融合引导注意力的中文长文本摘要生成

郭 哲[1,2],张智博[1],周炜杰[1],樊养余[1],张艳宁[2]
(1. 西北工业大学电子信息学院,陕西西安 710129;
2. 西北工业大学空天地海一体化大数据应用技术国家工程实验室,陕西西安 710129)

摘 要: 当前基于深度学习的中文长文本摘要生成的研究存在以下问题:(1)生成模型缺少信息引导,缺乏对关键词汇和语句的关注,存在长文本跨度下关键信息丢失的问题;(2)现有中文长文本摘要模型的词表常以字为基础,并不包含中文常用词语与标点,不利于提取多粒度的语义信息. 针对上述问题,本文提出了融合引导注意力的中文长文本摘要生成(Chinese Long text Summarization with Guided Attention,CLSGA)方法. 首先,针对中文长文本摘要生成任务,利用抽取模型灵活抽取长文本中的核心词汇和语句,构建引导文本,用以指导生成模型在编码过程中将注意力集中于更重要的信息. 其次,设计中文长文本词表,将文本结构长度由字统计改变至词组统计,有利于提取更加丰富的多粒度特征,进一步引入层次位置分解编码,高效扩展长文本的位置编码,加速网络收敛. 最后,以局部注意力机制为骨干,同时结合引导注意力机制,以此有效捕捉长文本跨度下的重要信息,提高摘要生成的精度. 在四个不同长度的公共中文摘要数据集LCSTS(大规模中文短文本摘要数据集)、CNewSum(大规模中国新闻摘要数据集)、NLPCC2017和SFZY2020上的实验结果表明:本文方法对于长文本摘要生成具有显著优势,能够有效提高ROUGE-1、ROUGE-2、ROUGE-L值.

关键词: 自然语言处理;中文长文本摘要生成;引导注意力;层次位置分解编码;局部注意力
基金项目: 国家自然科学基金(No.62071384);陕西省重点研发计划项目(No.2023-YBGY-239)
中图分类号: TP391.11 **文献标识码:** A **文章编号:** 0372-2112(2024)12-3914-17
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230429

Chinese Long Text Summarization with Guided Attention

GUO Zhe[1,2], ZHANG Zhi-bo[1], ZHOU Wei-jie[1], FAN Yang-yu[1], ZHANG Yan-ning[2]
(1. *School of Electronics and Information, Northwestern Polytechnical University, Xi'an, Shaanxi* 710129, *China*;
2. *National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Northwestern Polytechnical University, Xi'an, Shaanxi* 710129, *China*)

Abstract: Current research on Chinese long text summarization based on deep learning has the following problems: (1) summarization models lack information guidance, fail to focus on keywords and sentences, leading to the problem of losing critical information under long-distance span; (2) the word lists of existing Chinese long text summarization models are often word-based and do not contain common Chinese words and punctuation, which is not conducive to extracting multi-grained semantic information. To solve the above problems, a Chinese long text summarization method with guided attention (CLSGA) is proposed in this paper. Firstly, for the long text summarization task, an extraction model is presented to extract the core words and sentences in the long text to construct the guided text, which can guide the generation model to focus on more important information in the encoding process. Secondly, the Chinese long text vocabulary is designed to changing the text structure from words statistics to phrases statistics, which is conducive to extracting richer multi-granularity features. Hierarchical location decomposition encoding is then introduced to efficiently extend location encoding of long text and accelerate network convergence. Finally, the local attention mechanism is combined with the guided attention mechanism to effectively capture the important information under the long text span and improve the accuracy of summarization. Experimental results on four public Chinese abstract datasets with different lengths, LCSTS, CNewSum, NLPCC2017 and SFZY2020, show that our proposed method has significant advantages over long text summarization and can effectively improve the value of ROUGE-1, ROUGE-2 and ROUGE-L.

Key words: natural language processing; Chinese long text summarization; guided attention; hierarchical location

收稿日期:2023-05-12;修回日期:2024-11-12;责任编辑:王天慧

decomposition coding; local attention

Foundation Item(s): National Natural Science Foundation of China (No.62071384); Key Research and Development Project of Shaanxi Province (No.2023-YBGY-239)

1 引言

随着互联网的发展,中文长文本摘要生成任务在数字图书馆、医患对话、新闻推荐等领域的需求越加广泛[1]. 通过提取文本的核心内容来生成摘要,可以有效帮助人们理解大量数据中所蕴含的关键信息. 以Transformer[2]为基础的Seq2Seq(Sequence-to-Sequence)体系结构[3]已经在包括中文文本摘要生成在内的自然语言处理任务中取得了令人瞩目的结果,但Transformer中自注意操作与序列长度的平方成正比,这就导致大多数模型限制输入文本的长度,因而无法处理长文本的摘要生成. 同时,现有的中文文本摘要模型的词表常以字为基础,并不包含中文常用词语与标点[1,4],在长文本跨度下很难有效联系上下文,从而严重限制了中文长文本自动摘要模型性能的发挥.

虽然在英文长文本摘要生成领域已经出现了一些成熟的方法,如Longformer[5]、BigBird[6]等,通过结合局部注意力、全局注意力、随机注意力等多种注意力机制,突破对输入文本长度的限制,有效提高了长文本摘要生成的准确性,但上述方法由于缺少信息引导,缺乏对关键词汇和语句的关注,存在输入文本中关键信息丢失的问题. 同时,由于中英文语种和语法不同、中英文词表亦不同,英文长文本摘要生成方法应用在中文长文本摘要生成时需要额外的资源准备[4];而从头开始预训练一个中文长文本模型所耗费的资源又是十分巨大的,并且模型训练往往需要在大规模GPU上进行[7]. 虽然以ChatGPT为代表的大模型通用性很强,对多种自然语言处理任务都有处理能力,但在本文所关注的中文长文本摘要生成任务上,受限于中文语料数据的采集以及大模型训练和部署的困难程度,相比于目前的微调小模型范式性价比较低.

针对上述挑战,本文提出了一个融合引导注意力的中文长文本摘要生成方法,称为CLSGA(Chinese Long text Summarization with Guided Attention). CLSGA从摘要撰写的语义合理性机理出发,构建了包含抽取模型和生成模型的引导生成架构,有效缓解了长文本跨度下词汇的远距离依赖问题. 其中抽取模型通过灵活抽取长文本中的核心词汇和语句来构建引导文本,用以指导生成模型在编码过程中将注意力集中于更重要的信息,并依据该引导信息约束生成模型,实现长文本摘要的可控解码.

本文所提CLSGA的创新性具体表现在:

(1)构建了CLSGA网络,在生成模型中引入局部注意力机制,同时训练抽取模型构建引导信息,生成模型融合引导信息进行训练,能有效捕捉长文本跨度下的重要信息,约束生成摘要与源文本的偏差,提高了摘要生成的能力.

(2)构建了适应于中文长文本摘要模型的词表,基于该词表进行分词,可将文本结构由字改变为词组,有利于提取更加丰富的多粒度特征;利用层次位置分解编码技术对长文本位置进行有效编码,加速网络收敛.

(3)在4个不同长度的中文摘要数据集LCSTS(大规模中文短文本摘要数据集)[8]、CNewSum(大规模中国新闻摘要数据集)[9]、NLPCC2017[10]和SFZY2020①上的实验结果表明,与现有流行方法相比,本文方法对于长文本摘要生成具有显著优势.

2 相关工作

文本摘要生成方法按照处理文本的长度可以分为短文本摘要生成和长文本摘要生成.

2.1 短文本摘要生成

在短文本摘要生成方面,文献[11]通过对文本的每一句话与参考摘要的匹配程度进行排序,根据排序结果训练摘要抽取模型. 文献[12]通过在BERT[13](Bidirectional Encoder Representation from Transformers)[13]中训练句向量,根据句向量的分类结果用于短文本的摘要抽取. 文献[14]提出了BART(Bidirectional and Auto Regressive Transformers)算法,在Transformer结构上利用多种形式的噪声训练提高了模型的生成能力. 文献[15]提出了用于中文自然语言生成的非对称式编解码网络CPT(Chinese Pretrained unbalanced Transformer),通过更深的编码器加强模型的理解,使用更浅的解码器加快模型的生成速度;他们也公开发布了相应的中文BART预训练模型. 文献[16]通过GSUM(Guided neural abstractive SUMmarization)方法抽取文本关键词和关键句,对文本内容进行补充编码,以提高生成摘要的流畅度与忠诚度. 文献[17]提出了融合上下文信息和关键信息的中文文本摘要生成方法,有效提高了摘要的总结性和连贯性. 但是,上述方法都存在输入文本长度受限的问题,随着输入长度的增长,自注意力的二次计算复杂性成为瓶颈.

2.2 长文本摘要生成

在长文本摘要生成方面,针对Transformer自注意机制的二次计算问题,近期的工作大多着手于修改自

① http://cail.cipsc.org.cn/

注意力机制[18],提出各种类型的注意力变体,以此降低二次平方的计算复杂度. 文献[5]提出了Longformer方法,该方法提出的局部注意机制与全局注意力机制将二次依赖复杂度降低为线性,有效提高了输入文本的长度. 但是,Longformer中的局部注意力机制由于缺少信息引导,在过长的文本跨度下会丢失一定的语义信息,并且调整注意力窗口的超参大小也是一项复杂的工作. 文献[6]提出了BigBird方法,通过结合局部注意力、随机注意力、全局注意力等多种注意力机制,突破对输入文本长度的限制. 但该方法面临着与Longformer同样的问题,同时上述改进模型仅针对英文长文本数据集,其预训练权重在中文长文本处理领域难以获取.

此外,也有学者通过知识蒸馏或分层架构的方式将大型模型压缩为小型模型,以减少内存占用和计算量,用于长文本的摘要生成. 文献[19]使用分层循环神经网络(Recurrent Neural Network,RNN)结构对长文本进行摘要生成. 文献[20]使用分层Transformer结构处理多文档的长文本摘要生成. 文献[21]使用局部注意力与内容选择分层结构来生成摘要. 文献[22]将提取的文本片段作为潜在变量,在解码过程中赋予动态注意力权重,用以应对长文本摘要生成任务. 文献[23]将文档结构纳入注意力分数的计算,并由此引入文档的层次化结构,以缓解模型的计算压力. 文献[24]首先生成涵盖输入文本突出信息的摘要视图,在预训练模型指导下构建能量函数,将这些视图组合成最终的摘要. 文献[25]提出了基于强化学习的抽取式长文本摘要生成模型MemSum,在任何给定的时间步长内都包含有关当前提取历史的信息,以构建多步情景马尔可夫决策过程的摘要提取框架. 文献[26]基于多目标优化差分演化以及加权和方法,提出了基于分解的多目标差分进化的文本摘要生成结构,有效提高了运行效率. 然而,目前支持大跨度文本输入的模型基本都存在长文本跨度下前后语义信息丢失的问题,无法达到Transformer自注意力的效果;而通过抽取长文本内容再送入生成模型优化生成的方法,则极容易存在因遗漏关键信息而造成的文本生成错误.

