文章编号：1003-0077(2020)05-0056-08

基于案件要素指导的涉案舆情新闻文本摘要方法

韩鹏宇[1,2]，高盛祥[1,2]，余正涛[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 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：涉案舆情新闻文本摘要任务是从涉及特定案件的舆情新闻文本中，获取重要信息作为其简短摘要，因此对于相关人员快速掌控舆情态势具有重要作用。涉案舆情新闻文本摘要相比开放域文本摘要任务，通常涉及特定的案件要素，这些要素对摘要生成过程有重要的指导作用。因此，该文结合深度学习框架，提出了一种融入案件要素的涉案舆情新闻文本摘要方法。首先构建涉案舆情新闻摘要数据集并定义相关案件要素，然后通过注意力机制将案件要素信息融入新闻文本的词、句子双层编码过程中，生成带有案件要素信息的新闻文本表征，最后利用多特征分类层对句子进行分类。为了验证算法有效性，在构造的涉案舆情新闻摘要数据集上进行实验。实验结果表明，该方法相比基准模型取得了更好的效果，具有有效性和先进性。

关键词：涉案舆情摘要；案件要素；双层编码；多特征分类

中图分类号：TP391　　**文献标识码**：A

Case-involved Public Opinion News Summarization with Case Elements Guidance

HAN Pengyu[1,2], GAO Shengxiang[1,2], YU Zhengtao[1,2], HUANG Yuxin[1,2], GUO Junjun[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, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: The summary task of the public opinion news on a judical case is to obtain important information on public comments on the case in a short summary. Compared with the task of text summarization in open domain, this kind of summary usually involves specific case elements that are of great guiding effect in the process of summary generation. Therefore, a case-related news text summarization method is proposed based on deep learning framework. First, a dataset of the public opinion news summary is collected, and the case elements are defined. Then, through the attention mechanism, the case element information is integrated into the double-layer coding process of words and sentences in the news text to generate the news text representation that contains the case element information. Finally, the multi-feature classification layer is used to classify the sentences. Experiments are conducted on the public opinion news summary datasetand show that the proposed method has better performance than the base model.

Keywords: summary of grievances involving cases; case elements; two-layer encoding; multi-feature classification

0　引言

涉案舆情是指与司法案件相关的互联网舆情。与一般的新闻舆情相比，涉案舆情具有敏感性、特殊性，极易诱发群体性事件，产生严重的社会不良影响。因此快速准确地获取该类舆情信息，掌控其发展动向，对于防范和化解舆情风险至关重要。文本摘要任务，通过提取能够概括文本语义的核心句，可缩减新闻文本长度，帮助用户在大量的舆情数据中

收稿日期：2019-09-16　**定稿日期**：2019-10-30

基金项目：国家重点研发计划(2018YFC0830105，2018YFC0830101，2018YFC0830100)；云南省高新技术产业专项(201606)

获取舆情事件关键信息。

涉案舆情新闻文本摘要可以看作是特定领域的文本摘要任务。开放域的文本摘要任务通常采用抽取式或生成式两类方法。前者按照一定的选择机制，从原文中寻找与中心思想最接近的一条或多条句子作为摘要，具有流畅度高、可读性强的优点。而后者则要求计算机模拟人类阅读原文的过程，在理解全文的基础上，逐词生成新的摘要。随着深度学习研究的不断发展，基于深度学习框架的摘要生成方法已成为当下研究的热点，本文工作主要聚焦于基于神经网络的抽取式文本摘要方法。

目前基于深度学习的抽取式摘要方法通过神经网络来自动提取文本特征，然后对句子进行分类或者排序[1-3]。这些方法通常关注开放域文本摘要任务，不能很好地利用领域知识来指导摘要的生成过程。如表 1 所示，在涉案舆情新闻文本摘要中，因为舆情新闻文本通常与特定案件相关，所以在新闻文本摘要中包含有案件名、案发地、涉案人员、案件描述等案件要素。从表 1 可以看出，“榆林、产妇、坠楼、护士、家属”等案件要素关键词，都在摘要中出现，因此在新闻文本编码过程中融入案件要素，对于涉案舆情新闻文本摘要具有重要的指导作用。

表 1　涉案舆情新闻事例

正文：榆林产妇坠楼瞬间监控画面曝光。护士没拉住？马某某丈夫延先生称警方介绍该案调查进展时，提及马某某跳下去时，曾有一位护士试图拉马某某，但是没有拉住。坠楼细节揭露长达 21 分钟！榆林产妇坠楼最新消息：陕西通报产妇坠楼事件：产妇之死与医院诊疗无关！但是在最新曝光的产妇坠楼瞬间监控画面中，为什么从产妇坠楼到施救竟达 21 分钟？家属指责医院施救不及时……（部分）

