面向低资源场景、基于Mamba-Transformer混合模型的高效抽取式摘要

Nisrine Ait Khayi独

立研究员，孟菲斯大学

arXiv:2603.01288v1 [cs.CL] 1 Mar 2026

摘要

长文档的抽取式摘要受二次复杂度制约，往往被迫截断，从而限制了其在资源受限环境中的部署。我们提出首个用于抽取式摘要的Mamba–Transformer混合模型，将预训练Transformer的语义优势与状态空间模型的线性时间处理相结合。借助Mamba无需截断即可处理完整文档的能力，我们的方法在保持较强摘要质量的同时保留上下文。该架构包括：（1）用于句子级语义的Transformer编码器；（2）用于高效捕获句间依赖的Mamba状态空间模型；（3）用于句子相关性预测的线性分类器。在低资源条件下，跨新闻、议论文和科学领域，我们的方法取得了：（1）相较BERTSUM和MATCHSUM的大幅提升，包括在ArXiv上ROUGE-1提高 +0.23，以及在所有数据集上具有统计显著性的改进（$p < 0.001$）；（2）跨领域的一致优势，在最长文档上最为显著；（3）在训练数据有限时仍表现稳健。我们提出了首个用于摘要的混合Transformer–状态空间架构，在低资源场景中展现出显著的ROUGE提升与效率增益；以及（4）在新闻摘要（CNN/DailyMail）上快24–27%。

1 引言

文本摘要是自然语言处理（NLP）中的一项关键任务，旨在生成包含原始文档中最重要信息的简洁文本。新闻摘要对于数字时代的信息消费已变得至关重要，其方法涵盖从抽取式方法（如（Dong等，2025年；Le和Luu，2024年）到生成式技术（Bao等，2025年；Zhang等，2024年）。近期的深度神经网络方法取得了出色的性能，但需要大量计算资源和大规模训练数据（Zhang等，2023年；An等，2022年；Sun和Li，2021年）。然而，这些进展面临严峻挑战，尤其是在资源受限的环境中。具体而言，基于Transformer的摘要模型严重依赖自注意力机制，并且相对于序列长度存在二次计算复杂度，从而被迫进行截断。因此，训练和推理都需要大量计算资源，同时还需要创建成本高昂的超大规模标注数据集。为了在保持有效语义理解的同时克服这些计算障碍，我们首次提出一种用于抽取式摘要的Mamba-Transformer混合架构。在该方法中，预训练 Transformer编码器生成丰富的句子级表示。随后，这些表示经由Mamba的线性复杂度状态空间模型处理，以在不进行截断的情况下处理全文文档并捕捉句间依赖。该设计保留了完整的上下文信息，降低了推理成本，并且对科学论文等超长文档尤为有效；在这类文档上，Transformer往往无法保留全部上下文。本文的主要贡献如下：

1. 我们提出了首个用于抽取式摘要的 Mamba-Transformer 混合模型。

2. 我们展示了线性复杂度扩展，且无需截断，从而能够高效地对长文档进行摘要。

3. 我们提供了跨领域的低资源评估，并在新闻、议论文和科学数据集上进行了全面实验。

4. 在 Arxiv 上实现了 +0.23 的 ROUGE-1 提升，取得显著增益（p < 0.001），并使新闻推理速度提升 24-27%。

5. 我们展示了该方法在长文档场景下的稳健性。最长文档上的增益最大，证明了该架构在序列长度和上下文保留至关重要的场景中的优势。

2 相关工作

2.1 基于 Transformer 的自然语言处理中的计算挑战

基于 Transformer 的模型虽然取得了出色的自然语言处理性能，但其计算复杂度呈二次增长，限制了它们在资源受限环境中的部署（Varis 和 Bojar，2021年）。近期的高效方法包括剪枝（Gordon 等人，2020年）、知识蒸馏（Zhu 等人，2022年）和量化（in 等人，2024年）。我们采用量化感知微调，在资源受限的条件下开展实验。

