[文章编号] 1003－4684(2026)04-0070-09

基于 mT5 的融合注意力机制与 LSTM 的生成式文本摘要方法研究

刘 冯，熊才权

（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 为解决传统的 mT5 模型在提取文本关键信息方面存在局限性，无法准确捕获长文本中的所有关键内容，导致生成的摘要信息缺乏全面性和准确性等问题，提出一种基于 mT5 的融合注意力机制与 LSTM 的生成式文本摘要方法（mT5-ATT-LSTM）。该方法首先利用 mT5 的编码器获取输入文本的初始词向量表示；然后引入注意力机制增强词向量表示，使模型能够关注输入文本中的关键信息；接着通过引入 LSTM 来捕捉长程依赖关系；最后通过 mT5 的解码器将信息转换为目标文本生成摘要。在公开的中文新闻短文本数据集 LCSTS 上对本方法进行实验，本方法的 ROUGE-1、ROUGE-2、ROUGE-L 分别为 37.71、24.61、34.51，相比传统 mT5 模型 ROUGE 值分别提升了 2.1%、3.3%、2.1%，通过对比和消融实验验证了该方法的有效性。

[关键词] mT5；注意力机制；LSTM；生成式文本摘要

[中图分类号] TP18 [文献标识码] A

文本自动摘要技术分为抽取式摘要与生成式摘要两大类。早期，抽取式摘要方法占据主导地位。这类方法依赖于统计技术，如隐式马尔可夫链[1]、条件随机场[2]、TextRank[3] 和概率潜在语义分析[4]等，确定文档中哪些句子或短语最为重要，进而将其提取为摘要的一部分。抽取式摘要方法的一个明显的局限性是无法创造性地补充源文本中未直接包含的信息。

随着深度学习技术的发展，生成式文本摘要迎来了新的发展机遇。人们采用 Seq2Seq[5]、Transformer[6] 等神经网络模型，通过大规模语料库训练成功提升了生成式文本摘要性能。但这些方法仍存在一系列挑战，包括处理大规模文本、提升生成文摘质量、提高模型可解释性等。近年来，大语言模型（如 GPT-3[7]、BERT[8]、T5[9]）的崛起显著推动了文本摘要技术的发展，采用了“预训练＋微调”的架构，通过学习海量文本数据，可以生成更丰富更高层次的语义表示，极大提升了各文本摘要的质量。T5 模型作为 Transformer 代表，其独特"Text-To-Text Transfer Transformer"思想不仅在多任务学习中表现出色，还为生成式任务提供了统一框架。mT5[10]（multilingual Text-to-Text Transfer Transformer）是 T5 的多语言版本，通过在大规模多语言语料库上的训练，使其具备更好的跨语言泛化能力，特别在涉及多语种文本摘要中更有价值。

基于 mT5 的文本摘要的研究有很多，针对多语言摘要任务，Abadi 等人[11] 利用 mT5-base 模型和 PPO 算法，成功应用于波斯语文本摘要任务，采用 ROUGE（Recall-Oriented Understudy for Gisting Evaluation）评估，其中 ROUGE-1、ROUGE-2、ROUGE-L 分别是 53.17、37.12、44.13，取得了显著的 ROUGE 评分，开启了文本摘要研究的新思路；Hasan 等人[12] 利用目前最大的抽象摘要数据集 XL-Sum 对 mT5 模型进行微调，并通过实验表明 XL-Sum 在多语言训练中提高了 ROUGE 值，并在低资源语言上表现出具有竞争力的性能；Calizzano 等人[13] 利用 mT5 模型在三种不同的训练情景下微调，取得了较好的文本摘要效果，并提出名为 WikinewsSum 的新数据集，填补了多语言文本摘要任务的研究空白。针对特定语言摘要任务，Yang[14] 针对匈牙利语文本摘要任务，提出了首个基于 mBART 和 mT5 模型的匈牙利语生成式摘要工具，填补了该领域的研究空白；Mukherjee[15] 提出了针对孟加拉语文本摘要的深度学习模型，该模型使用

[收稿日期] 2024－04－01
[基金项目] 湖北省科技计划项目（2021BLB171）；国家自然科学基金（61902116）
[第一作者] 刘 冯（1999－），男，湖北武汉人，湖北工业大学硕士研究生，研究方向为自然语言处理文本摘要算法。
[通信作者] 熊才权（1966－），男，湖北鄂州人，工学博士，湖北工业大学教授，研究方向为人工智能、辩论模型、智能决策。

了 LSTM-RNN 编码器和解码器模型以及基于 Transformer 的多语言 mT5,为资源匮乏的自然语言处理提供了新的解决方案。在模型优化和改进方面,Gryaznov 等人[16]使用 mT5 模型和新闻语料库进行了新闻标题预测实验,认为调整权重更新次数能显著提高文本摘要的准确性,但在一定阶段后性能增长逐渐减缓,表明模型可能接近饱和状态;Chi 等人[17]提出了改进的多语言文本处理模型 mT6,对 mT5 进行了优化。通过引入新的预训练任务和部分非自回归目标,mT6 在跨语言迁移能力上表现更强,在八个多语言基准数据集上的评估结果也验证了其性能提升。针对特定任务应用,Galeshchuk[18]改进了 mT5 模型在乌克兰语新闻摘要任务中的性能,提出了新的乌克兰语数据集,并在不同参数规模的 mT5 模型上进行微调,以提高摘要质量。基于 mT5 的文本摘要研究在多个方面取得了显著进展,但仍然存在一些问题,如不能完全捕捉关键信息和内容覆盖不够全面等。

