province (2018FB104)

收稿时间: 2020-02-10; 修改时间: 2020-04-11, 2020-05-26; 采用时间: 2020-06-26

auxiliary nodes to enhance the relationship between sentences. Multiple features are used to calculate the relationship between different nodes. Then, graph convolutional neural network is used to learn this sentence association graph, and the sentence is classified to obtain the candidate summary sentence. Finally, the sentence is deduplicated and ranked to obtain the case-related public opinion summarization. Experiments are performed on the case-related public opinion summary dataset. The results show that the method achieves better results than the benchmark model, indicating that both the composition method and the graph convolution learning method are effective.

Key words: case-related public opinion summarization; graph convolution network; case elements; sentence association graph

案件舆情是指与案件相关的互联网舆情,与一般的新闻舆情相比,案件舆情具有敏感性、特殊性,有着更大的社会影响.案件舆情摘要能够从案件相关新闻文本中摘取重要信息,从而简化新闻文本长度,帮助用户在大量的舆情数据中获取舆情事件的关键信息,对于案件舆情的监控与及时处理有着重要的作用.

案件舆情摘要本质上是一种特定领域的多文档摘要任务,在多文档摘要的研究中,关键问题是对句子的重要性进行评价,并以此抽取摘要句子.传统方法有基于统计的摘要方法[1−4]、基于主题模型的摘要方法[5−7]和基于图的摘要方法[8−11]等.基于统计的方法一般通过词频、句子位置、句子相似度等这类特征来评价句子的重要程度,然后通过一定的策略选取重要句子得到摘要,其中具有代表性的方法有基于词频-逆文档频率(TF-IDF)的统计方法[1].Hong 等人[4]提出了一种简单的多文档摘要方法,用词的概率作为输入,然后选择平均词概率较高的句子作为摘要.基于主题模型的方法一般采用狄利克雷分布(LDA)的方法得到文本簇中预设数量的主题,然后采用不同的算法计算句子和主题的相似度来得到摘要句.例如:刘娜等人[6]引入主题重要性的概念,将 LDA 建立的主题分成重要和非重要两类,并使用词频、位置等统计特征和 LDA 特征一起计算句子权重;吴仁守等人[7]提出一种方法将新闻事件划分为多个不同的子主题,在考虑时间演化的基础上同时考虑子主题之间的主题演化,最后将新闻标题作为摘要输出.还有很多研究者提出了一些基于图的方法[8−11],将文本表征成一张图,图中使用句子或其他单元作为顶点,用边连接两个有相似性或者关联关系的顶点,使用各种方法计算句子相似度或关联关系来构建边.典型的有 Mani 等人在 1997 年最早使用图模型进行多文档摘要任务的研究[8].Mihalcea 等人在基于 PageRank 算法的基础上,提出了一种基于图排序的 TextRank 模型[9].Li 等人[10]利用主题和句子之间的关系,将主题模型集成到图排序中.Yasunaga 等人[11]提出一种图卷积的多文档摘要方法,统计句子中出现的动名词组合数、位置信息等特征来进行构图,然后用图卷积的方法对句子进行分类.

基于统计的摘要方法虽然实现简单且有一定效果,但对于句子的打分一般都是比较孤立的,忽略了文本结构信息、尤其是句子与句子之间的关联关系.基于主题模型的方法一般针对没有特定主题的多文档摘要任务,不适合主题明确的案件舆情摘要.基于图的方法虽然可以较好地表征句子间的关联关系,但构图方法一般是通用方法,不涉及特定要素或关键词之间的关联关系.

以上方法无论是基于统计、主题模型和图模型的,多是通用领域的无监督多文档摘要方法.针对案件舆情这一特定领域问题,需要更好地考虑案件主题的相关信息以及跨文档句子之间的关联关系.同一案件相关的多篇新闻文本构成一个文本簇,具有与特定案件相关的主题,这一主题可以通过一些案件要素来进行表征.如表 1 所示,在“奔驰女车主维权案”中,案发地、涉案主体、案件描述:“西安、奔驰 4s 店、女车主、利之星、发动机漏油、消费者维权”等关键词就是该案件的案件要素,代表其主题信息.可以看出:这些案件要素贯穿于多篇新闻文本,共现于和案件主题相关的句子当中,并且集中出现在参考摘要中,对于句子关系的表征和摘要生成的准确性都有着重要的作用.又因为句子都是词的集合,因此在抽取句子形成摘要的过程中,需考虑异构的句子关联图特征:借鉴基于统计的方法,引入词节点来得到句子的特征表示,借助案件要素节点来加强与案件主题相关的句子间的关联关系,然后再学习这些关系来对句子的重要性进行评价.在如何对图进行学习方面,借鉴 Yao 等人提出的一种基于图卷积的文本分类方法[12]使用两层图卷积神经网络来对图中节点的特征进行学习,可以很好地学习到图中的结构信息.针对以上分析,本文探索在句子关联图中用词节点和案件要素节点强化句子间关联关系的表征,研究使用图卷积的方法预测句子的得分,然后经过去重和重排序进而得到摘要.

