计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(10):229−237 [doi: 10.15888/j.cnki.csa.009973] [CSTR: 32024.14.csa.009973]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

多视角事件重构的摘要生成①

孙　锐

(南京信息工程大学 软件学院, 南京 210044)
通信作者: 孙　锐, E-mail: 202212490309@nuist.edu.cn

摘　要: 在当前互联网信息多元分布的背景下, 单文档信息抽取的传统范式已难以满足用户对事件全局认知的需求. 针对多源文本数据中信息冗余与观点碎片化的问题, 本文提出基于过滤机制的多维度文本摘要生成模型 (FM-MDSG), 该框架通过 3 阶段创新架构实现跨源信息的结构化融合, 首先采用微调 RoBERTa 模型构建层次化语义表征, 捕获输入文本的上下文依赖. 其次, 设计双层过滤机制, 同步执行基于注意力权重的显著性检测与领域自适应的冗余抑制, 筛选出信息密度优化的语义单元. 最后, 构建知识增强的 ERNIE 解码器, 通过动态门控策略实现多层级语义特征的协同生成. 在 CSL 数据集上的实验表明, 该模型 ROUGE-1/2/L 的 *F* 值分别达到 55.37%、47.28% 和 49.56%, ROUGE-L 较经典基线模型提升 6.8 个百分点. 消融实验进一步验证, 过滤机制通过噪声抑制带来 9.22% 的 ROUGE-1 性能增益. 该模型实现了对异构来源证据的系统性整合, 能够在开放域场景下重构多视角观测的完整事件范式.

关键词: 文本生成; 信息抽取; 过滤机制; 多维度文本; 知识增强

引用格式: 孙锐.多视角事件重构的摘要生成.计算机系统应用,2025,34(10):229–237. http://www.c-s-a.org.cn/1003-3254/9973.html

Multi-perspective Event Reconstruction for Summary Generation

SUN Rui

(School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: In the context of the current multi-dimensional distribution of Internet information, the traditional paradigm of single-document information extraction no longer meets the demand for comprehensive event understanding. To address issues of information redundancy and fragmented viewpoints in multi-source text data, a multi-dimensional text summary generation model (FM-MDSG) based on a filtering mechanism is proposed. The proposed framework enables a structured fusion of cross-source information through a three-stage architecture. First, a hierarchical semantic representation is constructed by fine-tuning the RoBERTa model to capture contextual dependencies of the input text. Second, a two-layer filtering mechanism is designed to simultaneously perform saliency detection based on attention weights and domain-adaptive redundancy suppression, thus extracting semantic units with optimized information density. Finally, a knowledge-enhanced ERNIE decoder is employed, utilizing a dynamic gating strategy to enable the collaborative generation of multi-level semantic features. Experiments on the CSL dataset demonstrate that the proposed model achieves ROUGE-1/2/L *F*-scores of 55.37%, 47.28%, and 49.56%, respectively, ROUGE-L representing an improvement of 6.8 percentage points over classic baseline models. Further ablation studies confirm that the filtering mechanism contributes to a 9.22% performance gain in ROUGE-1 through noise suppression. The proposed model enables the systematic integration of heterogeneous source evidence and supports the reconstruction of complete event paradigms from multi-perspective observations in open-domain scenarios.

Key words: text generation; information extraction; filtering mechanism; multi-dimensional text; knowledge enhancement

① 基金项目: 国家自然科学基金 (62473201); 江苏省自然科学基金 (BK20231142)
收稿时间: 2025-02-19; 修改时间: 2025-03-18; 采用时间: 2025-04-27; csa 在线出版时间: 2025-08-28
CNKI 网络首发时间: 2025-08-29

1 引言

移动互联网技术的快速演进与智能终端的广泛渗透, 使得多源异构信息的实时获取成为常态[1]. 然而, 突发事件的动态性与复杂性导致新闻报道呈现指数级增长, 不同媒体源在报道角度、叙事逻辑和信息粒度上的差异加剧了信息整合的难度[2]. 传统单文档摘要方法虽能缓解信息过载, 但其单一来源的局限性导致生成内容难以覆盖事件的多维特征, 显著降低了用户对事件全局的认知效率[3]. 多源信息冗余不仅延长了用户决策路径, 还增加了关键信息被噪声遮蔽的风险. 这一矛盾在重大公共事件场景中尤为突出, 亟待通过技术创新实现深层次信息融合[4].

基于深度学习的自动摘要技术近年来取得显著突破, 预训练语言模型通过自监督学习捕获大规模语料的语义规律, 为生成任务奠定基础[5]. 多文档摘要领域的主流方法聚焦于跨文档语义对齐与冗余控制, 其中层次化注意力机制和对比学习策略被证明能有效提升生成内容的连贯性[6].

然而, 现有技术对突发事件的多维度特征仍缺乏系统性建模能力, 导致生成摘要难以满足多维认知需求[7]. 首先, 信息融合维度单一, 多数研究依赖浅层语义相似度计算, 忽视事件参与者交互模式与时空演变等高阶特征的联合建模[8]. 其次, 噪声抑制策略存在效率瓶颈, 跨源输入中大量重复片段与低信息密度内容直接影响生成质量. 最后, 媒体立场偏差未充分消解, 生成内容易受信源主观倾向影响[9]. 此外, 静态特征融合机制难以适应突发事件不同阶段的信息密度与语义焦点变化, 进一步制约了模型的实际应用价值[10].

