计算机科学
COMPUTER SCIENCE

结合预训练的多文档摘要研究

丁一，王中卿

引用本文

丁一，王中卿. 结合预训练的多文档摘要研究[J]. 计算机科学，2024，51(6A)：230300160-8.

DING Yi, WANG Zhongqing. Study on Pre-training Tasks for Multi-document Summarization[J]. Computer Science, 2024, 51(6A): 230300160-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300160

结合预训练的多文档摘要研究

丁 一 王中卿
苏州大学计算机科学与技术学院 江苏 苏州 215006
(1959001912@qq.com)

摘 要 新闻文本摘要任务旨在从庞大复杂的新闻文本中快速准确地提炼出简明扼要的摘要。基于预训练语言模型对多文档摘要进行研究,重点研究结合预训练任务的具体模型训练方式对模型效果提升的作用,强化多文档之间的信息交流,以生成更全面、更简练的摘要。对于结合预训练任务,提出对基线模型、预训练任务内容、预训练任务数量、预训练任务顺序的对比实验,探索标记了行之有效的预训练任务,总结归纳了强化多文档之间的信息交流的具体方法,精炼提出了简明高效的预训练流程。在公开新闻多文档数据集上进行训练和测试,实验结果表明预训练任务的内容、数量、顺序对ROUGE值都有一定提升,并且整合三者结论提出的特定预训练组合对ROUGE值有明显提升。

关键词:新闻;摘要;预训练;多文档;信息交流
中图分类号 TP391

Study on Pre-training Tasks for Multi-document Summarization

DING Yi and WANG Zhongqing
School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China

Abstract News summarization aims to quickly and accurately extract a concise summary from the complex news text. This paper studies the multi-document summary based on the pre-training language model, focusing on the effect of model training methods combined with pre-training tasks on improving model performance, and strengthening information exchange between multiple documents to generate more comprehensive and brief summaries. For combined pre-training tasks, this paper conducts comparative experiments on the baseline model, pre-training task content, pre-training task quantity, and pre-training task order, explores and marks effective pre-training tasks, summarizes the specific methods to strengthen the information exchange between documents, and refines and proposes a concise and efficient pre-training process. Through training and testing on the public news multi-document dataset, experimental results show that the content, quantity, and order of the pre-training tasks have a certain improvement on the ROUGE value, and the specific pre-training combination proposed by integrating the conclusions of the three has a significant increase in the ROUGE value.

Keywords News, Summarization, Pre-training, Multi-document, Information exchange

1 引言

随着在线出版物数量的迅速增长,通过多文档摘要从多篇新闻中获取信息成为一个有意义的方法。多文档摘要有助于提高读者的阅读效率,同时提供更多客观的信息,帮助读者更好地了解相关事件。

现有的多文档摘要方法经常直接将单文档摘要模型应用在多文档摘要任务上。然而,单文档摘要模型难以捕获跨文档关系,并且无法处理输入文档之间的高度冗余和矛盾。因此,有必要采用有效的方法来利用输入文档之间的关系,从而提升模型效果。在过去的论文中,各种图神经网络(Graph Neural Networks)、编码器-解码器结构(Encoder-Decoder Structure)、预训练语言模型(Pre-trained Language Models)、层次网络(Hierarchical Networks)被提出用以解决该问题。

多文档摘要仍然存在一些挑战和问题需要解决:如何提高自动摘要的准确性和可读性,如何增强文本之间的信息交流,都需要解决。因此,本项目仍需要进一步深入研究和探索,完善已有的方法,并探索更有效的解决方案。

为更好地解决上述问题,本文将在预训练语言模型的基础上,对特定数据集的输入文档之间的关系进行连续不同的预训练,来让模型更好地理解文档间关系,以达到保留文档的重要部分的同时,又降低了冗余程度的效果。

本项目基于预训练语言模型对多文档摘要进行研究,重点针对结合预训练任务的模型训练方式对模型效果提升的作用,强化多文档之间的信息交流。对于如何结合预训练任务,提出了多样化的文本处理方式,并对本项目进行任务内容、数量、顺序的对比实验。

通过在公开新闻多文档数据集Multi-News上进行训练和测试,尝试让模型在ROUGE值上获得有效的提升。实验结果表明,本文提出的方法可以有效地提高多文档新闻摘要的性能。

论文通过构造不同预训练任务的形式,构造适用于多文档摘要任务的预训练模型。本文主要有3个主要意义。1)提高模型性能:预训练任务可以帮助模型在特定任务上获得更

通信作者:王中卿(wangzq@suda.edu.cn)

好的性能。通过预训练，模型可以学习到更多的语言模式和知识，从而在后续的多文档摘要任务上表现得更好。2）解决数据稀疏问题：在多文档摘要任务中，可能会遇到数据稀疏的问题，即某些特定的语言模式或知识在训练数据中出现的次数较少，导致模型难以学习到这些模式或知识。通过预训练任务，模型可以在更大的语料库上进行学习，从而缓解数据稀疏的问题。3）提高模型的泛化能力：预训练任务可以帮助模型学习到更广泛的语言模式和知识，从而提高模型的泛化能力。这意味着模型可以更好地处理在训练数据中未出现过的新的语言模式或知识。

合理的预训练任务组合可以解决多文档摘要中存在的一些问题。关于任务数量，过多的预训练任务可能会导致模型的训练变得复杂和困难，可能需要更多的计算资源和时间。我们旨在从任务数量的研究中找到模型效果和计算资源的平衡点。关于任务顺序，如果先进行的预训练任务与摘要任务关联性较低，那么模型可能需要更多的时间和数据来适应摘要任务。我们旨在从任务顺序的研究中探索合理的预训练任务顺序，以提高模型的学习效率和性能。