针对Seq2Seq体系结构不擅长分析文本中长距离关系的问题,有学者基于图神经网络(Graph Neural Networks,GNN)模型以捕捉跨句子的依赖关系. 文献[27]提出了一种基于统一语义图的长文本摘要生成框架,利用图结构改进文档表示和摘要生成过程. 文献[28]使用多元图来考虑不同的句子关系. 文献[29]提出了层次感知CNN,通过潜在结构树来学习层次化的文档结构. 文献[30]提出了一种通过捕获高阶交叉句关系来进行长文本摘要生成的超GNN. 基于GNN的文本摘要生成方法大都通过建模跨句关系获取语句的长距离依赖,但却较难从不同角度对句子的相互作用进行融合;另外,该类方法属于抽取式摘要技术,而本文主要关注更符合人类对文本压缩理解的基于Transformer结构的生成式摘要技术.

综上,现有的文本摘要生成方法存在输入文本长度受限的问题,或在长文本跨度下由于缺少信息引导而丢失关键语义信息. 因此,如何搭建高效清晰的网络结构,结合生成式与抽取式的优点,突破现有网络对输入文本的长度限制,对中文长文本生成合理有效的摘要,是本文所要解决的主要问题.

3 CLSGA网络架构

为解决中文长文本的摘要生成问题,本文提出了CLSGA,其网络结构如图1所示. CLSGA包含抽取模型与生成模型两部分,抽取模型负责将源文本划分为若干子句,再将子句集合送入句向量编码网络生成句向量,最后对句向量进行特征变换抽取引导信息. 生成模型则分别对源文本及引导信息进行分词、字词嵌入和位置编码嵌入,而后分别送入编码模块进行编码. 解码时根据引导信息的约束并基于局部注意力机制生成摘要,使输出内容与源文本具有更小的偏差度,最终实现CLSGA.

3.1 中文词表设计

合理的中文词表是中文长文本摘要生成的一个重要前提. 谷歌发布的中文BERT词表[13],共包含21 128个常用字符,全部以字为单位,现有的中文短文本摘要模型如BART、CPT等均采用该词表. 但对于中文长文本来说,以字为单位的词表存在一个核心问题,即分词结果很长,从而影响模型的执行效率. 此外,在长文本摘要生成过程中,还存在一定的归纳偏置问题[31],其错误累计发生概率随着生成摘要的增长而增大. 针对上述问题,本文重新设计了适用于中文长文本的词表,引入中文常用词语,不仅可以缩短长文本序列的长度,增强词义的确定性,加快模型的处理速度,同时还能够在一定程度上缓解归纳偏置现象.

在英文摘要领域通常采用sub-word units[32]作为词表的基本单位,词表采用"\#\#+字符"的形式表达,这是因为英文字母的表达多种多样,通过"\#\#"将单词分开,可以区分不同时态的同义词,有效降低词表的大小. 然而对于中文文本来说,该词表结构并非是必要的. 针对上述问题,本文构建了适用于中文长文本的词表,其设计示意图如图2所示. 具体设计流程为:首先保留中文短文本摘要模型BART的词表前13 317字符,将13 317之后的"\#\#+字符"替换为jieba①常用词,共40 000个,最后加入中文常用标点符号,最终得到的词表大小为53 321.

① https://pypi.python.org/pypi/jieba/

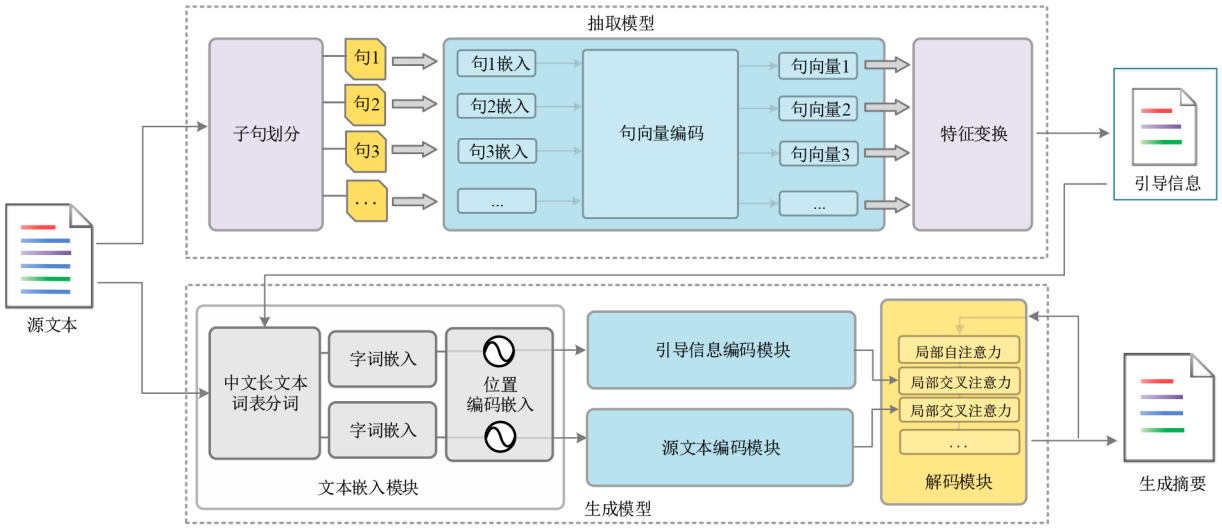


图 1　CLSGA 网络结构图

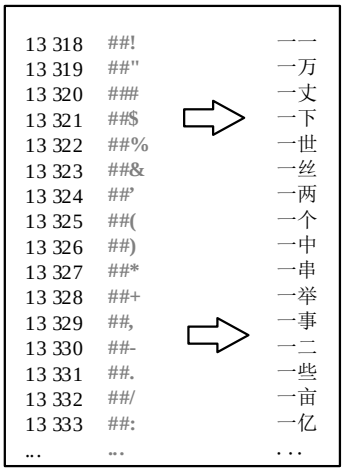


图 2　中文长文本词表设计示意图

CLSGA 网络模型基于所构建的适用于中文长文本的词表,以字词为基本单位对中文长文本进行分词操作. 具体操作过程分为两步,首先进行预分词,然后再进行原字符的分词操作. 尽管新添加的词汇并未进行预训练,但这些词汇都是由经过预训练的字加权来初始化的,并且在实际分词过程中仍有部分文字是按照字来进行切分,因此,该部分字依然能够利用预训练过程中所学到的知识,在模型训练过程中实现快速拟合,后续实验也证明了该策略能够有效提升模型的性能.

3.2　抽取模型

对长文本先进行信息抽取再凝聚生成,符合人类对长输入文本进行摘要生成的基本思路[22]. 通过识别并抽取文本中的重要信息,有助于将源文本输入减少至期望的预设范围内,从而有效克服模型无法处理超过限定范围长文本输入的问题. 此外,利用抽取重要信息引导生成,能缓解先抽取并仅利用抽取信息再生成带来的信息丢失问题. 基于上述分析,本文综合抽取式与生成式的优缺点,设计了抽取并引导生成的长文本摘要生成架构.

3.2.1　模型结构

本文所构建的抽取模型的结构如图 3 所示,首先将长文本划分为若干子句,为了实现更小的模型权重和更快的句子抽取速度,采用将上述子句送入冻结权重的 BERT 的方式,生成文本编码向量;然后进行平均池化,生成句向量;接着通过一个全连接网络对句向量进行降维操作,最后送入门控线性膨胀残差网络[33],实现特征变换. 门控线性膨胀残差网络共有八层,前五层为膨胀率递增的膨胀门卷积结构,膨胀率分别为 1、2、4、8、16,逐级学习细粒度到粗粒度的特征. 后三层为细粒度的精调,膨胀率均为 1. 最后的分类通过 sigmoid 函数将向量映射至 0~1,通过设定阈值的方式,将子句进行 0~1 分类. 为了更加详细地描述抽取模型的结构和运算过程,给出了伪代码算法描述,如算法 1 所示,描述了将长文本划分为子句,使用冻结权重的 BERT 生成文本编码向量,经过平均池化和全连接网络降维,接着通过门控线性膨胀残差网络进行特征变换,最终通过 sigmoid 函数进行分类的算法流程.

3.2.2　训练过程

具体训练时,引导信息通过贪婪搜索算法[34],抽取源本文中与摘要 ROUGE[35]分数最佳匹配的关键句子集,再通过图排序算法 TextRank[36]在关键句中抽取关键词. 关键句用$\{S_1,S_2,\cdots,S_u\}$表示,u为关键句数目;关键词子集用$\{\omega_1,\omega_2,\cdots,\omega_v\}$表示,$v$表示关键词数量. 为了最大限度降低训练与预测之间不一致,本文基于训练完成的抽取模型和抽取算法相结合,来生成模型引导信息所需的训练数据. 通过设置合理的阈值,保留原始抽取结果,同时将训练数据中大于阈值的句子也作

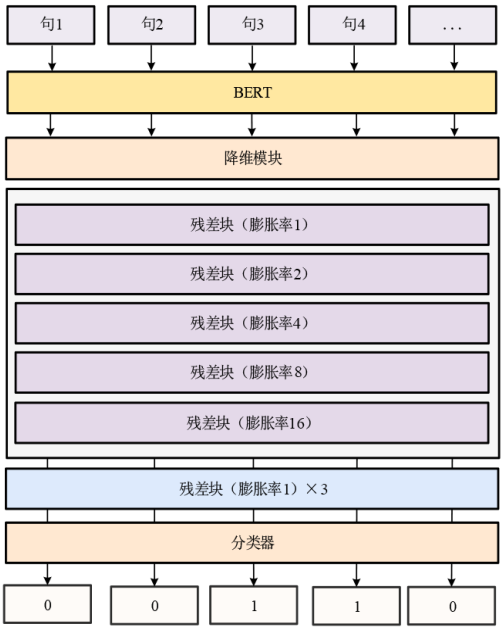


图3 抽取模型结构示意图

算法1 抽取模型

输入:中文源文本S;

输出:分类结果及引导信息;

Step1:划分源文本为子句:子句列表$\{\text{Sen}_1,\text{Sen}_2,\cdots,\text{Sen}_i\}(i=1,2,\cdots,N)$=划分文本为子句($S$);

Step2:使用冻结权重的BERT生成文本编码向量:

For $i=1$ to N

对子句列表中的每个子句Sen_i,文本编码向量Vect_i=BERT_frozen(Sen_i);

End

Step3:平均池化生成句向量:句向量SenVect=平均池化(文本编码向量列表Vect);

Step4:全连接网络降维:降维向量DecVect=全连接网络(句向量SenVect);

Step5:门控线性膨胀残差网络:

For $j=0$ to 4

膨胀卷积层=膨胀卷积层(降维向量DecVect,膨胀率=2^j);

End

For $j=5$ to 7

精调层=精调层(膨胀卷积层);

End

Step6:分类:输出向量=sigmoid (精调层);

分类结果=分类(输出向量,阈值);

引导信息=输出向量.

为关键句,由此构建生成模型的训练数据集. 抽取模型的训练过程采用能够衡量目标和输出之间的二值交叉熵,具有0~1分类能力的BCE Loss(Binary Cross Entropy Loss)[37]作为损失函数,描述为

$$L_{\text{bce}}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\lg\left(p_i\right)+\left(1-y_i\right)\lg\left(1-p_i\right)\right] \tag{1}$$

其中,N为一个批次内的样本总数;y_i为第i个样本的类别;p_i为第i个样本的概率值.