因此, 为了解决上述问题, 本研究提出基于过滤机制的多维度文本摘要生成模型 (FM-MDSG), 其核心创新在于构建覆盖事件全生命周期的多维表征与动态融合框架. 第 1 阶段采用预训练语言模型 RoBERTa 为编码层主干, 通过层次化神经网络提取文档内语义关联与跨文档实体互操作特征. 第 2 阶段设计双通路过滤架构, 结合可微分阈值机制动态识别冗余内容, 同时利用对抗训练策略抑制立场相关噪声. 第 3 阶段部署知识增强型解码器, 集成 ERNIE 3.0 的先验知识库, 通过多头注意力权重动态融合多维度特征, 实现突发事件发展脉络的重构. 本文的主要贡献如下.

(1) 提出基于字词与句子双粒度的可微分过滤机制, 提高文本生成的质量.

(2) 注入 ERNIE 解码架构, 实现事件发展逻辑链的因果推理, 使生成摘要具有时序连贯性.

(3) 提出面向全生命周期的事件表征范式, 突破了传统方法依赖单一语义相似度度量的局限性.

本文第 2 节讨论相关工作. 第 3 节提出具体算法模型. 第 4 节介绍仿真实验和评价的对比分析. 最后, 第 5 节对本文进行总结并讨论未来需要开展的工作.

2 相关工作

2.1 生成式摘要生成

在文本摘要技术领域, 大多数方法基于从非结构化文档中提取信息. 然而这种方法容易导致句子大小写分析中的非残差抽象问题. 为解决这一问题, Mohan 等人[11]提出了一种基于 Lattice 抽象的内容摘要 (Labs-CS) 方法, 通过减少 Intra 子集群来沉淀句子, 改善了非结构化文档的处理效果.

抽取式和生成式摘要方法各存在一些不足, 为结合两者的优点, Yan 等人[12]提出了一种基于 K-means 的融合方法. 面对隐性数据集无效的问题. Kwon 等人[13]提出一种多任务学习方法, 通过在编码器-解码器语言模型的微调过程中添加对比目标, 反映了学习过程中有关显著和偶然对象的信息. 为了同时从句子中捕获单词关系和结构信息. Guan 等人[14]提出了一种抽象句子摘要 (SWSum) 结构到单词动态交互模型, 通过动态交互机制, 有效提升了摘要的准确性和连贯性. 此外, 针对文本摘要任务中常见的冗余、缺乏词汇分配和不相关结果等问题, Abd 等人[15]引入了一种基于注意力的序列到序列模型. 该模型通过改进注意力机制, 能够更有效地捕捉关键信息, 生成高质量的摘要.

2.2 抽取式摘要生成

为了解决海量 Web 文本中上下文理解困难的问题, 学术界提出了多种创新性方法. Yadav 等人[16]提出了一种分层注意力指针堆叠去噪的变分自动编码器模型 (SDVAE) 有效捕获多义词的语义信息. 在阿拉伯语文本摘要领域, Alselwi 等人[17]提出了一种基于图结构的抽取式摘要技术, 该方法结合词嵌入和 PageRank 算法进行特征提取和句子排序, 实现了对阿拉伯语文本的有效摘要生成.

为进一步完善本地上下文信息的捕获与集成, Wang 等人[18]设计了一种长文本抽取式摘要模型, 能够同时捕获原始文本的本地主题信息和文档的层次结构信息.

然而, 当前抽取式文本摘要技术仍存在若干局限性, 包括上下文特征保留不足、特征提取能力有限以及处理分层和组合信息的能力有待提升等.

针对上述挑战, Gangundi 等人[19]提出了 RoBERTa-BiLSTM-CNN-Attention 抽取式文本摘要模型 (RBCA-ETS). 该模型采用 RoBERTa 生成上下文嵌入, 并通过并行连接的 CNN 和 BiLSTM 层提取文本特征, 有效提升了摘要生成的质量. 此外, 现有研究普遍忽视了主题极性在摘要生成中的重要性, 这可能导致摘要无法全面涵盖主题的各个维度. 为解决这一问题, Monir 等人[20]通过在基于主题和方面情感分析之前评估生成的摘要, 为抽取式摘要生成提供了新的研究方向.

如表 1 所示, 现有方法中, 尽管各种技术在不同方面取得了显著进展, 但仍存在一些共性问题. 许多方法在处理长文档时表现不佳, 尤其是在捕获全局上下文信息方面存在不足. 此外, 现有方法在处理多维度文本时, 往往缺乏足够的适应性, 导致摘要生成的质量参差不齐.

表 1 文献中的主要贡献

文献	方法简介	目的	性能	数据集
文献[11]	基于格抽象的内容摘要	正确且一致的格式显示摘要信息	Precision: 92.5%, Recall: 90.1%, Accuracy: 94%	Reuters
文献[12]	K-means结合Cw2vec	根据主要思想的强度, 对文章不同部分进行针对性总结	ROUGE-1平均改善: 55%	UC、Gagiword
文献[13]	Transformer加入对比模块	突出积极部分并附带消极部分的摘要	ROUGE-1/2/L比原始BART基础高出1.15/0.93/1.28 points	CNNDM、XSum
文献[14]	FrameNet结合多层基于图的动交互层	增强基于图方法中单词关系或结构信息的相关性	ROUGE-1: 42.25, ROUGE-2: 23.81, ROUGE-L: 39.14	Gigaword、DUC 2004
文献[15]	Bi-LSTM编码器结合注意力机制	解决长句大型源文档中要点表示不足的问题	ROUGE-1: 47.37%, ROUGE-2: 22.57%, ROUGE-L: 45.85%	CNN/Daily Mail
文献[16]	SDVAE结合双向注意力机制	解决语境化句子覆盖面不足及冗余问题	Accuracy: 98.30%	Twitter sentiments-text
文献[17]	词嵌入和PageRank	处理语言中复杂的形态联系问题	Precision: 65.2%	EASC语料库
文献[18]	LSTM-Minus	处理长文档及不同部分主题信息多样性的问题	ROUGE-1: 46.49%, ROUGE-2: 20.52%, ROUGE-L: 42.06%	PubMed
文献[19]	RoBERTa结合并行连接的CNN和BiLSTM	完善上下文理解、单一架构、注意力机制及句子提取依赖	ROUGE-1: 27.12%, ROUGE-2: 11.71%, ROUGE-L: 21.31%	CNN/Daily Mail
文献[20]	主题提取结合情感分类	生成准确传达用户情绪及关键点的摘要	*F*1-score: 92.4%	Twitter sentiments-text