2 相关工作

本文的多文档摘要与生成式自动摘要和多文档摘要有关，以下将介绍这两方面的相关工作。

2.1 生成式自动摘要

现在的自动摘要技术主要分为两种：抽取式和生成式。在生成式自动摘要中，目标是生成新的摘要文本，而不是从原文中提取句子或短语。生成式自动摘要更接近人类写作摘要的方式，并对不同的应用场景具有更好的适应性。由于生成式自动摘要可以理解文本的含义和上下文，并利用生成的过程产生更加连贯和自然的文本，它的摘要内容往往更加言简意赅。生成式自动摘要将任务转换为序列对序列（sequence-to-sequence，seq2seq）问题，其中编码器将源文档中的标记序列 $x=[x_1,\cdots,x_n]$ 映射到连续表示序列 $z=[z_1,\cdots,z_n]$，然后解码器以自回归的方式逐个标记生成目标摘要的标记序列 $y=[y_1,\cdots,y_m]$，因此建立条件概率模型 $(y_1,\cdots,y_m\mid x_1,\cdots,x_n)$。Rush 等[1]是生成式自动摘要的先驱，使用基于注意力机制的摘要（Attention-Based Summarization，ABS）在 Gigaword 数据集上取得 ROUGE-1 的最高分。Nallapati 等[2]使用循环神经网络（Recurrent Neural Network，RNN）改善了生成式自动摘要的效果。See 等[3]通过指针生成网络（Pointer-generator Network，PTgen）来从源文本中复制单词，并通过覆盖机制（Coverage Mechanism，Cov）增强了对已摘要的词的追踪能力。Celikyilmaz 等[4]在解码器上使用了分层注意力机制，并在模型训练中融入了强化学习。Paulus 等[5]提出深度强化模型（Deep Ieinforced Model，DRM），该模型利用内部注意力机制解决生成短语重复的问题。Zhang 等[6]提出了一种基于预训练语言模型的生成式自动摘要方法，该方法使用 Transformer 模型进行预训练，在 CNN/Daily Mail 数据集上产生了更加流畅的文本。Dong 等[7]提出一种用于自然语言理解和生成的统一语言模型预训练方法，称之为 GPT-2，并证明在大规模预训练语言模型上进行微调将进一步提升模型效果。Micheli 等[8]提出了 GPT-3 模型，并通过在大规模文本语料库上进行训练，使模型能够完成多种自然语言处理任务。Zhang 等[9]在 GPT-2 的基础上进行微调和训练，将模型应用在聊天机器人领域，将模型命名为 DialoGPT，为后续 ChatGPT 的出现奠定了基础。

2.2 多文档摘要

多文档摘要考虑多篇文档中信息的相互关系和重复，旨在从不同时间、不同视角编写的文档中总结出更全面、更准确、更丰富的摘要。但由于它试图解决潜在的多样性和冗余信息，模型需要具备更强的分析输入文档及识别和合并一致信息的能力。多文档摘要具有广泛的现实世界应用，包括对新闻、科学出版物、电子邮件、产品评论、医学文档、讲座反馈、百科文章的摘要任务。

在早期的多文档摘要研究中，研究者主要会使用传统的信息抽取和信息检索技术来生成摘要，例如基于词频抽取的方法[10]、聚类[11]、图[12]和潜在语义分析[13]。随着深度学习技术的发展，越来越多的研究者开始使用深度学习模型来生成多文档摘要。随着计算能力的显著提高和越来越多公共数据集的发布，MDS 逐步运用了更深层次和更复杂结构的神经网络，具有更好效果且更具鲁棒性的模型从而不断涌现。

Mao 等[14]为多文档摘要提出了一种强化学习框架，它通过端到端的学习统一了单文本生成和最大边际相关性（Maximal Marginal Relevance-Based，MMR）。通过在基准多文档摘要数据集上的验证，证明了框架在处理大搜索空间和冗余方面的卓越性能。Liu 等[15]在 Transformer 的基础上，以分层方式对文档进行编码，并通过共享信息的注意力机制获取跨文档关系，增强了摘要的流畅性和准确性。Alexander 等[16]收集并归纳了第一个大规模多文档新闻数据集 Multi-News，该数据集相较过往新闻数据集在规模和质量上存在明显的提升，对促进多文档新闻摘要的发展起到重要作用。Arumae 等[17]提出了一个基于问答奖励机制的抽取式文本摘要模型，奖励机制促进了摘要的流畅性，并在面对重要问题时作为文本替代回答。Zhang 等[18]认为抽取式摘要比生成式摘要更为可靠，但抽取式摘要需要句子级别的标签，作者通过提出 HIBERT 框架在 NYT50 数据集上实现了 SOTA。Xu 等[19]提出一种关注关键词并进行语义匹配的多文档摘要方法，通过查询扩展和关键词筛选技术提高匹配度，在 ROUGE 评价指标上表现出色。MA 等[20]撰写了一篇关于基于深度学习技术的多文档摘要的综述性论文，综述了近年来基于深度学习的多文档摘要方法，并分析各种方法的优缺点和应用场景。

3 方法

本文研究同一主题下的多篇新闻文本生成文本摘要的任务。首先，在新闻数据集的处理上，采用多样化的处理方式，包括 BERT-style 遮掩方式、全遮掩方式、关键词提取和文本拼接，最终达到增大数据集，进而提高模型的泛化能力，同时避免模型过拟合原始数据集的效果。其次，在 T5 的基础上，采取针对新闻文本的进一步预训练，包括不同的预训练内容，不同的预训练数量，不同的预训练顺序。最后，使用进一步预训练的模型进行多文档摘要任务，并通过对比实验归纳结论。结合两轮预训练的具体模型框架图如图 1 所示。