2.2 状态空间模型作为高效替代方案

Mamba 状态空间模型（Gu 和 Dao，2023年）具有线性 O(L) 复杂度，因此在序列建模方面颇具吸引力。近期研究展示了其在文本排序（Xu 等人，2025年）和机器翻译（Pitorro 等人，2024年）中的潜力，并且能以更少的资源达到与 Transformer 相当的性能。然而，它们在抽取式摘要中的有效性仍未得到充分探索。

2.3 抽取式摘要

基于 Transformer 的方法主导着抽取式摘要。BERTSUM（Liu，2019年）确立了句子级抽取范式，而 MATCHSUM（Zhong et al.，2020年）引入了语义匹配。这些方法在低资源环境中缺乏竞争力。近期方法如 COLO（An etal.，2022年）和 DiffuSum（Zhanget al.，2023年）取得了优异表现，但继承了二次复杂度，限制了其在低资源环境中的适用性。近期研究通过解耦表示学习（Yuanetal.，2023年）和释义改写（Tang etal.，2023年）来应对低资源场景。我们的贡献有所不同：从架构层面提升计算效率，提出首个混合方法，在保持竞争力的同时实现线性复杂度。

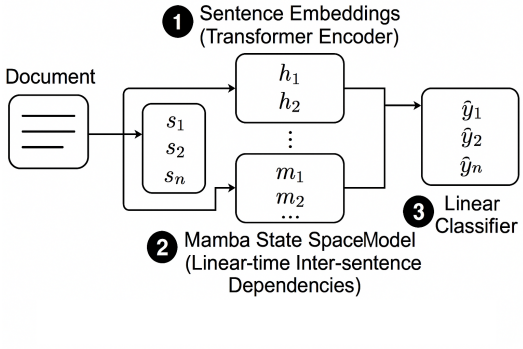


图 1：所提出的用于抽取式摘要的 Mamba–Transformer 混合架构概览。

3 方法

图 1 展示了我们提出方法的架构。

3.1 问题定义

我们将抽取式摘要表述为文档级句子分类任务。给定由n 个句子组成的文档 $D = \{s_1, s_2, \ldots, s_n\}$，我们的目标是在考虑文档上下文中句间关系的同时，将每个句子分类为与摘要相关（$y_i = 1$）或不相关（$y_i = 0$）。

3.2 架构概述

我们的方法由三个关键组件构成，用于在句子序列级处理文档。

3.2.1 使用 Transformer 进行句子编码

对于每篇文档 $D = \{s_1, s_2, \ldots, s_n\}$，我们首先应用一个预训练 Transformer（BERT），为每个句子独立生成语义嵌入 [CLS]：

$$h_i = \text{Transformer}(s_i) \in \mathbb{R}^d$$

其中，h_i 表示句子 s_i的句子级语义编码。

3.2.2 使用 Mamba 进行文档级序列处理

接着，我们构建句子嵌入序列 $H = [h_1, h_2, \ldots, h_n] \in \mathbb{R}^{n \times d}$，并通过 Mamba 状态空间模型处理该序列，以捕捉文档中的句间依赖：

$$M = \text{Mamba}(H) = [m_1, m_2, \ldots, m_n] \in \mathbb{R}^{n \times d}$$

数据集	文档平均长度（词元）	每篇文档平均句子数	正向（%）
CNN/DailyMail	781 29.4		18.2
DebateSum	643	24.1	22.7
ArXiv	1,247	41.8	16.9

表 1：数据集统计信息。

此步骤使模型能够捕捉文档结构和句子关系，同时保持相对于句子数量的线性计算复杂度 $\mathcal{O}(n)$。

3.2.3 句子相关性的二分类

最后，我们对 Mamba 输出序列中的每个位置应用线性分类器，以预测句子级别的相关性：

$$\hat{y}_i = \sigma(W \cdot m_i + b)$$

其中，m_i 是句子 i的 Mamba 编码，$W \in \mathbb{R}^d$ 是权重矩阵，b 是偏置项，σ 是 Sigmoid函数。这将生成文档中每个句子的相关性概率。

4 实验设置

4.1 数据集

受图形处理器限制，我们在三个不同的数据集上评估了 Mamba-Transformer 模型，每个数据集使用 200 篇文档。这些数据集代表了现实中的低资源场景，在此类场景中，大规模标注和计算资源均受到限制。**数据集**：