基于以上问题,单独使用注意力机制可能无法有效解决文本摘要任务中的所有问题,特别是在处理具有复杂语义结构的文本时,生成的摘要可能缺乏连贯性和一致性。同样,传统的 LSTM 模型在捕捉文本序列中的长程依赖关系方面表现不佳,导致生成摘要可能存在语义不连贯或表达不流畅的问题。本文提出了一种称为 mT5-ATT-LSTM 的文本摘要方法,该模型是对 mT5 模型的改进,在 mT5 模型的编码器后引入了注意力机制和 LSTM 网络,旨在优化其在中文文本生成式摘要任务中的表现。该方法通过注意力机制增强模型对输入文本不同部分的关注,并借助 LSTM 网络捕捉文本序列中的长程依赖关系。这两个元素的结合旨在提升模型对输入文本的理解和表达能力,从而使 mT5 的解码器能够更有效地生成准确的中文文本摘要。本文通过对比实验验证了 mT5-ATT-LSTM 相较于传统 mT5 模型的提升。如果直接使用 mT6-ATT-LSTM,虽然可能进一步提升性能,但由于 mT6 自身的改进,难以明确区分性能提升的具体来源是由于 mT6 还是 ATT-LSTM 的贡献。

1　mT5-ATT-LSTM 文本摘要方法

mT5-ATT-LSTM 文本摘要方法分为三个阶段,第一阶段涉及中文文本的输入,通过 mT5 的编码器获得初始的词向量表示。在第二阶段,引入注意力机制以增强词向量表示,此注意力机制目的在于引导模型关注输入文本中的不同部分,从而提升模型的性能。第三阶段将经过注意力机制加强的词向量表示输入到 LSTM 中,这是一种循环神经网络的变体,其设计目的在于捕捉序列数据中的长程依赖关系,这对于更好地理解文本的上下文信息至关重要。最后,经过注意力和 LSTM 处理的表示被输入到 mT5 的解码器,将输入信息转换为最终的文本摘要。整个流程的目标在于通过引入注意力机制和 LSTM 提升输入文本的表示能力,以使 mT5 的解码器更为有效地生成精准的文本摘要。其整体结构的图形说明如图 1 所示。

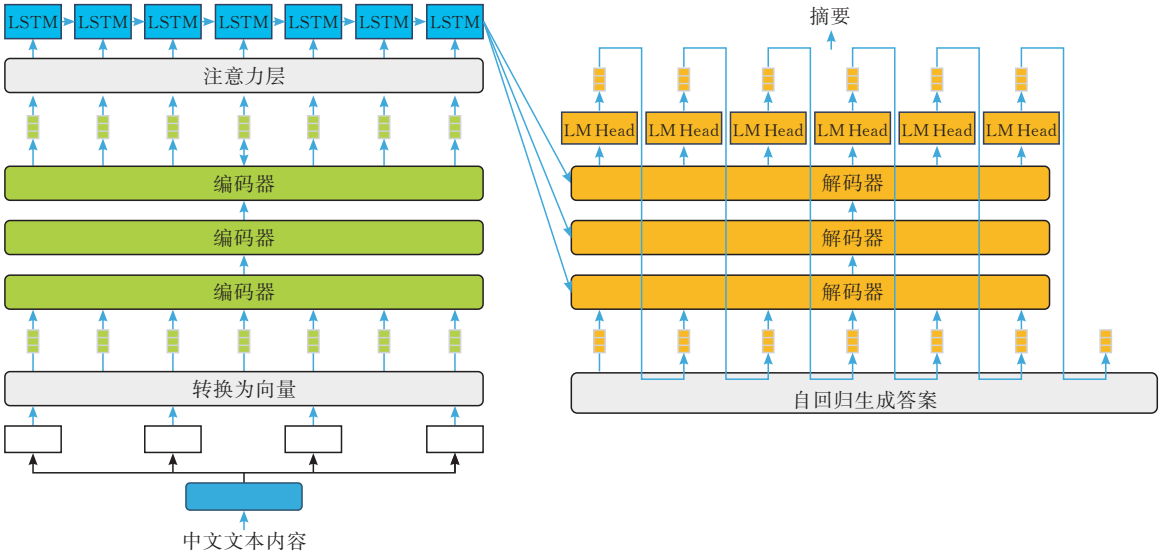


图 1　mT5-ATT-LSTM 模型结构图

1)文本内容输入到 mT5 的编码器层。数据集为 LCSTS 数据集,需要去除 HTML 标签、特殊字符等非文本内容,还需要进行分词、停用词去除等,具体公式如(1)—(4)所示。

$$d_{\text{clean}} = \text{remove_html_tags}(d) \tag{1}$$

$$d_{\text{clean}} = \text{remove_special_characters}(d) \tag{2}$$

$$d_{\text{tokenized}} = \text{tokenize}(d_{\text{clean}}) \tag{3}$$

$$d_{final} = \{w_i \mid w_i \notin stopwords\} \tag{4}$$

最后将数据集划分为训练集、测试集和验证集。数据格式是{"id":id,"summary":summary,"content":content}。

假设输入的原始文本为 $X = \{x_1, x_2, \cdots, x_n\}$，其中 x_i 表示第 i 个单词，其对应的词向量表示为 v_i，则整个序列的向量表示为 $\{v_1, v_2, \cdots, v_n\}$，可以使用预训练的词嵌入模型来实现。为了在模型中添加关于单词在序列中的位置信息，需要添加位置编码。位置编码是一个与位置相关的固定向量，它被添加到每个输入词向量中，其位置编码的方式用式(5)(6)表示：