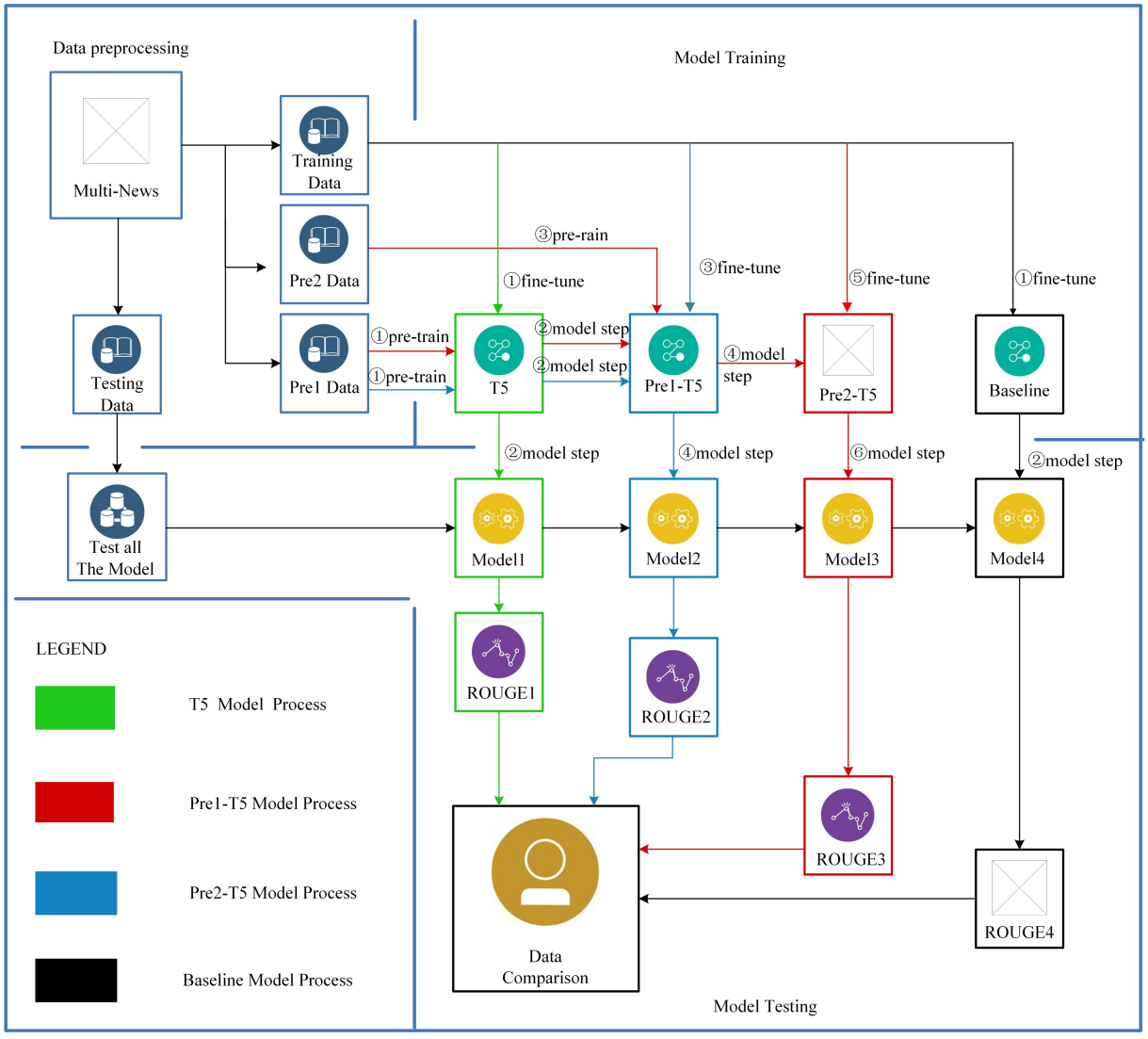

图 1　结合两轮预训练的基础模型框架图

Fig. 1　Framework of the basic model combined with two rounds of pre-training

3.1　基于预训练的生成模型

T5(Text-to-Text Transfer Transformer)是一种基于Transformer架构的预训练语言模型。它提出一种将各种NLP任务转换为Text-to-Text任务的统一模型框架,以此评估不同的模型结构、预训练目标函数和无标签数据集的影响。T5采用了预训练-微调的方法,将一个大型语料库训练成模型,然后在任务特定的数据集上进行微调。通过对比实验,在模型结构方面采用了以Denoising为目标函数的encoder-decoder Transformer结构。Transformer采用了自注意力机制,并在机器翻译和语言建模方面取得了最先进的结果,后来也被成功应用在了SDS任务上。在预训练目标函数方面,肯定了BERT-style预训练目标的有效性。在无标签数据集方面,实验验证了选取经过数据清洗且包含一定领域数据的数据集对下游任务更有效。在本文的实验中,将对T5模型进行针对指定数据集的额外预训练任务,而后将其应用到下游多文档摘要任务中。

本论文只选取T5模型作为各项预训练任务的测试对象。由于论文的目标是构造适用于多文档摘要任务的预训练模型,而这些预训练任务理论上应该可以迁移到其他的生成模型,如BART等。这是因为这些预训练任务主要是通过改变模型的输入和输出来训练模型学习特定的语言模式和知识,而这些语言模式和知识是通用的,不仅限于特定的模型。BART等模型和T5都是基于Transformer的生成模型,它们都使用了Transformer的自注意力机制来捕获输入序列中的长距离依赖关系,并且都使用了生成模型的架构来生成输出序列。

3.2　基础预训练任务

在数据处理部分,对源文档中的标记序列 $x=[x_1,\cdots,x_n]$ 使用〈t〉进行标注,并用 t 表示标注后的序列。对目标摘要的标记序列 $y=[y_1,\cdots,y_m]$ 使用〈g〉进行标注,并用 g 表示表示。在对数据进行掩码操作时,分别进行了两种遮掩方式:BERT-style遮掩方式和全遮掩方式。