案件要素：

案件名：榆林产妇坠楼事件

案发地：**榆林**、医院

涉案人员：**产妇**、**护士**、**家属**

案件描述：**坠楼**

关键词：马某某、**榆林**、**瞬间**、**产妇**、**监控**

摘要：最新情况**榆林产妇坠楼瞬间监控**画面曝光。**护士**没拉住？**坠楼细节**揭露长达 21 分钟！**家属**指责施救不及时。

针对涉案舆情新闻文本摘要任务，本文提出融入案件要素和多分类特征的文本摘要模型。首先采用基于神经网络的词、句子双层编码来提取文本特征，并将案件要素进行编码，通过注意力机制融入词、句子双层编码过程中，获得案件要素指导的文本语义表征。最后通过文本信息、当前句子位置、当前句子显著性和新颖性等多特征来对句子进行分类，获得摘要句。

综上所述：本文创新点包括以下几个方面：

（1）提出采用融入案件要素和多特征分类的文本摘要方法。据我们所知，本文是首次提出涉案舆情新闻文本摘要任务并在该方向上进行探索。

（2）在收集的涉案舆情新闻摘要数据集上进行了实验，实验结果表明我们的方法相比基准模型取得了更好的性能。

1　相关工作

抽取式摘要方法一般可以分为两大类[4]，无监督的和有监督的。早期无监督方法主要有基于特征的和基于图排序等思想的方法。其中基于特征的方法主要考虑使用词频、句子位置[5]、关键词相似度[6]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是基于词频—逆文档频率（TF-IDF）指数的统计方法[7]。基于图的模型可以有效地表示文档结构，因此在自动摘要中也得到广泛的应用。Mihalcea R 等人在基于 PageRank 算法[8]的基础上，提出了一种基于图排序的 TextRank 模型[9]，将文档中的每一句话看成是一个节点、句子之间的相似度作为边，以此来构建句子关联图，再根据 PageRank 求解每个句子的重要程度，然后选择得分较高的句子组成摘要。在这类思想基础上的改进有很多，例如，Sankarasubramaniam 等人利用维基百科等外部知识，结合图模型来处理文本摘要任务[10]。

随着深度学习的兴起，基于神经网络的方法在抽取式摘要任务上也得到了很多应用。对于文档中每一句话的分类可以建模为一个句子分类模型，其核心思想是：为原文中的每一个句子分配一个二分类标签：1 表示该句子是摘要句，0 表示该句子不是摘要句。最后选择标签为 1 的句子组成文档摘要。文献[11]用 CNN 对句子进行压缩，变成稠密向量，然后将各个句子送入一个 LSTM，再利用基于 Attention 的 LSTM 对每一句话进行分类。Nallapati 提出 SummaRuNNer 的文本分类模型[12]，并达到了当时的最佳性能。SummaRuNNer 采用词和句子两层编码器结构，每一层都采用双向 RNN 表示，最终可以得到每一个句子是否为摘要句的判断结果。在融入主题信息或事件要素方面的研究中，Wang L 提出一种主题信息融合的生成式摘要

方法[13]，通过 LDA 获取主题信息，融入摘要生成中，增加了对主题信息的关注程度。王振超等提出了一种以事件作为基本语义单元的生成式摘要方法[14]，通过对事件聚类反映篇章的主题分布，然后指导多语句压缩生成摘要。

研究表明，在抽取摘要的时候，融入主题信息[13]或者事件信息[14]等外部特征能够很好地提升摘要的效果。但这些方法大都是针对开放域的文本生成摘要，而针对涉案舆情新闻文本摘要任务的研究还是比较少的。因此，如何有效利用领域知识来指导摘要的生成，成为了本文研究的重点。

2 基于案件要素指导的抽取式摘要方法

本文提出一种基于融合案件要素和多分类特征的文本摘要方法。其核心是充分利用案件要素等领域知识和文档、句子关系、句子位置等多分类特征来对句子进行分类，得到摘要句。本文模型参考了 IBM 团队的 SummaRuNNer[12]，并在其基础上进行改进，具体结构如图 1 所示。

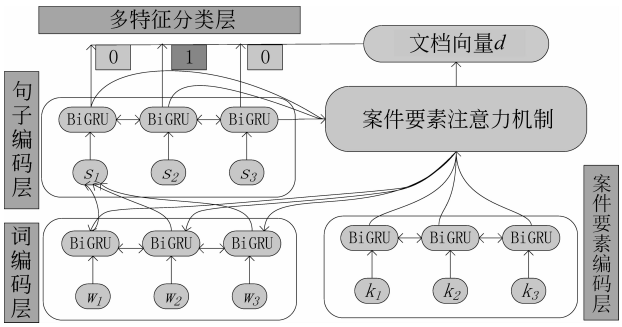

图 1 基于案件要素指导的抽取式摘要模型

模型包括 5 个主要部分，分别是案件要素编码层、案件要素注意力机制、词编码层、句子编码层和多特征分类层。本节分别对上述部分进行介绍。

2.1 案件要素

涉案舆情新闻文本与普通新闻文本不同的是，前者一般都有与特定案件相关的主题，可以选择使用一些案件要素作为关键词来表征这一主题信息。通过对中国裁判文书网中案件要素构成的分析，同时综合考虑新闻文本的舆情特点，我们定义了“案件名、案发地、涉案人员、案件描述”四个要素作为案件要素。