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{5}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{6}$$

然后每个单词的词向量表示与其位置编码相加，形成了输入到 Transformer 模型的初始表示，公式如(7)所示。

$$\text{Input}_{\text{transformer}} = \{v_1 + \text{PE}_1, v_2 + \text{PE}_2, \cdots, v_n + \text{PE}_n\} \tag{7}$$

在 Transformer 的每个编码层中，输入被输入到多头自注意力机制中，自注意力机制允许每个单词关注其他所有单词，并根据它们的相关性对自身进行加权，加权过程可以通过公式(8)计算得到。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

其中，Q、K、V 分别表示查询向量、键向量和值向量；d_k 是查询和键向量的维度。

在自注意力机制之后，输出被输入到一个残差连接和层归一化的子层中，以减少训练过程中的梯度消失和梯度爆炸问题，公式如(9)所示。

$$\text{LayerNorm}(\text{x} + \text{Attention}(Q, K, V)) \tag{9}$$

最后，输出被输入到一个前馈神经网络中进行非线性转换，公式如(10)所示。

$$FFN(X) = ReLu(W_2(ReLu(W_1 x + \boldsymbol{b}_1)) + \boldsymbol{b}_2) \tag{10}$$

其中，W_1 和 W_2 是权重矩阵；$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 是偏置向量；$ReLu$ 表示修正线性单元激活函数。

2)编码器层结果引入注意力机制。经过 mT5 的编码器层处理后，得到了编码结果，记为 $\{h_1, h_2, \cdots h_n\}$，其中 h_i 是第 i 个单词的编码表示。为了计算每个单词之间的注意力权重，需要首先定义三个向量：查询向量 $\boldsymbol{q}_i$、键向量 $\boldsymbol{k}_i$ 和值向量 $\boldsymbol{v}_i$，这些向量可以后编码结果 $\{h_1, h_2, \cdots h_n\}$ 线性映射得到，其中查询向量和键向量用于计算单词之间的相关性，而值向量用于加权求和以生成加权后的表示，公式如(11)(12)所示。

$$\boldsymbol{q}_i = W_q * h_i, \forall i \tag{11}$$

$$\boldsymbol{k}_i = W_k * h_i, \forall i \tag{12}$$

其中，W_q 和 W_k 是权重矩阵。

计算注意力分数需要使用查询向量和键向量之间的点积，然后通过 softmax 函数获得标准化的注意力权重，公式如(13)(14)所示。

$$e_{ij} = q_i * k_j, \forall i, j \tag{13}$$

$$a_{ij} = soft\max(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{j/=1}^{n} \exp(e_{ij/})} \tag{14}$$

根据注意力权重对值向量进行加权求和，得到加权后的表示，公式如(15)所示。

$$Attention(h_i) = \sum_{j=1}^{n} a_{ij} v_j \tag{15}$$

最后经过注意力机制加权后的表示被用作下一步的输入，用于后续的处理。

3)经过注意力后的加权表示传递到 LSTM 模型。在经过编码器层后，得到了每个单词的表示 h_i，其中 h_i 是一个向量，表示包含了输入文本的丰富语义信息，然后引入注意力机制来对这些表示进行加权，以便模型可以更好地关注输入文本中的重要部分。加权的过程公式如(16)－(18)所示。

$$e_{ij} = \text{score}(h_i, h_j) \tag{16}$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} \tag{17}$$

$$c_i = \sum_{j=1}^{n} a_{ij} h_j \tag{18}$$

其中，e_{ij} 表示 h_i 和 h_j 之间的相关性得分，通过某种函数计算得到。

然后通过 softmax 函数将这些得分转换为注意力权重 a_{ij}，用来对 h_j 进行加权求和，得到加权表示 c_i。加权表示 c_i 被作为 LSTM 模型的输入，用于处理序列数据并捕捉长期依赖关系，对于每个时间步 t，LSTM 模型的输入是加权表示 c_t，并根据序列数据的特点，逐步更新内部的隐藏状态 h_t。

4)LSTM 模型处理序列数据。LSTM 根据上一个时刻隐层状态 h_{t-1} 和当前输入 x_t 计算出当前时刻的遗忘门 f_t、输入门 i_t 和临时细胞状态 $\hat{c}_t$ 相关数据，公式如(19)－(21)所示。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{19}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{20}$$

$$\hat{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{21}$$

其中，σ 是激活函数 sigmoid，tanh 是双曲正切激活函数；W 和 b 分别代表门的权重矩阵和偏置向量。

根据遗忘门 f_t、记忆门 i_t 和临时细胞状态 $\hat{c}_t$ 计算出当前细胞状态 c_t，公式如(22)所示。

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{22}$$

其中:运算符 ⊙ 表示矩阵乘法。

进而结合输出门 o_t 和当前的细胞状态 c_t 计算出当前隐藏层状态 h_t,公式如(23)(24)所示。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{23}$$

$$h_t = o_t \odot \tanh(c_t) \tag{24}$$

LSTM 包括门控操作、更新记忆单元、输出隐藏状态,通过这些操作,LSTM 模型可以有效地处理序列数据,并捕捉序列中的长程依赖关系,这使得它在许多序列建模任务中表现出色,如文本摘要等。