本文的主要贡献总结如下:

1) 提出在案件舆情领域进行多文档摘要的研究探索,创新性地引入案件要素信息来指导摘要句的抽取;

2) 提出一种基于案件要素句子关联图卷积的摘要模型,融入案件要素节点、词节点,并构造异构图来对文本簇进行建模,有效利用了文本语义特征、句子与案件要素之间的关联关系等特征;

3) 与多种多文档摘要方法进行比较评估,在收集的案件舆情摘要数据集上进行了实验,验证了本文方法的有效性.

Table 1 Case analysis of case-related public opinion

表 1 案件舆情实例分析

案件名	奔驰女车主维权
案件要素	西安,奔驰4s店,女车主,利之星,发动机漏油,消费者维权
新闻文本1	…“奔驰女车主维权”视频刷屏,事情本身的是非曲直,自然要等到检测机构的调查结论出炉才能有定论,但这个事情本身之所以能够引入检测机制,是与奔驰车主的维权方式分不开的…(部分)
新闻文本2	漏油了,良心不能“漏”!多家媒体关注奔驰女车主维权.这两天,一则女车主在4S店内哭诉的视频在网上广泛流传.视频中,在西安利之星奔驰4S店,一辆价值60多万元的CLS奔驰轿车,还未开出4S店,就发生了发动机漏油的事情.事情经过几天的发酵,很多网友为车主发声,支持车主的维权行为…(部分)
共现要素	奔驰、4s店、漏油、女车主、维权
参考摘要	本周“西安奔驰女车主坐在引擎盖上哭诉维权”事件继续在网络上发酵,引发各方的广泛关注…在西安利之星奔驰4S店,购买了一辆价值60多万元的CLS奔驰轿车,但新车还未开出4S店,就发现车辆发动机存在漏油问题.由此,女主展开了一场艰难的消费者维权拉锯战…(部分)

注:新闻文本 1:报道时间:2019 年 4 月 16 日;报道出处:https://www.sohu.com/a/308317749_428290

新闻文本 2:报道时间:2019 年 4 月 14 日;报道出处:https://www.sohu.com/a/307849655_428290

1 模型结构

本文提出一种基于图卷积的案件舆情摘要方法,融合句子、词和案件要素共同构建跨文档的句子关联图,再用图卷积的方法得到每个句子的重要性得分,经过去重和重排序得到文本摘要.模型部分参考了 Yao 等人 2019 年在文本分类领域有关图卷积的相关工作[12],将其应用于多文档摘要领域,并进行了改进,具体结构如图 1 所示(图中展示了一个案件对应的文本簇的核心处理过程,圆角矩形节点表示句子,矩形节点表示词,菱形节点表示案件要素,圆形节点表示句子的分类),其中,$S_{1.2}$ 表示第 1 个文本中的第 2 个句子,W 表示词,C 表示案件要素.

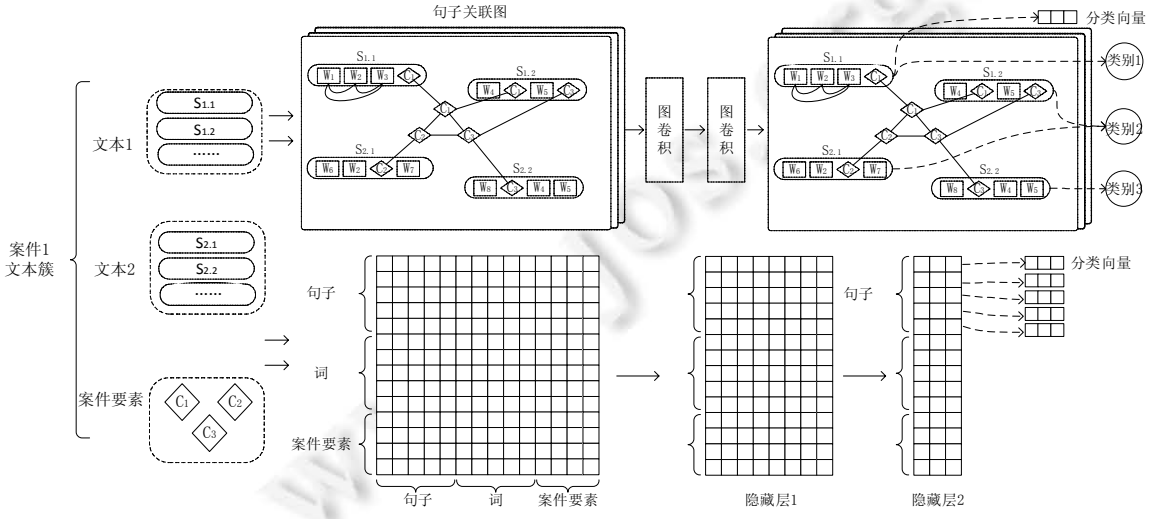

Fig.1 Case-related public opinion summarization method based on graph convolution of sentence association graph with case elements

