

文章编号: 1003-0077(2025)01-0001-15

中文大模型生成式摘要能力评估

王俊超¹, 樊可汗¹, 霍智恒^{1,2}

(1. 先进计算与智能工程(国家级)实验室, 河南 郑州 450001;

2. 郑州大学 网络空间安全学院, 河南 郑州 450002)

摘要: 从传统的纸带机到当今大语言模型时代, 自动文本摘要技术发展经历了多次质的飞跃并不断提升。但在中文摘要方面, 由于其语言特点及叙述方式, 机器生成的摘要难以与人工撰写的相媲美。如今, 众多国产开源大模型均加强了对中文语料的训练并展示出较为优秀的成果。为了评估这些开源大模型在中文摘要任务上的实际表现, 该文筛选 ChatGLM2-6B、Baichuan2-7B 和 InternLM-7B 等中文大模型作为研究对象, 在中文摘要数据集上采用不同提示词生成零样本和少样本摘要, 通过自动评估和人工比对的方法详细分析了它们在自动文本摘要任务上的表现及其不足之处。评估结果表明, ChatGLM2-6B 和 Baichuan2-7B 通过零样本的方法通常能够总结出语句通顺叙述详尽的摘要, 但在凝练程度上仍有不足; 而少样本的方法可以使大模型生成更为精炼的摘要, 但对重点信息的把握程度明显下降。此外, 大模型也存在陷入重复、出现幻觉、与事实矛盾等问题。

关键词: 自动文本摘要; 大语言模型; 能力评估

中图分类号: TP391

文献标识码: A

Benchmarking the Performance of Chinese Large Language Models for Abstractive Text Summarization

WANG Junchao¹, FAN Kehan¹, HUO Zhiheng^{1,2}

(1. Laboratory for Advanced Computing and Intelligence Engineering, Zhengzhou, Henan 450001, China;

2. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan 450002, China)

Abstract: Automatic text summarization has undergone transformative advancements into the current era of large language models (LLM). In recent times, numerous domestic open-source large models, having intensified training on Chinese corpora, demonstrate remarkable progress. To assess the performance of Chinese open-source large language models on the task of abstractive text summarization, this paper selects ChatGLM2-6B, Baichuan2-7B, and InternLM-7B as subjects. These models are tested on two Chinese summarization datasets using various prompts for both zero-shot and few-shot summarization. The results show that ChatGLM2-6B and Baichuan2-7B typically generate coherent, detailed summaries via zero-shot, albeit lacking conciseness. On the other hand, few-shot method allows LLMs to generate brief summaries, but results in the omission of many key information. Additionally, the models exhibited issues such as repetition loops, hallucinations, and factual inconsistency.

Keywords: automatic text summarization; Large Language Model; benchmark

0 引言

自动文本摘要 (Automatic Text Summarization) 作为一种应对信息过载的关键工具, 旨在通过压缩文本内容以提炼出文档中最为核心且易于理解的信息。该技术可大致划分为两类: 抽取式摘要

(Extractive Summarization) 和生成式摘要 (Abstractive Summarization)。其中, 抽取式摘要是直接从原文中挑选重要的句子并组合成摘要; 生成式摘要则涉及对文本的深层次理解, 利用自然语言生成技术构造摘要文本。除此之外, 混合式方法是这两种方法的结合, 通常情况下先抽取文本中最关键的信息, 加以筛选、串联后再通过生成式的方法

得到摘要。

从 20 世纪传统纸带机问世,到如今人工智能技术逐渐成为人们生活中的必需品,自动文本摘要技术也在不断更新。从最早的基于特征统计的方法来提取原始文档中的有效信息,再到通过组合优化、机器学习、神经网络等方法选取句子、生成摘要,到当前通过向大语言模型 (Large Language Model, LLM) 输入简短的提示词 (Prompt) 和原始文档,就可以获得对应的摘要 (图 1)。近期的研究主要聚焦于特定领域的文本摘要上,以科学论文摘要^[1]、代码摘要^[2]、对话摘要^[3]、多模态摘要^[4] 为主,这些领域由于文本结构的特殊性而受到学术界的密切关注。总的来说,自动文本摘要技术仍面临挑战,目前阶段,利用大语言模型生成的自动摘要虽展现出一定前景,但仍存在一些阶段性局限和待克服的问题。

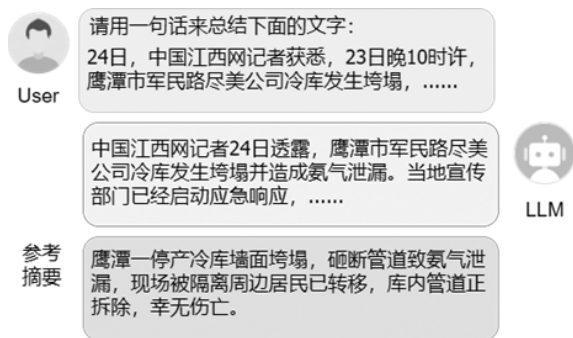


图 1 通过提示词使 LLM 生成摘要

2022 年 11 月底,美国 OpenAI 公司发布了聊天型生成式预训练模型 ChatGPT。随着 ChatGPT 问世,大语言模型所展现的非同一般的能力逐渐被世人所认可。经过大规模非结构化文本语料的训练,和千亿级训练参数的优化,这种通用型的大语言模型可以被应用于多种自然语言处理的下游任务,同时具有较强的自然语言理解和生成能力。

在此之后,国内的一些大语言模型,如 ChatGLM^①、讯飞星火^②、通义千问^③、百川大模型^④等陆续发布,加强了在中文数据上的训练,提升了对中文的语义理解,也受到了学者们的高度关注,并促使他们展开相应的评估工作。

相较于此前的预训练模型 (Pretrained Language Models, PLM),大语言模型在上下文理解和语言生成方面的表现更为卓越,这与自动文本摘要任务的需求高度契合。而中文语言凝练程度更高,且富有吸引力,在叙述方式上可能为了突出结果而改变原文的主谓宾结构,使得中文摘要任务成为一项更

具挑战性的自然语言生成任务。

本文工作主要包括:①使用多个零样本提示词及少样本提示词,通过重复性实验对现有的 3 个开源中文大模型进行了评估,发现即使表面上看似含义相近的提示词,也会对摘要结果产生影响,不同的模型对摘要的侧重点也有区别;②针对实验的结果,讨论了在现有中文大模型在自动文本摘要任务上的表现及其不足之处;③提供了大量的实验数据,可以为今后的研究工作提供参考、建立实验基础。