注: ROUGE-1/2/L均指ROUGE-1/2/L的*F*值.

因此, 针对多源文本数据中信息冗余与观点碎片化的问题, 本文以基于过滤机制的多维度文本摘要生成模型 (FM-MDSG) 为框架, 利用微调 RoBERTa 模型、双层过滤机制和知识增强的 ERNIE 解码器, 从而提高摘要生成的准确性和信息密度. 同时, 为了适配真实互联网环境中的实时性原则, 提出了最优双层过滤机制, 通过同步执行显著性检测与冗余抑制, 确保在高效筛选关键信息的同时, 满足实时性需求. 该框架不仅能够有效解决传统单文档信息抽取范式的局限性, 还能在多源异构数据中实现信息的系统性整合, 从而为用户提供更加全面和准确的事件全局认知.

3 模型设计

3.1 整体设计

本文将所设计的模型命名为基于过滤机制的多维度文本摘要生成模型 (FM-MDSG), 其结构如图 1 所示. FM-MDSG 模型分为 3 个主要部分, 分别负责输入文本的处理、语义特征的提取与优化以及摘要的生成, 有效解决了多源文本数据中的信息冗余与观点碎片化问题.

首先, 输入文本处理模块, 运用微调后的 RoBERTa 模型, 对多源文本实施层次化语义编码. 借助多层 Transformer 编码器, RoBERTa 模型可捕获文本的上下文依赖关系, 生成高质量的语义表征. 这种层次化编码方式, 有助于识别和区分不同来源文本的关键信息, 降低信息冗余, 增进对不同来源文本关联性的理解, 整合分散观点, 构建更连贯、完整的事件描述. 其次, 双层过滤机制不仅关注单个句子或词组的信息密度, 还考虑了跨文档的语句一致性. 通过动态调整权重矩阵, 保留关键信息并去除干扰内容, 确保生成的摘要能够覆盖多个视角, 减少观点的碎片化. 最后, 知识增强的 ERNIE

解码器利用多头注意力机制, 对优化后的语义特征进行解码, 生成连贯且信息丰富的文本摘要. 通过集成 ERNIE 的先验知识库, 解码器能够在生成过程中更好地理解语义关系, 避免生成冗余信息.

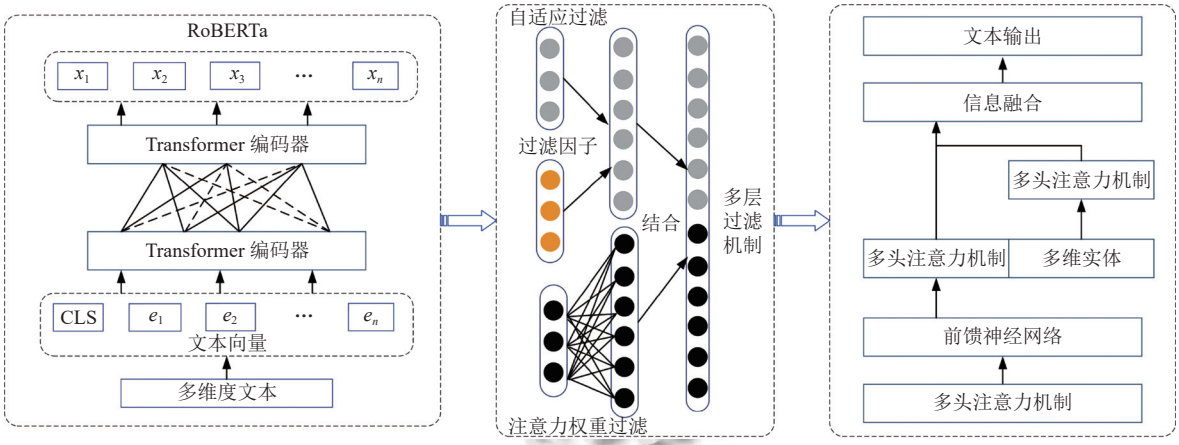

图 1 模型整体结构图

3.2 向量嵌入层构建

在 FM-MDSG 模型中, 向量嵌入层的构建作为文本处理的基础环节, 其核心目标是通过微调的 RoBERTa 模型对多源文本进行层次化语义编码, 从而捕获文本的上下文依赖关系, 为后续的过滤和生成模块提供高质量的语义表征. RoBERTa 模型作为 BERT 的改进版本, 通过动态掩码策略和更大规模的预训练数据, 显著提升了语义表征的能力. 在本文模型中, RoBERTa 模型被进一步微调, 以适应多源文本数据的特性.