3.3 生成模型

生成模型包含3个部分,分别为文本嵌入模块、编码模块和解码模块. 输入源文本数据,经过文本嵌入模块转换为模型能够识别的矩阵向量$\boldsymbol{X}_{\text{src}}=\left\{\boldsymbol{x}_{\text{src}}^1,\boldsymbol{x}_{\text{src}}^2,\cdots,\boldsymbol{x}_{\text{src}}^n\right\}$. 同时对长文本抽取引导信息并进行文本嵌入,从而得到引导矩阵向量$\boldsymbol{X}_g=\left\{\boldsymbol{x}_g^1,\boldsymbol{x}_g^2,\cdots,\boldsymbol{x}_g^m\right\}$,其中$\boldsymbol{x}_{\text{src}}^n,\boldsymbol{x}_g^m\in\mathbb{R}^h$,$h$是隐藏层大小. 编码模块同时对源文本与引导文本进行语义建模,得到各自的编码隐向量. 解码模块则分别对引导编码$\boldsymbol{X}_g$与源文本编码$\boldsymbol{X}_{\text{src}}$解码,得到解码信息,经过全连接分类后给出摘要信息.

3.3.1 文本嵌入模块

文本嵌入包含位置编码嵌入与字词嵌入. 为了解决长文本位置的编码问题,在有限资源的情况下,本文通过引入层次位置,分解编码技术①,将BART的位置编码嵌入长度从512扩展为8 192. 即已有训练好的中文BART位置编码向量为$\boldsymbol{p}_s^1,\boldsymbol{p}_s^2,\cdots,\boldsymbol{p}_s^k$,在此基础上构造长文本模型所需要的编码向量,记为$\boldsymbol{p}_l^1,\boldsymbol{p}_l^2,\cdots,\boldsymbol{p}_l^n$,其中$n>k$,本文中$k$=512,$n$最大为8 192. 则长文本模型的编码向量式为

$$\boldsymbol{p}_l^n=\lambda\frac{\boldsymbol{p}_s^i-\lambda\boldsymbol{p}_s^1}{1-\lambda}+(1-\lambda)\frac{\boldsymbol{p}_s^j-\lambda\boldsymbol{p}_s^1}{1-\lambda} \tag{2}$$

其中,i为BART位置编号,j为由i层次分解得到的位置编号,$i,j<k$;n为扩展后的位置编号,并有$n=(i-1)\times k+j$,理论上$n\leqslant k^2$;λ为超参数,且有$\lambda\in(0,1)$,$\lambda\neq0.5$,文中实验设置0.4. 超参数$\lambda\neq0.5$是为了区分(i,j)和(k,i),当$n\leqslant k\leqslant512$时,$\boldsymbol{p}_l^n=\boldsymbol{p}_s^k$,即位置向量保持不变,因此,编码嵌入能使新的编码向量与已经训练好的中文短文本模型BART兼容. 对于字词嵌入操作,将新添词语的字词嵌入矩阵,并通过词语中每个字的嵌入矩阵的加权平均进行初始化,接着在微调训练过程中训练至最佳,并采用权值共享策略,尽可能减小网络框架结构的复杂度.

3.3.2 编码模块

为了应对中文长文本任务的挑战,本文构建了多层堆叠连接自上而下的局部注意力结构,用于在全部长文本尺度上构建高层次语义特征映射,并使用BART[15]预训练模型作为局部注意力结构的初始化. Longformer[5]方法为了提高输入文本的长度,也使用了局部注意机制. Longformer方法在计算自注意力时只关

① https://kexue.fm/archives/7947

注当前字/词固定窗口大小附近的内容,虽然能有效降低计算复杂度,但是由于其选择的特定字/词有限且缺少信息引导,在过长的文本跨度下依然会丢失一定的语义信息. 为解决局部注意力结构因上下文距离过长而产生的语义丢失问题,本文引入了引导注意力机制,通过提取长文本中的引导信息来进行二次编码,从而有效补充可能丢失的重要信息,同时加强对长文本上下文的关注程度,约束摘要的生成.

编码模块包含两个略微不同的子编码模块,如图4所示,分别对源文本和引导文本进行编码. 每个子编码模块包含多层重复堆叠且共享参数的编码器,每个编码器由局部自注意力模块(Local-Self-Attention)和前馈神经网络(Feed-Forward)构成. 两个编码器分别对源文本与引导文本单独编码,源文本与引导文本经过局部自注意力与前馈神经网络迭代计算后得到编码信息. 每一层编码器的具体计算过程为

$$\begin{cases} \boldsymbol{X}_{\text{src}} = \text{LN}\left(\boldsymbol{X}_{\text{src}} + \text{LocalSelfAttention}\left(\boldsymbol{X}_{\text{src}}\right)\right) \\ \boldsymbol{X}_{\text{src}} = \text{LN}\left(\boldsymbol{X}_{\text{src}} + \text{FeedForward}\left(\boldsymbol{X}_{\text{src}}\right)\right) \end{cases} \tag{3}$$

$$\begin{cases} \boldsymbol{X}_{g} = \text{LN}\left(\boldsymbol{X}_{g} + \text{LocalSelfAttention}\left(\boldsymbol{X}_{g}\right)\right) \\ \boldsymbol{X}_{g} = \text{LN}\left(\boldsymbol{X}_{g} + \text{FeedForward}\left(\boldsymbol{X}_{g}\right)\right) \end{cases} \tag{4}$$

其中,$\boldsymbol{X}_{\text{src}}$和$\boldsymbol{X}_{g}$分别为源文本编码与引导文本编码;LN代表层归一化.

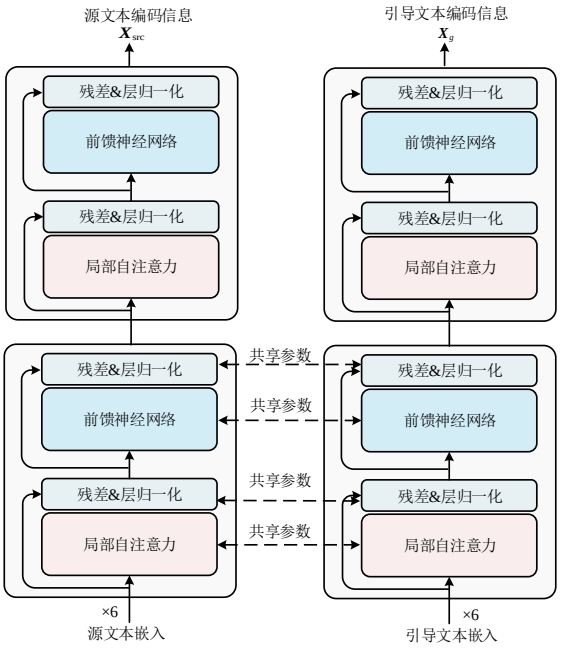


图4　编码模块结构

3.3.3　解码模块

解码模块同样采用由多层解码器重复堆叠的解码结构,每个解码器包含两个局部交叉注意力模块,通过两次解码引导文本与源文本的编码,加强对长文本上下文的解释,具体结构如图5所示. 在第一次解码过程中,先对引导信息进行解码,首先与来自编码模块的引导矢量序列$\boldsymbol{X}_{g}$进行交互注意力,让模型将焦点注意到关键词或关键句,从而提取长文本上下文中的重要信息. 在第二次解码过程中,与来自编码模块的源文本矢量序列$\boldsymbol{X}_{\text{src}}$进行交互注意力,解释源文本信息中的高层次语义信息,捕捉长文本跨度下的焦点信息. 解码模块通过引导注意力,明确指引网络模型应该提取哪些原文本中的强特征句子集,并获取最终的解码输出. 在该架构下,模型能够根据引导信息对源文本的不同序列做不同的局部自注意力.

每一层解码器可以描述为

$$\begin{cases} \boldsymbol{Y} = \text{LN}\left(\boldsymbol{Y} + \text{LocalSelfAttention}\left(\boldsymbol{Y}\right)\right) \\ \boldsymbol{Y} = \text{LN}\left(\boldsymbol{Y} + \text{LocalCrossAttention}\left(\boldsymbol{Y}, \boldsymbol{X}_{g}\right)\right) \\ \boldsymbol{Y} = \text{LN}\left(\boldsymbol{Y} + \text{LocalCrossAttention}\left(\boldsymbol{Y}, \boldsymbol{X}_{\text{src}}\right)\right) \\ \boldsymbol{Y} = \text{LN}\left(\boldsymbol{Y} + \text{FeedForward}\left(\boldsymbol{Y}\right)\right) \end{cases} \tag{5}$$

其中,$\boldsymbol{Y}$代表解码信息;LocalCrossAttention(·)表示局部交叉注意力模块.

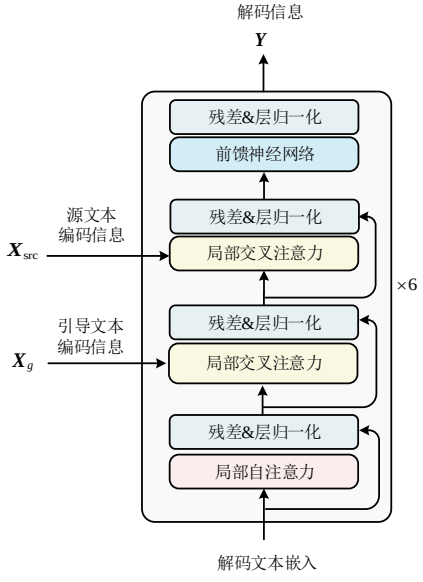


图5　解码模块结构

3.3.4　训练过程

生成模型在训练时,其计算摘要的过程可以描述为

$$p\left(\boldsymbol{Y}_{t} \middle| \boldsymbol{Y}_{1:t-1}, \boldsymbol{X};\ \theta\right) = \text{softmax}(\boldsymbol{E}) \tag{6}$$

其中,t为时间步;θ为训练参数. 模型在解码时间步骤t时将最后一层的解码器输出$\boldsymbol{Y}_{t}$,并做线性变换至词汇表大小V的logits向量$\boldsymbol{E}$,其中$\boldsymbol{E} \in \mathbb{R}^{|V|}$,然后经过softmax

变换输出概率值,进而根据概率值选择下一个词语并最终输出摘要. 参数θ则通过每个解码步骤t的最小化交叉熵来训练. 最小化交叉熵公式为

$$\mathcal{L}_g(\theta)=-\frac{1}{T}\sum_{t=1}^{T}\ln p\left(\boldsymbol{Y}_t|\boldsymbol{Y}_{1:t-1},\boldsymbol{X};\ \theta\right) \tag{7}$$

其中,T为解码时间步总步数. 训练时,Transformer的并行训练机制能够同时计算每一步的解码流程,因此,T等于参考摘要的长度.

4 实验

4.1 数据集

本文选择4个不同长度的中文文本摘要数据集对所提方法进行实验验证,分别为短文本摘要数据集LCSTS[8]、中等长度摘要数据集CNewSum[9]、NLPCC2017[10],以及长文本摘要数据集SFZY2020.