1 相关工作

本文的工作是评估大语言模型在生成式文本摘要任务上的表现,因此,本节将对大语言模型、自动文本摘要技术和大模型时代的生成式摘要技术的发展进行简要介绍。

1.1 大语言模型

目前,自然语言处理 (Nature Language Processing, NLP) 中的语言模型 (Language Model) 已经进入了第四发展阶段^[5]——大语言模型时代。而 LLM 也被认为是自然语言处理领域正在经历的一次范式转变^[6] (Paradigm Shift),其将 NLP 下游任务的相关研究也推向了一个崭新的高度。通常,LLM 是指包含千亿级别以上参数的语言模型,这些参数经过海量的文本数据进行训练。研究人员发现,随着训练样本规模的增加,LLM 的性能也将逐步提高,并且在少样本、零样本场景下,都表现出与传统监督学习相当的性能^[7]。与上个发展阶段的一些较小规模的 PLM 相比,这种 LLM 在预训练期间获得了某些意料之外的能力,被称为涌现能力^[8] (Emergent Abilities),如世界知识、任务泛化、语义理解、复杂推理等,这些能力使得大模型能够更好地理解和生成,甚至可以如常人般进行算术推理和常识推理^[9]。

大模型的背后离不开庞大的算力支持和昂贵的计算成本。于是,大模型轻量化工作也逐渐展开。如减少模型的存储空间,并提高计算效率,这也是现阶段需要考虑的重要问题之一。作为普通爱好者,更需要小型化的大模型来解决一些简单的现实问题。因

① ChatGLM: <https://chatglm.cn/>

② 讯飞星火: <https://xinghuo.xfyun.cn/>

③ 通义千问: <https://tongyi.aliyun.com/qianwen/>

④ 百川大模型: <https://www.baichuan-ai.com/home>

此,在尽可能减少大模型性能损失的情况下,对大模型进行轻量化是问题的关键。其中,缩小参数规模也是一种可行的办法。主流的开源模型如 LLaMa^[10]、Alpaca^[11]、ChatGLM^[12]等均有 7B 或 6B 的参数模型可供使用,可在一些小型计算机上进行部署。除此之外,还可以使用量化技术(Quantization)^[13]、知识蒸馏(Distillation)^[14]、模型修剪(Pruning)^[15]等方法。

1.2 自动文本摘要

自动文本摘要从生成的方法进行分类,可以分为抽取式和生成式,以及将两种方式进行混合的混合式。虽然方法不同,但自动文本摘要的本质在于对输入文本的降维处理,需要清楚地理解输入文本的内容,包括字、词、短语、句子、段落和篇章等多粒度的结构,再过滤掉不重要以及重复的信息,最终通过抽取式或是生成式的技术获得摘要,这对模型的语义理解能力和文本生成能力有着非常高的要求。

此前在 PLM 时代,一些较为出众的,专注于自然语言生成以及自动摘要的 PLM 被提出,包括 BART^[16]、PEGASUS^[17]、Z-code++^[18],成为当时最为出众的成果。此后的优秀文本摘要模型还包括了 SimCLS^[19]、SummaReRanker^[20]、BRIO^[21]等。这些方法较好地解决了 Seq2Seq 模型的曝光偏差(Exposure Bias)问题和目标函数与 ROUGE^[22](Recall-Oriented Understudy for Gisting Evaluation)评价指标不一致的问题。前者是指 Seq2Seq 模型在测试阶段需要以自回归的方式生成摘要,因此生成过程容易造成错误累加;后者是因为在计算过程中,目标函数的计算是基于局部的句子,而 ROUGE 这类评价指标则计算参考摘要和模型生成的摘要整体上的相似性。于是,强化学习^[23]、对比学习^[21]思想成为上述生成式摘要问题的主要解决方案,通过构建多个候选摘要对其进行打分,最终选择评分最高的摘要结果。

1.3 大模型时代的生成式摘要

在大模型时代,输入简短的提示词和原始文档就可以获得摘要。Wang 等人^[24]提出了一种“思维链摘要(SumCoT)”技术,通过思维链(Chain of Thought)的提示方式先整合新闻文档中的元素(Element)信息(如时间、地点、人物等),诱导 LLM 逐步生成摘要,此方法使得 LLM 将原始文档中更多细粒度的这些元素整合到最终摘要中。Zhang 等

人^[25]利用 ChatGPT 并分别使用直接生成、先提取后生成,以及使用与参考摘要最为相似的句子子集(Oracle Sentences)生成的三种方法得到摘要结果,之后还使用了 FactCC^[26](Factual Consistency Checking Model)评估摘要生成的可信度。Liu 等人^[27]继续沿用此前与 BIRO 类似的对比学习的思想,提出了将 LLM 生成的伪参考摘要(Quasi-reference Summaries)作为参考的两种学习策略。其一是使用 BART 生成候选摘要,LLM 作为参考,利用对比学习进行优化;其二是让 ChatGPT 和 GPT-4 反馈不同 LLMs 生成的候选摘要和经过对比学习优化后的候选摘要的好坏,以证明对比学习的有效性。

大模型生成摘要的质量并不能一概而论,至少它能模仿人类的思路对文章中的内容进行提取和概括。虽然大模型可能会自动生成一些不遵循于原文或不符合事实的文本,这个问题被称为“幻觉”(Hallucination),而且在传统的自动评估指标(如 ROUGE)下,LLM 与微调模型相比没有明显的优势,但人类评估结果表明,与微调模型的结果相比,人类倾向于更喜欢 LLM 生成的结果^[28-30]。Zhang 等人^[29]基于事实一致性的角度,重新构建了对比数据集(事实一致与非事实一致),让现有的 23 个 LLM 进行摘要评估,实验表明 LLM 更加倾向于从输入文章中逐字逐句地提取文本,即使它们与事实不一致。Pu 等人^[30]在单新闻、多新闻、对话、源代码和跨语言摘要 5 个摘要测试集上评估了 GPT-3、GPT-3.5 和 GPT-4 的性能。最终得出以下结论: LLM 在人类评估方面具有优势,GPT-4 在生成摘要方面能够极大地减少“幻觉”的产生,在多文档的新闻摘要数据集(Multi-News^[31])上尤为明显。Wang 等人^[32]在中英跨语言对话摘要数据集 XSAMSum^[33]上对 ChatGPT 等 6 个模型进行了评估,发现 ChatGPT 和 GPT-4 最初倾向于生成带有详细信息的冗长摘要,但可以在交互式提示下进一步平衡信息和简明化,显著提高跨语言摘要的性能。武俊宏等人^[34]使用 ChatGPT 在中文对话摘要数据集 CSDS 上进行了测试,并与该数据集上的 SOTA 模型进行了对比。实验结果表明,ChatGPT 在中文对话摘要上的性能并不如当前最优秀的模型,而且生成摘要的长度明显远长于参考摘要。在给定长度限制之后,则会降低生成摘要的质量。上述工作主要在闭源 LLM(如 ChatGPT)上展开,对现有开源 LLM 工作没有进行详细的验证。

