Computer Science and Application 计算机科学与应用, 2024, 14(3), 10-19
Published Online March 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.143053

Hans汉斯

基于T5-PEGASUS-PGN模型的中文新闻文本摘要生成方法

曹一平*，张胜男

沈阳工业大学软件学院，辽宁 沈阳

收稿日期：2024年2月7日；录用日期：2024年3月5日；发布日期：2024年3月13日

摘　要

针对预训练模型训练任务与下游摘要生成任务存在差异、生成文本存在重复内容造成摘要可读性差的问题，基于T5-PEGASUS和指针生成网络，提出了一种自动摘要模型——T5-PEGASUS-PGN。首先利用T5-PEGASUS获取最符合原文语义的词向量表示，然后借助引入覆盖机制的指针生成网络，生成高质量、高可读的最终摘要。在公开的长文本数据集NLPCC2017的实验结果表明，与PGN模型、BERT-PGN等模型相比，结合更贴合下游摘要任务的预训练模型的T5-PEGASUS-PGN模型能够生成更符合原文语义、内容更加丰富的摘要并且能有效的抑制重复内容生成，同时Rouge评价指标Rouge-1提升至44.26%、Rouge-2提升至23.97%以及Rouge-L提至34.81%。

关键词

生成式摘要模型，预训练模型，PGN，Coverage机制

A Method of Generating Chinese News Text Summaries Based on the T5-PEGASUS-PGN Model

Yiping Cao*, Shengnan Zhang

School of Software, Shenyang University of Technology, Shenyang Liaoning

Received: Feb. 7th, 2024; accepted: Mar. 5th, 2024; published: Mar. 13th, 2024

Abstract

To address the challenges of differences between the training tasks of pretrained models and the

*通讯作者。

文章引用: 曹一平, 张胜男. 基于 T5-PEGASUS-PGN 模型的中文新闻文本摘要生成方法[J]. 计算机科学与应用, 2024, 14(3): 10-19. DOI: 10.12677/csa.2024.143053

downstream summary generation tasks, as well as the poor readability caused by repeated content in the generated texts, an automatic summary model called T5-PEGASUS-PGN is proposed based on T5-PEGASUS and pointer generation networks. This model first utilizes T5-PEGASUS to obtain the most semantically consistent word vector representation. Then, with the help of the pointer generation network that applies the coverage mechanism, high-quality and readable final summaries are generated. Experimental results on the public long-text dataset NLPCC2017 show that compared with models such as PGN and BERT-PGN, the T5-PEGASUS-PGN model, which combines a pretrained model that fits the downstream summary task better, can generate summaries that are more consistent with the original text semantics, contain richer content, and effectively suppresses repeated content generation. At the same time, we have raised the Rouge-1 metric to 44.26%, the Rouge-2 metric to 23.97%, and the Rouge-L metric to 34.81%.

Keywords

Abstractive Summarization Model, Pre-Trained Language Model, Pointer Generator Network (PGN), Coverage Mechanism

1. 引言

近年来，计算机科技的飞速进步推动了互联网上文本数据的爆炸式增长，新闻内容更是以惊人的速度持续涌现，无时不刻不在更新。用户如何在有限的时间内获取到海量信息的核心内容成了亟需解决的问题，文本摘要算法应运而生。

自动文本摘要主要分为抽取式和生成式两种技术路线[1]。抽取式摘要认为文中每一个句子都可能是文章的摘要，根据概率进行排序，然后选择概率最高的句子来组成摘要；生成式摘要则使用机器学习模型去理解文章的意思从而生成新的摘要。

Luhn [2]最早提出文本摘要的概念，基于统计学原理，对文中的高频词进行统计，对高频词所在句子进行综合打分，将得分最高的句子视为摘要。机器学习的快速崛起推动了文本摘要技术飞速发展。

2014 年 Google Brain [3]提出了 Seq2Seq 模型，即序列到序列模型。编码器和解码器是 Seq2Seq 模型的核心部分，Seq2Seq 模型可以根据给定序列去推理生成另外一个序列。

针对 Seq2Seq 模型处理长序列时早期信息容易被覆盖的问题，Dzmitry Bahdanau 等人[4]将注意力机制引入到 Seq2Seq 模型当中，使模型可以关注到上下文信息，在机器翻译任务中取得显著效果。