给定输入文本序列 $X=\{x_1,x_2,\cdots,x_n\}$, 其中 x_i 表示第 i 个词, RoBERTa 模型通过多层 Transformer 编码器对输入序列进行编码, 生成上下文相关的语义表征, 如式 (1) 所示:

$$H = RoBERTa(X) \tag{1}$$

其中, $H=\{h_1,h_2,\cdots,h_n\}$表示输入序列的语义表征向量, $h_i \in R^d$ 是第 i 个词的 d 维向量表示. 通过多层 Transformer 的堆叠, 文本的局部和全局语义信息得以捕获, 从而为后续处理提供丰富的上下文特征.

为了适应多源文本数据的特性, 本文对 RoBERTa 模型进行进一步微调. 微调过程采用多任务学习策略, 结合文本分类和摘要生成任务, 优化模型的语义表征能力, 如式 (2) 所示:

$$L_{ft} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{sum}} \tag{2}$$

其中, L_{cls} 是文本分类任务的损失函数, L_{sum} 是摘要生成任务的损失函数, λ_1 和 λ_2 是权重系数.

3.3 双层过滤机制构建

在 FM-MDSG 模型中, 双层过滤机制的设计旨在从字词和句子两个层面分别对嵌入向量中的无用信息进行滤除, 从而提升模型对社交媒体文本的处理能力. 社交媒体文本的特点在于其不规则性, 如拼写错误、网络流行语、单字表达等, 这些特点使得传统的自然语言处理方法在字词层面和句子层面的处理效果存在显著差异. 因此, 本研究设计了基于字词层面和句子层面的两种过滤机制.

过滤机制的核心思想是通过权重矩阵对嵌入向量进行动态调整, 保留关键信息, 去除干扰内容. 给定嵌入向量 $H \in R^{n\times d}$, 其中 n 为序列长度, d 为向量维度, 首先通过变换操作生成权重矩阵 W_{Weight} 如式 (3) 所示:

$$W_{\text{Weight}} = f_{\text{Transform}}(H) \tag{3}$$

其中, $f_{\text{Transform}}$ 表示对 H 的变换操作, W_{ij} 表示权重矩阵中的元素. 为增强权重矩阵的表达能力, 本文采用 tanh 函数对其进行激活, 生成激活后的权重矩阵 $W_{\text{Activated}}$, 如式 (4) 所示:

$$W_{\text{Activated}} = \tanh\left(W_{\text{Weight}}\right) \tag{4}$$

激活后的权重矩阵 $W_{\text{Activated}}$ 与嵌入向量 H 通过 Hadamard 积结合, 生成过滤因子 F_{Filter} 如式 (5) 所示:

$$F_{\text{Filter}} = W_{\text{Activated}} \otimes H \tag{5}$$

从字词层面出发, 通过结合目标字词及其上下文信息, 生成包含上下文语义的权重矩阵, 如图 2 所示. 给定目标字词 $h_i \in H$, 其上下文信息 h_i^* 通过文本卷积操

作提取, 如式 (6) 所示:

$$h_i^* = \sum_{h_i \in N(h_i)} f_{\text{Conv}}(h_i) \tag{6}$$

其中, $N(h_i)$表示h_i的邻居信息, f_{Conv}为文本卷积函数. 为了适应社交媒体文本中词语长度不固定的特点, 本研究采用多尺度卷积核对文本进行卷积, 生成多尺度卷积特征 C_1, C_2, C_3, 并将其拼接后映射为综合特征, 如式 (7) 所示:

$$C_k = [c_{k1}, c_{k2}, \cdots, c_{kd}] \tag{7}$$

其中, c_{kj}为第 k 个卷积核的卷积结果. 激活后的权重矩阵$W_{\text{Activated-}f1}$如式 (8) 所示:

$$W_{\text{Activated-}f1} = \tanh\left(W_{f1}\left[C_1 \| C_2 \| C_3\right]\right) \tag{8}$$

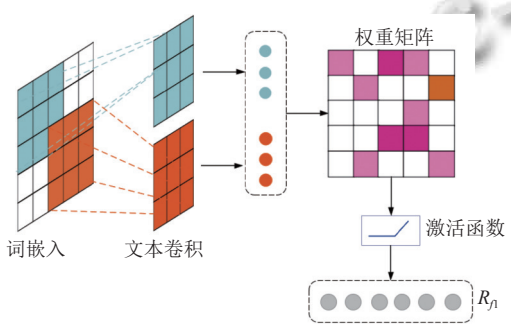

图 2 字词过滤结构

过滤因子$F_{\text{Filter-}f1}$和输出向量R_{f1}分别如式 (9) 和式 (10) 所示:

$$F_{\text{Filter-}f1} = W_{\text{Activated-}f1} \otimes H \tag{9}$$

$$R_{f1} = W_{\text{Filter-}f1}\left[H \| F_{\text{Filter-}f1}\right] + b_{\text{Filter-}f1} \tag{10}$$

在句子层面上, 通过门控循环单元 (GRU) 和评分矩阵提取全局句子特征, 如图 3 所示.

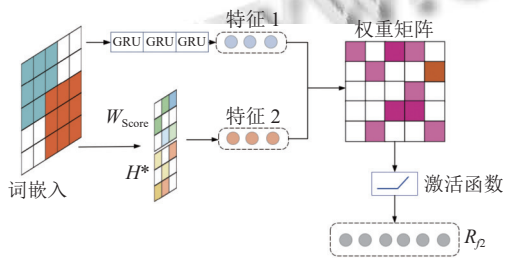

图 3 句子过滤结构

首先通过 GRU 提取句子特征 $Feature_1$, 如式 (11):

$$Feature_1 = GRU(H) \tag{11}$$

其中, GRU 通过逐时间步处理输入序列 H, 生成一个隐藏状态序列, 该隐藏状态序列看作是对整个句子的语义表示.