BERT-style遮掩方式是随机选择标记序列中15%的标记,然后将其中80%的标记使用[$MASK$]进行替换,10%的标记使用标记序列中的随机标记进行替换,10%的标记保持不变。使用BERT-style方式处理的源文档标记序列和目标摘要标记序列分别使用 $\hat{t}$ 和 $\hat{g}$ 表示。

全遮掩方式是将标记序列中的20%用[$MASK$]进行替换,使用全遮掩方式处理的源文档标记序列和目标摘要序列分

别使用 t 和 g 表示。两种遮掩方式的具体处理如图 2 所示。

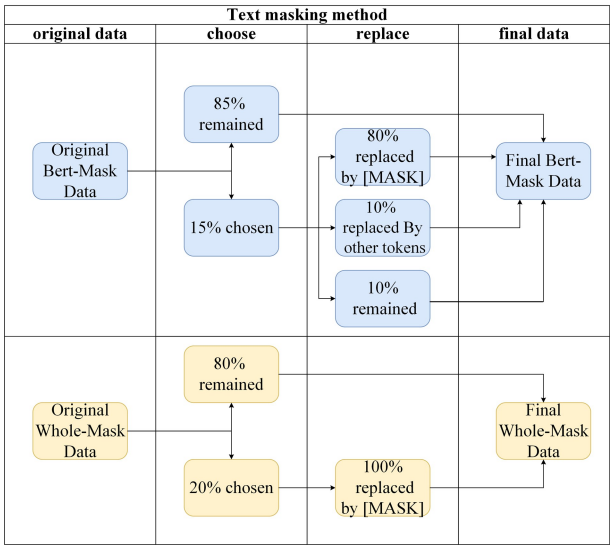

图 2 文本遮掩方式

Fig. 2 Text masking method

为了让模型学习文档间的关系，提取多文档中各篇文档作为预训练数据集。由于各个标记序列中包含的文档数量不一定一致，将对数据进行一定的特殊处理。

在使用的新闻文本中，经过统计可知所有标记序列至少包含 2 篇文档，且超过 4 篇文档的比例偏低，具体数统计见 4.1 节实验数据部分。因此，对前两篇文档进行全部提取，提取结果分别用 t_1 和 t_2 表示；对第三篇文档，若存在第三篇文档，则保留，否则使用最接近且存在的文档代替，对第四篇文档的处理同理，提取结果分别用 t_3 和 t_4 表示。

我们还采用了词频-逆向文件频率(Term Frequency-Inverse Document Frequency，TF-IDF)方法提取源文档中的关键词作为数据集，以达到增强模型对关键词的提取能力。TF-IDF 是一种用于信息检索与文本挖掘的常用加权技术。在这种统计方法中，字词的重要性与它在文件中出现的次数成正比，但与它在语料库中出现的频率成反比。经过关键词提取的前两篇源文档标记序列用 t_{1k} 和 t_{2k} 表示。在具体的预训练任务中，将对上述数据集进行拼接，例如 $\hat{t}g$ 表示 $\hat{t}$ 与 g 拼接后的数据集，并用 $\hat{t}g2t$ 表示以 $\hat{t}g$ 为输入数据且以 t 为输出数据的具体预训练任务。

3.3 额外的预训练任务

本小节将对预训练任务进行介绍，单独的预训练内容主要概括为以下 9 种：$\hat{t}2t$，$g2t$，$\hat{t}2\hat{g}$，$\hat{tg}2g$，$\hat{g}t2t$，$\hat{t}g2t$，$t_1 2g$，$t_1 2t_2$，$t_{1k}2g$，并将预训练内容分为 4 类，分别是自生成任务、辅助生成任务、单文档生成任务、关键词生成任务。我们通过结合预训练任务来鼓励文本与文本之间的信息交换，旨在比对单独预训练内容对模型效果提升的优劣。具体的训练内容如表 1 所列。

1)源文档去噪($\hat{t}2t$)，其中对以 BERT-sytle 方式遮掩的源文档标记序列进行去噪任务，以培养模型对源文档的认知能力。

2)目标摘要生成源文档($g2t$)，其中用目标摘要标记序列生成源文档标记序列，以培养模型认知目标摘要和源文档之间的反向关系。

3)遮掩源文档生成遮掩目标摘要($\hat{t}2\hat{g}$)，其中使用 BERT-sytle 方式遮掩的源文档标记序列生成 BERT-sytle 方式遮掩的目标摘要标记序列，通过残缺文本生成任务培养模型的文本生成能力。

4)源文档辅助目标摘要去噪($\hat{tg}2g$)，使用完整的源文档标记序列辅助 BERT-style 方式遮掩的目标摘要标记序列进行去噪任务，以培养模型运用源文档进行文本生成的能力。

5)目标摘要辅助源文档去躁($\hat{g}t2t$)，使用完整的目标摘要标记序列辅助 BERT-style 方式遮掩的源文档标记序列进行去噪任务，以培养模型运用目标摘要进行文本生成的能力。

6)全遮掩目标摘要辅助源文档去躁($\hat{t}g2t$)，使用全遮掩的目标摘要标记序列辅助 BERT-style 方式遮掩的源文档标记序列进行去噪任务，以强化模型利用源文档的能力。

7)第一篇源文档生成目标摘($t_1 2g$)，使第一篇源文档标记序列生成目标摘要标记序列，增强模型对第一篇源文档的理解程度，同理将对第二、三、四篇源文档进行相关实验。