LCSTS、CNewSum和NLPCC2017数据集中的具体内容均为中文新闻文本及摘要. 其中LCSTS数据集由哈尔滨工业大学基于新闻媒体在微博上发布的新闻及摘要整理并发布. CNewSum数据集由字节跳动公司基于现有公开的中文新闻,人工提取摘要并发布. NLPCC2017为NLPCC2017摘要数据集,是比赛任务3中所使用的数据集. SFZY2020则为中国法律智能技术评测机构于2020年发布的法律文书及摘要数据集. 四个数据集所包含的摘要和正文的平均长度以及文档个数如表1所示. 由表1可知,LCSTS、CNewSum、NLPCC2017和SFZY2020中的正文平均长度逐渐增长. 实验过程中也将按照该数据集的顺序来逐步验证本文所提方法对长文本摘要生成的有效性. 对于长文本摘要数据集SFZY2020,为了进一步对比不同方法对长文本摘要生成的性能,将该数据集按照正文长度的不同,划分为3个子集,其正文平均长度分别为1 936字、2 463字和3 341字,3个子集的具体情况如表2所示.

表1 不同长度数据集中摘要、正文平均长度及文档个数统计对比

数据集	摘要平均长度/字	正文平均长度/字	训练集个数	验证集个数	测试集个数
LCSTS	18	104	2 400 591	10 666	1 106
CNewSum	35	714	275 596	14 400	14 400
NLPCC2017	44	990	48 000	2 000	2 000
SFZY2020	280	2 587	10 825	1 353	1 353

4.2 实现细节与评价指标

实验平台的操作系统为ubuntu16.04,NVIDA 2080 Ti GPU. 在Pytorch基础上使用Transformers工程库搭建整个网络框架,生成模型使用中文短文本BART预训练模型作为初始化权重,训练时使用AdamW优化器进行梯度更新,初始学习率为2×10^{-4},weight decay为0.001,warm up为500. 最大编码位置嵌入设为8 192,最大解码位置嵌入为1 024,局部注意力窗口为512. 编码器与解码器各6层,隐向量大小为768,前馈神经网络大小为3 072. 在测试过程中使用beam search算法[38]进行下一个字词的查找,beam size设为4. 抽取模型训练时使用AdamW优化器进行梯度更新,初始学习率为2×10^{-5},分类阈值设置为0.3,降维全连接大小为384维,为了提高模型泛化性能,生成训练引导数据及测试时分类阈值设置为0.2.

表2 SFZY2020数据集不同长度子集摘要、正文平均长度及文档个数统计

SFZY2020子集	摘要平均长度/字	正文平均长度/字	训练集个数	验证集个数	测试集个数
子集1	236	1 936	2 332	290	290
子集2	278	2 463	5 614	682	682
子集3	321	3 341	2 879	381	381
共计	280	2 587	10 825	1 353	1 353

实验选取ROUGE-1(R-1)、ROUGE-2(R-2)和ROUGE-L(R-L)三个客观评价指标进行评测[35],各评价指标采用F值计算,该F指标从准确率和召回率计算得到. 其中ROUGE-N基于参考摘要与生成摘要的n元词(n-gram)的匹配情况来评价,具体计算方法为

$$\text{ROUGE-N}=\frac{\sum_{s\in rs}\sum_{(n\text{-gram})\in s}\text{Count}_m(n\text{-gram})}{\sum_{s\in rs}\sum_{(n\text{-gram})\in s}\text{Count}(n\text{-gram})}\times100\% \tag{8}$$

其中,rs为参考摘要集合;s为rs中的某条摘要;分子表示参考摘要与生成摘要同现的相同n-gram的最大数量;分母为参考摘要中n-gram个数. ROUGE指标综合评判生成摘要n-gram的精确率与召回率,越大越好.

为了进一步评价生成摘要的事实一致性,使用文摘事实性指标FactCC[39],将源文本和生成摘要作为输入,经过训练的模型,输出生成摘要的事实一致性标签,用于验证生成摘要在源文本中的事实一致性.

4.3 相关工作对比实验结果

本文分别在LCSTS、CNewSum、NLPCC2017和SFZY2020数据集上与现有流行的中文文本摘要基准模型进行比较. 表3给出了在LCSTS数据集上与现有9种流行方法的实验对比结果,其中长度一栏指源文本的截断长度,模型描述中的Oracle指测试时以抽取算法的抽取结果作为引导信息时的指标,也就是理论上模型所能达到的最高指标. W代表关键词引导信息,Lead-n指抽取源文本前n句作为摘要的生成指标,ours表示本文所构建的文本摘要生成模型. 在对LCSTS数据集训练时,只使用关键词作为引导,因为LCSTS中正文平均长度过短,因此实验过程中直接对正文使用TextRank抽取算法进行关键词的抽取.

表3　LCSTS数据集上的摘要生成对比结果

模型	词表	长度	R-1 ↑/%	R-2 ↑/%	R-L ↑/%	FactCC ↑/%
Lead-1	—	—	27.20	14.41	23.33	41.88
CLSGA(W,Oracle)	words	128	48.44	35.89	45.75	74.27
RNN-context[8]	—	—	26.80	16.10	24.10	41.81
CopyNet[40]	—	—	35.00	22.30	32.00	53.90
Transformer[2]	char	—	38.92	25.83	36.41	61.12
BERT-abs[34]	char	—	40.84	27.07	36.79	62.49
BART[14]	char	128	41.42	27.64	38.06	65.03
CPT[15]	char	128	42.23	29.08	39.15	67.15
Longformer[5]	char	128	43.14	30.36	39.68	68.64
DYLE[22]	char	128	43.29	30.87	40.12	67.96
HEGEL[30]	char	128	43.37	30.89	40.25	69.39
BART	words	128	46.92	**34.67**	**44.39**	72.13
CLSGA(ours,W)	words	128	**47.01**	34.45	44.31	**72.67**

从表3中可以看出,CLSGA在中文短文本摘要数据集LCSTS上的ROUGE分数均领先Transformer、BERT、CPT和BART等中文短文本摘要模型,比上述模型中最好的CPT在R-1、R-2、R-L和FactCC指标上分别提高了4.78%、5.37%、5.16%和5.52%. 与长文本摘要生成模型Longformer、DYLE和HEGEL的对比结果也展示了本文方法的优势. 此外,本文还测试了将BART词表按照本文所构建的中文词表进行扩展的结果,结果如表3倒数第二行所示,在LCSTS数据集上的ROUGE指标亦获得了显著的提升,R-1相比以字为词表的BART提高了5.50%,R-2和R-L分别提高了7.03%和6.33%,涨幅甚至超过了本文的CLSGA方法. 这是因为大规模的数据集有效拟合了新词表,并且LCSTS的摘要长度较为短小,以词生成可能导致更少的错误. 虽然基于中文新词表的BART在R-2和R-L指标上略高于本文的CLSGA方法,但CLSGA在R-1指标和文摘事实性指标FactCC上依然表现最优,表明CLSGA生成的摘要与源文本更加一致,更能有效表达源文本的关键信息.

表4和表5分别给出了在两个中等长度数据集CNewSum和NLPCC2017上与现有多种流行方法的实验对比结果. 模型描述中的Oracle是理论上模型所能达到的最高指标,W表示采用关键词作为引导文本,S表示采用关键句作为引导文本. CNewSum数据集正文平均长度为714,相比于LCSTS的104长了许多.

从表4的结果可以看出,本文提出的CLSGA在3个ROUGE指标和事实性指标FactCC上均高于所对比的10种摘要生成方法,R-1、R-2、R-L和FactCC指标分别达到了52.09%、32.85%、45.42%和76.34%. 从表5的结果可以看出,在正文平均长度为990的NLPCC2017数据集上,本文提出的CLSGA方法采用关键句进行引导

表4　CNewSum数据集上的摘要生成对比结果

模型	词表	长度	R-1 ↑/%	R-2 ↑/%	R-L ↑/%	FactCC ↑/%
Lead-3	—	—	30.43	17.26	25.33	47.47
TextRank	—	—	24.04	13.07	20.08	37.74
CLSGA(W,Oracle)	words	1 024	51.76	32.89	44.94	75.06
CLSGA(S,Oracle)	words	1 024	54.55	35.22	47.07	78.61
Pointer Generator[41]	—		25.70	11.05	19.62	40.61
NeuSum[42]	—	—	30.61	17.36	25.66	48.97
Transformer-ext[2]	char	512	32.87	18.85	27.59	52.26
BRET-ext[34]	char	512	34.78	20.33	29.34	53.30
Transformer-abs[2]	char	512	37.36	18.62	30.62	56.04
BERT-abs[34]	char	512	44.18	27.37	38.32	70.69
CLSGA(ours,W)	words	1 024	51.52	32.35	44.62	74.52
BART[14]	char	512	51.55	32.39	44.76	74.83
CPT[15]	char	512	51.96	32.78	45.03	75.25
Longformer[5]	char	1 024	51.99	32.80	45.09	75.39
DYLE[22]	char	1 024	52.02	32.81	45.13	75.76
HEGEL[30]	char	1 024	52.03	32.83	45.17	75.79
CLSGA(ours,S)	words	1 024	**52.09**	**32.85**	**45.42**	**76.34**

表5　NLPCC2017数据集上的摘要生成对比结果

模型	词表	长度	R-1 ↑/%	R-2 ↑/%	R-L ↑/%	FactCC ↑/%
Lead-3	—	—	31.11	17.08	25.84	49.78
CLSGA(W,Oracle)	words	1 024	55.29	37.88	50.04	83.46
CLSGA(S,Oracle)	words	1 024	58.98	41.75	53.72	89.41
BART[14]	char	512	54.31	37.12	49.37	82.44
CPT[15]	char	512	54.39	37.25	49.48	82.63
Longformer[5]	char	1 024	54.41	37.12	49.43	82.55
DYLE[22]	char	1 024	54.47	37.22	49.68	82.98
HEGEL[30]	char	1 024	54.45	37.21	49.74	83.07
CLSGA(ours,W)	words	1 024	54.47	37.16	49.45	82.58
CLSGA(ours,S)	words	1 024	**55.93**	**38.96**	**50.85**	**84.92**

后的R-1相对CPT提升了1.54%,R-2提升了1.71%,R-L提升了1.37%,FactCC提升了2.29%. 与3个长文本摘要生成方法Longformer、DYLE和HEGEL的指标结果相比,也有一定的提升. 由结果可知,CLSGA可以在更大的上下文跨度上关注更重要的信息,而短文本摘要模型仅能在512字符长度阶段进行学习,因此损失了大量的有用信息. 同时,以关键句作为引导信息比以关键词作为引导信息具有更佳的表现,这是由于句子比词语具有更丰富的语义信息,能够更加准确地定位不同上下文的焦点信息,从而给予模型更多的指导.

表6给出了在长文本摘要数据集SFZY2020上的实验对比结果. 由于BART及CPT模型仅能在512字符长度阶段进行学习,为了更充分地与所提出的CLSGA进行对比,本文将上述模型以重复复制位置编码的方式

进行最大输入长度的扩展. 在RTX2080Ti GPU上分别将BART扩展至1 024和2 048长度;对于CPT模型,由于RTX2080Ti GPU的显存不足以支撑2 048长度输入CPT的模型,因此,将CPT最大扩展至1 024长度. 结果表明:BART和CPT模型的生成指标均随着输入文本长度的增加而提高,证明了在长文本跨度下输入文本长度对摘要生成的影响力. 然而,BART和CPT模型在有限资源下的最大可输入长度是固定的,而SFZY2020数据集正文平均长度在2 400以上,因此上述模型无法在SFZY2020数据集上获得良好的生成. 而本文的CLSGA框架,由于可以有效缩短源文本的序列长度,因此在2 048输入条件下得到了最佳的表现,R-1、R-2、R-L和FactCC指标比表现最好的输入本文长度2 048下的BART分别提升了7.58%、12.47%、11.34%和8.79%,比4种对比的长文本摘要生成方法的R-1、R-2、R-L和FactCC指标平均提升5.99%、8.28%、6.84%和6.8%. 此外,本文还对长文本模型MemSum分别使用字符和词组两种分词模式进行训练,并给出生成摘要的指标结果. 从对比结果可以看出,虽然仅对长文本进行抽取拼接生成也能够取得较好的结果,但还是难以达到本文采用的抽取再生成模型的摘要生成效果.