PGN [5] (指针生成网络)在 Seq2Seq + attention 的基础上，增加一层指针概率的计算，使模型根据概率自由选择生成新词或从原文中复制单词，避免了 OOV(未登录词)问题的出现。同时在指针生成网络中可以加入覆盖机制，避免注意力过分聚焦某一个单词，从而生成重复的单词或短语。

谭等人[6]在将预训练语言模型 BERT 与 PGN 相结合，提出了分阶段的生成式模型 BERT-PGN，第一阶段通过 BERT 模型获取含有多维语义特征的词向量，第二阶段通过 PGN 模型进行摘要生成，从而获得了贴合原文语义的摘要。

自动文本摘要中的新闻摘要与其他摘要类型的区别在于其时效性、语言风格、信息选取和目的与用途等方面的特点。同时，采用与下游任务更相近的预训练模型可以显著提高模型整体性能。

因此，本文将 T5-PEGASUS 和指针生成网络(PGN)融合，提出了一种改进的针对中文新闻文本摘要生成的模型——T5-PEGASUS-PGN。该模型首先通过 T5-PEGASUS 预训练模型获取富含原文语义特征的文本词向量，生成最接近原文语义的原始摘要，然后通过指针生成网络生成最终的新闻摘要。

2. PEGASUS 模型的相关工作

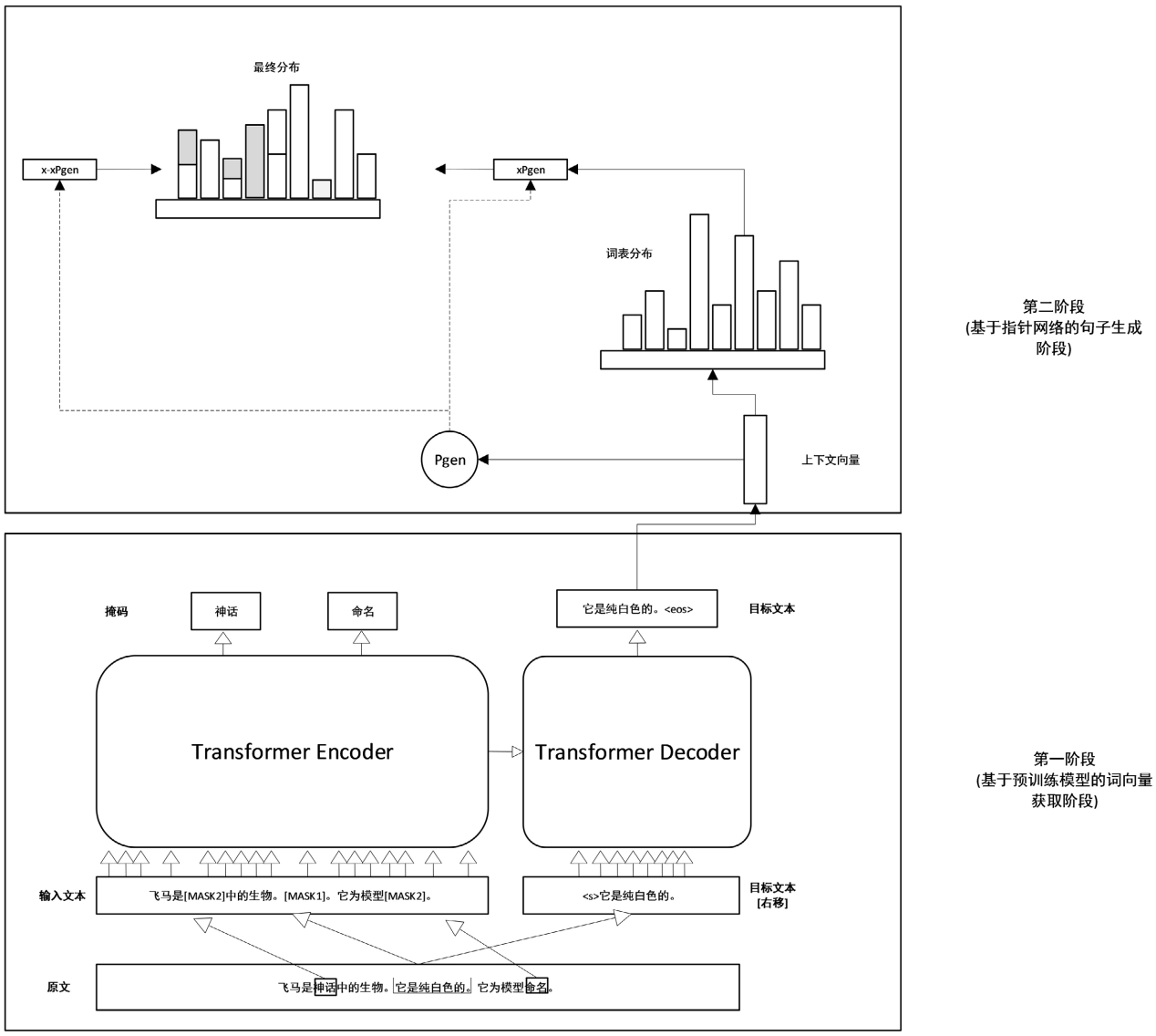

Figure 1. T5-PEGASUS-PGN Model
图 1. T5-PEGASUS-PGN 模型

2020 年谷歌提出了以 Transformer [7]为核心的 PEGASUS [8]预训练模型。PEGASUS 是一种专门为摘要任务设计的预训练模型，其预训练任务构建更加接近摘要任务。与其他通用预训练模型相比，PEGASUS 模型在摘要抽取的效果上表现更佳。PEGASUS 模型包含 GSG 和 MLM 两种预训练任务。MLM 任务分别使用不同的 MASK 对输入文本进行遮掩。编码器部分负责对遮掩部分进行恢复。GSG 任务主要是通过对文本中的重要的句子进行遮掩，解码器部分负责恢复被遮掩的句子。Yang 等人[9]通过微调数据集大小，将 PEGASUS 模型应用在解答数学应用题中，发现训练过程中只需要少量的样本，模型就能取