图 1 基于案件要素句子关联图卷积的案件舆情摘要模型

模型包括 3 个主要部分,分别是融合案件要素的句子关联图模块、基于图卷积的句子分类模块、摘要生成

模块.下面分别对以上部分进行介绍.

2 融合案件要素的句子关联图构建方法

2.1 案件要素

案件舆情文本摘要可以看作特定领域的摘要问题,同一案件相关的多篇新闻构成一个文本簇,这些文本具有相同的案件相关信息.通过对案件本身和新闻舆情的特点进行分析,定义了一些案件要素来表征案件的主题信息,包括“案件名、案发地、涉案人员、案件描述”这 4 个要素作为案件要素,具体实例见表 2.

Table 2 Case elements

表 2 案件要素表

要素名	关键词
案件名	南京摩托车飙车案
案发地	江苏、南京、高速公路
涉案主体	史某、史学伟
案件描述	飙车、危险驾驶、时速299

注:报道时间:2018 年 3 月 3 日;报道出处:https://www.thepaper.cn/newsDetail_forward_2015942

表 2 中以南京摩托车飙车案为例,“案发地”包括案发的城市地区和案发的具体场所,例如“江苏、南京、高速公路”等.“涉案主体”不仅仅局限于受害人、嫌疑人和其代称,还包括关键证人,相关家属等所有与案件相关人员.“案件描述”是指发生的是什么事情以及一些其他案件关键词,例如“飙车、危险驾驶”等.通过对每一个案件构建一组案件要素,来表征案件相关信息.共构建了 50 组案件要素.

2.2 关联图构建方法

本节引入词节点来得到句子的特征表示、句子间的关联关系,借助案件要素节点来加强与案件主题相关的句子间的关联关系.使用词频-逆文档频率(TF-IDF)、互信息(PMI)、同属关系、包含关系等方法来计算边的权重,构建了一个包含句子、词和案件要素这 3 种节点的句子关联图:

$$G=(V,E) \tag{1}$$

$$V=\{S,W,C\} \tag{1}$$

其中,集合 V 表示图中节点的集合,由 3 部分构成:句子集合 S、词集合 W 和案件要素集合 C.

- 句子集合 $s=\{s_1,s_2,\ldots,s_l\}$里共有 l 个句子,是不同文本簇的所有文档经过去除特殊字符、分句、去除短句子等预处理之后的句子总和.其中,s_2 表示第 2 个句子,l 表示句子集合的大小;
- 词集合 $w=\{w_1,w_2,\ldots,w_m\}$是对所有文本簇使用 jieba 分词工具进行分词以及去停用词等操作后得到的词表,其中,m 表示词表大小;
- 案件要素集合 $c=\{c_1,c_2,\ldots,c_n\}$共有 n 个案件要素,包括所有不同案件的案件要素,其中,c_2 表示第 2 个案件要素.E 表示图中边的集合:$E=\{(v_i,v_j)|v\in V\}$,其中,v_i 表征图中第 i 个节点.

因为图中有 3 种节点,所以图的邻接矩阵 $\boldsymbol{A}$ 由 9 个分块矩阵构成,见公式(3).其中,$\boldsymbol{A}_{SS}$ 表示句子和句子节点的关系矩阵,$\boldsymbol{A}_{SW}$表示句子和词节点的关系矩阵, $\boldsymbol{A}_{SC}^{T}$ 表示句子和案件要素关系矩阵的转置:

$$\boldsymbol{A}=\begin{vmatrix}\boldsymbol{A}_{SS} & \boldsymbol{A}_{SW} & \boldsymbol{A}_{SC}\\ \boldsymbol{A}_{SW}^{T} & \boldsymbol{A}_{WW} & \boldsymbol{A}_{WC}\\ \boldsymbol{A}_{SC}^{T} & \boldsymbol{A}_{WC}^{T} & \boldsymbol{A}_{CC}\end{vmatrix} \tag{3}$$