为了补充 GRU 可能遗忘的信息, 构建评分矩阵W_{Score}, 并将其与嵌入向量H^*相结合, 生成句子特征 $Feature_2$, 如式 (12) 和式 (13) 所示:

$$H^* = W_{\text{Transform}}(H) \tag{12}$$

$$Feature_2 = \left[\alpha_i \beta_j\right]_{n \times n} \tag{13}$$

其中, α_i为H^*中的元素, β_j为评分矩阵中的元素. 线性变换$W_{\text{Transform}}$将嵌入向量 H 映射到一个新的特征空间, 以便更好地与评分矩阵结合. 评分矩阵 $Feature_2$ 用于评估句子的重要性, 从而补充 GRU 可能遗漏的信息.

因此, 激活后的权重矩阵$W_{\text{Activated-}f2}$和过滤因子$F_{\text{Filter-}f2}$, 如式 (14) 和式 (15) 所示:

$$W_{\text{Activated-}f2} = \tanh\left(W_{f2}\left(Feature_1 + Feature_2\right)\right) \tag{14}$$

$$F_{\text{Filter-}f2} = W_{\text{Activated-}f2} \otimes H^* \tag{15}$$

其中, 将 GRU 提取的句子特征 $Feature_1$ 和评分矩阵生成的句子特征 $Feature_2$ 相加, 并通过线性变换W_{f2}和 tanh 激活函数生成激活后的权重矩阵$W_{\text{Activated-}f2}$.

最后, 输出向量R_{f2}计算如式 (16) 所示:

$$R_{f2} = W_{\text{Filter-}f2}\left[H \| F_{\text{Filter-}f2}\right] + b_{\text{Filter-}f2} \tag{16}$$

其中, 通过一个线性层将过滤因子$F_{\text{Filter-}f2}$与原始嵌入向量 H 拼接后映射到最终的输出向量R_{f2}. $b_{\text{Filter-}f2}$是偏置项.

通过R_{f1}和R_{f2}, 本研究实现了从字词和句子两个层面对嵌入向量的过滤. 双层过滤机制的输出为两个过滤单元的输出向量的加权和, 如式 (17) 所示:

$$R = \lambda_1 R_{f1} + \lambda_2 R_{f2} \tag{17}$$

其中, λ_1和λ_2为权重系数, 用于平衡字词层面和句子层面过滤的贡献. 通过双层过滤机制, 模型能够更全面地捕获文本的语义信息, 如算法 1 所示.

算法 1. 双层过滤机制算法

输入: Embedding matrix set H with shape $n \times d$, multi-scale kernel sizes set *kernel_sizes*, GRU hidden dimension parameter *hidden_dim*, fusion weight parameters λ_1, λ_2.
输出: Filtered embedding set *H_final*.

```
1. //对于嵌入矩阵 H 中的每个位置 i
2. For each position i in H:
3.    //如果该位置的字词已处理, 则跳过
3.    IF word_processed_flags[i] == True Then:
4.       Continue
```

```
5.  End IF
6.  //获取位置 i 的上下文
7.  context_window = get_neighbors(H, i, kernel_sizes)
8.  //通过多尺度卷积核生成多尺度卷积特征
9.  C_multi = Concat(
10.     Conv(context_window, kernel_sizes[0]),
11.     Conv(context_window, kernel_sizes[1]),
12.     Conv(context_window, kernel_sizes[2]))
13.  //标记该位置的字词已处理
14.  word_processed_flags[i] = True
15. End For
16. Initialize sentence_processed_flag = False
17. IF sentence_processed_flag == False Then:
18.  //通过 GRU 提取句子特征
19.  gru_features = Equation(11)
20.  //构建评分矩阵
21.  H_proj = Equation(12)
22.  score_matrix = Equation(13)
23.  //计算句子特征
24.  sentence_features = matrix_multiply(H_proj, score_matrix)
25.  W_f2 = tanh(linear_transform(gru_features + sentence_features))
26.  //生成过滤因子 F_sentence
27.  F_sentence = elementwise_multiply(W_f2, H_proj)
28.  R_sentence = linear_layer(Concat(H, F_sentence))
29.  //标记句子已处理
30.  sentence_processed_flag = True
31. End IF
32. //最终输出 H_final 为字词层面和句子层面过滤的加权和
33. H_final = λ1×R_word + λ2×R_sentence
34. Return H_final
```

3.4 知识增强型解码器

知识增强型解码器的主要目标是结合 ERNIE 模型输出的 token 和多维实体, 从而生成高质量的文本摘要. 该解码器由多个解码层构成, 每个解码层包含多头注意力机制和前馈神经网络, 用于对输入信息进行处理和转换.

解码器的输入包括经过双层过滤机制处理后的文本向量 R、ERNIE 模型的 token 输出 $\{w_1^o, w_2^o, \cdots, w_n^o\}$以及多维实体输出$\{e_1^o, e_2^o, \cdots, e_n^o\}$. 为将这些信息整合到统一的特征空间中, 文本向量 R 首先进行线性变换, 如式 (18) 所示:

$$R' = W_R R + b_R \tag{18}$$

其中, W_R是可学习的权重矩阵, b_R是偏置项. 随后, R'与 token 输出和多维实体输出进行拼接, 如式 (19) 所示:

$$I = \left[R' \| \left\{w_1^o, w_2^o, \cdots, w_n^o\right\} \| \left\{e_1^o, e_2^o, \cdots, e_n^o\right\}\right] \tag{19}$$

在每个解码层中, 多头注意力机制被用于捕捉输入信息之间的依赖关系. 多头注意力机制的计算公式如式 (20) 所示:

$$MultiHead(Q,K,V) = Concat(head_1, \cdots, head_h)W^O \tag{20}$$

其中, $head_i = Attention(QW_i^Q, KW_i^k, VW_i^v)$, W_i^Q、W_i^k、W_i^v和W_i^O是可学习的权重矩阵.