8)第一篇源文档生成第二篇源文档($t_1 2t_2$)，使用完整的第一篇源文档标记序列生成第二篇源文档标记序列，充分利用前两篇文档之间的关联。

9)第一篇源文档关键词生成目标摘要($t_{1k}2g$)，使用第一篇源文档的关键词标记序列生成目标摘要标记序列，强化模型对第一篇源文档中的关键词的注意力，减少生成摘要的冗余程度，同理将对第二篇源文档关键词进行生成目标摘要任务，并与此任务进行对比。

表 1 预训练内容

Table 1 Pre-training contents

预训练内容	分类	输入数据	输出数据
$\hat{t}2t$	自生成任务	$\hat{t}$	t
$g2t$	自生成任务	g	t
$\hat{t}2\hat{g}$	自生成任务	$\hat{t}$	$\hat{g}$
$\hat{tg}2g$	辅助生成任务	$\hat{tg}$	g
$\hat{g}t2t$	辅助生成任务	$\hat{g}t$	t
$\hat{t}g2t$	辅助生成任务	$\hat{t}g$	t
$t_1 2g$	单文档生成任务	t_1	g
$t_2 2g$	单文档生成任务	t_2	g
$t_3 2g$	单文档生成任务	t_3	g
$t_4 2g$	单文档生成任务	t_4	g
$t_1 2t_2$	单文档生成任务	t_1	t_2
$t_{1k}2g$	关键词生成任务	t_{1k}	g
$t_{2k}2g$	关键词生成任务	t_{2k}	g

4 实验

4.1 实验数据

本文数据集来源于 Multi-News[21]，它是一个大规模的多文档数据集，其中包含大量的新闻文章和对应的人类撰写的摘要。它旨在评估模型生成摘要算法的性能，并且可以作为其他研究的基线数据集。该数据集经过精心构建和预处理，保证了数据的多样性和专业性。

Multi-News 数据集规模为 56216 条，其中 80%作为训练集，10%作为验证集，10%作为测试集。数据集为了反映现实世界的情况——对于一个新的或特殊的事件，通常只有几篇新闻报道，因此使每个示例的文档数量有所不同。数据集中

每个示例的文档数量统计如 表 2 所列。

表 2　每个示例的文档数量统计(按频数)

Table 2　Statistics on the number of documents per sample(by frequency)

文档数	频数	文档数	频数
2	23894	7	382
3	12707	8	209
4	5022	9	89
5	1873	10	33
6	763		

4.2　评价方法

本文使用 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[22]作为标准摘要评价指标。ROUGE 是自然语言领域常用的文本摘要评价指标,它通过比较生成的摘要与人工标准摘要的重叠部分来评估摘要质量。ROUGE 包括 ROUGE-N(n-gram 重叠)、ROUGE-L(最长公共子序列)和 ROUGE-W(加权 n-gram 重叠)等多个版本,本文实验选取 ROUGE-N 和 ROUGE-L 两个指标进行评测。

ROUGE 计算公式中的召回率(Recall)表示生成的摘要中与标准摘要重叠部分的占比,而准确率(Precision)表示生成摘要中和标准摘要重叠的部分占生成摘要的比例,具体计算公式如式(1)－式(4)所示。

$$ROUGE\text{-}N=\frac{\sum_{S\in Y}\sum_{gram_n\in S}Count_{\text{match}}(gram_n)}{\sum_{S\in Y}\sum_{gram_n\in S}Count(gram_n)} \tag{1}$$

其中,n 表示 n-gram 的长度,Y 表示标准摘要,$Count_{\text{match}}(gram_n)$表示生成摘要和标准摘要同时出现 n-gram 的个数,$Count(gram_n)$表示标准摘要中出现 n-gram 的个数。

$$R_{\text{lcs}}=\frac{LCS(X,Y)}{m} \tag{2}$$

$$P_{\text{lcs}}=\frac{LCS(X,Y)}{n} \tag{3}$$

$$F_{\text{lcs}}=\frac{(1+\beta^2)R_{\text{lcs}}P_{\text{lcs}}}{R_{\text{lcs}}+\beta^2P_{\text{lcs}}} \tag{4}$$

其中,X 为表示标准摘要,Y 表示生成摘要,m 表示 X 的长度,n 表示 Y 的长度,$LCS(X,Y)$表示 X 和 Y 的最长公共子序列,β 是一个超参数。

ROUGE 是一种相对的评估方法,它通过衡量生成的摘要与标准摘要的相似度来评估摘要生成模型的质量。

4.3　参数设置

实验模型的相关数据如表 3 所列。

表 3　参数设置

Table 3　Parameter settings

参数	值
Batch size	4
Hidden size	1024
Layer	12
Parameters	220×10^6
train epochs	5
optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	1×10^{-2}
drop out	0.5

4.4　对比实验

本节将介绍 3 种基准模型及本实验模型来验证所提方法的有效性。

1)LEAD-3[23]:该算法的基本思想是通过选择文本中的前 3 句话来生成摘要。该算法涵盖了摘要生成的基本要素,包括信息的紧凑性和重要性。它通过简单的文本选择方法来生成摘要,不需要花费大量的计算资源。由于该算法的简单性和高效性,它已经成为文本摘要生成的基本算法之一。在进行评测时,LEAD-3 算法的效果往往比其他算法更加稳定。

2)PageRank[24]:该算法将句子作为图节点,基于节点间的链接关系和各自的权重来评估句子的重要性。它通过每篇文档之间的相关性及与其他文档的关系来计算权重,进而得出重要性排名。PageRank 算法的结果作为选择文档片段的依据,为生成多文档摘要提供有力的支持。

