基于大语言模型的文本摘要生成方法

孙其航[1] 荆晓远[2*] 杜杰宾[1]
SUN Qihang JING Xiaoyuan DU Jiebin

摘 要 摘要生成是人工智能领域的重要研究方向，直接影响自然语言处理的实用价值和效果。传统方法过分关注文本序列化信息，忽视内部逻辑结构，导致生成摘要连贯性不足，难以满足实际需求。针对这一问题，文章提出了一种基于大语言模型的新型摘要生成方法，并构建了专门的指令数据集和 1 000 条高质量评测样本。实验结果表明，该方法在 ROUGE-1、ROUGE-2 和 ROUGE-L 三个评价指标上均优于现有同规模最先进模型。通过 DeepSeek 大语言模型的进一步验证，充分证明了所提方法在提升摘要质量和逻辑性方面的显著优势，为自然语言处理任务提供了新的解决方案。

关键词 摘要生成；大语言模型；下游任务微调；DeepSeek；自然语言处理

doi：10.3969/j.issn.1672-9528.2025.08.001

0 引言

伴随信息技术的快速发展，学术界正经历一场前所未有的技术革新浪潮。以计算机视觉领域的国际顶级会议 CVPR 为例，其发文量从 2016 年的 633 篇增至 2024 年的 2 719 篇，增长了约 4.3 倍。这庞大的论文数据量，使研究者面临前所未有的阅读压力。使用摘要生成工具辅助阅读已成为研究者的普遍选择。然而，现有方法在信息量与简洁性之间难以实现理想平衡，致使读者获取关键信息仍需投入大量时间。此外，大多数传统摘要生成方法仅关注序列化特征提取，难以深入挖掘研究论文的核心内容。在这类复杂文献中生成既准确又简洁连贯的摘要内容，已成为摘要生成领域面临的关键挑战。本文针对摘要生成不准确、不连贯等问题，提出了基于大语言模型进行文本摘要生成任务的方法，提高了摘要生成的准确性与连贯性。

摘要生成工作主要可分为抽取式和生成式两大类。抽取式摘要是从原文中直接抽取词语、短语或句子，然后通过排序和重组这些内容来生成文章的摘要。在经典的抽取式方法中，Mihalcea 等人 [1] 提出的 TextRank 算法尤为突出。这一方法受 PageRank 算法 [2] 启发，将文本中的单词或句子表示为图中的节点，并根据句间相似度构建节点之间的连接关系，从而形成文本的图结构，基于此进行摘要抽取。Wong 等人 [3] 则提出了一种学习式方法，将句子特征划分为表面特征、内容特征、相关特征和事件特征四类，旨在通过融合多种特征来提升摘要生成效果。Bichi 等人 [4] 开发了一种创新的基于图的单文档抽取式方法，通过改进传统 PageRank 算法，更好地捕捉了文本中的关键信息。随后的研究中，深度学习技术被引入到文本摘要领域。例如，Nallapati 等人 [5] 提出的 SummaRuNNer 模型采用循环神经网络构建序列模型，专为抽取式摘要设计，展现出较高的性能和良好的可解释性。此外，Liu 等人 [6] 开发了一种基于 BERT 的新型文档级编码器，通过堆叠多层句间 Transformer 结构，有效识别并提取关键信息。

虽然抽取式摘要生成技术能够大致表达文档内容，但其局限于原文信息，无法生成超出文档的叙述和推理。这一限制促使研究者转向生成式摘要生成技术的探索。在神经序列到序列（seq2seq）模型推动了抽象式文本摘要发展的背景下，See 等人 [7] 针对该类模型面临的挑战，提出了结合指针 - 生成器网络（pointer-generator network）和覆盖机制（coverage mechanism）的方法，从而显著提升了文本摘要的质量。Liu 等人 [8] 则提出了一种基于对抗训练的抽象式文本摘要方法，该方法能够生成更加抽象、可读且多样的摘要。针对现有方法难以产生类似人类书写风格的抽象式摘要这一问题，Yang 等人 [9] 提出了一个新颖的层次化人类认知启发的深度神经网络模型（hierarchical human-like deep neural network for ATS, HH-ATS）。该模型通过模拟人类的粗略阅读、主动阅读与后编辑 3 个阶段，生成更符合人类阅读习惯的摘要。