共有 6 种边,每种边的定义和计算见公式(4):

$$A_{ij}=\begin{cases}\{1,0\}, & v_i\in S,v_j\in S:v_i \text{ 和}v_j\text{属于同一文档时}\boldsymbol{A}_{ij}=1\text{;否则,}\boldsymbol{A}_{ij}=0\\ TF\text{-}IDF, & v_i\in S,v_j\in W\\ \{1,0\}, & v_i\in S,v_j\in C:v_j\text{在句子}v_i\text{中出现时}\boldsymbol{A}_{ij}=1\text{;否则,}\boldsymbol{A}_{ij}=0\\ PMI, & v_i\in W,v_j\in W\\ \{1,0\}, & v_i\in W,v_j\in C:v_i=v_j\text{时}\boldsymbol{A}_{ij}=1\text{;否则,}\boldsymbol{A}_{ij}=0\\ \{1,0\}, & v_i\in C,v_j\in C:v_i\text{和}v_j\text{属于同一案件时}\boldsymbol{A}_{ij}=1\text{;否则,}\boldsymbol{A}_{ij}=0\end{cases} \tag{4}$$

其中,$\boldsymbol{A}_{ij}$表示第 i 和第 j 两个节点之间边的权值.这 6 种关系的具体计算方法是:

(1) 对于句子与句子关系矩阵 $\boldsymbol{A}_{SS}$,使用同属关系来计算:当一个句子和另一个句子同属于一个文本时,在它们之间连接一条边;

(2) 对于句子与词关系矩阵 $\boldsymbol{A}_{SW}$:使用词频-逆文档频率(TF-IDF)的方法来计算词节点 w_j 和句子节点 s_i 之间边的权重,见公式(5):

$$TF\text{-}IDF(s_i,w_j)=TF(s_i,w_j)*IDF(s_i) \tag{5}$$

其中,s_i 表示第 i 个句子节点,w_j 表示第 j 个词节点,TF 表示词在句子中的词频,IDF 表示词在所有文本中出现的频率.当一个像“的”这样的高频词在所有文本中出现的频率越多,其 IDF 值就越低.通过在句子和大量词之间构建关联关系,可以用词来表征句子的特征,同时也在所有句子之间构建了一层关联关系;

(3) 对于句子与案件要素关系矩阵 $\boldsymbol{A}_{SC}$,使用包含关系来计算:当一个案件要素出现在某个句子中时,在它们之间连接一条边;

(4) 对于词与词关系矩阵 $\boldsymbol{A}_{WW}$:使用互信息(PMI)来计算两个词节点之间边的权重,见公式(6):

$$PMI(w_i,w_j)=\log_2\frac{p(w_i,w_j)}{p(w_i)p(w_j)} \tag{6}$$

其中,w_i 和 w_j 表示第 i 和第 j 个词节点,两个词的相关性越大,其 PMI 值也就越大.当 PMI 的值小于 0 时,表示两个词相关性为负,也就是互斥的,此时,两个词之间边权重为 0;

(5) 对于词与案件要素关系矩阵 $\boldsymbol{A}_{WC}$:案件要素会出现和某一个词相同的情况,当案件要素和某一个词恰好相同时,在它们之间连接一条权重为 1 的边;

(6) 对于案件要素与案件要素关系矩阵 $\boldsymbol{A}_{CC}$,使用同属关系来计算:当一个案件要素和另一个案件要素同属一个案件时,在它们之间连接一条边.

通过以上方法,可以构建一个融合案件要素的句子关联图.下一步,在此基础上使用图卷积的方法得到每个句子的重要性评价.

3 图卷积层

图卷积网络(GCN)是一种在图上学习的神经网络,可以直接处理图,并利用图的结构信息.图卷积网络具有强大的学习能力,研究表明:两层的 GCN 即可以得到很好的学习效果,过多的层数可能导致节点之间更加趋同.因此,在本文实验中也采用两层的 GCN.

在第 2.2 节构造的句子关联图 G 中,节点总数 $size=l+m+n$.因为每一个节点在进行图卷积的时候,既要包含周围节点的特征,又要包含自身的特征,所以每个节点还应该有一个连接到其自身的闭环,还需要将邻接矩阵 $\boldsymbol{A}$ 对角线上元素初始化为 1,即 $\boldsymbol{A}_{ij}=1$,最后构成一个大小为 $size\times size$ 的图的邻接矩阵 $\boldsymbol{A}$:

$$\boldsymbol{A}_{ij}\in\mathbb{R}^{size\times size} \tag{7}$$

令图的度矩阵为 $\boldsymbol{D}$,表示每一个节点和多少个其他节点相连,其中,度矩阵对角线上元素为

$$\boldsymbol{D}_{ij}=\sum\nolimits_j\boldsymbol{A}_{ij} \tag{8}$$

根据公式(7)和公式(8),可以得到可以进行图卷积操作的规范化的矩阵 $\tilde{\boldsymbol{A}}$:

$$\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{9}$$

将节点的特征矩阵 $\boldsymbol{X}$ 初始化为一个和邻接矩阵 $\boldsymbol{A}$ 一样大小的单位矩阵,相当于使用 *one-hot* 向量表示节点的特征.