对此,本文总结了上述工作所使用的大模型提示词(表 1)。当下,自动文本摘要技术的现实需求仍然强烈,需选择合适的评估方法对中文开源 LLM 的中文文摘能力进行合理的评估。

表 1 零样本摘要提示词

工作	提示词	数据集
Wang ^[21] 等人	Summarize the above article;	CNN/ DailyMail
	Summarize the above article in one sentence;	BBC XSum
Tam ^[7] 等人	• [input] • The summary of [input] is • Summarize: [input]	CNN/ DailyMail
Liu ^[24] 等人	Summarize the above article in three sentences.	CNN/ DailyMail

2 评估

2.1 模型

随着大模型时代的到来,以及越来越多的开源模型得到了 NLP 学术界的认同,国内的众多研究者也竞相公布了各自研究团队的大模型成果。

本文试图选取影响力和受众较多的大模型作为研究对象,以提升本次评估实验的实用性和可靠性。根据 2023 年 9 月的 ModelScope^①、互链高科^②网站上的下载排行榜,选取了 ChatGLM2-6B^③、Baichuan2-7B^④模型,并结合了众多大模型公布的测试中结果相当优异的 InternLM-7B^⑤模型,下面将对这些大模型做简单介绍。

2.1.1 ChatGLM2-6B

清华数据挖掘团队(Tsinghua University Data Mining, THUDM)在 2023 年 3 月推出了 ChatGLM-6B 双语对话语言模型,并在 2023 年 6 月 25 日公布了第 2 代 ChatGLM2-6B,通过在 1.4 万亿标识符(Token)的数据上进行预训练,结合人类偏好对齐训练(Reinforcement Learning with Human Feedback, RLHF)。在众多的中英双语测试数据集上,较初代模型实现了显著的性能提升。

2.1.2 Baichuan2-7B

Baichuan 大模型是百川智能公司在 2023 年 9 月 6 日发布的第 2 代模型 Baichuan2,分别有 7B、

13B 和 53B(闭源)参数的模型。模型在第一代的基础上提升了训练数据,包含了 2.6 万亿标识符,技术、商业、娱乐类别的语料较多。在中英文通用、数学、代码、医疗、法律的基准数据集上的得分均有明显的提高。

2.1.3 InternLM-7B

InternLM-7B 是上海人工智能实验室等机构在 2023 年 7 月发布的轻量级模型,在综合考试的数据集上与国内外代表性 7B 开源模型相比均有一定程度上的领先。针对 7B 模型,对上下文窗口进行了扩展,提升至 8K,增强了对长文本理解、复杂推理、撰写代码等方面的能力。

三个模型的对比如表 2 所示。

表 2 模型部分参数信息

	ChatGLM2-6B	Baichuan2-7B	InternLM-7B
参数量	6 243 584 000	7 505 973 248	7 321 948 160
词表大小	65 024	125 696	103 168
上下文长度	32 768	4 096	8 192
Transformer 层数	28	32	32
词嵌入层大小	4 096	4 096	4 096

2.2 方法

2.2.1 零样本摘要

对比近期通过大模型生成摘要的部分工作(表 1),可以看出零样本(Zero-Shot)提示词的设计并不困难。对于中文摘要,本文通过采用不同的中文提示词生成摘要,发现即使表面上看似含义相近的提示词,也会对摘要结果产生影响。由于大模型的生成结果往往具有多样性,需要进行重复性实验以得到更准确和稳定的结果。此外,根据提示工程

① ModelScope; <https://www.modelscope.cn/models>

② 互链高科; <https://www.aliendao.cn/>

③ ChatGLM2-6B; <https://github.com/thudm/chatglm2-6b>

④ Baichuan2-7B; <https://www.modelscope.cn/models/baichuan-inc/Baichuan2-7B-Chat/summary>

⑤ InternLM-7B; https://www.modelscope.cn/models/Shanghai_AI_Laboratory/internlm-chat-7b/summary

的理念,设计提示本身就是注重细节的过程,也需要多次实验来优化提示词。

于是,为了尽可能地研究出提示词与摘要结果之间的联系,本文采用策略性方法来筛选出最优提示词,如图 2 中示例。以下是具体步骤:

(1) 对于每个中文大模型,选择使用表 3 所示的 n 个不同提示词 $p = \{p_1, p_2, \dots, p_n\}$, 使用每个

提示词对数据集的前 100 份数据分别生成 m 个候选摘要结果 $c_i = \{c_{i_1}, c_{i_2}, \dots, c_{i_m}\}, i \in [1, n]$ 。接着对这些候选摘要要进行打分 $\text{ROUGE}(c_{i_j}), j \in [1, m]$ 。并且针对每个提示词 p_i , 计算对应候选摘要分数的平均值 $\text{SCORE}(p_i)$ 。