5) LSTM 的输出经过 mT5 的解码器层。LSTM 模型的输出隐藏状态 h_t 被用作 mT5 的解码器层的输入,以便生成最终的文本摘要,其中解码器层通过多个 Transfomer 层将这些表示转换为目标文本摘要。在解码器层,输入的隐藏状态 h_t 结果自注意力机制和其他 Transformer 层的操作,逐步被转换为目标文本序列,它会根据当前的上下文信息,预测下一个单词,然后将预测的单词作为输入,继续预测下一个单词,直到生成完整的文本摘要。在这个过程中,LSTM 的输出隐藏状态提供了上下文信息,帮助解码器层更好地理解输入文本并生成相应的摘要。

6)生成摘要。解码器层的初始输入经过自注意力机制,使得模型可以根据当前的上下文信息关注不同部分的重要性,并进行相应的加权,自注意力机制的计算方式与编码器层中的相同,包括计算注意力分数和应用 softmax 函数,这个过程可以使得模型在生成每个单词时考虑到输入序列中其他单词的信息。自注意力机制之后,解码器层的输出被输入到一个前馈神经网络中进行非线性转换,前馈神经网络的计算方式与编码器层中的相同,包括两个全连接层和激活函数。经过自注意力机制和前馈神经网络之后,解码器层将生成下一个单词的概率分布。可以使用 softmax 函数将这个分布转换为概率,然后从中采样得到下一个单词。该过程可以通过公式(25)表示。

$$P(y_i \mid y < i, x) = soft\max(W_o * h_i) \tag{25}$$

其中,y_i 是第 i 个生成的单词;$y<i$ 是已生成的前 $i-1$ 个单词的序列;x 是输入文本序列;h_i 是解码器层的输出隐藏状态;W_o 是权重矩阵。

重复以上步骤,解码器层能够将 LSTM 的输出逐步转换为目标文本序列,生成相应的文本摘要。整个过程中,模型能够利用注意力机制关注重要部分,结合上下文信息生成连贯、准确的摘要内容。

2　实验评估指标

采用经典的 ROUGE 指标作为对生成的摘要客观评价的依据。ROUGE[19]指标是一种广泛应用于自动摘要任务的评价方法,目的在于度量生成的摘要与参考摘要之间的相似性。ROUGE-1(1-gram)表示单个词重叠情况,反映摘要的词汇准确性和内容涵盖程度;ROUGE-2(2-gram)表示相邻两个词的组合的重叠情况,反映摘要的短语准确性和连贯性,ROUGE-L(最长公共子序列)表示最长公共子序列的长度,反映生成摘要与参考摘要的结构和顺序的一致性。分别从不同的角度对生成的摘要与参考摘要的重叠情况进行了评估。具体评估公式如下:

$$\text{ROUGE-}N = \frac{\sum_{S\in\{\text{Ref}\}}\sum_{\text{N-gram}\in S}\text{Count}_{\text{match}}(\text{N-gram})}{\sum_{S\in\{\text{Ref}\}}\sum_{\text{N-gram}\in S}\text{Count}(\text{N-gram})} \tag{26}$$

$$\text{ROUGE-}L = \frac{(1+\beta^2)R_{\text{lcs}}P_{\text{lcs}}}{R_{\text{lcs}}+\beta^2 P_{\text{lcs}}} \tag{27}$$

$$R_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{\text{len}(m)} \tag{28}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{\text{len}(n)} \tag{29}$$

总体而言,ROUGE 是一套全面而灵活的评估标准,用于客观地衡量自动文本摘要系统的性能。通过考虑不同级别的文本结构和重叠,ROUGE 能够提供多方面的评估,能够客观地评估生成摘要的准确性、连贯性和内容涵盖度。本文使用这些指标来衡量所提出的文本摘要模型的效果,并与其他方法进行对比。

3　实验验证与结果分析

3.1　数据来源及处理

在本章节中,实验数据集 LCSTS[20](Large Scale Chinese Short Text Summarization)是一个中文短文本摘要数据集,用于自然语言处理和文本摘要的研究。该数据集主要用于短文本自动摘要任务,即从一个或多个句子的输入中生成一个简洁的摘要,概括文本的主要内容。它被广泛应用于短文本新闻摘要任务,它源自新浪微博,拥有超过 200 万条数据。该数据集分为三个主要部分,为研究人员提供了多样性的样本,用于模型的训练、验证和测试。这个数据集含有 2400591 条的训练集,10666 条的验证集和 1106 条的测试集。

3.2　参数设置与优化

本文的实验使用 GPU(A100)进行训练,实验环境是 Windows10、CUDA 11.1、Python 3.8、Pytorch 1.9.0。参数设置如表 1 所示。

在本研究中,选用了 LCSTS 数据集,并通过对 mT5模型的每个编码器层后引入了注意力机制和

LSTM，以优化原始的摘要生成过程。在训练过程中，对模型进行了 10 个轮次的训练，使其逐步适应摘要任务的生成。Pytorch 实现了本模型并进行了训练和评估。在数据集方面，将输入文本的最大长度设定为 150，以确保模型能够有效处理输入数据。目标摘要文本的最大长度被限定为 32，而每次模型更新时的批处理样本数量（batch_size）则被设置为 48，该设置为设备允许范围内的最大值。在文本处理中，由于序列长度不同，采用了填充（padding）的策略，使用－100 作为填充标记的 ID。

表 1 实验参数设置

参数类型	参数值
num_train_epochs	10
batch_size	48
max_source_length	150
min_source_length	32
steps_per_epoch	None
pad_token_id	－100
learning_rate	1e-4