3)LSTM[25]:该算法是一种基于循环神经网络的深度学习技术,在多文档摘要中被广泛应用。LSTM 算法通过对文本序列进行建模,能够捕捉文本中长期的依赖关系和短期的依赖关系,从而解决普通循环神经网络难以处理长序列问题和训练时梯度消失的问题。在多文档摘要中,LSTM 可以用于从多篇文章中提取关键信息,并生成一个简洁的摘要。通过对文本的长期和短期依赖关系的建模,有效地提取文档中的重要信息,在多文档摘要任务中取得了良好的效果。

4)T5＋预训练任务:利用多文档的优势,强化文档与文档之间的联系。在 T5-base 模型的基础上,通过采用针对多文档的预训练任务,进一步优化模型的多文档摘要生成能力。

4.5　实验结果及分析

4.5.1　基准模型对比实验结果

对于 4.4 节所提出的基准模型对比实验进行结果分析。具体评测结果如表 4 所列。

表 4　基线模型对比实验

Table 4　Comparative experiments of baseline model

Model	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	29.48	10.12	14.62
PageRank	35.75	12.35	17.25
LSTM	37.86	13.10	20.23
T5＋$\hat{t}2t$	**39.81**	**14.07**	**22.20**

由表 4 可以得出以下结论:

1)本文提出的结合预训练的方法优于 LEAD-3,PageRank,LSTM 和 T5 4 种方法,说明本文提出的方法在提升多文档摘要生成效果方面是成功的。

2)相比其他的抽取式摘要方式,使用 T5 预训练语言模型进行生成式摘要提升效果明显,表明了生成式摘要方法对提升句间连贯性和信息丰富性有效,更适用于多文档摘要。

4.5.2　预训练内容对比实验结果

对 3.3 节所提出的预训练内容进行对比实验结果分析。具体评测结果如表 5 所列。由表 5 可以得出以下结论:

1)在主要的 9 种预训练内容中,提升效果明显的有$g\hat{t}2t$,$t_1 2g$,$t_2 2g$,$t\hat{g}2g$,$\hat{t}2t$ 5 种,说明这些预训练任务可以有效提升模型对文档间关系的理解,进而使摘要更加简洁明了且内容更加完整,在多文档摘要生成任务中可以优先选择使用。

2)在自生成预训练任务即 $\hat{t}2t$,$g2t$,$\hat{t}2\hat{g}$ 3 组预训练任务中,$\hat{t}2t$ 提升效果最显著,因此在自生成预训练任务中可以优先考虑 $\hat{t}2t$。

3)在辅助预训练任务即 $t\hat{g}2g$,$g\hat{t}2t$,$\hat{t}g2t$ 3 组预训练任

务中，以 BERT-style 式遮掩的辅助任务优于全遮掩的辅助任务，其中 $g\hat{t}2t$ 任务最优，可优先选择。

4）在单文档摘要预训练任务即 $t_1 2g$，$t_2 2g$，$t_3 2g$，$t_4 2g$ 4 组预训练任务中，文本生成效果逐渐下降，这证明模型更依赖于位置靠前的文档进行生成。因此，在单文档摘要预训练任务中选择 $t_1 2g$ 和 $t_2 2g$ 更有利于多文档摘要生成任务。

5）在关键词摘要预训练任务即 $t_{1k} 2g$ 和 $t_{2k} 2g$ 中，$t_{1k} 2g$ 和 $t_{2k} 2g$ 的文本生成效果相近。这说明对于关键词生成任务，在进行具体训练时，应结合具体数据集进行分析。

表 5 预训练内容对比实验

Table 5 Comparative experiments of pre-training content

Model	ROUGE-1	ROUGE-2	ROUGE-L
T5+$\hat{t}2t$	39.81	14.07	22.20
T5+$g2t$	39.62	14.01	22.25
T5+$\hat{t}2\hat{g}$	39.58	13.83	22.17
T5+$t\hat{g}2g$	39.82	13.97	22.21
T5+$g\hat{t}2t$	**40.24**	**14.27**	**22.49**
T5+$\hat{t}g2t$	39.79	14.06	22.18
T5+$t_1 2g$	40.03	14.13	22.40
T5+$t_2 2g$	39.91	14.09	22.26
T5+$t_3 2g$	39.78	14.04	22.12
T5+$t_4 2g$	39.65	13.90	22.19
T5+$t_1 2t_2$	39.41	13.85	22.14
T5+$t_{1k} 2g$	39.39	13.84	22.18
T5+$t_{2k} 2g$	39.47	13.87	22.16

在 4 类预训练任务中，取 4 类任务的 ROUGE-L 的平均值进行分析，自生成任务、辅助生成任务、单文档生成任务和关键词生成任务的平均值分别为 22.21，22.30，22.22 和 22.17。其中，辅助生成任务的效果明显好于其余 3 类任务，自生成任务和单文档生成任务的效果相当，关键词生成任务差于其余任务。因此，若考虑结合单独预训练任务的模型训练，可以优先进行辅助生成任务，并且尝试使用自生成任务和单文档生成任务对模型进行测试优化。结合上述所有结论，当结合单独预训练进行训练时，预训练测试顺序建议采用 $\hat{g}t2t$，$t_1 2g$，$\hat{t}2t$，这样的顺序既考虑了训练类型的全面性，又考虑了具体任务的有效性，可广泛应用于各数据集进行测试。

4.5.3 预训练数量对比实验结果

这部分将对预训练数量对比实验进行介绍并对实验结果进行分析。在上述的单独预训练内容中选取 4 项预训练内容进行了数量对比实验，连续的预训练任务之间用“—”进行分隔。实验内容主要包括以下 4 种：$g2t$，$g2t-\hat{t}2t$，$g2t-\hat{t}2t-t_1 2g$，$g2t-\hat{t}2t-t_1 2g-\hat{g}t2t$。