特定案件的舆情新闻文本，除了具有特定案件相关的主题信息外，同一案件相关的不同文章往往也会有不同的侧重点，因此我们采用 TextRank[9]算法，对每个文档同时提取出 5 个最重要的关键词，用来表征不同新闻文本所具有的不同的主题信息，将关键词也作为一部分引入到案件要素中。

综上，通过构建案件要素，并且联合发掘到的文章中的其他关键词来指导模型生成摘要。其中案件要素及其构成实例如表 2 所示。

表 2 案件要素及其构成实例

要素名	关键词
案件名	榆林产妇坠楼事件
案发地	榆林、医院
涉案人员	产妇、护士、家属
案件描述	坠楼
关键词	马某某、榆林、瞬间、产妇、监控

表 2 中，本文以榆林产妇坠楼事件为例，“案发地”包括案发的城市地区和案发的具体场所，例如，“榆林、医院”等。“涉案人员”不仅仅局限于受害人与嫌疑人，而且包括关键证人、相关家属等所有与案件相关人员。“案件描述”指发生的是什么事情，例如杀人、跳楼等。“关键词”指同一个案件下不同新闻文本所对应的不同的关键词。

2.2 词、句子双层编码

新闻文本往往是超过几百字长度的文本，为了较好地获得句子和文本的特征，选择词、句子双层编码器来将句子和文本来表征为向量。双向 GRU (BiGRU)通常要优于传统的 RNN，因此在词编码层和句子编码层，都采用基于 BiGRU 的神经网络，用以对句子、文档进行向量化语义表示。

编码层的输入是一篇含有 l 个句子的文本 $d=\{s_1,\ldots s_i,\ldots s_l\}$，其中 s_i 表示文档中第 i 个句子。每个句子由 m 个词组成 $s_j=\{w_1,\ldots w_i,\ldots w_m\}$。其中 w_i 表示句子中第 i 个词。在词编码层，将句子中每一个词的词向量按顺序送入一个由 BiGRU 单元构成的神经网络，得到词的隐层向量 $\{h_1^w,\ldots h_i^w,\ldots h_m^w\}$，$h_i^w$ 是句子中第 i 个词的隐层向量表示。这一阶段的每个步骤中，前向 GRU 基于当前输入 w_i 和先前隐状态向量 h_{i-1}^w 计算当前的隐层向量 $\overrightarrow{h_i^w}$。我们还从 w_m 到 w_1 反向运行第二个 GRU 来生成后向隐层向量表示 $\overleftarrow{h_i^w}$，如式(1)、式(2)所示。

$$\overrightarrow{h_i^w}=\overrightarrow{\mathrm{GRU}}(w_i,h_{i-1}^w) \tag{1}$$

$$\overleftarrow{h_i^w}=\overleftarrow{\mathrm{GRU}}(w_i,h_{i+1}^w) \tag{2}$$

最后，我们通过拼接前向隐层向量 $\overrightarrow{h_i^w}$ 和后向隐层向量 $\overleftarrow{h_i^w}$ 得到句子 s_i 的新表示 h_i^s，如式(3)所示。

$$h_i^s = [\overrightarrow{h_i^w}, \overleftarrow{h_i^w}] \tag{3}$$

相似的，文档中每一个句子的 h_i^s 表示，又作为句子编码器的输入。句子编码器同样采用一个 BiGRU 结构的神经网络，每一个 GRU 单元输入的是当前句子编码和上一 GRU 单元句子的隐层表示 h_{i-1}^s。拼接双向句子隐层向量后最终得到文档的编码向量 $\boldsymbol{d}$，如式(4)所示。

$$\boldsymbol{d} = \tanh\left(\boldsymbol{W}\frac{1}{l}\sum_{i=1}^{l} h_i^s + \boldsymbol{b}\right) \tag{4}$$

其中，$\mathbf{W}$ 和 $\boldsymbol{b}$ 是参数，l 是文档中句子数。

2.3　案件要素注意力机制

在案件要素编码层，我们将一个新闻文本对应的案件要素构成一个集合 $k = \{k_1, \cdots, k_i, \cdots, k_n\}$ 作为输入，其中 n 为案件要素的总数。我们采用与词编码层一样的词向量对 $\boldsymbol{k}_i$ 进行表示。将 k 通过一层 BiGRU 变换，得到的输出作为注意力向量 $\boldsymbol{q}$。具体结构如图 2 所示。