- CNN/DailyMail（Hermann 等人，2015年）：带有多句摘要的新闻文章
- DebateSum（Roush 和 Balaji，2020年）：需要理解议论文结构的辩论证据
- ArXiv（Cohan 等人，2018年）：包含摘要的科学论文，非常适合评估可扩展性

预处理：我们采用贪心 ROUGE-2 最大化方法生成句子级二元标签。为提升数据集多样性，我们基于 TF-IDF 表示进行 K-means 聚类（500 个簇），然后随机选取 200 篇文档。文本使用 NLTK 进行切分，使用 HuggingFace 进行分词，并截断至 128 个词元。文档按文档级别以 120/40/40（训练/验证/测试）划分，以防止数据泄漏。

4.2 实施细节

我们的架构由 BERT-base-uncased 编码器（1.1 亿参数，768 维嵌入）组成，其后接 Mamba-130M 状态空间模型（NF4 量化，$d_{\text{state}} = 16$，$d_{\text{conv}} = 4$）以及一个将 768 维映射到 1 维并采用 Sigmoid 激活函数的线性分类器。训练使用 Adam 优化器（学习率 1×10^{-5}）、二元交叉熵损失、批大小为 1 并采用梯度累积，共训练 2 个轮次。我们应用了 0.2 的丢弃法，以及最大_范数 $= 1.0$ 的梯度裁剪。数据预处理使用 BERT 分词器，每个句子的最大长度为 128 个词元，采用 ROUGE-2 贪心标注进行监督，且不对文档进行截断。实验在 NVIDIA A100 图形处理器（40GB）上进行，使用 PyTorch 2.6.0+cu124 和 HuggingFace Transformers 4.35.0，随机种子为 $\{42, 123, 456\}$。评估使用 ROUGE-1、ROUGE-2 和 ROUGE-L F1 分数，并进行配对 t 检验显著性检验。

5 实验结果

5.1 总体性能

在三个领域中，MAMBA-BERT 的 ROUGE 分数始终优于纯 Transformer 基线。在 CNN/DailyMail（新闻）上，其得分达到 0.61/0.51/0.53（ROUGE-1/2/L），显著高于 BERTSUM（+0.17 ROUGE-1）和 MATCHSUM（+0.21 ROUGE-1）。在 DebateSum（议论文）上，提升幅度较小，但表现稳定。最大的提升出现在 ArXiv（科学）上，MAMBA-BERT 的 ROUGE-1 比 BERTSUM 高出 +0.23，比 MATCHSUM 高出 +0.56。统计显著性检验证实，这些提升并非随机波动所致；在所有领域的 ROUGE-1 指标上，MAMBA-BERT 均显著优于 BERTSUM：CNN/DailyMail（新闻）（$t = 5.503$, $p < 0.001$, Cohen's $d = 1.395$,$n = 40$）、DebateSum（$t = 9.6$, $p < 0.001$, Cohen's $d = 2.702$, $n = 40$）和 ArXiv（$t = 28.09$,$p < 0.001$, Cohen's $d = 6.183$, $n = 40$）。这些结果表明，在各数据集上，差异既具有统计显著性，也具有较大的实际效应量。我们尝试将 MemSum 作为近期的抽取式基线进行评估；然而，在我们的低资源设置（200 篇文档）下，该模型生成的结果不稳定且不合逻辑。因此，我们未将 MemSum 纳入定量比较，并指出这一局限

模型	ROUGE-1	ROUGE-2	ROUGE-L	时间（秒/样本）
CNN/DailyMail				
MAMBA-BERT	**0.61**	**0.51**	**0.53**	**13.7**
BERTSUM	0.44	0.28	0.33	18.0
MATCHSUM	0.40	0.18	0.27	18.6
DebateSum				
MAMBA-BERT	**0.57**	**0.54**	**0.56**	0.65
BERTSUM	0.50	0.35	0.40	0.52
MATCHSUM	0.46	0.34	0.37	0.12
ArXiv				
MAMBA-BERT	**0.73**	**0.64**	**0.64**	8.14
BERTSUM	0.50	0.26	0.27	3.67
MATCHSUM	0.17	0.04	0.12	0.45