为了进一步对比长文本模型MemSum、Longformer、DYLE、HEGEL、扩展为2 048长度的BART和本文的CLSGA对长文本数据摘要生成的性能,将上述方法在SFZY2020数据的3个子集(正文平均长度分别为1 936字、2 463字和3 341字)上进行生成摘要的指标对比,结果如表7所示.

由对比结果可知,在正文平均长度逐渐增加的3个子集中,本文的CLSGA在四个指标上均为最优. 随着子集正文长度的增加,CLSGA的优势愈加明显,子集3中CLSGA的R-1、R-2、R-L和FactCC指标比次优方法分别提升6.02%、8.49%、6.85%和6.86%,均高于SFZY2020数据集上的平均提升值. 通过对表3~7的实验结果分析可知,Longformer由于其选择的特定字/词有限且缺少信息引导,在过长的文本跨度下依然会丢失一定的语义信息. MemSum、DYLE和HEGEL等模型均为英文长文本摘要设计,虽然能通过构建多步马尔可夫决策、动态片段级注意力权重以及超GNN来捕捉高阶跨句子关系,进行长文本摘要生成,但是上述方法在用于中文文本摘要生成时,均使用BART中文预训练模型,由于该词表以字为基础,不包含中文常用词语与标点,因此在长文本跨度下很难有效联系上下文,导致部分语义信息的丢失. 与之相比,本文的CLSGA构建了包含抽取模型和生成模型的引导生成架构,有效缓解了长文本跨度下词汇的远距离依赖问题. 通过抽取长文本中的核心词汇和语句来构建引导文本,能够更加准确地定位不同上下文的焦点信息,从而给予生成模型更多的指导. 同时,CLSGA构建了适应于中文长文本摘要模型的词表,基于该词表进行分词,可将文本结构由字改变为词组,有利于提取更加丰富的多粒度特征. 因此,CLSGA能够对上下文进行更充分且有效的建模,在长文本跨度下发挥更重要的作用,对于长文本数据集SFZY2020的性能提升最大,且随着SFZY2020的

表6 SFZY2020数据集上的摘要生成对比结果

模型	词表	长度	R-1 ↑/%	R-2 ↑/%	R-L ↑/%	FactCC ↑/%
Lead-3	—	—	23.60	8.85	17.92	37.76
CLSGA(S,Oracle)	words	2 048	79.64	65.21	75.27	92.42
BART[14]	char	512	64.05	40.71	52.87	67.89
CPT[15]	char	512	64.44	41.31	53.16	74.75
BART	char	1 024	66.43	43.76	55.78	77.06
CPT	char	1 024	67.50	45.32	57.44	78.30
BART	char	2 048	69.65	48.97	60.82	80.79
MemSum[25]	char	4 096	71.24	53.13	65.33	82.35
MemSum[25]	words	4 096	71.37	53.26	65.38	82.79
Longformer[5]	char	4 096	71.21	53.10	65.27	82.37
DYLE[22]	char	4 096	71.28	53.19	65.34	82.69
HEGEL[30]	char	4 096	71.26	53.21	65.36	82.81
CLSGA(ours,S)	words	2 048	**77.23**	**61.44**	**72.16**	**89.58**

表7 SFZY2020不同子集的摘要生成对比结果

模型	SFZY2020数据集不同子集											
	子集1(正文平均长度1 936字)				子集2(正文平均长度2 463字)				子集3(正文平均长度3 341字)			
	R-1	R-2	R-L	FactCC	R-1	R-2	R-L	FactCC	R-1	R-2	R-L	FactCC
CLSGA(S,Oracle)	78.92	64.60	74.35	91.52	79.38	65.07	75.12	92.37	80.12	65.83	75.80	92.51
BART	69.16	47.83	59.73	79.99	69.51	48.62	60.72	80.68	70.02	49.10	61.03	81.10
MemSum[25]	70.98	52.70	64.72	81.97	71.15	52.98	65.27	82.18	71.32	53.34	65.38	82.41
Longformer[5]	71.07	53.01	65.03	82.12	71.16	53.05	65.21	82.24	71.25	53.16	65.32	82.48
DYLE[22]	71.17	53.10	65.17	82.46	71.25	53.14	65.25	82.65	71.31	53.24	65.38	82.86
HEGEL[30]	71.11	53.12	65.22	82.52	71.23	53.20	65.31	82.73	71.29	53.23	65.39	82.98
CLSGA(ours,S)	**76.85**	**61.31**	**72.08**	**89.07**	**77.14**	**61.39**	**72.14**	**89.51**	**77.34**	**61.83**	**72.24**	**89.84**

3个子集中正文长度的逐渐增加,优势更加明显.

为了更直观地对比不同方法对中文长文本摘要生成的性能,本文给出了输入文本长度分别为512、1 024和2 048的BART模型以及本文提出的CLSGA方法在中文长文本摘要数据集SFZY2020上的摘要生成结果示例,具体如表8~表10所示. 由表8可以看出,本文模型在“借款合同合法有效”和“判决被告偿还原告借款及利息”等关键信息处,相比不同输入长度下的BART模型均得到了更加准确的生成,与参考摘要完全一致. 在表9中与参考摘要中的“判决:一、租赁关系解除,被告应向原告返还涉案租赁土地;二、被告内向原告支付租金215 700元”. 判决信息相比,本文模型生成的判决

表8　示例1(源文本共3 400字符,参考摘要290字符)

正文:[…]原告中国农业银行股份有限公司鲁山县支行(以下简称中国农行鲁山支行)诉被告高国营、高冰申、高国锋借款合同纠纷一案,本院于2017年9月9日立案受理,依法组成合议庭,公开开庭进行了审理.原告中国农行鲁山支行的委托代理人郭振岗到庭参加诉讼.[…]判决如下:一、被告高国营偿还原告借款本息25 000元,及截至2017年2月23日止的部分利息868.94元,本息共计25 868.94元(利息、罚息、复利部分根据合同约定自2017年2月24日起,按天计算至还款之日止)[…]
参考摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告贷款及利息;保证人承担连带担保责任.被告未答辩.经审查,原告与被告签订的借款合同合法有效,被告未偿还该笔借款本金及利息,该行为实属违约,依法应承担违约责任,保证人在保证期间依法应承担连带责任保证义务.综上,依照《合同法》第六十条、第一百九十六条、第一百九十八条、第二百零五条、第二百零六条、第二百零七条、《担保法》第十二条、第十八条、《中华人民共和国民事诉讼法》第一百四十四条,参照中国人民银行《人民币利率管理规定》第二十条的规定,判决被告偿还原告借款及利息;保证人承担连带清偿责任;逾期付款,应加倍支付迟延履行期间的债务利息.
BART-512生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告贷款及利息;保证人承担连带清偿责任.被告未答辩.经审查,原告与被告签订的个人购房借款/担保合同合法有效,被告在借款期限届满后未按合同约定偿还借款及支付利息,已构成违约,应当承担违约责任,同时担保人应在最高额范围内对被告的借款本息承担共同还款义务.综上,依照《中华人民共和国合同法》第一百九十六条、第二百零六条,《中国人民银行关于审理民间借贷案件适用法律若干问题的规定》第二十九条第二款第(二)项,《担保法》《民事诉讼法》之规定,判决被告支付原告借款、利息、罚息;逾期付款,应加倍支付迟延履行期间的债务利息.
BART-1024生成摘要:原被告系借款合同纠纷.原告提出诉求:被告偿还原告借款本金及利息;保证人对上述款项承担连带担保责任.被告未答辩.经审查,原、被告间签订的个人借款和保证合同合法有效,被告逾期未还借款,应承担违约责任,同时担保人应在其保证范围内对被告的借款承担共同还款义务.因此,依照《中华人民共和国合同法》第一百零七条、第二百零五条,《中国人民银行关于审理民间借贷案件适用法律若干问题的规定》第二十六条第一款第(二)项、第三十九条第二款、第四十条,最高人民法院《关于适用〈民事诉讼法〉的解释》第九十条之规定,判决驳回原告的诉讼请求;如未按本判决指定的期间履行金钱义务,应加倍支付迟延履行期间的债务利息.
BART-2048生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息;保证人承担连带清偿责任.被告未答辩.经审查,原告与被告签订的个人贷款借贷合同与保证合同均合法有效,借款到期后,被告没有按期归还借款,已构成违约,应当承担违约责任,同时担保人应为该笔借款提供担保.综上,依照《中华人民共和国合同法》第一百九十六条、第二百零七条,《担保法》《最高人民法院关于审理民间借贷案件适用法律若干问题的规定》第二十九条第二款、第三十条及《民事诉讼法》之规定,判决被告支付原告贷款本息、罚息、复利、逾期付款违约金;保障人承连带保证责任并有权向借款人追偿.
CLSGA生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息;保证人承担连带清偿责任.被告未答辩.经审查,原告与被告签订的借款合同合法有效,被告未按合同约定偿还借款及支付利息,已构成违约,应当承担违约责任,同时保证人应对被告的借款本息承担连带清偿责任.综上,依照《中华人民共和国合同法》第六十条第一款、第一百九十六条、第一百九十八条、第二百零五条、第二百零六条、第二百零七条、《中华人民共和国担保法》第十二条、第十八条、《中华人民共和国民事诉讼法》第一百四十四条,《人民币利率管理规定》第二十条的规定,判决被告偿还原告借款及利息;保证人承担连带清偿责任;逾期付款,应加倍支付迟延履行期间的债务利息.
CLSGA生成引导信息:原告中国农业银行股份有限公司鲁山县支行(以下简称中国农行鲁山支行)诉被告高国营、高冰申、高国锋借款合同纠纷一案,本院于2017年9月9日立案受理,依法组成合议庭,公开开庭进行了审理.2、被告高冰申、高国锋对上述款项承担连带担保责任;被告高国营未作答辩.2、被告高冰申、高国锋对上述款项承担连带担保责任;借款到期后高国营未偿还该笔借款本金及利息,该行为实属违约,依法应承担违约责任,被告高冰申、高国锋在保证期间依法应承担连带责任保证义务.最后一笔贷款清偿时,“利随本清”的规定,该约定不违反法律法规的强制性规定,现原告请求被告高国营偿还原告借款本息25 000元,及截至2017年2月23日止的部分利息868.94元,本息共计25 868.94元(利息、罚息、复利部分根据合同约定自2017年2月24日起,按天计算至还款之日),被告高冰申、高国锋对该笔借款本息的清偿承担连带保证责任,于法有据,理由正当,本院予以支持.依照《中华人民共和国合同法》第六十条第一款、第一百九十六条、第一百九十八条、第二百零五条、第二百零六条、第二百零七条、《中华人民共和国担保法》第十二条、第十八条、《中华人民共和国民事诉讼法》第一百四十四条,参照中国人民银行《人民币利率管理规定》第二十条的规定,判决如下:一、被告高国营偿还原告借款本息25 000元,及截至2017年2月23日止的部分利息868.94元,本息共计25 868.94元(利息、罚息、复利部分根据合同约定自2017年2月24日起,按天计算至还款之日止).二、被告高冰申、高国锋在对上述还款义务承担连带清偿责任.如果未按照本判决指定的期间履行给付金钱义务,应当依照《中华人民共和国民事诉讼法》第二百五十三条的规定,加倍支付迟延履行期间的债务利息.