得令人满意的效果。追一科技基于 T5 模型的多国语言版本 mT5，针对中文特点优化了分词器，其预训练任务的构建参考了 PEGASUS 模型，最终开发出 T5-PEGASUS 模型。Zhang [10]等人将 Pkuseg 分词方法应用到 T5-PEGASUS 模型中，并在多个公开数据集上进行验证其优化的有效性。

大多数基于预训练模型进行提取特征的自动摘要方法存在预训练任务与下游摘要任务差异的问题，不能很好地还原新闻原意。本文提出的 T5-PEGASUS-PGN 模型基于摘要任务定制化的 PEGASUS 预训练模型，针对中文新闻文本，能够获取更丰富的上下文语义，包含更多的全文信息，更符合原文的语义，最终生成质量更高可读性更好的摘要。

3. T5-PEGASUS-PGN 模型

本文提出的两阶段自动摘要模型 T5-PEGASUS-PGN，模型结构如图 1 所示。首先通过预训练模型 T5-PEGASUS 获取新闻文本的语义表示，然后利用引入了 coverage 机制的指针网络生成摘要。模型在第一阶段利用预训练模型 T5-PEGASUS 获取新闻文本的词向量表示；在第二阶段，指针生成网络根据上一步获得的词向量表示进行摘要生成。同时，coverage 机制的引入，一方面模型能够有效抑制重复文本的生成，另一方面保留了模型生成原文中不存在的新词的能力，从而确保模型生成摘要的多样性和准确性，最终实现高质量的新闻摘要生成。

3.1. 词向量获取阶段

3.1.1. T5-PEGASUS 预训练模型

语言模型本质上是一种概率预测模型，它是建立在概率论、统计学、信息论和机器学习等技术之上的，其建模思想在自然语言处理领域广泛应用，在文本生成、文本分类、机器翻译等任务中发挥着关键性作用。

随着深度学习技术的发展，语言模型与深度学习相结合形成了如今的神经语言模型，这种模型经过大量数据样本的训练后，表现出强大的表示能力和学习能力，这标志着语言模型的发展进入了一个新的时期。预训练语言模型作为神经语言模型的后起之秀，弥补了训练过程中数据样本标注不足的缺点，正逐渐成为自然语言处理领域技术应用的基石。