在第 1 层图卷积网络中:

$$\boldsymbol{L}^{(1)} = ReLU(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_1) \tag{10}$$

其中,$\boldsymbol{L}^{(1)}$表示第 1 层的输出,$\tilde{\boldsymbol{A}}$ 是规范化的邻接矩阵,$\boldsymbol{X}$ 是特征矩阵,$\boldsymbol{W}_1$ 是参数矩阵,激活函数使用 *ReLU*.在第 2 层图卷积网络中,使用 *softmax* 进行分类,如公式(11)所示:

$$\boldsymbol{L}^{(2)} = softmax(\tilde{\boldsymbol{A}}\boldsymbol{L}^{(1)}\boldsymbol{W}_2) \tag{11}$$

采用交叉熵作为模型的损失函数:

$$loss = -\sum_{i\in S} y_i \ln \boldsymbol{L}_i^{(2)} \tag{12}$$

其中,s 是训练集中参与计算损失的所有句子,y_i 表示第 i 个句子的标签,$\boldsymbol{L}_i^{(2)}$ 表示第 i 个句子的预测结果.通过两次图卷积操作后,可以得到每一个句子节点的分类结果,表示每一个句子的得分.

4 摘要生成

摘要句既要反映文档的中心思想,又要具有低冗余性和一定的时序关系.通过前面的方法得到每一个句子评分之后,需要从中选取得分最高的几个句子,对其进行去重和排序,具体流程如下所示.

(1) 对于测试集中不同的文本簇,分别进行摘要生成;

(2) 对于一个文本簇,首先选取一个得分最高的句子加入候选摘要句集合中;

(3) 然后选取下一个句子和候选摘要句集合中的每一个句子计算相似性,其值若小于相似性阈值,则将该句子加入候选摘要句集合中;

(4) 重复第(2)步的操作,直到候选摘要句集合长度超过摘要预期长度;

(5) 最后再对候选摘要句集合中的句子按照文档的爬取顺序(代表文章发表的时序)以及句子在文档中出现的顺序排序,得到最终的多文档摘要.

5 实 验

5.1 数据集

本文针对 50 个案件,构造 50 组案件要素,使用爬虫程序从互联网上搜集相关新闻,对数据清洗去噪,得到 50 个文本簇.每个文本簇包含 10 篇文档.对每个文本簇人工撰写摘要,最终构建出案件舆情摘要数据集.见表 3.

Table 3 Dataset

表 3 数据集

	案件数	文档数	句子数	句子平均长度	摘要长度
训练集	30	300	7 292	49.5	186.4
验证集	10	100	3 014	50.3	209.5
测试集	10	100	2 827	44.3	185.1

5.2 评价标准

本文采用自动摘要任务中常用的一种评价指标 ROUGE 来作为介绍评价指标.ROUGE 是基于摘要中 n 元语法(n-gram)的共现信息来评价摘要的一种方法,包括 ROUGE-1,ROUGE-2 等.ROUGE-L 和 ROUGE-N 相似,是一种基于最长公共子序列的评价方法.ROUGE 值越高,说明摘要效果越好.例如,ROUGE-N 的一般计算方法见公式(13):

$$ROUGE\text{-}N=\frac{\sum_{s\in(Referemce\ Summaries)}\sum_{gram_n\in s}Count_{match}(gram_n)}{\sum_{s\in(Reference\ Summaries)}\sum_{gram_n\in s}Count(gram_n)} \tag{13}$$

其中,分子表示模型输出的摘要和参考摘要中同共现的 *n*-gram 的个数,分母则表示参考摘要中的 *n*-gram 个数.

5.3 实验设置

实验采用 2 层图卷积网络,特征矩阵每一行使用 *one-hot* 向量来初始化,第 1 层输出的节点特征向量为 200 维,第 2 层输出的节点分类向量为 10 维.*Dropout* 设置为 0.5,学习率设置为 0.02,训练轮次设置为 400,提前截至的容忍度设置为 12,摘要预期长度设置为 200.

本文共设置了 3 组对比实验和 1 个实例分析.