通过改变模型进行的预训练任务的数量，对比模型最终从源文档标记序列生成目标摘要标记序列的效果。主要探究 3 个方面的问题：1）预训练数量的增加对模型效果的影响；2）预训练数量的增加对模型效果提升速度的影响；3）在兼顾计算时间和模型效果的前提下，选择合理的预训练任务数量。具体评测结果如表 6 所列。

由表 6 可以得出以下结论：

1）本文提出的增加预训练数量的方法对模型效果的再次提升有明显帮助，预训练后的模型效果明显优于基线 T5 模型。

2）在选取的 4 种预训练任务 $g2t$，$\hat{t}2t$，$t_1 2g$，$\hat{g}t2t$ 上，从 $g2t$ 逐步添加数量构成 3 种连续训练任务 $g2t-\hat{t}2t$，$g2t-\hat{t}2t-t_1 2g$，$g2t-\hat{t}2t-t_1 2g-\hat{g}t2t$，模型效果逐步提升，直到超过所有单独预训练任务。

3）对于连续预训练任务，执行任务后的模型效果都优于构成连续预训练任务的任意一种单独预训练任务的模型效果。由此可见，预训练任务数量的提升对多文档摘要生成任务有明显帮助。

表 6 预训练数量对比实验

Table 6 Comparative experiments of pre-training quantities

Model	ROUGE-1	ROUGE-2	ROUGE-L
T5	39.58	13.81	22.14
T5+$g2t$	39.62	14.01	22.25
T5+$\hat{t}2t$	39.81	14.07	22.20
T5+$t_1 2g$	40.03	14.13	22.40
T5+$\hat{g}t2t$	40.24	14.27	22.49
T5+$g2t-\hat{t}2t$	40.12	14.10	22.33
T5+$g2t-\hat{t}2t-t_1 2g$	40.14	14.20	22.41
T5+$g2t-\hat{t}2t-t_1 2g-\hat{g}t2t$	**40.40**	**14.51**	**22.53**

4.5.4 预训练顺序对比实验结果

本节将介绍预训练顺序对比实验并进行实验结果分析。我们在预训练数量对比实验的基础上，探寻预训练顺序对模型效果的影响。

1）在预训练任务数量为 2 个时，进行 $t_1 2g-\hat{g}t2t$ 和 $\hat{g}t2t-t_1 2g$ 2 项预训练任务。在预训练任务数量为提升至 3 个时，考虑全部 6 种预训练任务拼接情况，进行 $t_1 2g-\hat{g}t2t-\hat{t}2t$，$t_1 2g-\hat{t}2t-\hat{g}t2t$，$g\hat{t}2t-t_1 2g-\hat{t}2t$，$g\hat{t}2t-\hat{t}2t-t_1 2g$，$\hat{t}2t-t_1 2g-\hat{g}t2t$，$\hat{t}2t-\hat{g}t2t-t_1 2g$ 6 项预训练任务。通过改变模型进行的预训练任务的顺序，对比模型的最终效果。主要探究 3 个方面的问题：(1)预训练顺序的改变对模型效果的影响；(2)探究可以显著提升预训练任务的适当组合；(3)在了解各项单独预训练任务效果的前提下，选择合适的排序方式使模型效果提升最大化。具体评测结果如表 7 所列。

表 7 预训练顺序对比实验

Table 7 Comparative experiments of pre-training sequences

Model	ROUGE-1	ROUGE-2	ROUGE-L
T5	39.58	13.81	22.14
T5+$\hat{t}2t$	39.81	14.07	22.20
T5+$t_1 2g$	40.03	14.13	22.40
T5+$\hat{g}t2t$	40.24	14.27	22.49
T5+$t_1 2g-\hat{g}t2t$	40.41	14.42	22.51
T5+$\hat{g}t2t-t_1 2g$	40.36	14.38	22.44
T5+$t_1 2g-\hat{g}t2t-\hat{t}2t$	40.44	14.42	22.54
T5+$t_1 2g-\hat{t}2t-\hat{g}t2t$	40.30	14.29	22.43
T5+$\hat{g}t2t-t_1 2g-\hat{t}2t$	40.29	14.39	22.41
T5+$\hat{g}t2t-\hat{t}2t-t_1 2g$	40.46	14.40	22.50
T5+$\hat{t}2t-t_1 2g-\hat{g}t2t$	40.06	14.21	22.28
T5+$\hat{t}2t-\hat{g}t2t-t_1 2g$	40.20	14.37	22.44

由表 7 可以得出以下结论：

1）本文提出的调换预训练顺序的方法在 3 项 ROUGE 值上都有进一步的提升，对模型的提升效果显著。

2）在预训练数量为 2 个时，$t_1 2g-\hat{g}t2t$ 和 $\hat{g}t2t-t_1 2g$ 的

实验结果相近,因此在这种情况下进行预训练任务顺序的调换不会显著影响模型的效果。

3)在预训练数量为 3 个时,两项任务组合 $t_1 2g-\hat{g}t2t-\hat{t}2t$ 和 $\hat{g}t2t-\hat{t}2t-t_1 2g$ 的结果相近且优于其余 4 种预训练任务,由此可见进行预训练任务组合的拼接是有效的,且它们的公共预训练任务组合 $\hat{g}t2t-\hat{t}2t$ 非常有效。

4)在预训练数量为 3 个时,按照单独预训练任务效果升序排序进行拼接组合而成的任务 $\hat{t}2t-t_1 2g-\hat{g}t2t$ 对模型的影响弱于其余模型。结合 3)的结论可得,在预训练任务顺序上,预训练任务组合的选择对模型效果是至关重要的,选择合适的预训练任务组合比选择某种特定的排列顺序更重要。