T5-PEGASUS 模型以 T5 模型的多国语言版本 mT5 为核心，然后根据中文分词特点将分词器替换为 BERT 模型的分词器，同时在分词过程中加入 jieba 分词，最后，参照 PEGASUS 模型的思路，利用庞大的中文语料库构建了预训练任务。

与 GPT [11]和 BERT [12]模型不同，T5-PEGASUS 模型采用完整的 Encoder-Decoder 结构。使用 Transformer 可以提取到更多的上下文信息，其原理是根据上下文进行字的表示，这解决了传统神经网络模型无法解决一词多意的问题。Transformer 的 Encoder 由多个 Encoder 层叠加而成，每个 Encoder 层有两个子层，第一个子层包含多头注意力机制，第二个子层由前馈全连接网络构成。自注意力机制的输入部分由三个向量组成，分别是 Query 向量(Q)、Key 向量(K)和 Value 向量(V)，这三个向量都来自同一个字。通过 Q 和 V 矩阵相乘表示字向量的相似度，然后通过键的维度 d_k 进行矩阵缩放，避免 *soft*max 函数在计算过程中出现梯度消失或爆炸的问题。最终的词向量表示由 *soft*max 函数得到。具体计算方式如下：

$$Attention\left(Q,K,V\right)=soft\max\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{1}$$

Transformer 的 Decoder 由多个 Decoder 层叠加组成，每个 Decoder 有三个子层，主要结构分别是带掩码的多头自注意力子层、多头注意力子层(编码器到解码器)以及前馈全连接子层。其中带掩码自注意力

层的具体计算公式如下：

$$Attention(Q,K,V)=soft\max\left(\frac{QK^T+M}{\sqrt{d_k}}\right)V \tag{2}$$

其中 M 矩阵的元素为 0 或者$-\infty$。M 全为 0 则表示正常的自注意力层，如果是$-\infty$，则 *soft*max 的结果为 0，即权重 0，表明信息被遮掩。

3.1.2. 相对位置编码

为了更好地捕捉语序关系，处理变长输入序列，T5-PEGASUS 模型采用相对位置编码。具体计算公式如下：

$$Attention(Q,K,V)=soft\max\left(\frac{Q(K+R_K)^T}{\sqrt{d_k}}\right)V \tag{3}$$

3.1.3. 门控线性单元 GLU

T5-PEGASUS 的全连接层的激活函数采用门控线性单元 GLU，既降低了梯度弥散又保留了其非线性能力。实现公式如下：

$$FFN(x)=(GeLU(xW_1)\otimes xW_2)W \tag{4}$$

当输入值较大或较小的情况下，使用 GELU 函数可以避免产生梯度消失的问题。同时，当输入值为负数时，避免神经元死亡的问题。

3.1.4. 伪摘要式训练任务

预训练任务与下游任务之间的相似性或近似程度，可以对模型的性能产生显著影响。在预训练过程中选择最重要的句子用掩码[MASK]将其屏蔽，T5-PEGASUS 负责将句子重建。预训练过程中模型通过计算每个句子与其他句子的 ROUGE1-F1 指标作为句子重要的判断依据，以此为参考选出最重要的句子，从而更加接近摘要。

3.2. 基于指针网络的句子生成阶

指针生成网络是通过引入注意力机制的 Seq2Seq 模型改进得到的，主要由编码器、解码器和指针生成器三部分组成。该模型可以根据实际情况灵活选择直接从原文中复制词语或者是生成新的词语。原文本经过预训练模型 T5-PEGASUS 得到语义丰富和具备上下文特征的词向量，随后传入到 BiGRU 编码器，生成隐层状态序列 h_i。在 t 时刻，上一时刻生成的词向量输入到单向 GRU 解码器可以通过解码得到当前状态序列 s_t。