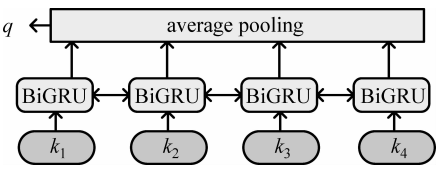

图 2　案件要素编码层

图 2 中，k_i 是第 i 个案件要素的词向量，将 k_i 按顺序送入一个 BiGRU 单元，得到每一个时间步的隐层状态 h_i^k，最终采用 average polling 计算所有 h_i^k 的平均值得到注意力向量 $\boldsymbol{q}$，如式(5)所示。

$$\boldsymbol{q} = \tanh\left(\boldsymbol{W}\frac{1}{n}\sum_{i=1}^{n} h_i^k + \boldsymbol{b}\right) \tag{5}$$

其中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是参数，n 是文档中句子数。

在词编码层得到句子向量的过程中，我们通过案件要素注意力向量 $\boldsymbol{q}$ 和每个词计算注意力权重得到 α_i^w，然后计算出当前句子的隐层向量 h_i^s，如式(6)、式(7)所示。

$$\alpha_i^w = \frac{\exp(h_i^w, q)}{\sum_{i=1}^{m}\exp(h_i^w, q)} \tag{6}$$

$$h_i^s = \sum_{i=1}^{m} \alpha_i^w h_i^w \tag{7}$$

其中，m 是当前句子的长度，h_i^w 是当前句子第 i 个词的隐层表示。

在句子编码层得到文档向量 $\boldsymbol{d}$ 的时候，我们首先通过案件要素注意力向量 $\boldsymbol{q}$，和每个句子计算权重得到 α_i^s，然后融合案件要素的信息和句子信息生成当前文档的向量表示 $\boldsymbol{d}$，如式(8)、式(9)所示。

$$\alpha_i^s = \frac{\exp(h_i^s, \boldsymbol{q})}{\sum_{i=1}^{l}\exp(h_i^s, \boldsymbol{q})} \tag{8}$$

$$d = \sum_{i=1}^{l} \alpha_i^s h_i^s \tag{9}$$

其中，h_i^s 是当前文档第 i 个句子的隐层向量，l 表示当前文档共有 l 个句子。

2.4　多特征分类层

常规的文本分类任务输入的是一个文本，判断分类的对象就是输入的对象，是对等的。但是文本摘要任务有着自己的特点，输入的是一篇文本，分类的对象却是文档中的句子，是不对等的，文本只是作为一个上下文信息而存在。因此在对句子分类的时候，必须考虑上下文信息对句子的影响。

首先，对于整个文本，我们考虑当前句子的显著性，通过句子向量 $\boldsymbol{h}_i^{\mathrm{T}}$ 和文本向量 $\boldsymbol{d}$ 求得句子显著性信息 salience，如式(10)所示。

$$\text{salience} = \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{W}_s\boldsymbol{d} \tag{10}$$

Nallapati 等[12]还提出考虑当前句子新颖度的思想。具体来说，如式(11)所示，s_j 是之前句子信息的加权和，随着遍历到不同的句子而动态变化，表示一种前文信息。

$$s_j = \sum_{i=1}^{j-1} \boldsymbol{h}_i \boldsymbol{P}(y = 1 \mid \boldsymbol{h}_i, \boldsymbol{s}_i, \boldsymbol{d}) \tag{11}$$

其中，$\boldsymbol{h}_i$ 是句子的隐层向量，$\boldsymbol{d}$ 是文本向量。在式(12)中，使用句子向量 $\boldsymbol{h}_i^{\mathrm{T}}$ 和 $\boldsymbol{s}_i$ 计算得到当前句子和前文的重复信息，减去重复的信息就可以得到当前句子的新颖性 novelty。

$$\text{novelty} = -h_i^T\mathbf{W}\tanh(s_i) \tag{12}$$

以往的文本摘要研究[14]表明，句子位置也是很重要的一个因素，在文章开头部分的句子成为摘要句的概率通常要大一些。

综上所述，在句子分类层，重新遍历句子序列，在每个时间步融合句子信息、句子显著性、句子新颖性、句子相对位置和绝对位置等多个特征来进行句子分类，如式(13)所示。

$$\begin{aligned} P(\boldsymbol{y}_i \mid \boldsymbol{h}_i, \boldsymbol{s}_i, \boldsymbol{d}) = \sigma(&\boldsymbol{W}_c\boldsymbol{h}_i && \#\text{当前句子信息} \\ &+\text{salience} && \#\text{当前句子显著性} \\ &+\text{novelty} && \#\text{当前句子新颖性} \\ &+\boldsymbol{W}_{\mathrm{ap}}\boldsymbol{P}_i^a && \#\text{绝对位置} \\ &+\boldsymbol{W}_{\mathrm{rp}}\boldsymbol{P}_i^r && \#\text{相对位置} \end{aligned} \tag{13}$$