表9 示例2(示例源文本共3 261字符,参考摘要300字符)

正文:[…]2015年12月30日,平庄村委会向原告发出告知书,告知根据上级主管部门要求对临时用地要及时清理,合同到期后自动终止无效,要求被告提前做好相关停产搬迁准备,并协商残值补偿事宜.被告称:被告承租上述房屋和土地用于再生资源利用分解,被告仅使用了一年原告的电,第二年起因电压负荷不足就不用了,且原告未向原告开具租金发票等,故被告延交房租具有合理理由.[…]依照《中华人民共和国合同法》第九十四条、第二百二十七条、第二百三十五条之规定,判决如下:一、原告昆山市周市镇平庄富民合作社与被告朱国付之间的租赁关系于2017年2月6日解除,被告应于本判决发生法律效力之日起十日内向原告返还涉案占地3 330平方米的租赁土地.[…]
参考摘要:原被告租赁合同纠纷一案.原告请求:1、解除租赁合同;2、被告立即腾空房屋并将该土地交付原告;3、被告支付结欠租金215 700元.被告辩称原告主张租金的面积有误,不同意解除合同并搬迁.经查明,原告为法律规定的其他组织,与被告签订租赁协议合法有效.租赁期间虽未届满,但被告经原告催要租金后仍未在合理期限内支付.被告既未提供其未实际使用原告提供电力的证据,也未就原告提供电力无法达到约定要求并就此与原告协商进行举证,故支持原告关于租金的主张.依《中华人民共和国合同法》第九十四条、第二百二十七条、第二百三十五条,判决:一、租赁关系解除,被告应向原告返还涉案租赁土地.二、被告内向原告支付租金215 700元.
BART-512生成摘要:原被告系租赁合同纠纷.原告提出诉讼请求:1、解除租赁协议;2、被告立即腾空迁让交付房屋并将该土地交付原告;3、被告支付结欠租金;被告辩称:原告所述属实,同意解除合同,但不同意返还租赁土地.法院认为,原告与被告之间的租赁关系合法有效,被告应当按照合同约定向原告履行交付租赁物的义务.被告在合同履行期间内,一直占用租赁房屋五间,并且未及时交付,已构成违约,应当承担违约责任.对于原告主张的租金数额,因双方合同中并未明确约定,故本院不予支持.依照《中华人民共和国合同法》第六十条、第九十四条的规定,判决:驳回原告的全部诉求.
BART-1024生成摘要:原告与被告租赁合同纠纷一案.原告提出诉求:1、判令被告立即腾空迁让交付房屋并将该土地交付原告2、被告支付结欠租金.被告辩称:不同意解除合同.经查明原被告签订的《平庄村废品堆场租赁协议》合法有效;原告依约履行了出租义务,被告未按照合同约定履行支付租金的义务;现原告要求被告返还涉案租赁土地,予以支持.根据《中华人民共和国合同法》第八条、第六十条第一款、第九十四条第(二)项、第二百一十二条、《最高人民法院关于民事诉讼证据的若干规定》第二条的规定,判决:1、解除原告和被告的《合作社》2、由被告于本判决生效之日起十日内将位于昆山市周市镇平庄富民合同社的房屋腾空交付给原告3、驳回原告其他诉讼请求.
BART-2048生成摘要:原被告系租赁合同纠纷.原告提出诉讼请求:解除租赁协议;被告立即腾空迁让交付房屋并将该土地交付原告;支付结欠租金.被告辩称:原告主张租金的面积计算有误,根据约定的欠付租金金额是正确的,但是原告实际出租给我的土地面积不足,要求对临时用地要及时清理,合同到期后自动终止无效,被告提前做好相关停产搬迁准备,并协商残值补偿事宜.本院认为,原告作为法律规定的其他组织与被告之间签订的租赁调解协议合法有效,双方均应遵守.现被告未向原告开具租金发票等,故被告延交房租具有合理理由.依照《合同法》《民事诉讼法》的规定,判决:一、被告将废品堆场租赁场地腾空、交付给原告.二、驳回原告的其他诉求.
CLSGA生成摘要:原被告系租赁合同纠纷.原告提出诉讼请求:1、解除租赁合同;2、被告腾空迁让房屋并将该土地交付原告;3、被告支付结欠租金;4、诉讼费用由被告承担.被告辩称:不同意解除合同,不同意搬迁.原告主张租金的面积计算有误.本院认为,原告作为法律规定的"其他组织"与被告之间签订的租赁协议合法有效,双方均应遵守.现双方约定的租赁期间虽未届满,但被告经原告催要租金后仍未在合理期限内支付,原告据此主张租赁协议解除、被告返还租赁土地及支付结欠租金的诉请有事实和法律依据,本院予以支持.现租赁协议已到期,被告理应无条件返还占有的土地并自行负责相关建筑物的拆除、清场和搬迁活动.依照《合同法》的规定,判决:原告与被告之间的租赁关系解除,被告向原告返还租赁土地.被告向原告支付租金.
CLSGA生成引导信息:[…]本院认为,原告作为法律规定的"其他组织"与被告之间签订的租赁协议合法有效,双方均应遵守.现双方约定的租赁期间虽未届满,但被告经原告催要租金后仍未在合理期限内支付,原告据此主张租赁协议解除、被告返还租赁土地及支付结欠租金的诉请有事实和法律依据,本院予以支持,结合被告签收原告诉状的时间,本院认定原被告之间的租赁关系自2017年2月6日解除.关于被告欠付原告租金的金额,双方合同手写部分明确约定"多收300平方米土地租金,免收15千瓦电力租金,按合同收费",现被告既未提供其未实际使用原告提供电力的证据,也未就原告提供电力无法达到约定要求并就此与原告协商进行举证,故原告按照合同约定标准要求被告支付结欠租金215 700元,本院予以支持.另一方面,现租赁协议已到期,被告理应无条件返还占有的土地并自行负责相关建筑物的拆除、清场和搬迁活动,且审理期间相关房屋因涉及违章搭建已被行政部门依法拆除,故根据在案证据、结合本案实际情况,本院对被告赔偿损失的意见不予采纳.依照《中华人民共和国合同法》第九十四条、第二百二十七条、第二百三十五条之规定,判决如下:一、原告昆山市周市镇平庄富民合作社与被告朱国付之间的租赁关系于2017年2月6日解除,被告应于本判决发生法律效力之日起十日内向原告返还涉案占地3 330平方米的租赁土地.二、被告朱国付应于本判决发生法律效力之日起十日内向原告昆山市周市镇平庄富民合作社支付租金215 700元.

信息判决:原告与被告之间的租赁关系解除,被告向原告返还租赁土地. 被告向原告支付租金. 相对来说更加精准和全面. 在表10中,本文模型生成的摘要中则准确包含了"保证人承担连带清偿责任"这一重要信息,而不同输入长度下的BART模型均未注意到该内容.

同时,本文也给出了CLSGA的生成引导信息,由三个示例的引导信息可以看出,所生成的引导信息几乎完全覆盖了参考摘要中的核心内容,这使得生成模型在生成摘要的过程中,能够被引导信息有效约束,从而将注意力集中于文本中更重要的内容. 同时生成模型还可以继续从源文本中搜索引导信息中不包含的信

表10　示例3(示例源文本共1 690字符,参考摘要290字符)

正文:[…]曹栋智与上海悦瑞三维科技股份有限公司劳动合同纠纷一审民事判决书上海市松江区人民法院民 事 判 决 书(2018)沪0117民初338号原告:曹栋智,男,1976年11月7日生,汉族,住江苏省通州市.被告:上海悦瑞三维科技股份有限公司,住所地上海市松江区.[…]据此,依照《中华人民共和国劳动合同法》第二条、第二十六条的规定,判决如下:驳回原告曹栋智的诉讼请求.案件受理费10元,减半收取5元,由原告曹栋智负担(已付).如不服本判决,可在判决书送达之日起十五日内,向本院递交上诉状,并按对方当事人的人数提出副本,上诉于上海市第一中级人民法院.审判员杨 名 二〇一八年一月二十五日书记员[…]
参考摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告贷款及利息;保证人承担担保还款责任.被告未答辩.经审查,原告与被告签订的小额联保借款合同与小额贷款联保协议书合法有效,被告应及时偿还借款,借款逾期后,保证人未能及时还款应承担连带清偿责任,故原告的诉请符合法律规定.故依照《中华人民共和国合同法》第四十四条、第六十条、第一百九十六条、第一百九十七条、二百零五条、第二百零六条、第二百零七条,《担保法》第十八条、第二十一条、第三十一条,《中华人民共和国民事诉讼法》第一百四十四条之规定,判决被告偿还原告借款及利息;保证人承担连带清偿责任;逾期付款,应加倍支付迟延履行期间的债务利息.
BART-512生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息、罚息;保证人承担连带清偿责任.被告未答辩.经审查,原告与被告签订的小额贷款联保协议书合法有效,被告应当按照合同的约定履行偿还借款本息的义务,同时联保小组成员应对上述款项承担共同还款义务.综上,依照《中华人民共和国合同法》第四十四条、第六十条第一款、第一百九十六条,第二百零七条,《中国人民银行关于适用<担保法>若干问题的解释》第十八条,最高人民法院《关于民事诉讼证据的若干规定》第二条之规定,判决被告还款;逾期付款,应加倍支付迟延履行期间的债务利息.
BART-1024生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息、罚息;保证人互相承担担保还款责任.被告未答辩.经审查,原告与被告签订的小额联保借款、小额贷款联保协议书合法有效,被告逾期未还借款,应当承担违约责任,同时联保小组成员应对被告的借款本息承担连带清偿责任;综上,依照《中华人民共和国合同法》第一百九十六条、第二百零七条,《中国人民银行关于适用简易程序审理民事案件的若干规定》第二十九条第一款、第三十条第二款之规定,判决驳回原告的诉讼费用;逾期付款,应加倍支付迟延履行期间的债务利息.
BART-2048生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息、罚息;保证人互相承担担保还款责任.被告未答辩.经审查,原告与被告签订的个人借据与小额贷款联保协议书均合法有效,被告应当按照合同约定偿还借款与利息及罚息,同时联保人应对被告的借款本息承担连带清偿义务.综上,依照《中华人民共和国合同法》第一百九十六条、第二百零七条,《担保法》第十八条,《最高人民法院关于审理民间借贷案件适用法律若干问题的规定》第二十九条第二款、第三十一条,参照《中华人民共和国民法通则》第八十四条、一百零八条规定,判决驳回原告的其他诉讼要求;逾期付款,应加倍支付迟延履行期间的债务利息.
CLSGA生成摘要:原被告系借款合同纠纷.原告提出诉讼请求:被告偿还原告借款及利息;保证人承担担保还款责任.被告未答辩.经审查,原告与被告签订的小额联保借款合同与小额贷款联保协议书均合法有效,被告未按合同约定偿还借款及支付利息,已构成违约,应当承担违约责任,同时保证人应对被告的借款本息承担连带清偿责任.故依照《合同法》第四十四条、第六十条、第一百九十六条、第一百九十七条、二百零五条、第二百零六条、第二百零七条,《担保法》第十八条、第二十一条、第三十一条,《民事诉讼法》第一百四十四条之规定,判决被告偿还原告借款及利息、罚息;保证人承担连带清偿责任;逾期付款,应加倍支付迟延履行期间的债务利息.
CLSGA生成引导信息:原告中国邮政储蓄银行股份有限公司宾县支行(以下简称邮储银行)与被告王凤祥、李淑兰、李淑清、孙洪霞、郑喜有借款合同纠纷一案,本院受理后,依法组成合议庭,于2017年8月8日公开开庭进行了审理.原告诉称:要求被告王凤祥、李淑兰偿还在邮储银行借款本金40 000元,按合同约定支付利息及罚息至给付时止;由被告李淑清、孙洪霞、郑喜有互相承担担保还款责任;被告王凤祥、李淑兰、李淑清、孙洪霞、郑喜有未出庭,无答辩.现邮储银行要求被告王凤祥、李淑兰偿还在邮储银行借款本金40 000元,按合同约定支付利息及罚息至给付时止;由被告李淑清、孙洪霞、郑喜有互相承担担保还款责任;本院认为,原告邮储银行与被告王凤祥、李淑兰签订的小额联保借款合同及与李淑清、孙洪霞、郑喜有签订的小额贷款联保协议书依法成立,合法有效.本案中,王凤祥、李淑兰在邮储银行借款属实,王凤祥、李淑兰应及时偿还借款.借款逾期后,保证人李淑清、孙洪霞、郑喜有对王凤祥、李淑兰未能及时还款应承担连带清偿责任.故依照《中华人民共和国合同法》第四十四条、第六十条、第一百九十六条、第一百九十七条、二百零五条、第二百零六条、第二百零七条,《中华人民共和国担保法》第十八条、第二十一条、第三十一条,《中华人民共和国民事诉讼法》第一百四十四条之规定,判决如下:一、被告王凤祥、李淑兰于判决生效之日起十日内偿还原告中国邮政储蓄银行股份有限公司宾县支行借款本金40 000元,按合同约定给付原告中国邮政储蓄银行股份有限公司宾县支行借款利息、罚息至给付之日止;二、被告李淑清、孙洪霞、郑喜有对上款承担连带清偿责任.