对于输入 I, 查询矩阵 Q、键矩阵 K 和值矩阵 V, 使用式 (21)–式 (23) 计算:

$$Q = W_Q I + b_Q \tag{21}$$

$$K = W_K I + b_K \tag{22}$$

$$V = W_V I + b_V \tag{23}$$

其中, W_Q、W_K、W_V是可学习的权重矩阵, b_Q、b_K、b_V是偏置项. 此外, 多头注意力机制的输出经过前馈神经网络进行进一步的非线性变换, 如式 (24) 所示:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_1 \tag{24}$$

解码器的训练损失函数采用交叉熵损失, 并结合 ERNIE 模型训练的损失函数, 共同被用于指导模型的训练, 如式 (25) 所示:

$$L = \alpha L_{\text{cross-entropy}} + \beta L_{\text{ERNIE}} \tag{25}$$

其中, α 和 β 是权重系数, 用于平衡交叉熵损失和 ERNIE 模型损失的贡献. 通过知识增强型解码器, ERNIE 模型提取的语义和知识信息被充分利用, 结合双层过滤机制处理后的文本向量, 高质量的文本摘要被生成.

4 实验与分析

4.1 实验环境

本研究的实验环境配置如表 2 所示.

表 2 实验环境

硬件	配置参数	软件	配置参数
CPU	AMD Ryzen 9 5900X	操作系统	Windows 10 64 位
GPU	NVIDIA RTX 4080 Ti	开发语言	Python 3.9
内存	DDR4 32 GB 3200 MHz	加速库	CUDA 11.4
存储	512 GB NVMe SSD+2 TB HDD	开发框架	PyTorch 1.10

实验选用 AdamW 优化器进行参数更新, 初始学习率设置为 3E–5, 权重衰减系数配置为 0.02 以实现正则化约束. 训练过程中固定批处理尺寸为 64, 共进行 200 个训练周期以充分拟合数据分布. 为防止梯度异常值干扰, 设置梯度裁剪阈值为 2.0, 通过动态调整梯度

幅度保障训练的稳定. 此外, 编码模块基于 RoBERTa-Large-Chinese 预训练模型进行微调, 该架构包含 10 层动态掩码 Transformer 结构, 每层隐藏维度提升至 1 024 以增强表征能力. 微调阶段采用分层学习率策略, 底层参数学习率设置为 1E–5, 顶层参数学习率则调整为 5E–5, 同步解冻最后 3 层 Transformer 参数实现针对性特征适应. 解码器维持 5 层堆叠设计, 每层集成 6 头稀疏注意力机制, 前馈网络隐藏层维度保持 1 024, 激活函数采用 GELU 以优化梯度传播.

4.2 实验数据集

本研究采用 CSL 数据集作为实验数据来源. CSL 数据集是一个广泛用于中文自然语言处理任务的数据集, 涵盖丰富的文本信息, 适用于摘要生成任务[21]. 为确保实验的准确性和可靠性, 本研究对数据进行了严格的筛选和预处理. 具体而言, 本研究从 CSL 数据集中选取了 15 000 条数据作为训练集, 3 000 条数据作为验证集, 以及 2 000 条数据作为测试集.

在数据预处理阶段, 本研究特别关注了数据中的噪声问题. 由于原始数据中可能存在拼写错误、格式不一致或无关字符等噪声, 这些噪声会对模型的训练和性能产生负面影响. 因此, 本研究采用去噪技术, 包括文本清洗、格式标准化以及去除无关字符等, 以提高文本质量.

4.3 评价指标选取

本研究对本文提出的基于过滤机制的多维度文本摘要生成模型 (FM-MDSG), 采用 ROUGE 评价指标中的 ROUGE-*N* 和 ROUGE-L 对生成的摘要进行性能评估. ROUGE-*N* 通过计算生成摘要与参考摘要之间共享的 *n*-gram 数量来评估文本的召回率. ROUGE-L 则基于最长公共子序列 (LCS) 评估生成摘要与参考摘要的语义相似性.

4.4 实验结果分析

实验对比不同数量候选摘要的生成质量后发现, 对比学习策略对摘要生成效果有显著提升. 如图 4 所示, 当候选摘要数量为 8 个时, 其生成质量较 4 个候选摘要的场景有明显改善, 具体表现为信息更完整、语义更连贯以及观点更丰富. 然而, 当候选摘要数量继续增加时, 生成质量的提升幅度逐渐减弱, 表明候选摘要数量存在性能瓶颈.

为全面评估提出 FM-MDSG 算法的有效性, 本文设计并执行了一系列消融实验, 实验结果如表 3 所示.