$$+\boldsymbol{b}) \qquad \#\text{偏置向量} \tag{13}$$

其中，$\boldsymbol{h}_i$ 是第 i 个句子的隐层向量，$\boldsymbol{W}_c\boldsymbol{h}_i$ 是当前的句子信息，salience 表示第 i 个句子在整篇文章中的显著性信息。novelty 表示第 i 个句子和前面句子相比所具有的新颖度信息。$\boldsymbol{W}_{ap}\boldsymbol{P}_i^a$ 和 $\boldsymbol{W}_{rp}\boldsymbol{P}_i^r$ 分别表示第 i 个句子绝对位置和相对位置的信息。

最终的句子被分为两种标签，是摘要句和不是摘要句，针对这样的二分类的问题，选择交叉熵作为损失函数，如式(14)所示。

$$l(\boldsymbol{W},\boldsymbol{b}) = -\sum_{d=1}^{N}\sum_{i=1}^{l}((y_i^d \log P(y_i^d \mid h_i^d, s_i^d, \boldsymbol{d}_d) + (1-y_i^d)\log(1-P(y_i^d \mid s_i^d, s_i^d, \boldsymbol{d}))) \tag{14}$$

其中，N 表示文档的数量，l 表示每个文本的句子数，角标 d 表示第 d 篇文档，$\boldsymbol{d}_d$ 表示第 d 篇文档的向量表示。

3 实验

3.1 数据集

本文针对 30 组案件要素，从互联网上搜集相关新闻，构建了涉案舆情新闻文本数据集。然后对每一篇新闻文档和案件之间的关系进行分析和校对，得到每篇文档和一个案件对应的关系。我们对文本标题逐条进行人工修改校对，以修改后的内容作为参考摘要。数据集相关信息如表 3 所示。

表 3 数据集相关信息

	文本数	句子数	文本长度	摘要长度
训练集	17 434	15.38	776	26.78
验证集	1 000	16.46	776	29.63
测试集	1 000	14.42	722	28.69

与大多数的摘要数据集一样，数据集也是只包含人工书写的参考摘要。但是对于基于神经网络的句子分类模型，需要标记每个句子的分类标签。为了解决这个问题，Cao 等人曾提出的是一种基于整数线性规划(ILP)的方法来解决这个问题[15]。我们使用一种与 Svore 等人提出的相似的方法[16]，通过文档中的句子和人工摘要的 ROUGE 评分，来寻找一个得分最高的句子组合作为摘要句。由于寻找一个全局最优句子组合的计算代价过高，因此我们采用一种贪婪搜索的方法来寻找最优组合。首先选取一个评分最高的句子加入到摘要集合内，然后在摘要集合中一次添加一个句子，观察集合的 ROUGE 得分是否上升。若上升，则将新句子加入到集合中，直到遍历完所有剩余的句子。最终，这个集合中的句子都标记为 1，其余的标记为 0，用这样的数据来作为标记数据。

3.2 评价标准

本文采用自动摘要任务中常用的一种内部评价指标 ROUGE[17] 来作为评价指标。ROUGE 是基于摘要中 n 元语法(n-gram)的共现信息来评价摘要，包括 ROUGE-1，ROUGE-2 等。ROUGE 值越高说明效果越好，具体计算方法如式(15)所示。

$$\text{ROUGE-N} = \frac{\sum_{s\in(\text{Referemce Summaries})}\sum_{\text{n-gram}\in s}\text{Count}_{\text{match}}(\text{n-gram})}{\sum_{s\in(\text{Referemce Summaries})}\sum_{\text{n-gram}\in s}\text{Count}(\text{n-gram})} \tag{15}$$

其中，分子表示模型输出的摘要和参考摘要中同时出现 n-gram 的个数，分母则表示参考摘要中出现的 n-gram 个数。

3.3 实验设置

3.3.1 预训练词向量

使用 gensim 中的 word2vec 开源工具，联合搜狗实验室的全网新闻数据集[18] 和本文的涉案舆情新闻文本数据集来训练词向量。其中词向量维度为 100 维，词表大小为 382 753。

3.3.2 实验参数设置

本文词编码层、句子编码层与案件要素编码层均采用 BiGRU，隐层向量均设置为 200 维。为了避免网络过拟合，Dropout 设置为 0.3。同时为了避免梯度爆炸采用梯度裁剪技术，设置最大梯度为 10。其他参数设置如下：批次大小为 16，优化器选择 adam，初始学习率为 0.001。