通过回顾上述研究，可以清晰地看到，各项研究均致力于生成更加接近人类水平的摘要。这种追求不仅体现在对抽象式摘要的探索中，也反映在对生成式摘要技术的不断优化上。近年来，大语言模型在自然语言处理领域展现出了前所未有的潜力，特别是在自然语言理解与生成方面表现尤为突

1. 吉林化工学院信息与控制工程学院 吉林吉林 132022
2. 广东石油化工学院计算机学院和省市共建石化装备智能安全广东省重点实验室 广东茂名 525000

[基金项目] 国家自然科学基金项目“基于类不平衡深度特征学习的石化动设备故障信号分类研究”（62176069）；广东省自然科学基金项目“软件缺陷预测关键技术和新方法研究”（2023A1515012653）

出。以 ChatGPT 为代表的大型预训练语言模型，凭借其强大的自动生成、编辑和完善文本的能力，为摘要生成任务带来了革新性的进展。随着 Llama、DeepSeek-R1、ChatGLM 等多种大语言模型的开源发布，研究人员获得了更加丰富的工具和数据支持，这些资源极大推动了摘要生成技术的发展。因此，大型语言模型不仅在参数规模和模型架构上取得了显著提升，更重要的是能够通过自注意力机制深入捕捉文本中的长距离依赖关系，从而生成更加连贯、准确和自然的摘要。这种能力的提升为研究人员提供了更强大的实验平台，使得在抽象式摘要、多模态摘要等前沿方向上的探索变得更加可行。朱丹浩等人 [10] 探索了使用大语言模型进行法律文本摘要生成的可行性，取得了很好的结果。随着大语言模型技术的不断进步，我们有理由相信，未来摘要生成技术将迎来更多突破，为自然语言处理和人工智能领域带来更大的应用价值。

1 模型构建方法

1.1 方法总体框架

本研究主要由 3 个核心部分构成，其框架设计如下：

首先是数据集的构建与优化模块。本部分基于深度学习发展的实际需求，着重打造高质量、多样化的数据集。据统计，在深度学习算法工程师的工作中，数据处理占据了总时间的 90%。因此，本研究特别注重数据清洗、标注及增强，以确保数据质量，为后续模型训练奠定坚实基础。

其次是模型微调与优化模块。本研究选取了 Llama 3.2 3B-base 作为基线模型，基于构建的高质量数据进行了针对性的指令微调训练。通过这种方式，显著提升了该模型在摘要凝练任务上的表现力，使其能够更好地适应具体应用场景。

最后是模型评估与验证模块。本研究采用了双管齐下的评估策略：一方面运用经典的 ROUGE（recall-oriented understudy for gisting evaluation）指标进行自动化评价，另一方面引入 DeepSeek R1 70B 作为评价器，模拟专家对生成摘要质量进行打分。这种多维度、全方位的评估体系能够更全面地反映模型的实际性能，为结果分析提供了坚实保障。本文的总体框架如图 1 所示。

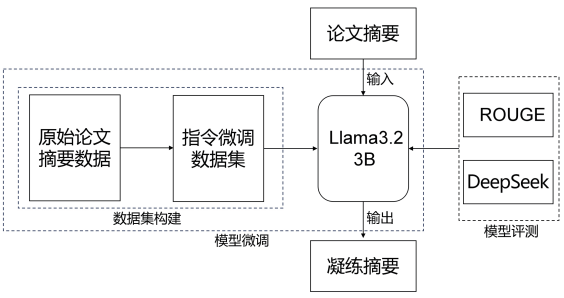

图 1 方法总体框架图

1.2 数据集构建部分

在深度学习领域，最核心的基础工作之一就是高质量数据集的构建。本研究专注于论文摘要凝练任务，为此，构建了一个大规模、高质量的数据集。具体来说，该数据集包含 67 万条摘要 - 标题对的训练样本，以及 1 000 条独立的测试样本。数据来源于知名的开源学术平台 arXiv。首先，利用 arXiv 提供的开放 API 接口，批量获取了大量论文的元数据，这些数据包括论文标题、摘要、作者信息以及出版时间等内容。在获取原始数据后，进入关键的数据处理阶段：通过 Python 脚本对爬取到的数据进行系统化的清洗和整理，去除与摘要凝练任务无关的冗余信息。随后，将处理后的数据按照指令微调的标准格式进行重构，最终形成如下结构：