5)对比预训练数量为 2 个和 3 个时的实验结果,预训练顺序的调整对模型提升更加显著,因此调整预训练顺序优先于不断对预训练数量进行提高。

6)在进行预训练顺序选择时,建议优先在预训练数量为 2 个时,测试并选择出合适的预训练任务组合,而后在预训练数量为 3 个时,在固定的预训练任务组合上进行进一步的预训练,在预训练任务数量不断上升时以此类推。

4.6　预训练流程

本节将根据实验结果和结论提出系统的预训练流程。

1)首先,控制预训练任务数量为一个,对 4 类预训练任务进行测试,优先推荐 $\hat{g}t2t$,$t_1 2g$,$\hat{t}2t$ 3 项预训练任务。

2)其次,控制预训练任务数量为两个,选择 1)中对模型提升效果明显的预训练任务,进行预训练任务组合测试,优先推荐 $\hat{g}t2t-\hat{t}2t$ 这项预训练任务组合。

3)最后,控制预训练任务数量为 3 个,选择 2)中最有效的预训练任务组合,与 1)中的预训练任务进行组合,进一步提升预训练效果。

结束语　本文提出了一个结合预训练的多文档摘要方法,并对预训练任务的内容、数量和顺序进行探讨,进一步提出了系统的预训练流程。该方法的特色在于模型通过结合预训练,缓解了多文档间的矛盾关系,强化了多文档间的文本交流。在公开的新闻数据集上进行实验,结果表明本文所提出的方法在模型效果和性能上均有一定的提高。

在未来的工作中,我们会选取更多的生成模型,不断验证预训练任务的有效性;还将对更多的数据集使用本文的方法,当前的实验只在一个数据集上进行了测试,这可能会导致过拟合或者不足以全面反映模型的性能;希望在未来研究效果更好、泛用性更强的预训练框架,从而加快模型预训练速度,并生成更加准确、连贯的摘要。

参 考 文 献

[1] RUSH A M,CHOPRA S,WESTON J. A neural attention model for abstractive sentence summarization[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. 2015:379-389.

[2] NALLAPATI R,ZHOU B W,SANTOS CN D,et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]// Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016:280-290.

[3] SEE A,LIU P J,MANNINGC D. Get to the point:Summarization with pointer-generator networks[C]// Association for Computational Linguistics. 2017:1073-1083.

[4] CELIKYILMAZ A,WANG X,HUANG Q Y. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation[C]// Conference on Computer Vision and Pattern Recognition. 2019:6669-6638.

[5] PAULUS R,XIONG C M,SOCHERR. A deep reinforced model for abstractive summarization[J]. arXiv preprint,arXiv:1705.04304,2017.

[6] ZHANG H,XU J,WANG J. Pretraining-based natural language generation for text summarization[J]. arXiv:1902.09243,2019.

[7] DONG L,YANG N,WANG W,et al. Unified language model pre-training for natural language understanding and generation[J]. Advances in Neural Information Processing Systems,2019,32.

[8] MICHELI V,FLEURET F. Language models are few-shot butlers[J]. arXiv:2104.07972,2021.

[9] ZHANG Y,SUN S,GALLEY M,et al. Dialogpt:Large-scale generative pre-training for conversational response generation[J]. arXiv:1911.00536,2019.

[10] RADEV D R,JING H,STYŚ M,TAM D. Centroid-based Summarization of Multiple Documents.[C]// Information Processing and Management. 2004:919-938.

[11] WAN X J,YANG J W. Multi-document Summarization Using Cluster-based Link Analysis.[C]// Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008:299-306.

[12] MANI N,BLOEDORN E. Multi-Document Summarization by Graph Search and Matching.[C]// Proceedings of the 14th National Conference on Artiicial Intelligence and 9th Innovative Applications of Artiicial Intelligence Conference. 1997:622-628.

[13] HAGHIGHI R,VANDERWENDE L. Exploring Content Models for Multi-document Summarization[C]// Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009:362-370.

[14] DEVLIN A,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2019:4171-4186.

[15] LIU Y H,OTT M,GOYAL N,et al. RoBERTa:A Robustly Optimized BERT Pretraining Approach[J]. arXiv:1907.11692,2019.

[16] MAO Y,QU Y,XIE Y,et al. Multi-document summarization with maximal marginal relevance-guided reinforcement learning[J]. arXiv:2010.00117,2020.

[17] ADFORD L,WU J,CHI L D,et al. Language Models are Unsupervised Multitask Learners[C]// OpenAI Blog. 2020:1,8,9.

[18] FABBRI A R,LI I,SHE T,et al. Multi-news:A large-scale multi-document summarization dataset and abstractive hierarchical model[J]. arXiv:1906.01749,2019.

[19] ARUMAE K,LIU F. Guiding extractive summarization with question-answering rewards[J]. arXiv:1904.02321,2019.

[20] ARUMAE K, LIU F. Guiding extractive summarization with question-answering rewards[J]. arXiv: 1904. 02321, 2019.

[21] ZHANG X, WEI F, ZHOU M. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization[J]. arXiv: 1905. 06566, 2019.

[22] ALEXANDER R F, LI I, SHE T W, et al. Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model[C]// Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019: 1074-1084.

[23] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]// Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL. 2004: 74-81.

[24] XU J C, GAN Z, CHENG Y, et al.. Discourse-Aware Neural Extractive Text Summarization[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5021-5031.

[25] DING Y, YAN E, FRAZHO A. PageRank for ranking authors in co-citation networks[J]. Journal of the American Society for Information & Technology, 2014, 60(11): 2229-2243.

[26] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM Network: A Machine Learining Approach for Precipitation Nowcasting[C]// Advances in Neural Information Processing Systems 28: Annual Conference. 2015: 802-810.

DING Yi, born in 2002, undergraduate. His main research interests include natural language processing and multi-document summarization.

WANG Zhongqing, born in 1987, Ph.D, associate professor. His main research interests include natural language processing, information extraction and emotional analysis.