$$\text{SCORE}(p_n) = \frac{1}{m} \sum_{j=1}^m \text{ROUGE}(c_{i_m}) \quad (1)$$

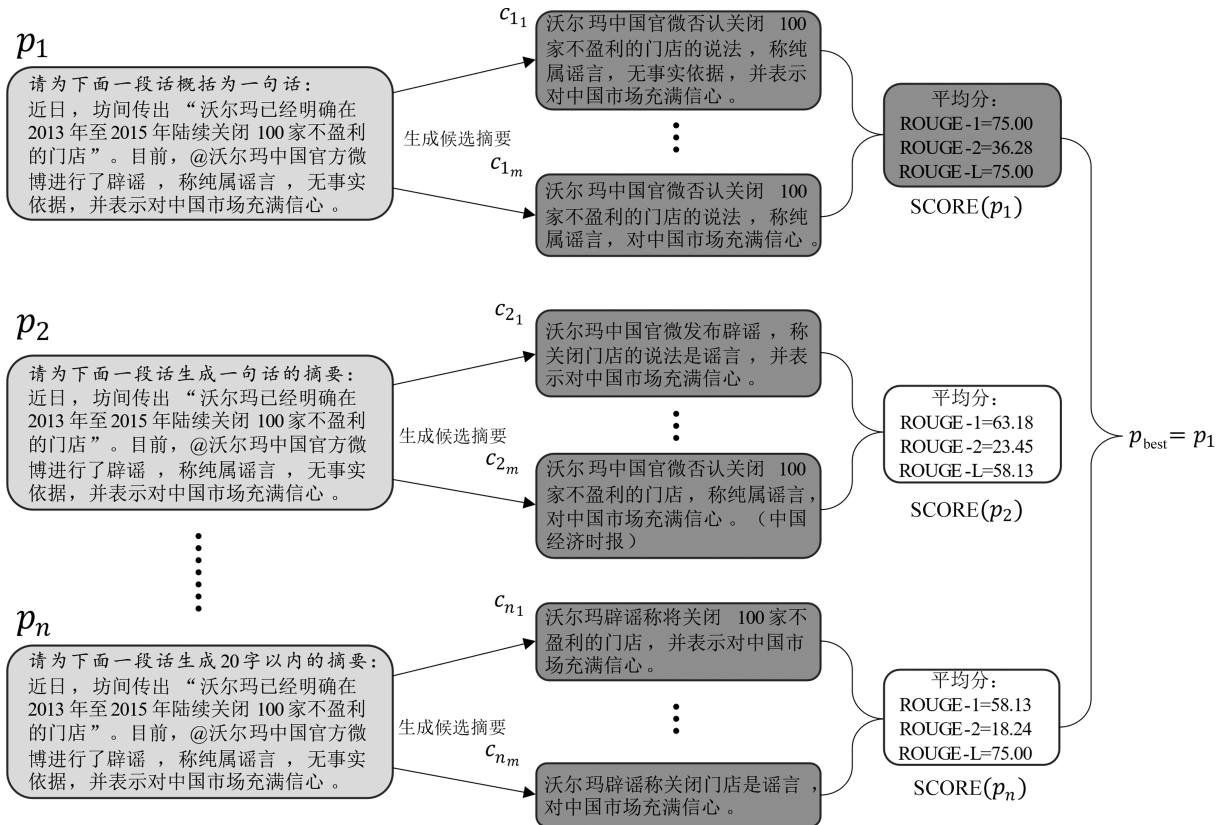


图 2 零样本摘要方法示例

表 3 本文采用的零样本提示词

数据集	提示词
NLPCC2017	p_1 : 请用一句话来总结下面的文字:
	p_2 : 请将下面一段话概括为一句话:
	p_3 : 请将下面的一段话生成一句话的摘要:
	p_4 : 请将下面的一段话生成 70 字以内的摘要:
LCSTS	p_1 : 请用一句话来总结下面的文字:
	p_2 : 请将下面一段话概括为一句话:
	p_3 : 请为下面的一段话生成一句话的摘要:
	p_4 : 请为下面的一段话生成 20 字以内的摘要:
	p_5 : 请为下面的一段话生成一个简短标题:

(2) 对于每个大模型,使用可以达到最大分数的最优提示词 p_{best} , 即

$$p_{\text{best}} = \text{Argmax}(\text{Score}(p)) \quad (2)$$

(3) 最终通过最优提示词 p_{best} 生成完整数据集的摘要结果 c_{best} , 并使用每个大模型对应的 c_{best} 评估大模型生成摘要能力。

2.2.2 少样本摘要

除了上述的零样本提示词, 实验还利用了大模型的上下文学习(In-Context Learning)能力设计了一些少样本(Few-Shot)的提示词。大模型直接生成

对应文章的摘要可能并不能完美地达到预期目标, 于是需要通过一些“示例”样本数据让大模型在上下文中进行理解。具体的做法是: 先向大模型说明任务——“自动文本摘要”——通过“文章”来生成“摘要”, 并提示它将会提供一些“文章”与“摘要”对。图 3 是使用 Streamlit^① 工具网页搭建了 ChatGLM2 模型, 使用对话的方式与大模型进行交互, 并通过展示“文章”与“标题”样本对让大模型学习, 生成最终的摘要结果。

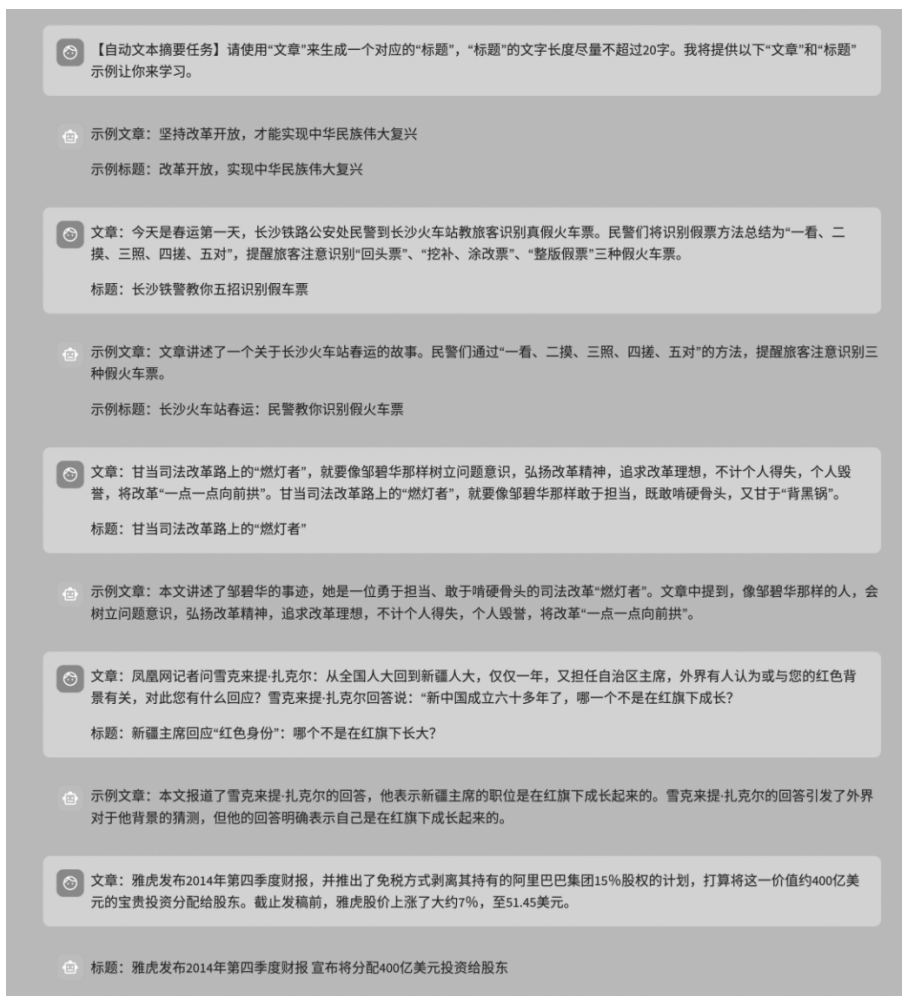


图 3 用 Streamlit 工具搭建基于提示词的少样本摘要方法(ChatGLM2 模型)

在经过与大模型多轮对话后, 大模型通常能够学习这些样本, 并且初步掌握对这些语料应当采用何种“缩放”的能力进行凝练。

当每个大模型通过少样本方法生成对应的摘要后, 可供探索和分析的方面将包括与大模型对话的推理时间、摘要长度、评估得分 3 个维度, 可以更加具体、全面地展现出大模型的综合能力。