在学习率和优化器的设置中，采用了 AdamW 优化器，其学习率被设定为 1e-4。在模型定义的部分，构建了 mT5ATT 模型，该模型融合了 T5 模型的编码器输出、BertSelf 注意力层和 LSTM 层。在前向传播中，通过这些层进行特征提取和加权处理。这样的模型设计是对特定任务需求的响应，通过层的堆叠和组合，目的在于提升模型对输入数据的表示能力。模型结构的选择和设计是模型优化的关键一环。在损失函数的设置中，采用了自定义的 CrossEntropyLoss 类，在前置实验中对损失函数的超参数进行了手工调整。

总体而言，通过对模型参数的优化和调整，本研究的目的在于通过微调和测试，提升 mT5 模型在特定任务上的性能。模型参数的优化涉及多个方面，包括优化器和学习率的选择，模型结构、损失函数和评估指标等。深入理解任务特性，灵活调整模型结构，并不断迭代模型的训练过程，是提升深度学习模型性能的关键步骤。

3.3 对比实验与实例分析

为了验证本章提出的模型生成摘要的有效性，将本章节实验模型与其他基线模型在 LCSTS 数据集上的实验结果进行对比，对比方式主要是比较 ROUGE-1、ROUGE-2 和 ROUGE-L 评分，对比实验结果如表 2 所示。

表 2 对比实验结果

模型	ROUGE-1	ROUGE-2	ROUGE-L
ASPM	32.80	16.80	32.80
TI-C-NHG	34.30	16.70	32.00
BERT	36.00	23.00	32.90
RoBERTa	36.20	22.90	33.00
mT5	36.93	23.82	33.80
DGRD	37.00	24.20	34.20
mT5-ATT-LSTM	37.71	24.61	34.51

对比实验柱形图如图 2 所示。

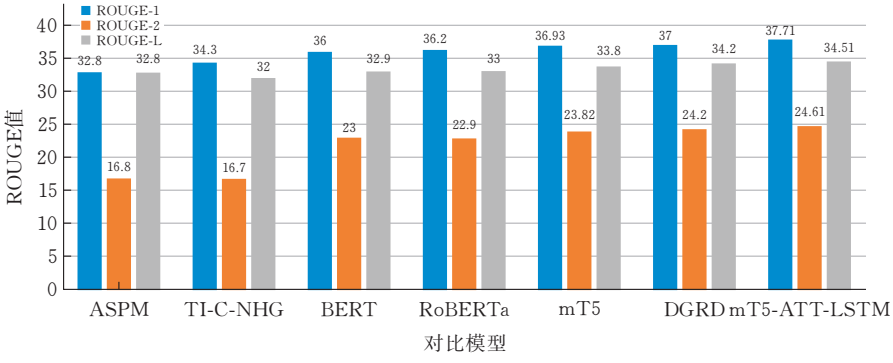


图 2 不同模型 ROUGE 指标对比（一）

对比的基线主要有：

1）ASPM[21]：一种融合义原的中文摘要生成模型，在 Pointer Network 模型的基础上引入义原向量，使得模型可以解决以词为单位对数据进行处理时出现的 OOV 问题。

2）TI-C-NHG[22]：该模型通过主题推理挖掘潜在的新主题，使主题理解成为句子级上下文的更广泛来源，并采用多层 Transformer 解码器构建的训练模型大大提高模型的并行能力，加快摘要生成的推理过程。

3）BERT[23]：Transformer 的双向编码器表示，其结构与 Transformer 模型的编码端相同。

4）RoBERTa[24]：预训练语言模型 RoBERTa 是 BERT 的改进版，与 BERT 相比 RoBERTa 有更多的参数，训练中设置了更大的 batch_size 并且使用了更多的训练数据。

5）mT5：mT5 是一种基于 T5 架构的多语言文本到文本转换模型，通过统一的文本生成方法处理多种语言的自然语言处理任务，展现了强大的通用性和灵活性。

6)DGRD[25]:一种基于注意力机制的 Seq2Seq 模型,增加了潜在结构向量来学习目标摘要的潜在语义信息。

在此次对比实验中,通过对主模型 mT5-ATT-LSTM 与 ASPM、TI-C-NHG、BERT、RoBERTa、mT5、DGRD 等对比模型在 ROUGE-1、ROUGE-2 和 ROUGE-L 指标上的性能进行详细分析,可以得出一系列深刻的观察和结论。

首先,从 ROUGE-1 指标的角度看,ASPM 呈现相对较低的性能,为 32.80,而 TI-C-NHG 随后有所提升,达到 34.30。BERT 和 RoBERTa 的 ROUGE-1 值进一步提升至 36.00 和 36.20,显示出预训练模型在捕获上下文语境方面的优越性。DGRD 在 ROUGE-1 上表现较佳,表明其对文本信息的深入理解和提取在短语级别上的效果较好,同时 mT5 的 ROUGE-1 值为 36.93 也表现不错。然而,主模型 mT5-ATT-LSTM 在 ROUGE-1 指标上超越了所有对比模型,达到 37.71,进一步验证了其在单词级别的生成准确性。

其次,在 ROUGE-2 方面,各模型的性能趋势与 ROUGE-1 基本一致,但相应数值更低。ASPM 的 ROUGE-2 值为 16.80,TI-C-NHG 为 16.70,BERT 为 23.00,RoBERTa 为 22.90,mT5 为 23.82,DGRD 为 24.20,而主模型 mT5-ATT-LSTM 在 ROUGE-2 上达到 24.61,显著领先于其他模型。ROUGE-2 指标更注重模型对文本中长距离依赖关系的捕捉,主模型通过整合 mT5 的全局理解和 LSTM 的序列建模,再辅以注意力机制,使其在生成摘要时更好地维持上下文的一致性,从而在 ROUGE-2 上表现卓越。