- 第 1 组对比实验对比了本文模型和 10 个基准模型的性能,其中包括未融入案件要素的消融实验:“句子+词+GCN”;
- 第 2 组对比实验研究了不同句子分类数目对生成摘要质量的影响,设置 2,5,10 和 20 等 4 种不同的分类数目,使用本文模型分别进行实验;
- 第 3 组实验研究了去冗余步骤中,不同相似度计算方法对摘要的影响,其中,rouge 方法阈值设置为 0.8、jaccard 方法阈值设置为 0.8、tf-idf 方法阈值设置为 0.8 和 word2vec 方法阈值设置为 0.9;
- 实例分析选取了针对“快递员遭投诉自杀”案件的摘要实例进行对比分析.

5.4 基准模型

本文共选择了 10 个基准模型,分别在案件舆情摘要数据集上进行实验,得到 ROUGE-1,ROUGE-2 和 ROUGE-L 这 3 种评分.基准模型包括 LEAD,Centroid,LexPageRank,TextRank,Submodular,ClusterCMRW,Query+MR,LDA,Manifold-Ranking 和“句子+词+GCN”,其中,部分代码由开源工具包 PKUSUMSUM 提供.

- LEAD 是一种依靠句子在文章中的位置来抽取摘要的方法,研究表明,文章中的重要信息很大概率会出现在文章开头部分;
- Manifold-Ranking[13]是一种类似 PageRank 的方法,利用流行排序进行多文档摘要;
- Query+MR 在 Manifold-ranking 模型的基础上增加了一个案件要素集合作为查询句,来对句子节点之间的权重进行调整,然后得到摘要;
- LDA 方法通过使用 LDA 对文本簇进行主题聚类,然后寻找含有主题信息最多的句子作为摘要;
- Centroid[14]是一种基于质心的多文档摘要方法,通过寻找中心词最多的句子来得到摘要;
- ClusterCMRW[15]是一种基于马尔科夫链和随机游走的多文档摘要方法,利用文档集中句子之间的链接关系来生成摘要;
- Submodular[16]利用次模函数的单调递减特性来抽取句子作为摘要;
- LexPageRank[17]和 TextRank[9]都是一种基于图的关键词提取算法,将句子视为节点,通过计算图中每个节点的得分,来选择得分最高的几个句子作为摘要;
- “句子+词+GCN”表示未融入案件要素的图卷积神经网络方法.

5.5 实验结果分析

第 1 组实验为了验证本文模型的有效性,与 10 个基准模型进行了对比实验,其中,和“句子+词+GCN”对比以验证融入案件要素的有效性.选取 ROUGE-1,ROUGE-2 和 ROUGE-L 这 3 种评分,实验结果见表 4.

根据表 4 的实验结果可以看出:

1) 在采用 ROUGE-1 的评价方法中,本文模型和其他基准模型相比,有 0.43~6.07 的提升,说明了本文模型的优越性;
2) 对比 TextRank,LexPageRank 和本文模型,虽然同为基于图的方法,但是图卷积比这两种方法具有显著的效果提升,充分说明了图卷积方法在多文档摘要任务上的优越性;

3) 对比“Manifold-Ranking”和“Query+MR”的结果可以看出,引入案件要素作为查询条件来指导摘要生成是有作用的;

4) 对比“句子+词+GCN”和本文模型的 ROUGE-1 和 ROUGE-2,本文模型分别提升了 3.37 和 2.92,说明在案件舆情领域,融合案件要素构建句子关联图的方法是有效的,能够很好地表征跨文档句子之间的关联关系,对于指导抽取出更贴近多文档主题的摘要句有着重要作用.

Table 4 Comparison of experimental results between our model and the baselines

表 4 本文模型与基准模型实验对比结果

模型	ROUGE-1	ROUGE-2	ROUGE-L
Centroid	30.50	8.66	18.34
LDA	31.29	12.33	19.55
Submodular	31.39	10.88	20.19
TextRank	31.40	8.11	15.95
Manifold-Ranking	31.72	5.54	13.45
LexPageRank	32.71	9.71	18.83
句子+词+GCN	33.20	14.10	22.86
Query+MR	34.88	13.72	21.44
ClusterCMRW	35.49	11.65	19.11
LEAD	36.14	12.22	25.06
本文模型	36.57	17.02	26.31

第 2 组实验研究了使用图卷积进行句子分类时,句子的不同分类数目对于摘要质量的影响.设置 2,5,10 和 20 等 4 种不同的句子分类数目,选取 ROUGE-1,ROUGE-2 和 ROUGE-L 作为评价指标,实验结果见表 5.

Table 5 Comparison experiments of different classification numbers

表 5 不同分类数目对比实验

分类数目	ROUGE-1	ROUGE-2	ROUGE-L
2	34.82	15.63	24.49
5	35.49	16.49	23.37
10	36.57	17.02	26.31
20	31.24	14.35	22.64

根据表 5 的实验结果可以看出:在句子分类数目为 10 的时候取得的摘要效果最好,分类数目较低会略微降低摘要质量,分类数目过高会严重降低摘要的质量.分析可能是因为分类数目的不同导致了句子分类准确率的不同.