3.3.2 对比实验设置

本文共设置了 3 组对比实验和 1 个实例分析，第一组实验对照本文模型和基准模型性能。第二组实验验证案件要素和关键词对模型性能提升的效果。第三组实验验证词、句子编码层融入案件要素的有效性。实例分析选取了 5 个基准模型和本文模型的摘要实例进行对比分析。

特别说明，GRUkey_Attn$_{all}$ 表示仅使用关键词作为注意力机制的实验，GRUcase$_{(-)}$_Attn$_{all}$ 表示仅使用案件要素而不用关键词作为注意力机制的实验，GRUcase_Attn$_{all}$ 表示使用案件要素和关键词一

起作为注意力机制的实验。GRUcase_Attn$_{word}$表示仅在词编码层引入案件要素注意力机制的实验，GRUcase_Attn$_{sent}$表示仅在句子编码层引入案件要素注意力机制的实验。

3.4　基准模型

本文选择了 5 个基准模型来进行实验。分别为：LexPageRank，Submodular，LEAD3，TextRank 和 SummaRuNNer。其中 LexPageRank 和 Submodular 由开源工具包 PKUSUMSUM 提供。

LEAD3 是一种依靠句子在文章中的位置来抽取摘要的方法，一般来说，文章的重要信息容易出现在文章开头的部分。LEAD3 抽取文章前三句作为摘要，虽然简单，但是很有效果。

Submodular[19] 利用次模函数的单调递减特性来抽取句子作为摘要。

LexPageRank[20] 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。

SummaRuNNer[12] 是一种利用神经网络训练句子分类模型的方法，利用大规模语料 CNN/DailyMail 进行训练，在英文摘要任务上，达到了当时的最优性能。

3.5　实验结果分析

第一组实验是本文模型和 5 个 baseline 在涉案舆情新闻文本数据集上的对比实验，其中本文模型 GRUcase_Attn$_{all}$在词、句子编码层都融入案件要素注意力机制，实验结果如表 4 所示。

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
LexPageRank	19.18	8.83	13.38
Submodular	19.38	8.83	13.38
LEAD3	19.62	8.81	10.04
TextRank	23.09	12.25	19.94
SummaRuNNer	25.53	14.11	18.52
GRUcase_Attn$_{all}$	**28.94**	**15.89**	**20.87**

针对系统输出过长的问题，使用系统摘要的前 80 个字作为评测数据，根据表 4 的实验结果可以看出：①在 ROUGE-1 的评价指标上，GRUcase_Attn$_{all}$比 LexPageRank 和 TextRank 分别高了 9.76 和 5.85，SummaRuNNer 比 LexPageRank 和 TextRank 分别高了 6.35 和 2.44，说明在自动摘要任务中，基于神经网络的方法更有效。②GRUcase_Attn$_{all}$和 SummaRuNNer 对比，ROUGE-1 提高了 3.41，ROUGE-2 提高了 1.78，ROUGE-L 提高了 2.35。③结果表明案件要素通过注意力机制融入文档编码中和分类中，可以较好地提高模型的摘要效果。

第二组对比实验为了测试案件要素在涉案舆情新闻文本摘要任务上的作用，包括 4 个实验。实验 GRU 表示不引入任何案件要素和关键词信息。实验结果如表 5 所示。

表 5　案件要素有效性实验对比结果

模型	ROUGE-1	ROUGE-2	ROUGE-L
GRU	25.53	14.11	18.52
GRUkey_Attn$_{all}$	26.21	14.52	18.54
GRUcase$_{(-)}$_Attn$_{all}$	27.38	15.43	19.86
GRUcase_Attn$_{all}$	28.94	15.89	20.87

根据表 5 的实验结果可以看出：①融入案件要素和关键词都使模型效果有一定的提升。②融入案件要素比仅融入关键词的效果要好，充分体现了案件要素对涉案舆情新闻文本摘要的指导作用。

第三组对比实验是为了测试词、句子编码层分别引入基于案件要素的注意力机制对模型效果的影响。实验结果如表 6 所示。

根据表 6 的实验结果可以看出：单独使用句子级注意力机制效果略优于词级注意力机制，因为该模型将案件要素关键词编码为注意力向量 $\boldsymbol{q}$。在模型上，$\boldsymbol{q}$ 和句子的隐层向量有着相似的地位。而且，在多特征分类层也都是对句子级别的信息进行处理。因此，该组对比实验表明本文提出的基于案件要素的注意力方法能够较好地将案件要素的信息融入摘要的生成中，使摘要的生成更接近特定案件相关的主题。

表 6　不同层融入案件要素注意力实验对比结果

模型	ROUGE-1	ROUGE-2	ROUGE-L
GRU	25.53	14.11	18.52
GRUcase_Attn$_{word}$	26.71	14.94	18.46
GRUcase_Attn$_{sent}$	28.04	15.01	20.25
GRUcase_Attn$_{all}$	28.94	15.89	20.87