第三,从 ROUGE-L 指标的角度来看,各模型的性能整体呈现出与 ROUGE-1 和 ROUGE-2 相一致的趋势。ASPM 在 ROUGE-L 上的表现相对较低,为 32.80,TI-C-NHG 为 32.00,BERT 为 32.90,RoBERTa 为 33.00,mT5 为 33.80,DGRD 为 34.20,而主模型 mT5-ATT-LSTM 在 ROUGE-L 上达到 34.51,继续保持其在整体性能上的领先地位。

从各个 ROUGE 指标的综合表现来看,主模型 mT5-ATT-LSTM 在文本摘要任务中的优越性能得到了全面验证。其相对于 ASPM、TI-C-NHG、BERT、RoBERTa、mT5 和 DGRD 的优势主要体现在更全局的文本理解、更准确的序列建模、更强大的注意力机制等方面。ASPM 对于中文语境中的长句子或复杂句子,模型可能表现不佳,因为它主要基于 Pointer Network 模型,而 Pointer Network 对于处理长序列可能存在一定的困难,引入的义原向量虽然有助于解决 OOV 问题,但可能需要额外的计算资源和内存来存储和处理这些向量,增加了模型的复杂度和训练成本。TI-C-NHG 主题推理虽然可以丰富模型对上下文的理解,但也可能导致模型对于一些特定主题的过度关注,从而影响了摘要生成的多样性和准确性。如果主题推理的效果不佳,模型可能会生成与输入文本不相关的摘要,降低了摘要的质量。BERT 和 RoBERTa 等预训练模型通过大规模无监督学习获得了上下文语境的深层次理解,但在生成摘要时可能受到固定长度的限制和信息丢失的问题。DGRD 在 ROUGE-1 和 ROUGE-2 上表现较好,同时 mT5 单模型各项指标都表现不错,但主模型在 ROUGE-2 上的领先地位表明其更好地处理了长距离依赖关系。总体而言,主模型 mT5-ATT-LSTM 的设计更加全面,综合考虑了全局理解、序列建模和长距离依赖关系的捕捉,因此在文本摘要任务中呈现出更为出色的性能。

表 3 是从实验结果中随机抽取的一条样例,这例新闻主要描述了水资源管理制度考核结果及相关措施的公布,具体摘要生成情况比对如下。

表 3　对比模型摘要生成实例

源文本	水利部水资源司司长陈明忠 9 月 29 日在国务院新闻办举行的新闻发布会上透露,根据刚刚完成的水资源管理制度的考核,有部分省接近了红线的指标,有部分省超过红线的指标,在一些超过红线的地方,将对一些取用水项目进行区域的限批,严格地进行水资源论证和取水许可的批准
标准摘要	部分省超过年度用水红线指标 取水项目将被限制批准
ASPM	据透露,部分省份接近或超过水资源管理红线指标。超出部分将严格限制取用水项目
TI-C-NHG	陈明忠透露,根据考核,有省份接近或超过水资源管理红线。超过部分将严格限制取用水项目
BERT	陈明忠透露,一些省份接近或超过水资源管理红线。将对超出部分取用水项目进行严格限制
RoBERTa	陈明忠透露,根据考核,有省份接近或超过水资源管理红线。超出部分将受到严格限制
mT5	水利部透露,根据水资源管理制度考核,部分省份超过红线指标,将对部分取水项目进行审批
DGRD	陈明忠透露,一些省份接近或超过了水资源管理红线。将严格限制超出部分的取用水项目
mT5-ATT-LSTM	陈明忠透露,根据最新水资源管理考核,一些省份接近或超过了红线。超出部分将严格限制取用水项目

在对上述模型进行分析比对时,可以看到它们在生成摘要方面的效果有所不同。首先,ASPM 模

型提供了简洁的摘要，提及了省份接近或超过了水资源管理红线，并将对取用水项目进行严格限制，但未提及“年度用水”。TI-C-NHG 和 BERT 的摘要与 ASPM 相似，都捕捉到了核心信息，但也未提及“年度用水”。RoBERTa、mT5 和 DGRD 的摘要提供了相似的信息，强调了省份接近或超过了水资源管理红线，并将对超出部分的取用水项目进行严格限制，但同样未提及“年度用水”。最后，主模型的摘要效果最佳，虽然也没有提到“年度用水”，但它提供了与标准摘要几乎相同的内容，包括了省份接近或超过了红线、对超出部分的严格限制等关键信息，与标准摘要保持了高度一致。因此，可以推断主模型在生成摘要任务上的效果最好，因为它能够更准确地捕捉到原文的核心内容，并生成与标准摘要最为相似的内容。

3.4 消融实验与分析

为了验证本章提出的模型生成摘要的有效性，将本章节实验模型与该模型的消融模型在 LCSTS 数据集上的实验结果进行对比，对比方式主要是比较 ROUGE-1、ROUGE-2 和 ROUGE-L 评分，对比实验结果如表 4 所示。

表 4 消融实验结果

模型	ROUGE-1	ROUGE-2	ROUGE-L
mT5(zero-shot)	13.66	6.74	12.46
mT5	36.93	23.82	33.80
mT5-LSTM	36.89	23.84	33.76
mT5-ATT	37.36	24.22	34.07
mT5-ATT-LSTM	37.71	24.61	34.51