第 3 组实验研究了不同相似度计算方法对摘要性能的影响,分别使用 rouge(0.8),jaccard(0.8),tf-idf(0.8)和 word2vec(0.9)等 4 种.其中,基于 word2vec 使用词向量+average pooling 来表示句子信息.选取 ROUGE-1, ROUGE-2 和 ROUGE-L 作为评价指标,实验结果见表 6.

Table 6 Comparison experiments of different similar computing methods

表 6 不同相似度计算方法对比实验

方法	ROUGE-1	ROUGE-2	ROUGE-L
rouge	35.69	15.23	22.47
jaccard	35.69	15.23	22.47
tf-idf	35.69	15.23	22.47
word2vec	36.57	17.02	26.31

根据表 6 的实验结果可以看出:前 3 种相似度计算方法得到的结果一致.可能的原因是:在本实验中,得分较高的几个句子之间的差异性是比较大的,这 3 种方法对句子差异性的敏感程度是相似的.Word2vec 的方法效果略好一点.

如表 7 的实例分析中,从测试集中选取了“快递员遭投诉自杀”案件,针对该案件的部分基准模型生成的摘要进行实例分析.

根据表 7 可以看出:

1) 对比 TextRank 和本文模型,本文结果在事件表述的完整性上有着较好的效果;
2) 对比 Centroid 模型结果,本文模型摘要更加贴近文本簇的中心思想;
3) 对比“句子+词+GCN”的结果可以看出本文模型在连贯性和可读性上有一定的优势.

Table 7 Example of summary comparison of “courier suicide”

表 7 “快递员遭投诉自杀”案摘要对比实例

对比模型	生成结果
案件名	快递员遭投诉自杀
参考摘要	又是一件和快递员有关的舆情事件,江苏常州的一名快递员杨军,在工作中遭到顾客的一次“恶意投诉”,杨军被公司处罚与书面检讨,激动的杨军说到,宁可不要工作,也要尊严,选择了吞安眠药自杀,幸运的是及时被人发现,送医院,暂时没有生命危险.虽然快递是一个服务行业,但发生这样的恶意投诉的事件还是引发了舆论的广泛讨论:快递公司是否应该建立健全相关机制,弥补管理漏洞,来保障快递员的权利
案件要素	江苏 常州 快递员 杨军 自杀 恶意投诉
TextRank	(详见《极昼》6 月 15 日报道《顺丰快递员遭投诉后自杀未遂:已接到公司道歉 以后不做快递员》)这些天,杨军的电话响个不停,全国各地的顺丰快递员在内部系统给他留言,表示支持,他也收获了同样多的质疑,有人说他性格偏激,“动不动就要死要活的”.因此在他看来,那名顺丰快递员因“恶意投诉”便试图轻生的做法,的确有些不值了,“真的为那种事儿犯不上.“不知道是心理作用还是怎么的,这几天投诉和处罚都少了
Centroid	中国网财经记者调查发现,在华丽光鲜的“快递第一股”外表下,近年来顺丰控股正遭遇公司创立以来少有的艰难时刻:服务水平下降、频繁遭遇投诉、公司利润下滑,而各种在 PPT 阶段显得创意十足的创新业务,在推出后却往往因执行力不足而呈现“雷声大雨点小”的境况,各种叠加的压力,使得这家曾以“服务好安全性高”而著称的公司有些心浮气躁.他给同在这个区域送快递的弟弟杨健帮忙.他已经离职了
GCN	“那一天,刚好在帮一名顾客取件,怠慢了另一位寄件的顾客,结果被当场投诉.”6 月 13 日消息,据报道,江苏常州一名顺丰快递员杨军,因顾客恶意投诉,遭到公司处罚.杨军在派送一个由河北廊坊发往本地的沙发四件套时,“由于运单上收件人电话号码少了一位数,我联系不上,于是我就按照运单上发件人的电话打过去询问.”“在尊严和工作面前,我可以不要工作”
本文模型	6 月 9 日,**杨军**“以死护尊严”,随后在家中吞下 40 粒安眠药,所幸的是,邻居和弟弟等及时赶到,将其送往医院抢救.6 月 13 日消息,据报道,江苏**常州**一名顺丰**快递员杨军**,因顾客**恶意投诉**,遭到公司处罚.同时,公司还让**杨军**写 500 字的书面检讨,**杨军**写道:“在尊严和工作面前,我可以不要工作.”目前,杨军脱离生命危险,已经出院.**杨军**是顺丰**快递**负责大件派送的员工,端午节期间,由于同事们放假,需要一个人负责平时 3 个人负责的配送区域

注:报道时间:2019 年 6 月 15 日;报道出处:https://www.sohu.com/a/320708729_120146415

6 结束语

针对案件舆情摘要任务,本文提出一种融合案件要素关联和句子关联的构图方法,有效地通过案件要素融入了案件主题信息,很好地表征了跨文档的句子关联关系.使用图卷积的方法充分学习到了图中的结构信息,抽取的摘要句和基准模型相比取得了一定的效果提升.