表 2：ROUGE 分数和推理时间。

统计显著性检验（$p < 0.001$，配对 t 检验，$n = 40$）：MAMBA-BERT 与 BERTSUM 在所有数据集上的 ROUGE-1 对比。

5.2 效率与长文档优势

在新闻摘要任务中，与基线模型相比，MAMBA–BERT 将推理时间缩短了 24–27%，同时保持线性复杂度扩展。在包含最长文档的 ArXiv 数据集上，我们的模型不仅能够处理完整输入而无需截断，还取得了最高的相对质量提升，证实了其适用于长文档摘要。

5.3 低资源鲁棒性

每个数据集的所有实验均仅使用了 200 篇文档。尽管受到这一限制，MAMBA–BERT 在不同领域中仍保持了出色的性能，展现了其在低资源条件下的有效性。

6 错误分析

对表现最差的案例进行人工检查后，发现了两种主要的失败模式：

- **过度选取无关细节（100%）：** 模型始终会选取不太重要的背景信息或过多的细节，而不是核心内容。

- **遗漏关键信息（75%）：** - 遗漏命名实体：缺少人名、组织或数字平台 - 遗漏主要事件：关键行动、决策或公告 - 遗漏上下文信息：解释性上下文和具有代表性的引语 - 遗漏归属细节：来源信息和年龄细节。

错误分析显示出明显的权衡：模型擅长识别语义相关内容，但在重要性排序和实体优先级排序方面表现不佳。这表明 MAMBA 的序列建模可能需要通过重视重要性的注意力机制或感知实体的机制加以增强。

7 结论、局限性与未来工作

我们提出首个用于抽取式摘要的 Mamba-Transformer 混合模型，将基于 Transformer 的语义编码与 Mamba 状态空间模型的线性时间处理相结合。该方法无需截断即可处理完整长度的文档，并在低资源环境中取得优异成果，在三个领域中均胜过 BERTSUM 和 MATCHSUM（ArXiv 上 ROUGE-1 为 +0.23；新闻上为 $p < 0.001$），同时将推理时间最多缩短 27%。配对 t 检验证实，这些提升具有统计显著性，并非偶然。我们的研究结果凸显了该方法的跨领域泛化能力和效率，表明其具有广泛的适用性。

尽管如此，该方法仍存在若干局限。模型未明确建模重要性排序和实体优先级排序，因此仍有空间进行重要性感知或实体感知方面的改进。受资源限制，我们的实验仅涵盖每个数据集中的 200 篇文档；开展更大规模的研究将有助于增强验证力度。此外，比较主要针对为受限设置选取的较早期基线，而在低资源场景下对较新的抽取式方法进行的初步测试未能取得有竞争力的结果。扩大基线覆盖范围、扩展训练数据规模，以及探索与生成式技术相结合，都是值得推进的方向。

我们希望这能为高效神经摘要中的混合架构开辟新方向，例如在科学文献、新闻编辑部和政策辩论等领域；在这些领域中，于有限资源下处理长文档至关重要。

参考文献

Chenxin An、Ming Zhong、Zhiyong Wu、Qin Zhu、Xuanjing Huang和Xipeng Qiu。2022年。Colo：一种基于对比学习的单阶段摘要重排序框架。载于《第*29*届国际计算语言学会议论文集》，庆州，大韩民国，第5783–5793页。

T. Bao、H. Zhang 和 C. Zhang。2025年。利用结构信息增强科学论文的抽象式摘要。专家系统与应用，261:125529。

Arman Cohan、Franck Dernoncourt、Doo Soon Kim、Trung Bui、Seokhwan Kim、Walter Chang 和 Nazli Goharian。2018年。一种用于长文档抽象式摘要的话语感知注意力模型。载于*2018*年北美计算语言学协会分会会议论文集，第 615–621 页。

Xiangyun Dong、Wei Li、Yuquan Le、Zhangyue Jiang、Junxi Zhong 和 Zhong Wang。2025年。Termdiffusum：一种用于法律文档抽取式摘要的术语引导扩散模型。载于第*31*届国际计算语言学大会论文集，阿布扎比，阿联酋，第3222–3235页。