消融实验对比柱形图如图 3 所示。

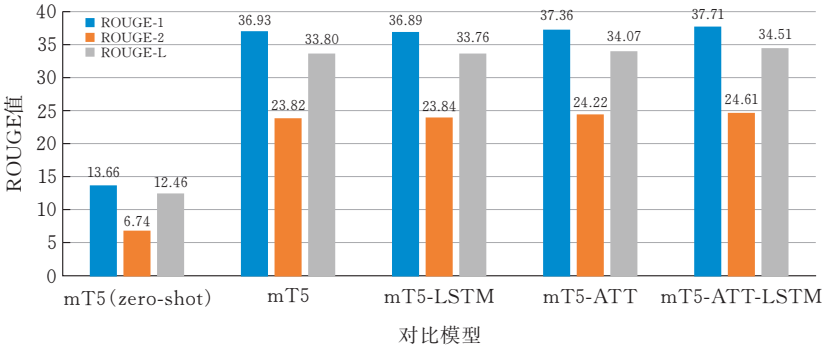


图 3 不同模型 ROUGE 指标对比(二)

在此次消融实验中，通过构建主模型 mT5-ATT-LSTM 以及一系列消融模型(mT5(zero-shot)、mT5、mT5-LSTM、mT5-ATT)来深入研究各个组件对于文本摘要性能的贡献。评估指标包括 ROUGE-1、ROUGE-2 和 ROUGE-L，通过对比各个模型在这些指标上的表现，揭示了不同组件的影响以及主模型整体的优越性。

首先，从 mT5(zero-shot)到 mT5 的过渡中，观察到在所有 ROUGE 指标上的显著性能提升。mT5(zero-shot)的 ROUGE-1、ROUGE-2 和 ROUGE-L 分别为 13.66、6.74、12.46，而引入预训练的 mT5 后，这些值分别提高至 36.93、23.82、33.80。这清晰地表明预训练模型的引入在文本摘要任务中取得了显著的性能提升，为后续的改进提供了坚实的基础。

其次，通过引入 LSTM 模块(mT5-LSTM)，看到在 ROUGE-1、ROUGE-2 和 ROUGE-L 上，相较于 mT5，性能的变化相对较小，分别为 36.89、23.84、33.76。这表明 LSTM 在当前的框架下对于文本摘要性能的提升相对有限。mT5 本身已经包含了对文本序列的深层次建模，而 LSTM 的引入并没有带来明显的性能提升，说明主模型已经在序列建模方面取得了较好的效果。

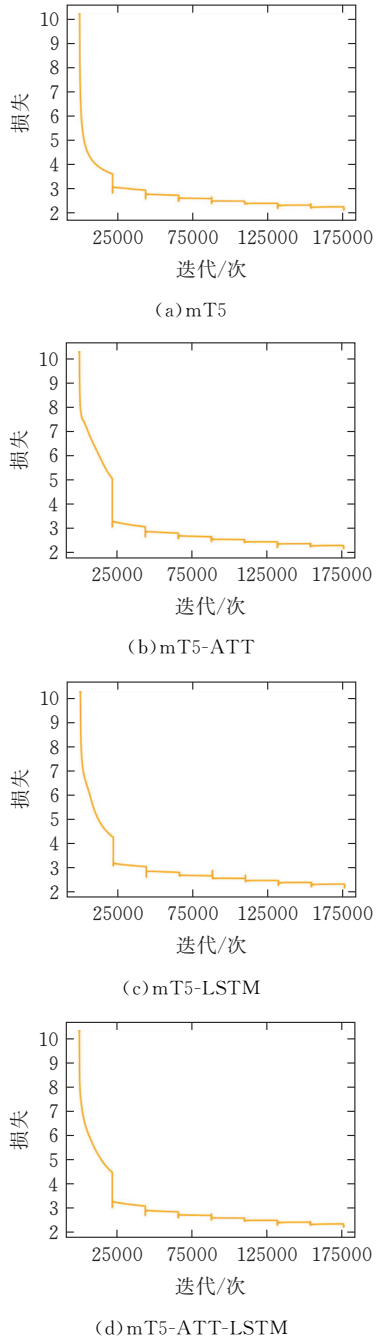
随后，通过引入注意力机制(mT5-ATT)，观察到在 ROUGE-1、ROUGE-2 和 ROUGE-L 上相对于 mT5，性能有了一定的提升，分别为 37.36、24.22、34.07。注意力机制的引入加强了对于文本中关键信息的关注，有助于更好地捕捉文本间的关联关系，进而提升生成摘要的质量。这进一步印证了注意力机制在文本摘要任务中的有效性。

最终，主模型 mT5-ATT-LSTM 在 ROUGE-1、ROUGE-2 和 ROUGE-L 上达到了 37.71、24.61、34.51的性能，进一步超越了其他模型。这说明通过综合考虑预训练模型、序列建模(LSTM)和关注机制(注意力)，主模型能够在文本摘要任务中取得最优异的性能。该模型的设计使其能够更全面地捕捉文本的重要信息，包括全局理解、序列建模和长距离依赖关系等多个方面，为文本摘要任务提供了强大的建模和生成能力。

消融实验学习曲线如图 4 所示。四个消融实验模型的损失图表现出了先快后慢的特点。这一趋势可以归因于模型训练过程中的多个因素相互作用。

初始阶段，模型的参数是随机初始化的，导致模型对数据的拟合能力较弱，损失迅速下降。随着训练的进行，模型逐渐接近最优参数位置，学习到的知识更加丰富和复杂，损失下降速度逐渐减慢。此外，学习率衰减策略、模型复杂度以及数据分布的复杂性等因素也影响了损失下降的速度。