注意力权重由解码器在 t 时刻使用上一时刻的隐层状态与编码器的输出加权求和后通过 *soft*max 运算得到。计算公式如下：

$$S_i^t=v^T\tanh(w_h h_i+W_s s_t) \tag{5}$$

v，W_h，W_s 是通过训练得到的参数。

$$a_i^t=soft\max(s_i^t) \tag{6}$$

将注意力权重矩阵与由编码器得到的隐层状态矩阵加权取平均值，得到上下文向量表示 h_t^*。

$$h_t^*=\sum_i a_i^t h_i \tag{7}$$

上下文向量 h_t^* 与上一时刻的解码序列 s_t 连接，作为当前时刻解码器的输入，经过线性变换后生成当前预测在字典上的概率分布 P_{vocab}，计算公式如下：

$$P_{vocab} = soft\max\left(V'\left(V\left[s_t, h_t^*\right]\right)\right) \tag{8}$$

指针网络通过参数 P_{gen} 来判断直接从原文中复制单词还是由模型自由生成新单词，计算公式如下：

$$P_{gen} = soft\max\left(w_h^T * h_i^* + w_s^T s_t + w_x^T x_t + b_{ptr}\right) \tag{9}$$

$S_i^t = v^T \tanh\left(w_h h_i + W_s s_t\right)$ 其中，W_h、W_s、W_x、b_{ptr} 是训练过程中的参数，x_t 为当前时刻输入到解码器中的序列。

进行文本生成时，通过如下公式计算扩展字典的概率分布：

$$P(w) = p_{gen} p_{vocab}(w) + \left(1 - p_{gen}\right)\sum a_i^t \tag{10}$$

为了解决摘要中包含重复文本的问题，计算中加入覆盖向量 c^t 作为惩罚系数，对重复生成的词进行惩罚，从而抑制重复内容的生成。c^t 的计算方式如下：

$$c^t = \sum_{t=0}^{t-1} a^t \tag{11}$$

在注意力机制中覆盖程度的向量表示为 c^t，将其纳入注意力机制的计算过程中，可以形成新的注意力分布，从而避免重复关注相同的位置，进而降低生成重复文本的可能性，计算公式如下：

$$e_i^t = v^T \tanh\left(W_h h_i + W_s s_t + W_c c_i^t + b_{attn}\right) \tag{12}$$

4. 实验与分析

4.1. 实验数据

本文采用的数据集为中文长文本新闻摘要数据集 NLPCC2017，包含训练集和测试集两部分。其中训练集包含 49,500 条新闻–摘要对，测试集包含新闻–摘要对 500 篇。正文平均字数 990 字，摘要平均字数 44 字，为长文本摘要数据集。

4.2. 评价指标

本文采用 Rouge [13]评价模型性能。Rouge 是用于评估机器自动摘要的度量指标。其核心思想是通过统计模型生成的摘要句子和标准摘要句子相同的 n-gram 比率来评价摘要生成质量。具体来说，Rouge 指标基于 n 元词(n-gram)的共现概率，即 n 元词在模型生成摘要和参考摘要中共现的概率，以此来评估摘要的质量。本文选取 Rouge-1、Rouge-2 和 Rouge-L 作为评价指标，对摘要的质量进行全面评估。Rouge-1、Rouge-2 和 Rouge-L 是 Rouge 指标的三种变体，它们分别基于不同的 n-gram 长度和不同的相似度计算方法。

4.3 对比实验

本文实验部分将 Seq2Seq + Attention 模型、指针生成网络 PGN、BERT-PGN、T5-PEGASUS-PGN 在 NLPCC2017 数据集上进行对比实验，通过 Rouge 指标综合评判本文提出模型的改进效果，从而验证改进有效性。

4.4 实验环境及参数设置

本文的实验环境如表 1 所示。

Table 1. Experimental environment
表 1. 实验环境

软硬件	配置/版本
操作系统	Windows 11
CPU	AMD Ryzen9 5900HX
GPU	NVIDIA GeForce RTX 3080
开发工具	PyCharm
Python	3.10
CUDA	11.7
torch	2.0.1
transformers	4.32.1

超参数设置的合适与否将直接影响模型的性能，因此需要根据以往经验以及不断试验，对超参数进行调整，才能最终确定合适的超参数。

经过多次迭代实验对超参数进行调整，最终确定最优化的参数设置如表 2 所示。

Table 2. Parameter setting
表 2. 参数设置

参数类型	参数值
迭代次数	500 k
字典大小	50 k
学习率	10e−4
batch_size (T5-PEGASUS)	32
batch_size (指针网络)	16
新闻最大长度	512
摘要最大长度	64
训练集:测试集	8:2
隐层(T5-PEGASUS)	768 维
隐层(指针网络)	256 维
注意力头(T5-PEGASUS)	12
训练时长	6 d 10 h