{"Instruction": " 用户指令 "; "Input": " 论文摘要 "; "Output": " 论文标题 "} 的形式。此外，为确保模型能够更好地适应实际应用场景，对部分样本进行了人工优化和校准，以进一步提升数据集的质量与实用性。

1.3 模型微调训练

Meta AI 作为全球领先的人工智能科技公司，其在人工智能领域的技术实力始终保持世界领先水平。基于其在多项评测基准中的卓越表现，本研究选择了 MetaAI 最新开源的 LLAMA 3.2-3B 模型作为基础模型。这一选择不仅得益于该模型在多个评测指标上的顶尖表现，还考虑到其 3B 参数规模能够更好地匹配本研究所使用的数据量和算力资源。为优化微调训练过程中的计算效率，本研究采用了 LoRA（low-rank adaptation）[11] 方法进行模型微调。LoRA 是一种高效的参数适应方法，其核心思想是通过保持预训练权重不变，仅调整两个低秩矩阵来实现对模型权重的更新近似。这种方法显著降低了微调所需的计算资源，同时保证了模型性能的提升。

1.4 模型评测部分

模型评测部分由传统方法与大模型评测方法两部分组成，传统方法主要采用文本摘要成领域常用的 ROUGE 评测方法。而大模型评测方法主要是模仿人类进行客观评价，采用的大模型为 DeepSeek R1 70B 详细介绍见实验部分。

2 实验

本研究实验设计包含两个主要部分。首先，采用传统的 ROUGE 评估指标，这一部分主要用于量化评估生成摘要与参考标题在文本内容上的相似度。其次，为更全面地评估模型性能，本研究还引入了基于大型语言模型的智能评测系统。通过这两部分实验，旨在从不同的维度对本文提出的摘要凝练方法进行综合评估，确保评价结果既有传统指标的客观性，又能结合先进模型的智慧判断，从而得出更为全面的结论。

2.1 ROUGE 评价方法

ROUGE 评价方法是在自然语言处理领域，特别是在文本摘要任务中最广泛应用的评价指标之一。该方法通过计算生成摘要与参考摘要之间的 n-gram 重合情况来量化评估摘要质量。具体来说，ROUGE 分为 3 种主要类型：ROUGE-1、

ROUGE-2 和 ROUGE-L，分别衡量生成摘要与标准摘要在 1 元、2 元，以及最长公共子序列上的相似度。其核心计算公式为：

$$\mathrm{ROUGE}-N=\frac{\sum_{S\in\{\mathrm{ReferenceSummaries}\}}\sum_{\mathrm{gram}_n\in S}\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)}{\sum_{S\in\{\mathrm{ReferenceSummaries}\}}\sum_{\mathrm{gram}_n\in S}\mathrm{Count}(\mathrm{gram}_n)} \quad (1)$$

该公式通过分子和分母进行比例计算，其中分子表示生成摘要中 n-gram 的数量，分母则代表标准摘要中 n-gram 的总数，这种设计能够有效评估生成内容与参考内容的一致性程度。在本研究中，构建了一个包含 1 000 条数据样本的评测数据集，用于对比评估多个先进模型的性能。具体来说，本文选取了 Llama 3.2-3B-base 模型与 ChatGLM-4-9B-Chat 模型进行对比实验。各模型在 ROUGE 指标下的详细得分数据如表 1 所示，根据表中数据可分析和比较不同模型在摘要生成任务中的表现差异。