最后，针对奔驰女车主维权案摘要生成的实例

分析，进一步验证案件要素在摘要生成过程中，具有重要的指导作用。其中该新闻文本如表7中第1行所示，相关案件要素如第2行所示。GRUcase_Attn$_{all}$和LEAD3对比，可以看出在本文方法里位置特征得到了很好的体现。和基准模型抽取的句子对比可以看出，因为本文模型将案件要素融入句子的分类中，所以本文模型能够较好地抓住文中的“陕西、西安、利之星4S店、车顶、发动机漏油”等案件要素。本文方法抽取出的句子比其他无监督的方法更贴近主题，更有总结性，而不是得到很多具体描述事件的句子。所以案件要素通过注意力机制融入文本编码和分类中，可以较好地提高模型的摘要效果。

表7 摘要对比实例

对比模型	生成结果
新闻文本	近日，陕西西安一女士花66万元买新奔驰，可车还没开出门就发现发动机漏油。此后，车就一直放在店里。在15天之内，经多次协商，该店的解决方案从退款、换车变成免费换发动机，女子不接受便坐在店内的车顶上要说法，有理有据、思路清晰的维权视频在网上引起热议。视频中，女车主坐在一辆红色轿车前引擎盖上，带着哭腔控诉西安利之星汽车有限公司，称距离自己签单提车才5分钟，都没开出门，发动机就开始了漏油。“我马上打电话给销售，对方说这是发动机没油了，你开来店里我给你加油。”把车子开回店里后，奔驰车就一直停在店内。这期间，车主和4S店交涉了三次。“第一次说要退款，后来又提出退款不方便改为换车，再又说换车也不方便改为补偿，我都答应了。”直到4月8日，15天后，利之星再一次推翻之前的解决办法，称目前的情况根据国家的“三包”规定只能免费更换发动机……（部分）
参考摘要	陕西西安66万新奔驰没开出门发动机就漏油，女车主无奈坐车顶维权
案件要素	**奔驰女车主坐车顶维权事件、陕西、西安、利之星4S店，奔驰女车主、发动机漏油坐车顶维权、车主、发动机、没开、换车、退款**
LexPageRank	店方工作人员劝其下车慢慢谈，但车主好声好气地回复道，“如果你们有人跟我谈，我就不会坐在这里丢脸了。“我马上打电话给销售，对方说这是发动机没油了，你开来店里我给你加油。我就不信这个天下没有说理的地方。这期间，车主和4S店交涉了三次。
Submodular	“我马上打电话给销售，对方说这是发动机没油了，你开来店里我给你加油。“如果是开出200公里一次长途出了问题，你跟我讲‘三包’，我同意，但我开车还没出这个门，一公里没开。把车子开回店里后，奔驰车就一直停在店内。
LEAD3	近日，陕西西安一女士花66万买新奔驰，可车还没开出门就发现发动机漏油。此后，车就一直放在店里。在15天之内，经多次协商，该店的解决方案从退款、换车变成免费换发动机，女子不接受便坐在店内的车顶上要说法，有理有据思路清晰的维权视频在网上引起热议。
TextRank	“我马上打电话给销售，对方说这是发动机没油了，你开来店里我给你加油。视频中，女车主坐在一辆红色轿车前引擎盖上，带着哭腔控诉西安利之星汽车有限公司，称距离自己签单提车才5分钟，都没开出门发动机就发生了漏油。“一公里都没开就换发动机，简直是无妄之灾！
SummaRuNNer	此后，车就一直放在店里。在15天之内，经多次协商，该店的解决方案从退款、换车变成免费换发动机，女子不接受便坐在店内的车顶上要说法，有理有据思路清晰的维权视频在网上引起热议。视频中，女车主坐在一辆红色轿车前引擎盖上…
GRUcase_Attn$_{all}$	近日，**陕西西安**一女士花66万买新**奔驰**，可车还没开出门就发现**发动机漏油**。视频中，**女车主**坐在一辆红色轿车前引擎盖上，带着哭腔控诉**西安利之星**汽车有限公司，称距离自己签单提车才5分钟，都**没开**出门**发动机**就发生了**漏油**。**车主**被逼无奈再次来到4S**店**，坐上**车顶**哭诉讨说法。

4 总结和展望

针对涉案舆情新闻文本，本文提出了一种融入案件要素的摘要方法，该方法通过构建相关的案件要素，并使用注意力机制将这些案件要素融入摘要句的分类过程。实验结果表明：针对涉案领域的新闻摘要任务，案件要素对于摘要句的提取有很好的指导作用，基于案件要素的注意力的方法也是有效的融入方法。在下一步的研究中，我们拟在领域知识及融合方式上进行探索，如尝试利用图卷积神经网络融入案件要素及要素关系等领域知识来提升摘要的生成质量。另外，在涉案舆情摘要任务上，探索同一案件多文档的案件摘要方法。

参考文献

[1] Kaikhah K. Automatic text summarization with neural networks[C]//Proceedings of the 2nd International IEEE Conference on Intelligent Systems. Varna, Bulgaria: IEEE Press, 2004, 1: 40-44.