息,对缺失信息加以完善,有效提升了生成摘要的丰富性与完整度.

此外,本文对生成摘要的事实一致性和信息完整性进行了人工评估.分别从NLPCC2017和SFZY2020的测试集中随机抽取100篇和50篇文章.然后,邀请3名相关领域的研究人员对每一篇文章进行打分,并最终根据打分结果评估生成摘要的事实一致性和信息完整性.其中,每个指标的分数范围在1~5,分数越高表明生成摘要的质量越好,表11给出了两个数据集的人工评估结果.从评估结果可以看出本文模型在事实一致性和信息完整性上均获得了最佳得分,其中事实一致性分数表明了模型根据引导信息能够生成与源文本事

实偏差更小的摘要，而信息完整性分数则表明了模型根据引导信息生成摘要的同时并未造成过多的信息丢失．人工评估的结果表明，本文模型在统计学意义上相比其他方法具有更好的表现．

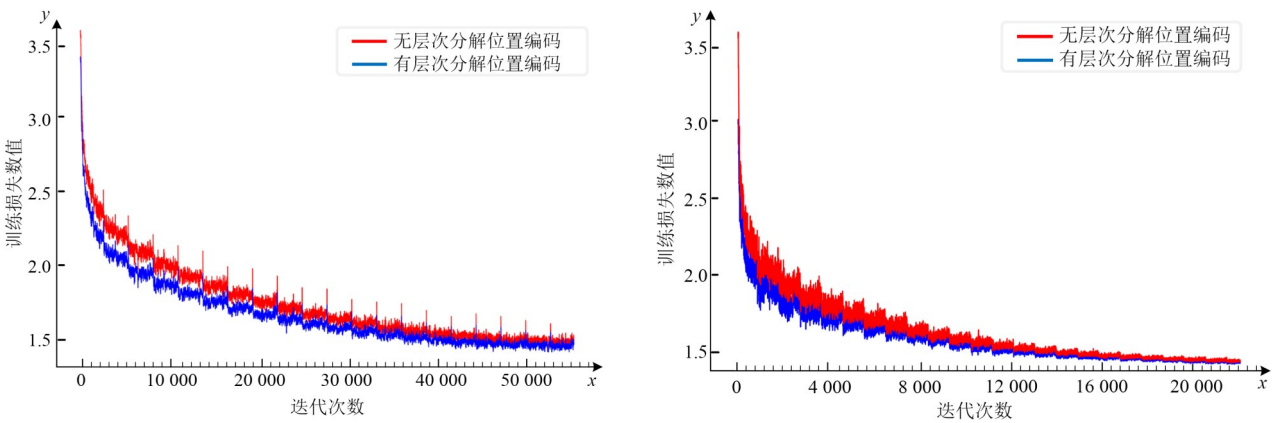
表 11 NLPCC2017及SFZY02020数据集生成摘要的人工评估结果

模型	长度	NLPCC2017事实一致性	NLPCC2017信息完整性	SFZY2020事实一致性	SFZY2020信息完整性
BART	512	4.11	3.59	4.36	4.61
BART	1 024	—	—	4.44	4.65
BART	2 048	—	—	4.62	4.77
CPT	512	4.18	3.63	4.39	4.66
Longformer[5]	4 096	—	—	4.62	4.78
DYLE[22]	4 096	—	—	4.66	4.82
HEGEL[30]	4 096	—	—	4.65	4.81
CLSGA	2 048	**4.32**	**3.95**	**4.68**	**4.89**

4.4 消融实验

为了评估本文所提CLSGA模型框架对中文长文本摘要生成的有效性，本节开展三项消融研究，分别验证层次位置分解编码、中文长文本词表与引导注意力机制的贡献程度．

首先分析验证本文所采用的层次位置分解编码技术的有效性，其基本作用是作为一个更佳的初始化参数，加速模型训练．CLSGA采用可学习的绝对位置编码，如果直接将BART的512位绝对位置编码进行多次复制至8 192，则不能得到一个最好的初始化结果．因此，本次消融实验采取包含层次位置分解编码的CLSGA方法，以及在CLSGA模型框架中删除层次位置分解编码，按照直接进行多次复制的初始化方法，分别进行训练，以对比两者的区别．在NLPCC2017和SFZY2020数据集中的训练损失曲线对比结果如图6所示．从图6的训练结果可以看出采用层次位置编码技术后模型可以更快地收敛至最佳状态，从而证明了层次位置编码的有效性．



(*a*) NLPCC2017数据集 (*b*) SFZY2020数据集

图6 不同位置编码方法的训练损失曲线对比图

接下来验证CLSGA方法中词表设计的有效性，分别从文本长度缩减、模型生成速度等方面进行综合评价．实验分别使用将CLSGA中词表更改为BART以字为基础的词表，与CLSGA方法进行对比实验，以验证本文设计的中文长文本词表的优势，在NLPCC2017数据集中词表分词的对比结果如表12所示．

由表12可知，首先，本文设计的以字词为基础的词表，可以大幅缩短文本的长度．以表12中正文1为例，摘要字分词长度为28，而CLSGA分词长度仅为21，正文长度缩短31.02%，参考摘要长度缩短25%．因此CLSGA中的词表设计可以大幅缩短序列长度．其次，CLSGA则可以有效保留词组，整段语句分词结构清晰，语义明确．第三，从摘要生成时间上来看，表12中正文1采用BART字表的生成时间为4.8 s，而CLSGA仅用时3.4 s．由于读取全部文本后进行自注意力计算的耗时是随着文本的增长而增加的，而经过以词分词后，文本的实际长度大幅缩短，从而有效减少了模型的运算时间．

接着分析CLSGA中的引导注意力机制，探讨其对

表 12　词表分词对比结果

正文 1:51 岁的付宗立是广西东兰县花香乡弄兰村的一个普通农民,家中有 88 岁的老父亲和 84 岁的老母亲,因为身体劳损,付宗立从去年开始就没有外出打工了.对于这个贫困家庭来说,经济压力陡增.从 2015 年 4 月开始,付家的低保也突然停发了… (3 348 字)	**正文 2**:一间斗室,一盏孤灯,一壶酒,一位智者,一位少年.少年功成,就要离别,智者举酒,有言相告.“功成之后莫轻使,持技欺人忘德行.”智者说.“不用武力,怎能伏魔,”少年问.“功夫有限,仁者无敌,”智者说… (2 620 字)
参考摘要:村支书停发低保户低保钱,花香乡弄兰村村民被逼成“侦探”. (28 字)	**参考摘要**:少年为救人将色狼打成重伤获刑,因此被迫放弃高考,现在工地做零工,受害女子未露面. (40 字)
字分词:村 支 书 停 发 低 保 户 低 保 钱 ,花 香 乡 弄 兰 村 村 民 被 逼 成 “侦 探”. (28 字)	**字分词**: 少 年 为 救 人 将 色 狼 打 成 重 伤 获 刑, 因 此 被 迫 放 弃 高 考, 现 在 工 地 做 零 工, 受 害 女 子 未 露 面. (40 字)
生成耗时: 4.8 s	**生成耗时**: 4.6 s
CLSGA 词组分词:村支书 停 发 低保户 低保 钱 , 花 香 乡 弄 兰 村 村民 被 逼 成 “侦探”. (21 字)	**CLSGA 词组分词**:少年 为 救人 将 色狼 打成 重伤 获刑, 因此 被迫 放弃 高考, 现在 工地 做 零工, 受害 女子 未 露面. 24 字
生成耗时: 3.4 s	**生成耗时**: 3.1 s
正文字分词长度: 3 346 字　**正文词组分词长度**: 2 308 字	**正文字分词长度**: 2 595 字　**正文词组分词长度**: 1 795 字

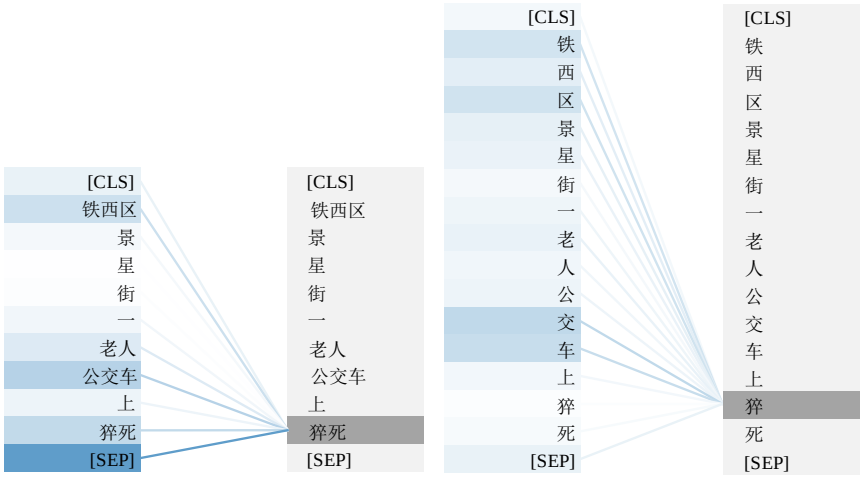
长文本上下文中的焦点信息关注度的贡献. 分别在 CNewSum、NLPCC2017 和 SFZY2020 数据集上对 CLSGA 去除引导注意力机制和 CLSGA 进行测试,结果如表 13 所示. 由实验结果可知,在引入引导注意力机制后,CLSGA 通过训练抽取模型构建以关键语句为引导本文的引导信息,并在生成模型中融合该引导信息进行训练,从而有效提高了对长文本上下文中焦点信息的关注度,约束生成摘要与源文本的偏差,最终提高了摘要生成的准确度. 在中等长度数据集 CNewSum 和 NLPCC2017 上的消融实验结果表明,增加引导注意力机制后,生成摘要的 R-1、R-2 和 R-L 指标均有所提升. 对于长文本摘要数据集 SFZY2020,由于关注了长文本跨度下的重要信息,因此,引导注意力机制的有效性更加显著,相比去除引导注意力机制,生成摘要的 R-1、R-2 和 R-L 指标分别提升了 2.1、2.42 和 2.42. 实验证明了引导注意力机制在长文本摘要生成的有效性.