4.5. 实验结果与分析

4.5.1. 摘要结果对比实验

T5-PEGASUS-PGN 与 baseline 模型的实验对比结果如表 3 所示。

Table 3. Comparison of rouge scores for different models
表 3. 不同模型 Rouge 分数对比

模型	Rouge-1	Rouge-2	Rouge-L
Seq2Seq + Attention	31.07%	18.74%	29.11%
PGN	35.85%	21.24%	31.79%
BERT-PGN	37.56%	21.96%	32.05%

续表

T5-PEGASUS	42.06%	21.43%	31.64%
T5-PEGASUS-PGN	44.26%	23.97%	34.81%

从表 3 中各模型 Rouge 分数对比可以看出，NLPCC2017 长文本摘要数据集上，PGN 模型相比于 Seq2Seq + Attention 模型在 Rouge-1，Rouge-2 以及 Rouge-L 上的分数分别提升至 35.85%、21.24%和 31.79%，说明基于指针网络和引入 coverage 机制的 PGN 模型得到了有效改进。BERT-PGN 模型相应的 Rouge-1 提升至 37.56%，Rouge-2 提升至 21.96%和 Rouge-L 提升至 32.05%，说明 BERT 预训练模型的引入，提高了 BERT-PGN 模型对上下文信息的理解能力，因此提高了模型的性能。而 T5-PEGASUS-PGN 模型，对应的 Rouge-1 提升至 44.26%，Rouge-2 提升至 23.97%和 Rouge-L 提升至 34.81%，这表明预训练任务更接近下游摘要任务的 T5-PEGASUS 预训练模型和指针生成网络的加入，改进了模型效果，使得模型对于文本语义理解更加准确，并且精准地概括原文主旨，从而生成质量更高的摘要。

4.5.2. coverage 机制分析

本文通过计算不同的 N-gram 片段的占比来验证 coverage 机制解决文本生成重复问题的有效性。其中 N-gram 表示一个句子中 N 个连续的词语。

由表 4 可知，本文提出的 T5-PEGASUS-PGN 相较于 Seq2Seq + Attention 模型，有效的减少了重复内容的生成。

Table 4. Analysis of coverage mechanism
表 4. coverage 机制分析

N-gram	Seq2Seq + Attention	T5-PEGASUS-PGN	标准摘要
1-gram	27%	22%	21%
2-gram	19%	7%	8%
3-gram	18%	2%	3%
4-gram	12%	2%	1%

4.5.3. Loss 值对比

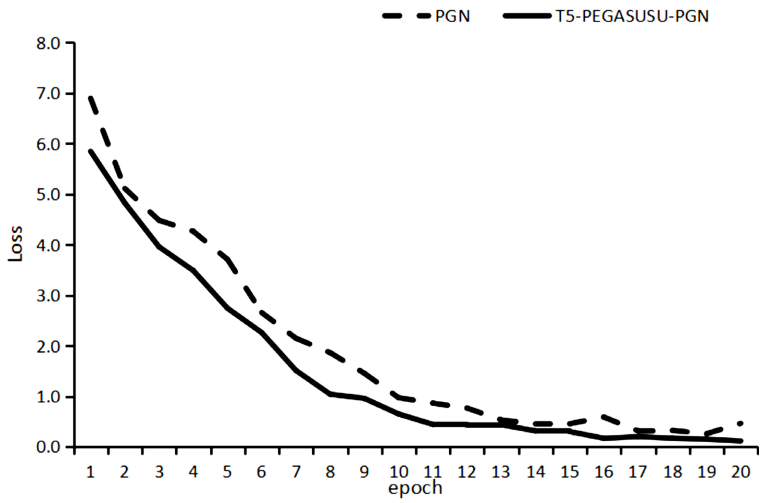

Figure 2. Model Loss convergence
图 2. 模型 Loss 收敛

本文将 T5-PEGASUS-PGN 模型与 PGN 模型的收敛情况对比，从另一个角度体现模型的改进效果，Loss 曲线图如图 2 所示。

由图 2 可知，因为采用更贴合下游摘要任务的 T5-PEGASUS，使得 T5-PEGASUS-PGN 模型可以更好地获取上下文信息以及更加丰富的语义信息，与 PGN 模型相比，T5-PEGASUS-PGN 模型收敛速度更快，Loss 值也相对更低。