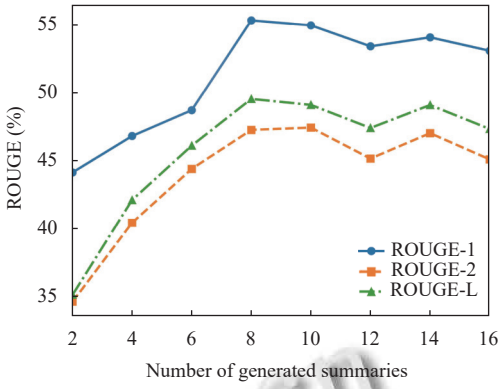

图 4 ROUGE 与摘要生成质量的指标

表 3 消融实验 (%)

移除的组件	ROUGE-1	ROUGE-2	ROUGE-L
RoBERTa	51.64	44.87	45.23
双层过滤机制	46.15	39.42	40.71
ERNIE	47.92	41.35	42.08
完整模型 (未移除任何组件)	**55.37**	**47.28**	**49.56**

如表 3 所示, 当 RoBERTa 组件被移除时, 模型的 ROUGE-1、ROUGE-2 和 ROUGE-L 指标分别被观测到下降至 51.64%、44.87% 和 45.23%, 表明该预训练模型对语义特征的提取具有重要支撑作用. 双层过滤机制的移除导致性能出现最大幅度衰减, 证实该模块在多文档信息筛选和冗余消除过程中发挥核心功能. ERNIE 组件的缺失则使指标降至 47.92%、41.35% 和 42.08%, 验证了其深层语义理解能力对生成质量的关键影响. 完整模型在所有组件保留时获得最优性能, 其中相较于双层过滤机制被移除的情况, ROUGE-1 指标提升达 9.22 个百分点. 实验数据表明, 算法各组件通过协同优化实现了性能增益, 特别是双层过滤机制对噪声抑制和信息整合的贡献最为显著.

为了评估 ERNIE 在不同层数下的性能, 本文进行了详细的对比实验, 结果如表 4 所示.

表 4 ERNIE 不同层数性能对比

层数	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	参数量 (M)	周期 (min)
4	47.82	40.15	41.03	82	38
8	52.16	44.73	45.91	164	65
16	54.89	46.95	48.32	328	104
24	**55.37**	**47.28**	**49.56**	**496**	**142**
32	55.41	47.31	49.63	662	189

表 4 实验结果表明 ERNIE 模型性能与层数呈非线性关系, 在 24 层时达到最优 ROUGE 指标. 当层数从 4 层增至 24 层时, ROUGE-L 提升 8.53 个百分点, 再扩展至 32 层后性能仅微增 0.07 个百分点, 显示出显

著的边际效应递减. 参数量与训练耗时呈现近似线性增长, 24 层模型相比 16 层模型参数量增长 51.2%, 而 ROUGE-L 提升仅 1.24 个百分点.

为进一步验证摘要生成的性能, 本文进行了评价指标的统计分析, 并与其他方法进行了比较. Seq2Seq[22] 作为基于编码器-解码器架构的经典序列生成模型, 其通过循环神经网络实现源序列到目标序列的端到端映射. ChatGLM-6B[23]是参数规模达 60 亿的通用预训练语言模型, 凭借大规模预训练获得了强文本生成能力. Transformer[24]采用标准的多头自注意力机制与位置编码, 建立全局依赖建模的基准框架. DNM[25]通过动态神经架构实现任务自适应参数调整.

图 5 实验数据显示, FM-MDSG 模型以 55.37% 的 ROUGE-1、47.28% 的 ROUGE-2 和 49.56% 的 ROUGE-L 实现了指标全面领先, 较经典 Seq2Seq 基线实现 6.8 个百分点的绝对提升. 该模型不仅超越参数量级更大的 ChatGLM-6B, 相较当前 DNM 模型也提升 2.31 个百分点.

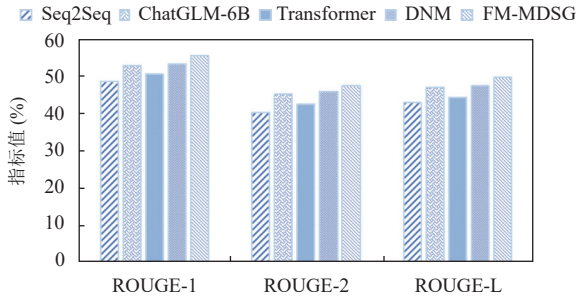

图 5 对比实验

此外本文还对 FM-MDSG 模型的参数量和推理复杂性进行了分析, 如表 5 所示. 可以看出, FM-MDSG 模型的参数量为 496M, 处于中等水平. 虽然参数量高于 Seq2Seq 和 Transformer, 但远低于大规模预训练模型 ChatGLM-6B. 在推理时间方面, FM-MDSG 模型的推理时间为 142 ms, 略高于 Transformer 和 DNM, 但显著低于 ChatGLM-6B. 这表明 FM-MDSG 模型在保持较高性能的同时, 具有相对较低的推理复杂性, 符合实际应用中的实时需求.

5 结论与展望

本文提出了一种基于过滤机制的多维度文本摘要生成模型 (FM-MDSG), 以应对多源异构信息冗余和观点碎片化的挑战. 实验结果表明, FM-MDSG 在 CSL 数据集上的 ROUGE-1/2/L 的 F 值分别为 55.37%、47.28% 和 49.56%, ROUGE-L 较经典基线模型提升了 6.8 个百分点. 消融实验进一步验证了双层过滤机制对噪声抑制和信息整合的重要作用, 带来了 9.22% 的 ROUGE-1 性能增益. 此外, ERNIE 的不同层数实验结果显示, 在 24 层时模型性能达到最优, 继续增加层数并未显著提升性能, 体现了模型的高效性与经济性.

表 5 ERNIE 不同层数性能对比

模型	参数量 (M)	推理时间 (ms)
FM-MDSG	**496**	**142**
Seq2Seq	50	80
ChatGLM-6B	6 000	300
Transformer	100	120
DNM	200	150

未来将引入用户反馈机制, 根据用户的偏好和需求动态调整摘要生成策略, 提升个性化服务水平. 此外, 随着自然语言处理技术的不断进步, FM-MDSG 可在更广泛的领域得到应用, 如新闻推荐、舆情分析等, 为用户提供更加精准的信息服务.