表 13　引导注意力机制的有效性对比

数据集	模型	词表	引导信号	长度	R-1 ↑/%	R-2 ↑/%	R-L ↑/%
CNewSum	CLSGA	words	×	1 024	51.17	32.14	44.96
	CLSGA	words	√IS	1 024	52.09	32.85	45.42
NLPCC2017	CLSGA	words	×	1 024	54.12	37.62	49.19
	CLSGA	words	√IS	1 024	55.93	38.96	50.85
SFZY2020	CLSGA	words	×	2 048	75.13	59.02	69.74
	CLSGA	words	√IS	2 048	77.23	61.44	72.16

注:“×”指去除引导注意力机制,“√IS”指采用关键句作为引导文本.

最后分析词注意力和字注意力两种词表的关注度对比,图 7 给出了 CLSGA 在两种词表下最顶层编码器(第六层源文本编码信息)对不同字或词的关注程度. 可以看出,以词为基础的关键词“猝死”可以有效被“铁西区”“老人”“公交车”等重要词汇关注到,而以字为基础的模型的注意力分布较为分散,语义理解层次较低. 通过消融实验验证了本文所提 CLSGA 模型中层次位置



(*a*) 词注意力　　(*b*) 字注意力

图 7　词表关注度对比示意图

分解编码、中文长文本词表与引导注意力机制的有效性,能更准确地捕捉长文本跨度下的重要信息,显著提高了长文本摘要生成的能力.

5 结论

本文提出了一种可灵活应用于中文长文本摘要生成任务的方法CLSGA. 该方法针对中文长文本摘要任务重新设计词表. 在降低文本长度与词语不确定性的同时,有效提升模型的性能. 此外,将局部注意力编解码结构与引导注意力相结合,使模型能够根据引导注意力机制着重关注焦点信息,在降低计算量与显存占用量的同时,对上下文进行充分有效的建模,使其在长文本跨度下发挥更重要的作用. 在4个不同长度的中文摘要数据集上进行实验,验证了本文方法对于长文本摘要生成的有效性. 同时,本文的方法还有一定的提升空间,例如对模型继续进行预训练,让其学习更多的知识,以及结合如高效注意力机制等更复杂的技术,进一步减少系统显存的空间占用. 未来工作主要关注多模态的中文长文本摘要生成,通过融合文本、图像、声音等多种数据模态,生成更加丰富和准确的摘要. 同时,通过领域适应性训练,进一步提高模型在特定领域(如法律、医疗、科技等)的摘要生成能力.

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.
LI J P, ZHANG C, CHEN X J, et al. Survey on automatic text summarization[J]. Journal of Computer Research and Development, 2021, 58(1): 1-21. (in Chinese)

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st International Conference on Neural Information Processing Systems, New York: Curran Associates Inc., 2017(30): 6000-6010.

[3] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 4(January): 3104-3112.

[4] 侯丽微, 胡珀, 曹雯琳. 主题关键词信息融合的中文生成式自动摘要研究[J]. 自动化学报, 2019, 45(3): 530-539.
HOU L W, HU P, CAO W L. Automatic Chinese abstractive summarization with topical keywords fusion[J]. Acta Automatica Sinica, 2019, 45(3): 530-539. (in Chinese)

[5] BELTAGY I, PETERS M E, COHAN A. Longformer: The long-document transformer[EB/OL]. (2020-04-10)[2023-05-12]. https://arxiv.org/abs/2004.05150v2.

[6] ZAHEER M, GURUGANESH G, DUBEY K, et al. Bigbird: Transformers for longer sequences[C]//The 34st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020(33): 17283-17297.

[7] 鲍宇, 黄书剑, 周浩, 等. 基于句法模板采样的无监督复述生成方法[J]. 中国科学: 信息科学, 2022, 52(10): 1808-1821.
BAO Y, HUANG S J, ZHOU H, et al. Unsupervised paraphrasing via syntactic template sampling[J]. Scientia Sinica (Informationis), 2022, 52(10): 1808-1821. (in Chinese)

[8] HU B T, CHEN Q C, ZHU F Z. LCSTS: A large scale Chinese short text summarization dataset[J]. Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015: 1967-1972.

[9] WANG D Q, CHEN J Z, WU X Z, et al. CNewSum: A large-scale summarization dataset with human-annotated adequacy and deducibility level[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 389-400.

[10] HUA L F, WAN X J, LI L. Overview of the NLPCC 2017 shared task: Single document summarization[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2017: 942-947.

[11] ZHONG M, LIU P F, CHEN Y R, et al. Extractive summarization as text matching[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 6197-6208.

[12] LIU Y. Fine-tune BERT for extractive summarization[EB/OL]. (2019-03-25)[2023-05-12]. https://arxiv.org/abs/1903.10318v2.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[14] LEWIS M, LIU Y H, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7871-7880.

[15] SHAO Y F, GENG Z C, LIU Y T, et al. CPT: A pre-trained unbalanced transformer for both Chinese language understanding and generation[EB/OL]. (2021-09-13)[2023-05-12]. https://arxiv.org/abs/2109.05729v4.

[16] DOU Z Y, LIU P F, HAYASHI H, et al. GSum: A general framework for guided neural abstractive summarization[C]//

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 4830-4842.

[17] 李志欣, 彭智, 唐素勤, 等. 融合上下文信息和关键信息的文本摘要[J]. 中文信息学报, 2022, 36(1): 83-91.
LI Z X, PENG Z, TANG S Q, et al. Fusing context information and key information for text summarization[J]. Journal of Chinese Information Processing, 2022, 36(1): 83-91. (in Chinese)

[18] KOH H Y, JU J X, LIU M, et al. An empirical survey on long document summarization: Datasets, models, and metrics[J]. ACM Computing Surveys, 2023, 55(8): 1-35.

[19] COHAN A, DERNONCOURT F, KIM D S, et al. A discourse-aware attention model for abstractive summarization ofLong documents[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies. Stroudsburg: Association for Computational Linguistics, 2018: 615-621.

[20] LIU Y, LAPATA M. Hierarchical transformers for multi-document summarization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5070-5081.

[21] MANAKUL P, GALES M. Long-span summarization via local attention and content selection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6026-6041.

[22] MAO Z M, WU C H, NI A S, et al. DYLE: Dynamic latent extraction for abstractive long-input summarization [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 1687-1698.

[23] CAO S Y, WANG L. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 786-807.

[24] FONSECA M, ZISER Y, COHEN S B. Factorizing content and budget decisions in abstractive summarization of long documents[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 6341-6364.

[25] GU N L, ASH E, HAHNLOSER R. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 6507-6522.

[26] WAHAB M H H, ABDUL HAMID N A W, SUBRAMANIAM S, et al. Decomposition-based multi-objective differential evolution for extractive multi-document automatic text summarization[J]. Applied Soft Computing, 2024, 151: 110994.

[27] WU W H, LI W, XIAO X Y, et al. BASS: Boosting abstractive summarization with unified semantic graph[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6052-6067.

[28] JING B Y, YOU Z Y, YANG T, et al. Multiplex graph neural network for extractive text summarization[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 133-139.

[29] QIU Y F, COHEN S B. Abstractive summarization guided by latent hierarchical document structure[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 5303-5317.

[30] ZHANG H P, LIU X, ZHANG J W. HEGEL: Hypergraph transformer for long document summarization[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 10167-10176.

[31] BENGIO S, VINYALS O, JAITLY N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 1171-1179.

[32] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[EB/OL]. (2016-09-26) [2023-05-12]. https://arxiv.org/abs/1609.08144.

[33] WU F, LAO N, BLITZER J, et al. Fast reading comprehension with ConvNets[EB/OL]. (2017-11-12) [2023-05-12]. https://arxiv.org/abs/1711.04352v1.

[34] 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). Stroudsburg: Association for Computational Linguistics, 2019: 3730-3740.
[35] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of Workshop on Text Summarization Branches Out. Stroudsburg: Association for Computational Linguistics, 2004: 74-81.
[36] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]//The 2004 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2004: 404-411.
[37] DE B P T, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.
[38] TILLMANN C, NEY H. Word reordering and a dynamic programming beam search algorithm for statistical machine translation[J].ComputationalLinguistics,2003,29(1): 97-133.
[39] KRYSCINSKI W, MCCANN B, XIONG C M, et al. Evaluating the factual consistency of abstractive text summarization[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 9332-9346.
[40] GU J T, LU Z D, LI H, et al. Incorporating copying mechanism in sequence-to-sequence learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 1631-1640.
[41] 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 forComputational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 1073-1083.
[42] ZHOU Q Y, YANG N, WEI F R, et al. Neural document summarization by jointly learning to score and select sentences[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 654-663.

作者简介

郭 哲 女,1984年3月出生于河南省洛阳市. 现为西北工业大学电子信息学院副教授、博士生导师. 在国内外发表学术论文60余篇. 主要研究方向为人工智能.
E-mail: guozhe@nwpu.edu.cn

张智博 男,1999年4月出生于黑龙江省佳木斯市. 主要研究方向为智能信息处理.
E-mail: billz@mail.nwpu.edu.cn

周炜杰 男,1997年6月出生于四川省万源市. 主要研究方向为智能信息处理.
E-mail: zhouweijie@mail.nwpu.edu.cn

樊养余 男,1960年4月出生于陕西省西安市. 现为西北工业大学电子信息学院教授、博士生导师. 获陕西省科学技术奖1项. 在国内外发表学术论文200余篇. 主要研究方向为多媒体信息处理.
E-mail: fan_yangyu@nwpu.edu.cn

张艳宁 女,1969年10月出生于陕西省市武功县. 现为西北工业大学副校长、教授、博士生导师. 获国家级、省部级科研奖励十余项. 在国内外发表学术论文300余篇. 主要研究方向为图像处理与计算机视觉.
E-mail: ynzhang@nwpu.edu.cn