2.3 数据集

本文采用的中文文本摘要数据集主要包括

NLPCC2017 Task3 数据集和 LCSTS 数据集, 它们都是近年来在该技术领域中被广泛使用的主流数据集, 实验结果和数据相对可靠。

2.3.1 NLPCC2017 数据集

NLPCC2017 数据集, 是 NLPCC (Natural Language Processing and Chinese Computing) 在 2017 年的 Task3^② 任务中所发布的数据集, 其文本

① Streamlit: <https://pypi.org/project/streamlit/>

② NLPCC2017: <http://tcci.ccf.org.cn/conference/2017/taskdata.php>

内容和参考摘要均来源于今日头条的新闻。该数据集囊括了 50 000 篇新闻文章,文本长度最短接近 100 字符,最长可达 2 万多字,而与之对应的参考摘要通常是总结为 60 个字符以内的句子。NLPCC 的文章相对更加丰富,在一些报道中,具备了详细的事件经过和细节。

2.3.2 LCSTS 数据集^[35]

LCSTS 数据集由哈尔滨工业大学整理,构建的是基于新闻媒体在微博平台发布的新闻摘要。该数据集由超过 200 万条新闻数据组成,每篇均为短新闻,大约包含 100 个字符,而对应的摘要约 20 个字符。相较于 NLPCC2017 数据集,该数据集的语料特征更接近于新闻快讯与新闻短标题,对事件概括和提炼方面的程度更为深入。

由于该摘要数据集中含有摘要数据质量的标签(质量标签为 0 至 5,标签值越大,则质量越高),本文将数据集进行了简要筛选,提取了质量标签大于或等于 3 的摘要数据共计 9 410 条,作为本次的实验数据集。

总之,上述两个新闻摘要数据集涵盖了政治、体育、经济、民生等多个类型的新闻内容,参考摘要更像新闻标题,通常会有语言精炼、凸显结果、增加吸引力的特点。受模型上下文窗口长度限制,本文仅从 NLPCC2017 摘要数据集提取出原数据集中原文长度在 4 000 字以下的部分文章,共 8 596 条数据进行评估。此外,为了评估 ChatGLM2-6B 处理长文本的性能,我们还将原文长度大于 10 000 字样本提取出来,共计 56 条数据,作为长文本评估实验数据集。数据集的统计情况如表 4 所示。

表 4 数据集分析情况

	NLPCC2017(4 千字以下)	NLPCC2017(1 万字以上)	LCSTS
平均字数(摘要)	45	48	18
最大/最小字数(摘要)	84/19	61/22	24/8
平均字数(原文)	1678	13 540	108
最大/最小字数(原文)	3997/82	22 313/10 032	152/76
数据集大小	8596	56	9410

2.4 环境及参数

本文基于 Python 环境,使用 Transformers 库加载 ChatGLM2-7B 模型,使用 ModelScope 库加载 Baichuan2-7B 和 Internlm-7B 模型,实验使用两块 16 GB 显存的 NVIDIA Tesla P100 GPU,系统使用 Ubuntu 20.04 LTS。在大模型生成的参数方面,模型默认的设置可以让模型生成更加丰富的摘要结果,但也会导致模型出现的错误概率增加。本次实验采用稍大的采样系数(Top-p,设置为 0.8)和较小的温度系数(Temperature,设置为 0.1),有助于增强摘要结果的稳定性,以达到较为准确的结果。

2.5 评估方法

自动文本摘要的评价主要采用 ROUGE^[22] 得分进行量化,该指标依据参考摘要和生成摘要之间的 n -gram 重叠度评估。常用的 ROUGE 算法包括了 ROUGE-N ($N = 1, 2$, 表示 n -gram 的长度、ROUGE-L(最长公共子序列)等。

ROUGE-N 的计算方法如式(3)所示。

$$R_{\text{ROUGE-N}} = \frac{\sum_{S \in \{\text{Ref}\}} \sum_{N_{n\text{-gram}} \in S} \text{Count}_{\text{match}}(N_{n\text{-gram}})}{\sum_{S \in \{\text{Ref}\}} \sum_{N_{n\text{-gram}} \in S} \text{Count}(N_{n\text{-gram}})} \quad (3)$$

其中, n -gram 表示了 n 元词, $\{\text{Ref}\}$ 表示了参考摘要, $\text{Count}_{\text{match}}$ 表示了摘要结果与参考摘要中同时出现的个数。

ROUGE-L 的计算方法如式(4)~式(6)所示,其中, A 是候选摘要,长度为 m , B 是参考摘要,长度为 n 。

$$P = \frac{\text{LCS}(A, B)}{m} \quad (4)$$

$$R = \frac{\text{LCS}(A, B)}{n} \quad (5)$$

$$F = \frac{(1 + b^2)RP}{R + b^2P} \quad (6)$$

这种基于字符串匹配的方法需要摘要结果与参考摘要实现尽可能的重叠,才能保证有较高的得分。而对于中文来说,许多单词表达了同样的意思但在机器生成摘要时可能用成了不同于参考摘要的词语,又或者是参考摘要字词凝练程度更高,这些都会对 ROUGE 评分产生影响。所以,实验中使用了 Chinese-Rouge 中文评估软件包^①,它在计算时先对文本进行预处理,切割为以“词”为粒度的基本单元,再使用上述公式计算最终得分。

FactCC^② 是 Kryscinski 等人^[23] 提出的基于事

① Chinese-Rouge: <https://pypi.org/project/rouge-chinese/>

② FactCC: <https://github.com/salesforce/factCC>

实一致性的自动评估指标。该方法对 CNN/Daily Mail 数据集进行了改写,创建了事实一致和不一致的二分类数据集,使用 BERT 模型对数据集进行了微调训练。测试时,可使用该模型将原文档和摘要结果作为输入,得到摘要是否忠实于原文档的标签。由于摘要结果大多为中文内容,本次将 FactCC 所使用的模型改为 Google 在 2019 年发布的中文预训练语言模型 bert-base-chinese^①。

此外,本次评估还记录了摘要生成的时间、摘要结果的长度等作为参考。

3 结果与讨论

本节记录了对第 2 节中所介绍的大模型、数据集和评估方法进行文本摘要任务的评估结果,并按照零样本摘要与少样本摘要的不同结果分开进行讨论分析。

3.1 零样本摘要

3.1.1 实验结果

根据表 3 所示的两个数据集样本分布情况,本

文采用了如表 4 所示的提示词。NLPCC 的摘要多为一句话的总结,以 50 字以内居多;LCSTS 的摘要平均字数仅有 18,更像一个简单的标题。实验通过设置提示词 p_4 以限定摘要字数,对于 LCSTS 数据集,增加了提示词 p_5 : 生成一个简短的标题。