表 1 ROUGE 得分

模型	ROUGE-1	ROUGE-2	ROUGE-L
Llama3.2 3B-base	0.16	0.07	0.15
Chatglm4 9B-Chat	0.24	0.13	**0.32**
Ours(3B)	**0.27**	**0.18**	0.26

2.2 大模型评价方法

随着大规模预训练语言模型（LLMs）在自然语言处理领域的迅速崛起，这些模型在多个任务中展现出了强大的应用潜力，包括在诸多场景下能够接近或超越人类水准的性能。其中，在众多应用场景中，摘要生成质量评价已成为一个重要研究方向。通过精心设计的提示引擎，大规模语言模型能够有效评估摘要的质量。与传统的人工评价方法相比，这种基于大模型的评价方式更加客观公正，因为其评估结果完全基于海量数据训练和固定的算法规则，能够有效避免人工评价中可能存在的个体偏好、情绪波动等主观因素。在本研究中，采用了国内发展最先进、能力最全面的大规模预训练语言模型——DeepSeek R1 进行摘要质量评估。由深度求索公司开发的 DeepSeek R1 作为目前为止最强大的国产大模型，其在多项任务中的性能已达到与国际顶尖水平相当的水准。这一模型具备了与人类专家不相上下的学术评估能力，能够模拟专业评价者的思维方式，为摘要质量的客观评估提供了可靠保障。下面通过该模型进行大规模评估实验来证明这一方法的有效性和可行性。

（1）提示词

请评估以下机器生成摘要的质量：

原文 : Block diagonalization (BD) is a linear precoding technique for the multi-user MIMO-OFDM systems that involves transmission of multiple data streams to each receiver such that no multi-user interference is experienced at any of the receivers. With perfect channel knowledge at the transmitter, BD can eliminate other users' interference at each receiver. But it is not practical for receiver to feedback the perfect CSIT to transmitter. In order to reduce the feedback overhead, a novel Limited Feedback and Non-Codebook based Block Diagonalization (LFNCBD) scheme is proposed in this paper. Simulation results show that the proposed scheme could achieve the asymptotical optimal BD sum-capacity performance with limited feedback overhead.

参考摘要 : A Limited Feedback-Based Block Diagonalization Scheme for Multi-User MIMO-OFDM Systems

模型生成摘要 : Limited Feedback and Non-Codebook based Block Diagonalization for Multi-user MIMO-OFDM Systems

请从以下几个方面评分 (1-5 分) 并给出简要解释：

① 准确性 - 摘要是否准确反映了原文的主要内容。

② 简洁性 - 摘要是否简明扼要。

③ 完整性 - 摘要是否包含了所有重要信息。

请以 JSON 格式返回评分和解释，格式如下：

```
{{
  "accuracy": {{"score": 4, "explanation": " 解释 "}},
  "conciseness": {{"score": 4, "explanation": " 解释 "}},
  "completeness": {{"score": 4, "explanation": " 解释 "}}
}}
```

（2）模型回答

<think>

好的，我现在要评估模型生成的摘要质量。首先，准确性方面，原文提到了 LFNCBD 方案，这个摘要里也包含了这些关键词，没有遗漏重要信息，所以准确性应该很高。

接下来是简洁性。参考摘要更短，而模型生成的稍长，但依然保持了简明，不啰嗦，所以简洁性也不错，虽然可能可以再精简一点。

关于完整性，原文中的关键点都涵盖在内，比如 Limited Feedback、Non-Codebook 和 Block Diagonalization 等，所以摘要是完整的，没有缺失主要信息。

</think>

```
{
  "accuracy": {
      "score": 4,
      "explanation": " 模型生成的摘要准确地反映了原文的主要内容，包括提出了基于有限反馈和非代码本的区块对角化方案，并适用于多用户 MIMO-OFDM 系统。"
  },
  "conciseness": {
      "score": 4,
      "explanation": " 摘要简洁明了，没有冗余信息，有效地概括了原文的核心内容。"
  },
  "completeness": {
    "score": 4,
```