[2] Joshi M, Wang H, McClean S. Generating object-oriented semantic graph for text summarization in mining intelligence and knowledge exploration [M]. New York, USA: Springer Press, 2014: 298-311.

[3] Hingu D, Shah D, Udmale S S. Automatic text summarization of wikipedia articles[C]//Proceedings of the 2015 International Conference on Communication, Information & Computing Technology (ICCICT). Singapore: IEEE Press, 2015: 1-4.

[4] Yao J, Wan X, Xiao J. Recent advances in document summarization[J]. Knowledge and Information Systems, 2017, 53(2): 297-336.

[5] Baxendale P B. Machine-made index for technical literature: An experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.

[6] Edmundson H P. New methods in automatic extracting[J]. Journal of the ACM (JACM), 1969, 16(2): 264-285.

[7] Yohei Seki. Sentence extraction by tf·idf and position weighting from newspaper articles [C]//Proceedings of the 3rd NTCIR Workshop on Research in Information Retrieval. Automatic Text Summarization and Question Answering . Tokyo: NII. 2002:55-59.

[8] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

[9] Mihalcea R, Tarau P. TextRank: Bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: ACL Press. 2004: 404-411.

[10] Sankarasubramaniam Y, Ramanathan K, Ghosh S. Text summarization using Wikipedia[J]. Information Processing & Management, 2014, 50(3): 443-461.

[11] Cheng J, Lapata M. Neural summarization by extracting sentences and words[J]. arXiv preprint arXiv: 1603.07252, 2016.

[12] Nallapati R, Zhai F, Zhou B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]//Proceedings of 31st AAAI Conference on Artificial Intelligence. San Francisco, California USA: AAAI Press, 2017. 3075-3081.

[13] Wang L, Yao J, Tao Y, et al. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization[J]. arXiv preprint arXiv:1805.03616, 2018.

[14] 王振超，孙锐，姬东鸿. 基于事件指导的多文档生成式摘要方法[J]. 计算机应用研究,2017, 34(2):343-346,356.

[15] Cao Z, Chen C, Li W, et al. Tgsum: Build tweet guided multi-document summarization dataset[C]//Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press. 2016:2906-2912.

[16] Svore K, Vanderwende L, Burges C. Enhancing single-document summarization by combining RankNet and third-party sources[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague. Czech Republic: EMNLP, ACL Press, 2007: 448-457.

[17] Lin C Y. Rouge: A package for automatic evaluation of summaries[J]. Text Summarization Branches Out, 2004: 74-81.

[18] Wang C, Zhang M, Ma S, et al. Automatic online news issue construction in web environment[C]//Proceedings of the 17th International Conference on World Wide Web. Beijing, China: ACM Press,2008: 457-466.

[19] Lin H, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions[C]//Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles CA USA: NAACL, ACL Press, 2010: 912-920.

[20] Erkan G, Radev D R. Lexpagerank: Prestige in multi-document text summarization[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: ACL Press, 2004: 365-371.

（下转第 73 页）

mation access[C]//Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, 2013.

[23] Duck G, Nenadic G, Brass A, et al. bioNerDS: Exploring bioinformatics' database and software use through literature mining[J]. BMC Bioinformatics, 2013, 14(1):194-194.

[24] Duck G, Nenadic G, Brass A, et al. Extracting patterns of database and software usage from the bioinformatics literature[J]. Bioinformatics, 2014, 30(17):601-608.

[25] Geraint D, Goran N, Michele F, et al. A survey of bioinformatics database and software usage through mining the literature[J]. PLoS ONE, 2016, 11(6):1-25.

[26] Toshihisa T, Yasunori Y. OReFiL: An online resource finder for life sciences[J]. BMC Bioinformatics, 2007, 8(1):1-81.

[27] Calle G D L, Miguel García-Remesal, Chiesa S, et al. BIRI: A new approach for automatically discovering and indexing available public bioinformatics resources from the literature[J]. BMC Bioinformatics, 2009, 10(1):320.

罗准辰(1984—)，博士，高级工程师，主要研究领域为科技信息智能挖掘服务技术、自然语言处理。
E-mail：zhunchenluo@gmail.com

叶宇铭(1990—)，通信作者，硕士，工程师，主要研究领域为科技信息智能挖掘服务技术。
E-mail：yuming_ye@163.com

赵赫(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：zhaohe1995@outlook.com

(上接第 63 页)

韩鹏宇(1995—)，硕士研究生，主要研究领域为自然语言处理、信息检索。
E-mail：hanpengyuice@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