本文先对每个数据集的前 100 份样本生成候选摘要。候选摘要的个数为 5,即生成 5 篇摘要后计算摘要的平均得分,重复性实验的目的在于提高实验结果的稳定性,并且确定相较于每个大模型更为合适的提示词。

对上述的多个提示词所生成候选摘要的平均分记录如表 5 所示。依照试验结果,在 NLPCC(4K 字以下)数据集中提示词 p_4 均取得了最高的 ROUGE 得分。而在 LCSTS 数据集中,每个模型所选择的提示词并不完全相同。按照第 2 节所提到的方法,生成完整数据集的数据如表 6 所示,并整理了相关数据,在 6 个维度上进行了归一化打分,形成图 4 所示的摘要评分雷达图,以供更好地表现各模型在生成摘要方面的质量及性能。

表 5 不同提示词摘要实验结果

数据集	模型	提示词	ROUGE-1	ROUGE-2	ROUGE-L	平均长度	时间消耗
NLPCC2017 (4 000 字以下)	ChatGLM2-6B	p_1	42.5	16.4	34.9	154	00 : 13 : 12
		p_2	43.2	16.2	35.2	159	00 : 13 : 39
		p_3	44.7	16.8	36.6	179	00 : 14 : 56
		p_4	47	18.1	38.8	192	00 : 15 : 45
	Baichuan2-7B	p_1	36.5	14.2	30.4	86	00 : 14 : 42
		p_2	33.8	13.1	27.9	68	00 : 12 : 51
		p_3	36	13.5	30	77	00 : 13 : 50
		p_4	42.2	16.3	34.7	120	00 : 19 : 11
	InternLM-7B	p_1	33.2	12.2	28.1	50	00 : 11 : 59
		p_2	34.3	13.4	29.3	52	00 : 12 : 20
		p_3	32.3	11.9	27.4	37	00 : 09 : 24
		p_4	34.3	13.4	29.3	41	00 : 09 : 52
NLPCC2017 (1 万字以上)	ChatGLM2-6B	p_1	32.5	10.5	28.6	1 362	01 : 30 : 47
		p_2	30.1	8.9	26	530	02 : 13 : 11
		p_3	30.3	8.5	25.5	992	03 : 05 : 45
		p_4	32.2	10	27.2	471	01 : 23 : 12

^① bert-base-chinese: <https://www.modelscope.cn/models/tiansz/bert-base-chinese>

续表

数据集	模型	提示词	ROUGE-1	ROUGE-2	ROUGE-L	平均长度	时间消耗
LCSTS	ChatGLM2-6B	p_1	46.5	16.2	41.6	72	00 : 03 : 53
		p_2	47.7	17.8	43.1	77	00 : 04 : 07
		p_3	51.3	18	46	88	00 : 04 : 39
		p_4	49	18.1	43.3	73	00 : 03 : 59
		p_5	35.2	11.7	30.8	32	00 : 01 : 58
	Baichuan2-7B	p_1	39.8	11.1	35.5	55	00 : 03 : 36
		p_2	36.8	11.1	32.4	47	00 : 03 : 07
LCSTS	Baichuan2-7B	p_3	39.6	11.3	34.7	48	00 : 03 : 11
		p_4	38.7	11.7	35	39	00 : 02 : 45
		p_5	31.8	9.1	26.9	25	00 : 01 : 50
	InternLM-7B	p_1	32	11.5	29.4	19	00 : 01 : 50
		p_2	33.7	12.9	30.8	29	00 : 02 : 33
		p_3	33.5	11.1	29.8	17	00 : 01 : 42
		p_4	33.2	11.4	29.8	21	00 : 02 : 01
		p_5	32.2	10.7	28.9	25	00 : 02 : 17

表 6 零样本摘要实验结果

	NLPCC2017(4 000 字以下)			LCSTS		
	ChatGLM2	Baichuan2	InternLM	ChatGLM2	Baichuan2	InternLM
ROUGE-1	46.2	40.5	35.2	42.1	36.5	30
ROUGE-2	18	14.7	14.6	14.8	12	11.7
ROUGE-L	37.4	33.2	30.4	37.1	31.9	27.3
FactCC	44.14	63.48	68.38	30.09	65.32	51.08
摘要平均长度	200	135	48	86	55	21
时间消耗	24 : 37 : 52	34 : 24 : 38	17 : 55 : 40	7 : 4 : 35	5 : 24 : 24	3 : 3 : 58

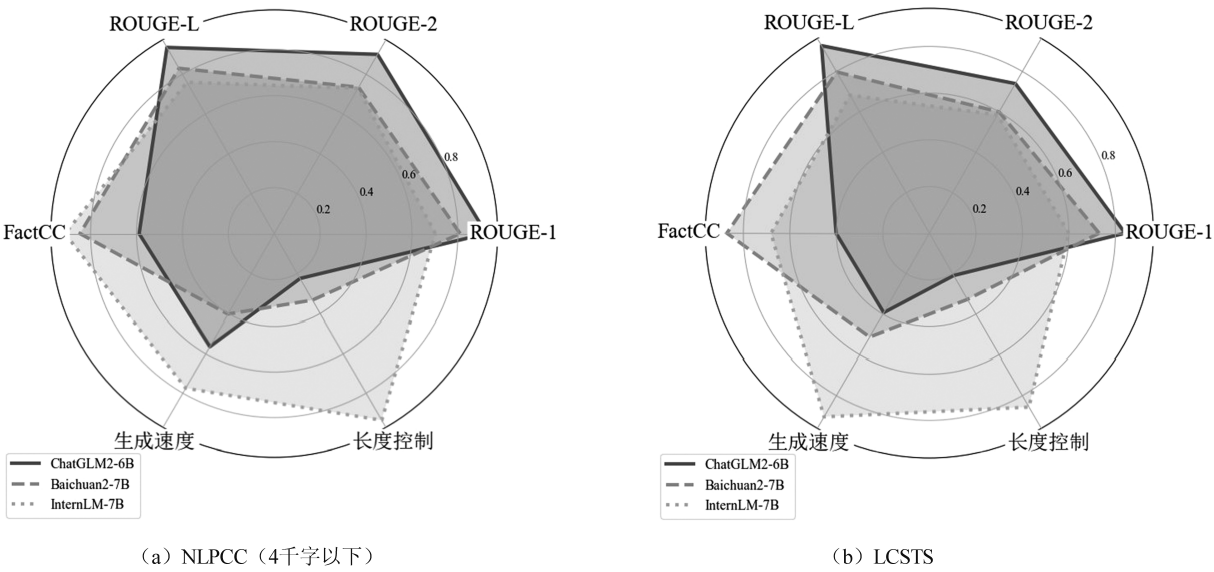


图 4 零样本摘要表现情况

3.1.2 结果分析

零样本摘要主要体现的是模型的泛化能力,是在模型未接触到测试集的完整数据样例、语言文字风格时,按照模型在预训练阶段所学习到的先验知识来对文本摘要进行预测。通过零样本摘要实验,可以展现出模型本身所具备的特点,观察实验结果及图 5,所见如下:

(1) ChatGLM2 模型生成的摘要在两个数据集上均获得了最好的 ROUGE 分数(表 6),并且推理速率($\approx$ 时间消耗/摘要长度)也占据优势。然而,从摘要长度上看,与理想的参考摘要相差甚远,分别相较两个数据集参考摘要的平均字数(表 3)超出了 344%和 378%,频繁冗余成为了 ChatGLM2 模型最大的问题;其次,事实性的评估指标明显低于其他两种模型,原因可能是摘要中的赘述过多而导致错误概率增加。

虽然受益于较长的上下文窗口,使得模型可以处理更长的文本,但在 NLPCC2017(1 万字以上)的长文本实验中,ChatGLM2 遇到了问题,容易陷入重复循环(Repetition Loops)。不仅使得生成的文本包含了大量重复的句子,使得摘要冗长且可读性差,还大大影响生成时间,造成无效等待。

(2) Baichuan2 模型所生成的摘要也十分具有参考价值,它所生成的摘要在得分上也非常具有竞争力,尤其是 FactCC 指标非常稳定。但其摘要长度,也超出参考摘要 200%左右。此外,实验发现,当 Baichuan2 模型遇到稍长文本时(NLPCC2017),推理速度明显下降,生成摘要的字数虽然少于 ChatGLM2,但时间消耗甚至明显超过后者。

(3) InternLM 模型生成的摘要长度十分近似于参考摘要的理想长度,且在 NLPCC 数据集上的事实性评估指标较为优秀,但 ROUGE 得分略逊色于前两种模型。从长度来看,该模型生成的摘要确实做到了言简意赅、短小精悍,但也缺少了许多关键信息导致 ROUGE 得分最低,甚至存在偏离原文本意的可能性。

实际上,使用 ROUGE 评分对大模型摘要进行评价时,可以认为 ROUGE 评分与摘要长度存在一定联系。因为在不限摘要长度的情况下,大模型将会尽可能多的捕捉到原文中的关键信息,进而可以与参考摘要有更多的 N 元词相匹配,使得 ROUGE 评分增加。但 FactCC 指标所得到的结果差异显著,这一现象可能是由于采用了较小的模型(BERT)来衡量大模型的结果,而前者的语义理解

能力相较于后者更弱。

对此,为了详细验证模型生成的摘要质量,我们再次进行了人工比对,所见如下:

(1) ChatGLM2 模型所生成的摘要偏向于将新闻事件叙述完整,有条有理,包含逻辑。摘要结果通常会描述新闻事件的整个过程,列举新闻中的一些分散信息。特别是原文中出现的数字信息,如时间信息、数量信息、人物的年龄信息、比赛结果比分信息等。如,在 NBA 体育新闻中,参考摘要通常会提供比赛结果的得分情况,以供读者较为直观地了解比赛信息,而 ChatGLM2 会将原文中的一些比赛细节也囊括其中,导致信息冗余。

(2) Baichuan2 模型所生成的摘要叙述也同样较为完整,但叙述风格与 ChatGLM2 有所差别。前者倾向于使用多个较短的句子总结原文中的不同信息点,在一定程度上控制了摘要字数。此外, Baichuan2 模型对 NLPCC2017 数据集(4 000 字以下)有 7 条数据未能生成对应摘要。原因主要为以下 3 点:一是模型认为原文中的信息尚存在缺失,如地点信息不完整;二是模型未能识别任务要求,可能是由于原文长度达到窗口临界值,模型已经遗忘提示词内容;三是原文中的信息与模型预训练时获得的历史知识矛盾,模型在纠正原文中的“错误”而并未生成摘要。

(3) InternLM 模型的摘要结果与数据中显示的相同,言简意赅,短小精炼。在两个数据集中,都能够使用非常恰当的字数对文本进行压缩,达到流畅可读、语言精妙、吸引读者的效果,并且保留了与参考摘要基本相近的含义,这一点是其他两个大模型未能达到的。但在摘要结果中发现,该模型也容易陷入重复,在 NLPCC2017 数据集(4 000 字以下)中出现了 27 条,而在 LCSTS 数据集中也有 4 条。甚至还有对原文内容进行了完整复述,篇幅明显超出平均范围。这可能是由于实验中设置的温度系数较低,且没有设置重复惩罚(Repetition Penalty)。

此外,在摘要结果中也存在事实错误,表 7 是 3 个大模型对一则“中国女排惜败意大利”的新闻进行的总结。其中, Baichuan2 和 InternLM 所生成的摘要都存在与事实不符的问题,都将比分信息总结错误,而 ChatGLM2 中并未总结出“2-3”的最终战果。出现这种结果通常是由于原文内部信息的复杂交互而造成模型在理解上出现偏差。

表 7 摘要事实不一致示例

部分原文	中国女排在大比分 2-1 领先的情况下没能顶住对手的反扑，鏖战五局 2-3 惜败仅得 1 分，五局比分为 15-25、25-22、25-22、18-25 和 12-15，力夺 2 分的意大利斩获大奖赛总决赛首胜。
参考摘要	中国女排五局鏖战 2-3 惜败意大利，遭大奖赛总决赛第二败
ChatGLM2	……中国女排派出较强阵容，但最终在五局 15-25、25-22、25-22、18-25 和 12-15 中输掉比赛，仅得 1 分。……
Baichuan2	2015 年世界女排大奖赛总决赛，中国女排在决赛中 3-2 惜败意大利，五局比分为 15-25、25-22、25-22、18-25 和 12-15。……
InternLM	女排大奖赛：中国女排 3-2 险胜意大利，获 1 分，五局比分为 15-25、25-22、25-22、18-25 和 12-15。

3.2 少样本摘要

3.2.1 实验结果

实验按照 2.2.2 节中以及图 3 所示的方法向大模型提供“示例文章/示例标题”样本。样本按照数量设置了 3-shot 和 5-shot，按照选择方法也设置了两，一种是固定样本(Fixed Shot)组，即在每次测试时，都使用相同的摘要样本；另外一种随机样本(Random Shot)组，即在每次提供示例样本时，使用随机样本对。测试考虑到大模型在长文本多轮对话的时间消耗和显卡算力资源的消耗相对增加，且上下文长度也更容易受到模型的限制等因素，我们仅对 LCSTS 数据集前 100 条数据进行了少样本摘要的测试，每项测试进行 5 次重复实验，然后取平均值。所选示例样本均采用非测试集数据。实验结果如表 8 所示。