参考文献

1 Pal S, Chang MG, Iriarte MF. Summary generation using natural language processing techniques and cosine similarity. Proceedings of the 21st International Conference on Intelligent Systems Design and Applications. Cham: Springer, 2021. 508–517.

2 Kayal P. Resiliency improvement in power distribution infrastructure employing distributed generation and switches—A review summary. Energy, Ecology and Environment, 2023, 8(3): 195–210. [doi: 10.1007/s40974-023-00272-x]

3 Morris JA, Boshoff CH, Schor NF, *et al*. Next-generation strategies for gene-targeted therapies of central nervous system disorders: A workshop summary. Molecular Therapy, 2021, 29(12): 3332–3344. [doi: 10.1016/j.ymthe.2021.09.010]

4 程五焰. 多源信息融合的生成式摘要研究 [硕士学位论文]. 武汉: 华中师范大学, 2022.

5 Rehman T, Sanyal DK, Chattopadhyay S. Research highlight generation with ELMo contextual embeddings. Scalable Computing: Practice and Experience, 2023, 24(2): 181–190. [doi: 10.12694/scpe.v24i2.2238]

6 Holttinen H, Kiviluoma J, Helistö N, *et al*. Design and operation of energy systems with large amounts of variable generation: Final summary report, IEA Wind TCP Task 25. VTT Technical Research Centre of Finland, 2021.

7 Guo Y, Qiu W, Leroy G, *et al*. Retrieval augmentation of large language models for lay language generation. Journal

of Biomedical Informatics, 2024, 149: 104580. [doi: 10.1016/j.jbi.2023.104580]
8 Jin HJ, Guo JL, Lin QS, *et al*. Comparative study of Claude 3.5-Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: Assessment of efficiency, accuracy, and quality. Frontiers in Digital Health, 2024, 6: 1456911.
9 Li JY, Tang TY, Zhao WX, *et al*. Pre-trained language models for text generation: A survey. ACM Computing Surveys, 2024, 56(9): 230.
10 Yehudai A, Carmeli B, Mass Y, *et al*. Genie: Achieving human parity in content-grounded datasets generation. arXiv:240114367, 2024.
11 Mohan GB, Kumar RP. Lattice abstraction-based content summarization using baseline abstractive lexical chaining progress. International Journal of Information Technology, 2023, 15(1): 369–378. [doi: 10.1007/s41870-022-01080-y]
12 Yan J, Zhou S. A text structure-based extractive and abstractive summarization method. Proceedings of the 7th International Conference on Intelligent Computing and Signal Processing (ICSP). Xi'an: IEEE, 2022. 678–681.
13 Kwon S, Lee Y. Enhancing abstractive summarization of implicit datasets with contrastive attention. Neural Computing and Applications, 2024, 36(25): 15337–15351. [doi: 10.1007/s00521-024-09864-y]
14 Guan Y, Guo SR, Li R. Structure-to-word dynamic interaction model for abstractive sentence summarization. Neural Computing and Applications, 2025, 37(9): 6567–6581. [doi: 10.1007/s00521-024-10970-0]
15 Abd Algani YM. A novel deep learning attention based sequence to sequence model for automatic abstractive text summarization. International Journal of Information Technology, 2024, 16(6): 3597–3603. [doi: 10.1007/s41870-024-01934-7]
16 Yadav M, Katarya R. Stacked denoising variational auto encoder model for extractive web text summarization. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2024, 48(4): 1501–1518. [doi: 10.1007/s40998-024-00751-9]
17 Alselwi G, Taşcı T. Extractive arabic text summarization using pagerank and word embedding. Arabian Journal for Science and Engineering, 2024, 49(9): 13115–13130. [doi: 10.1007/s13369-024-08890-1]
18 Wang T, Yang C, Zou MY, *et al*. A study of extractive summarization of long documents incorporating local topic and hierarchical information. Scientific Reports, 2024, 14(1): 10140. [doi: 10.1038/s41598-024-60779-z]
19 Gangundi R, Sridhar R. RBCA-ETS: Enhancing extractive text summarization with contextual embedding and word-level attention. International Journal of Information Technology, 2025, 17(2): 1127–1135. [doi: 10.1007/s41870-024-02192-3]
20 Monir E, Salah A. AraTSum: Arabic Twitter trend summarization using topic analysis and extractive algorithms. International Journal of Computational Intelligence Systems, 2024, 17(1): 227. [doi: 10.1007/s44196-024-00546-0]
21 Li YD, Zhang YQ, Zhao Z, *et al*. CSL: A large-scale Chinese scientific literature dataset. arXiv:2209.05034, 2022.
22 Anand A, Nair AR, Prasad K, *et al*. Advances in citation text generation: Leveraging multi-source Seq2Seq models and large language models. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise: ACM, 2024. 56–64.
23 Qi J, Yan S, Zhang Y, *et al*. RAG-optimized Tibetan tourism LLMs: Enhancing accuracy and personalization. arXiv:2408.12003, 2024.
24 Ghadekar P, Khanwelkar D, More H, *et al*. Transformer based text summary generation for videos. Proceedings of the 2024 International Conference on Current Trends in Advanced Computing (ICCTAC). Bengaluru: IEEE, 2024. 1–7.
25 Xu K, Liu B, Li JQ, *et al*. Automatic text summary generation method based on hybrid model DNM. Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Melbourne: IEEE, 2021. 637–642.

(校对责编: 王欣欣)