综上所述，预训练模型的引入为性能提升奠定了基础，而注意力机制的加入强化了对关键信息的关注，而 LSTM 在当前框架下贡献相对有限。主模型的成功设计充分发挥了各个组件的优势，为文本摘要任务中的高效生成提供了可行的解决方案。



(a)mT5　(b)mT5-ATT　(c)mT5-LSTM　(d)mT5-ATT-LSTM

图 4　消融实验学习曲线

4　结论

本实验研究了一种创新的文本摘要方法 mT5-ATT-LSTM，旨在提高文本摘要任务的性能。方法包括三个关键阶段：利用 mT5 的编码器得到词向量表示，引入注意力机制增强关键信息关注度，以及将加强表示输入到 LSTM 中捕捉长程依赖关系。最终，在 mT5 的解码器中生成文本摘要。整体而言，该方法通过充分融合预训练模型、注意力机制和 LSTM，实现了对输入文本更深层次的理解和准确的信息提取，通过对比实验以及消融实验，验证了算法的有效性。后续研究将进一步优化模型，并考虑诸如语料库、特定任务等额外因素对于文本摘要生成质量的影响。

[参　考　文　献]

[1] Conroy J M, O'leary D P. Text summarization via hidden markov models[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001: 406-407.

[2] Shen D, Sun J T, Li H, et al. Document summarization using conditional random fields[C]// IJCAI. 2007, 7: 2862-2867.

[3] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 404-411.

[4] Hennig L. Topic-based multi-document summarization with probabilistic latent semantic analysis[C]// Proceedings of the International Conference RANLP-2009. 2009: 144-149.

[5] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 27: 3104-3112.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[7] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[8] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.

[9] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[10] Xue L, Constant N, Roberts A, et al. mT5: A massively multilingual pre-trained text-to-text transformer

[J]. arXiv preprint arXiv:2010.11934, 2020.
[11] Abadi V N M, Ghasemian F. Enhancing Persian text summarization through a three-phase fine-tuning and reinforcement learning approach with the mT5 transformer model[J].Scientific Reports,2025,15 (01): 80.
[12] Hasan T, Bhattacharjee A, Islam M S, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages[J]. arXiv preprint arXiv: 2106.13822, 2021.
[13] Calizzano R, Ostendorff M, Ruan Q, et al. Generating Extended and Multilingual Summaries with Pre-trained Transformers[C]// Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022: 1640-1650.
[14] Yang Z G. Neural text summarization for Hungarian [J]. Acta Linguistica Academica, 2022, 69 (04): 474-500.
[15] Mukherjee A. Developing Bengali Text Summarization with Transformer Base model[D]. Dublin: National College of Ireland, 2022.
[16] Gryaznov A, Rybka R, Moloshnikov I, et al. Influence of the duration of training a deep neural network model on the quality of text summarization task[C]// AIP Conference Proceedings. AIP Publishing LLC, 2023, 2849(01): 400006.
[17] Chi Z, Dong L, Ma S, et al. mT6: Multilingual pretrained text-to-text transformer with translation pairs [J]. arXiv preprint arXiv:2104.08692, 2021.
[18] Galeshchuk S. Abstractive Summarization for the Ukrainian Language: Multi-Task Learning with Hromadske. ua News Dataset[C]// Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). 2023: 49-53.
[19] lin c y. ROUGE: A package for automatic evaluation of summaries[C] // ACL 2004: Text Summarization Branches Out, 2004: 74-81.
[20] Hu B, Chen Q, Zhu F. Lcsts: A large scale Chinese short text summarization dataset[J]. arXiv preprint arXiv:1506.05865, 2015.
[21] 崔卓,李红莲,张乐, 等. 一种融合义原的中文摘要生成方法[J].中文信息学报,2022,36(06):146-154.
[22] Li Z, Wu J, Miao J, et al. A topic inference Chinese news headline generation method integrating copy mechanism[J]. Neural Processing Letters, 2023, 55 (2): 1337-1353.
[23] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[24] Zhuang L, Wayne L, Ya S, et al. A robustly optimized BERT pre-training approach with post-training[C]// Proceedings of the 20th Chinese national conference on computational linguistics. 2021: 1218-1227.
[25] Ma S, Sun X, Lin J, et al. Auto encoder as assistant supervisor: improving text representation for Chinese social media text summarization [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melboume: ACL Press, 2018: 725-731.

Research on Generative Text Summarization Method Based on mT5 with Integrated Attention Mechanism and LSTM

LIU Feng, XIONG Caiquan

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: To address the limitations of traditional mT5 models in extracting key textual information, particularly their inability to accurately capture all critical content within lengthy texts, thereby resulting in generated summaries lacking comprehensiveness and accuracy, we propose a generative text summarization approach integrating attention mechanism and Long Short Term Memory (LSTM), denoted as mT5-ATT-LSTM. This approach first utilizes mT5 encoder to obtain initial word vector representations of input text. Subsequently, an attention mechanism is introduced to enhance word vector representations, enabling the model to focus on different aspects of the content. Then, LSTM is employed to capture long-range dependency relationships. Finally, mT5 decoder is utilized to transform the information into target text for summary generation. Experimental evaluations conducted on the publicly available Chinese news short text dataset LCSTS demonstrate the effectiveness of the proposed method. The ROUGE-1, ROUGE-2, and ROUGE-L scores achieved by this method are 37.71, 24.61, and 34.51, respectively. Compared to the traditional mT5 model, our approach yields improvements of 2.1%, 3.3%, and 2.1% in ROUGE scores, validating the efficacy of the proposed method through comparative and ablation experiments.

Keywords: mT5; attention mechanism; LSTM; generative text summarization

[责任编校: 闫 品]