"explanation": " 摘要包含了所有重要信息，包括技术名称、关键方法和应用场景。"
}
}

本研究中所采用的评估案例均来源于本文模型生成的样本。通过调用本地部署的 DeepSeek R1 70B 模型 API，采用批量评估的方式对这些样本进行了自动化质量评价。通过对各维度评估结果取算术平均值，得出了如表 2 所示的最终评价结果。通过对两部分评分结果的综合分析表明，本研究所开发的模型在性能上显著超越了基准模型。在特定实验条件下，本模型甚至优于参数规模远大于自身的其他模型，这充分验证了本研究所采用的方法论具有显著优势。

表 2 大模型评分

模型	准确性	简洁性	完整性
Llama3.2 3B-base	3.86	3.73	3.82
Chatglm4 9B-chat	4.38	4.17	4.43
Ours(3B)	4.13	4.11	4.08

3 结语

本研究提出了一种面向摘要生成任务的新方法，该方法基于 Llama3.2-3b 模型进行了优化。利用自建多样化数据集对该模型进行微调训练，以提升其在摘要生成任务中的表现。通过ROUGE指标和DeepSeek R1模型进行的双重评估表明，本研究提出的方法在相同规模下显著优于其他现有的模型。这些结果充分证明了本方法的有效性。然而，当前工作仍存在一些局限性。例如，模型在处理多文档摘要生成和长篇文本摘要时表现不够理想，这可能限制了其在实际应用中的适用范围。为此，未来计划将研究重点扩展至多文本摘要生成与长文本摘要领域，并进一步优化模型架构，以提升其处理复杂场景的能力。

参考文献：

[1] MIHALCEA R, TARAU P. TextRank: bringing order into text. in proceedings of the 2004 conference on empirical methods in natural language processing[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.Brussels:ACL,2004:404-411.

[2] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[EB/OL]. (2007-01-31)[2024-05-25].https://gwern.net/doc/technology/google/2007-klein-slides.pdf.

[3] WONG K F, WU M L, LI W J.Extractive summarization using supervised and semi-supervised learning[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Brussels:ACL,2008:985-992.

[4] BICHI A A, SAMSUDIN R, HASSAN R, et al. Graph-based extractive text summarization method for hausa text[J]. Public library of science one, 2023, 18(5): 285376.

[5] NALLAPATI R, ZHAI F F, ZHOU B W. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents[EB/OL].(2016-11-14)[2024-05-25].https://doi.org/10.48550/arXiv.1611.04230.

[6] LIU Y, LAPATA M. Text summarization with pretrained encoders[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Brussels: ACL, 2019: 3730-3740.

[7] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Brussels: ACL, 2017: 1073-1083.

[8] LIU L Q, LU Y, YANG M, et al. Generative adversarial network for abstractive text summarization[C/OL]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto:AAAI, 2018[2024-12-25]. https://aaai.org/papers/12141-generative-adversarial-network-for-abstractive-text-summarization/.10.1609/aaai.v32i1.12141.

[9] YANG M, LI C M, SHEN Y, et al. Hierarchical human-like deep neural networks for abstractive text summarization[J]. IEEE transactions on neural networks and learning systems, 2021, 32(6): 2744-2757.

[10] 朱丹浩，黄肖宇，李尧霖，等 . 基于大语言模型的法律文本的自动摘要方法 [J/OL]. 数据分析与知识发现 ,1-23[2025-02-17].http://kns.cnki.net/kcms/detail/10.1478.G2.20241013.1125.002.html.

[11] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL].(2021-10-16)[2024-05-12].https://doi.org/10.48550/arXiv.2106.09685.

【作者简介】

孙其航（1999—），男，山东泰安人，硕士研究生，研究方向：大语言模型领域微调、大模型 agent，email:qihang.sun@outlook.com。

荆晓远（1971—），通信作者（email:jingxy_2000@126.com），男，江苏南京人，博士，教授、博士生导师，研究方向：模式识别、计算机视觉、故障诊断。

杜杰宾（2000—），男，广东深圳人，硕士研究生，研究方向：故障诊断、类不平衡，email:dujiebin_29@126.com。

（收稿日期：2025-02-22 修回日期：2025-07-01）