在下一步的工作中,拟更多地去探索上下文关系、语义关系、篇章结构关系和逻辑关系等其他关系对摘要生成的作用.

References:

[1] Seki Y. Sentence extraction by tf/idf and position weighting from newspaper articles. In: Proc. of the 3rd NTCIR Workshop on Research in Information Retrieval. Automatic Text Summarization and Question Answering. 2002. 55−59.

[2] Timothy DR, Allison T, Blair-goldensohn S, *et al*. MEAD a platform for multidocument multilingual text summarization. In: Proc. of the Int’l Conf. on Language Resources and Evaluation. 2004.

[3] Nenkova A, Vanderwende L, McKeown K. A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In: Proc. of the 29th Annual Int’l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2006. 573−580.

[4] Hong K, Conroy JM, Favre B, *et al*. A repository of state of the art and competitive baseline summaries for generic news summarization. In: Proc. of the LREC. 2014. 1608−1616.

[5] Na L, Peng X, Ying L, *et al*. A topic approach to sentence ordering for multi-document summarization. In: Proc. of the 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2016. 1390−1395.

[6] Liu N, Lu Y, Tang XJ, *et al*. Multi-document summarization algorithm based on significance topic of LDA. Journal of Frontier of Computer Science and Technology, 2015,9(2):242−248 (in Chinese with English abstract).

[7] Wu RS, Liu K, Wang HL. An evolutionary summarization system based on local-global topic relationship. Journal of Chinese Information Processing, 2018,32(9):75−83 (in Chinese with English abstract).

[8] Mani I, Bloedorn E. Multi-document summarization by graph search and matching. In: Proc. of the 14th National Conf. on Artificial Intelligence and 9th Conf. on Innovative Applications of Artificial Intelligence. 1997. 622−628.

[9] Mihalcea R, Tarau P. Textrank: Bringing order into text. In: Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing. Barcelona: ACL Press, 2004. 401−411.

[10] Li C, Wei Z, Liu Y, *et al*. Using relevant public posts to enhance news article summarization. In: Proc. of the 26th Int'l Conf. on Computational Linguistics: Technical Papers (COLING 2016). 2016. 557−566.

[11] Yasunaga M, Zhang R, Meelu K, *et al*. Graph-based neural multi-document summarization. In: Proc. of the 21st Conf. on Computational Natural Language Learning. Vancouver, 2017. 452−462.

[12] Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 7370−7377.

[13] Wan X, Yang J, Xiao J. Manifold-ranking based topic-focused multi-document summarization. In: Proc. of the Int'l Joint Conf. on IJCAI. Morgan Kaufmann Publishers Inc., 2007. 2903−2908.

[14] Radev DR, Jing H, Styś M, *et al*. Centroid-based summarization of multiple documents. Information Processing & Management, 2004,40(6):919−938.

[15] Wan X, Yang J. Multi-document summarization using cluster-based link analysis. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2008. 299−306.

[16] Lin H, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions. In: Proc. of the Human Language Technologies: 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics. Los Angeles: NAACL, ACL Press, 2010. 912−920.

[17] Erkan G, Radev DR. Lexpagerank: Prestige in multi-document text summarization. In: Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing. Barcelona: ACL Press, 2004. 365−371.

附中文参考文献:

[6] 刘娜,路莹,唐晓君,等.基于 LDA 重要主题的多文档自动摘要算法.计算机科学与探索,2015,9(2):242−248.

[7] 吴仁守,刘凯,王红玲.一种基于局部—全局主题关系的演化式摘要系统.中文信息学报,2018,32(9):75−83.

韩鹏宇(1995－),男,硕士,主要研究领域为自然语言处理,文本摘要.

黄于欣(1983－),男,博士,CCF 学生会员,主要研究领域为自然语言处理,文本摘要,文本生成.

余正涛(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,信息检索.

郭军军(1987－),男,博士,讲师,CCF 专业会员,主要研究领域为自然语言处理,信息检索,机器翻译.

高盛祥(1977－),女,博士,副教授,CCF 专业会员,主要研究领域为自然语言处理,机器翻译,信息检索.