4.5.4. 不同 BeamSize 对 Rouge 的影响

模型采用集束搜索算法解码，其参数 BeamSize 会对最终生成的摘要产生影响，因此需要对比不同 BeamSize 下模型生成摘要的 Rouge 分数，从而确定最合适的 BeamSize，具体实验结果如表 5 所示。

Table 5. Comparison of Rouge scores for different Beamsizes
表 5. 不同 BeamSize 的 Rouge 分数对比

BeamSize	Rouge-1	Rouge-2	Rouge-L
1	41.11%	23.79%	33.03%
2	40.46%	22.81%	32.56%
3	**44.26%**	**23.97%**	**34.81%**
4	40.95%	24.25%	35.58%
5	42.37%	23.99%	33.41%

由上表可知，BeamSize 设置为 3 时，Rouge 分数均为最大值，另一方面 BeamSize 越大，搜索空间越大，会降低模型的解码速度，因此综合考虑，将 BeamSize 设置为 3。

5. 结语

本文提出了一种创新的中文新闻文本摘要模型 T5-PEGASUS-PGN。该模型利用更适用于摘要任务的 T5-PEGASUS 预训练模型来获取词向量，并将其与引入了 coverage 机制的 PGN 模型相融合，有效减少了未登录词和重复内容的生成，既提高了模型生成摘要的质量又提高了生成的摘要的可读性。实验结果表明，T5-PEGASUS-PGN 模型生成的摘要更贴近原文语义，更接近标准摘要，同时包含更少的冗余内容。

为了获取更高质量的新闻摘要，利用好新闻文本主题性强这一特点，我们计划挖掘新闻原文中隐含的主题信息并将其与注意力机制相结合，同时使用 Transformer 模型解决 RNN 只能串行计算而无法并行计算的问题，提高模型训练速度和生成摘要的速度；尝试引入外部新闻知识库，进一步提高模型对上下文的理解能力。

参考文献

[1] 李金鹏, 张闯, 陈小军, 等. 自动文本摘要研究综述[J]. 计算机研究与发展, 2021, 58(1): 1-21.

[2] Luhn, H.P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, **2**, 159-165. https://doi.org/10.1147/rd.22.0159

[3] Sutskever, I., Vinyals. O. and Le. Q.V. (2014) Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, **27**, 3104-3112.

[4] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science*, 1-16.

[5] See, A., Liu. P.J. and Manning, C.D. (2017) Get to the Point: Summarization with Pointer-Generator Networks. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Vancouver, July 2017, 1073-1083. https://doi.org/10.18653/v1/P17-1099

[6] 谭金源, 刁宇峰, 祁瑞华, 等. 基于 BERT-PGN 模型的中文新闻文本自动摘要生成[J]. 计算机应用, 2021, 41(1): 127-132.

[7] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. arXiv: 1706.03762.

[8] Zhang, J., Zhao, Y., Saleh, M., *et al*. (2020) PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. *Proceedings of the* 37*th International Conference on Machine Learning*, Vol. 119, 13-18 July 2020, 11328-11339.

[9] Yang, T.H., Lu, C.C. and Hsu, W.L. (2021) More than Extracting "Important" Sentences: The Application of PEGASUS. *International Conference on Technologies and Applications of Artificial Intelligence*, Taichung, 18-20 November 2021, 131-134. https://doi.org/10.1109/TAAI54685.2021.00032

[10] 张琪, 范永胜. 基于改进 T5 PEGASUS 模型的新闻文本摘要生成[J]. 电子科技, 2023, 36(12): 72-78.

[11] Radford, A., Wu, J., Child, R., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, **1**, 9.

[12] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[13] Lin, C.Y. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out. Stroudsburg*: *Association for Computational Linguistics*, Barcelona, July 2004, 74-81.