Mitchell Gordon、Kevin Duh 和 Nicholas Andrews。2020年。压缩 BERT：研究权重剪枝对迁移学习的影响。载于 自然语言处理表示学习研讨会第五届会议论文集，线上，第 143–155 页。Albert Gu 和 Tri Dao。2023年。 Mamba：采用选择性状态空间的线性时间序列建模。 计算研究资料库，arXiv:2312.00752。版本 1。Karl Moritz Hermann、Tomas Kocisky、Edward Grefenstette、Lasse Espeholt、Will Kay、Mustafa Suleyman 和 Phil Blunsom。2015年。教机器阅读和理解。载于 神经信息处理系统进展，第 1684–1692 页。Renren in、Jiangcun Du、Wuwei Huang、Wei Liu、Jian Luan、Bin Wang 和 Deyi Xiong。2024年。大语言模型量化策略的全面评估。载于 计算语言学协会年会论文集：*ACL 2024*年，第 12186–12215 页。Thang Le 和 Anh Tuan Luu。2024年。使用文本生成器进行抽取式摘要。载于*2024*年北美计算语言学协会分会年会：人类语言技术会议论文集（第一卷：长篇论文），第 157–174 页。Y. Liu。2019年。微调 BERT 以进行抽取式摘要。 计算研究资料库，arXiv:1903.10318。版本 2。Hugo Pitorro、Pavlo Vasylenko、Marcos Treviso 和 André Martins。2024年。状态空间模型在机器翻译中有多大效用？载于 第九届机器翻译大会论文集，第 1107–1124 页。Allen Roush 和 Arvind Balaji。2020年。DebateSum：大规模论证挖掘与摘要数据集。载于 第七届论证挖掘研讨会论文集，线上，第 1–7 页。Shichao Sun 和 Wenjie Li。2021年。通过对比学习缓解抽象式文本摘要中的曝光偏差。计算研究资料库，arXiv:2108.11846。版本 1。

Moming Tang、Chengyu Wang、Jianing Wang、Cen Chen、Ming Gao 和 Weining Qian。2023年。ParaSum：面向低资源抽取式文本摘要的对比释义改写。载于 知识科学、工程与管理，第 106–119 页。

Dusan Varis 和 Ondˇrej Bojar。2021年。序列长度即领域：Transformer模型中的基于长度的过拟合。载于*2021*年自然语言处理实证方法会议论文集，第8246–8257页。

Zhichao Xu、Jinghua Yan、Ashim Gupta 和 Vivek Srikumar。2025年。状态空间模型是强大的文本重排序器。载于第*10*届自然语言处理表示学习研讨会（*RepL4NLP-2025*），阿尔伯克基，新墨西哥州，第152–169页。

Ruifeng Yuan、Shichao Sun、Zili Wang、Ziqiang Cao 和 Wenjie Li。2023年。《分离上下文与模式：面向低资源抽取式摘要的解耦句子表示学习》。载于计算语言学协会 *ACL 2023*年会议论文集，第 7575–7586 页。

Haopeng Zhang、Xiao Liu 和 Jiawei Zhang。2023年。《DiffuSum：利用扩散增强生成的抽取式摘要》。载于计算语言学协会 *ACL 2023*年会议论文集，第 13089–13100 页。

Yusen Zhang、Nan Zhang、Yixin Liu、Alexander Fabbri、Junru Liu、Ryo Kamoi、Xiaoxin Lu、Caiming Xiong、Jieyu Zhao、Dragomir Radev、Kathleen McKeown 和 Rui Zhang。2024年。《多元视角的公平抽象式摘要》。载于*2024*年北美计算语言学协会分会会议论文集：人类语言技术（第 *1* 卷：长篇论文），墨西哥城，墨西哥，第 3404–3426 页。

Ming Zhong、Pengfei Liu、Yiran Chen、Danqing Wang、XipengQiu 和 XuanjingHuang。2020年。将抽取式摘要视为文本匹配。载于计算语言学协会第*58*届年会论文集，第6197–6208页。

Yichen Zhu、Ning Liu、Zhiyuan Xu、Xin Liu、Weibin Meng、Louis Wang、Zhicai Ou 和 Jian Tang。2022年。少教多学：论知识蒸馏中的不可蒸馏类别。载于神经信息处理系统。