表 8 少样本摘要实验结果

LCSTS(前 100 条数据)							
模型	分类	ROUGE-1	ROUGE-2	ROUGE-L	FactCC	摘要长度	时间消耗
ChatGLM2-6B	3-fixed shot	34.4	12.8	30.6	72.52	23	00 : 05 : 14
	3-random shot	30.9	10.3	27.2	70.63	23	00 : 05 : 09
	5-fixed shot	33.6	12.1	30.3	67.44	24	00 : 06 : 08
	5-random shot	31	10.9	27.5	47.87	26	00 : 06 : 00
Baichuan2-7B	3-fixed shot	32.6	10.2	29.3	68.04	30	00 : 15 : 34
	3-random shot	33.5	10.5	29.5	69.15	29	00 : 18 : 04
	5-fixed shot	33.7	11.2	30.2	69.15	27	00 : 26 : 38
	5-random shot	31.7	9.6	28.1	67.01	28	00 : 31 : 18
InternLM-7B	3-fixed shot	33.4	12.5	30.3	72.61	27	00 : 13 : 40
	3-random shot	33.8	12.7	30.7	68.74	18	00 : 13 : 29
	5-fixed shot	35.3	13.5	32.1	73.20	24	00 : 17 : 48
	5-random shot	33.8	13	30.1	68.93	20	00 : 17 : 35

3.2.2 结果分析

少样本摘要通过与 LLM 进行交互的方式向 LLM“投喂”数据，LLM 从上下文中理解“输入/输出”的样式，以提升模型的预测能力。

综合表 8 中的结果可以发现，通过向大模型展示少量的样本，的确能够让模型按照样本中的“输出”长度来控制结果。即使仅仅是 6B、7B 大小这样规模基础的大模型，其所具备的上下文学习能力依然可以让大模型初步掌握示例中的“输入/输出”格式，并据此进行泛化，使其具备生成“标题”的能力。

从评分上看，3 种模型所得到的摘要质量基本相当，事实性的评估指标也没有产生较大差距。由于 ChatGLM2-6B、Baichuan2-7B 生成的摘要大大缩减了摘要长度，导致其摘要结果可能没能抓住原文的重点，质量有不同程度的下降。最终，InternLM-7B 模型生成的摘要得到了最高的评分。

从时间上看，经过多轮对话后，摘要的生成也将占用更多的时间。而此次实验展现出 ChatGLM2-6B 在多轮对话中的处理速度上有较好的表现。

从样本选择方法上看，实验设计的固定样本和

随机样本在两个不同方法上所生成的摘要质量整体上相当,前者质量略高出一些,后者样本质量相对不稳定。但二者之间的摘要质量没有太大的区别,可以推测:大模型通过上下文可以学习到与样本数据相似的总结方式,生成与样本数据相似的摘要。两种方法对摘要数据的影响可能还需要后续更多的对比实验才能展现其差别。

经过人工对比摘要结果后发现,ChatGLM2-6B 与 Baichuan2-7B 的摘要结果明显对语句进行了凝练。在部分结果中,不同模型生成的摘要各具特色。ChatGLM2-6B 模仿 LCSTS 参考摘要中的语言风格将新闻总结成形如“主语:事件”格式的标题;Baichuan2-7B 为总结的标题加入了双引号来进行凸显;而 InternLM-7B 的结果与零样本的结果相比更加稳定,长短更加均衡,没有生成重复的内容。

此外,我们在 NLPCC 数据集上使用 ChatGLM2 模型做了一些少样本摘要的尝试,但模型并不能很好地控制长度。

4 总结与展望

4.1 总结

本文详细介绍了中文大模型与自动文本摘要任务在 NLP 领域中的进展,并展开对 ChatGLM2-6B、Baichuan2-7B、InternLM-7B 模型的介绍,以及使用它们生成摘要的对比实验。实验设置了多个维度对大模型进行评价:一是通过 ROUGE 和 FactCC 分数评估大模型生成摘要的质量;二是结合摘要长度这一指标评价生成摘要的冗余度;三是通过摘要生成时间衡量大模型的推理速度。

实验结果大致总结为以下 3 点:一是在零样本摘要中,大模型将会尽可能多地捕捉到原文中的关键信息,但随之而来的是冗余度的上升;少样本的方法可以使大模型输出更加简洁的摘要,但也容易丢失原文重点,导致得分下降严重。二是大模型对于提示词中所限制的摘要长度并不敏感,多数情况下会超出限制,需要通过多轮对话、提供少样本的方式加以控制。三是多轮对话占用的时间明显增加,ChatGLM2-6B 的处理速度相对较快,另外在单轮对话中,同一模型摘要生成长度基本与生成时间正相关。

4.2 面临的挑战

经过此次对大模型生成摘要的详细实验,并对大模型生成摘要结果进行人工分析后,我们认为目前大模型在摘要任务上面临的主要挑战包括以下:

(1) **指令理解** 在通过使用带有“一句话”或字数限制的提示词来使大模型生成摘要时,ChatGLM2-6B 和 Baichuan2-7B 并不能理解用户的指令,未能生成简短的摘要,而是按照模型在预训练期间所学习的先验知识、语言风格对文本进行总结。如果使用少样本的方式与大模型进行交互,可能会影响效率。

(2) **幻觉问题** 大模型在生成摘要时,有时会出现无中生有、偏离原文内容的现象。Baichuan2 模型对“史上署名人数最多论文出炉:5 154 名科学家合著论文测量希格斯玻色子质量”的一则新闻进行了总结。原文中并没有提及具体的撰写专家人数,Baichuan2 模型却给出了“作者数量达 264 人”的错误信息。

(3) **事实一致性** 本文在表 7 中展示了大模型在生成摘要时出现的事实错误个例。如果对原文中的信息进行复制(如 ChatGLM2),能够避免这种事实性错误,但其简洁性将下降。然而在 FactCC 的自动评估中,ChatGLM2 的得分反而最低,也有待深入研究。

(4) **长文本摘要** 本文的实验所采用的大模型都在一定程度上受制于上下文窗口,因此在 NLPCC2017 数据集上并没有进行完整的测试,ChatGLM2-6B 的得分情况以及生成的结果也不够理想。今后,随着大模型不断提升基础能力的同时,长文本摘要问题将不再受到此限制,但也面临其他一些挑战,例如,优化内存与计算效率,避免摘要的冗余、跑题,按照用户需求调整关注重点等。

4.3 展望

目前大语言模型已经在 NLP 下游任务中取得了显著优势,随着基础模型的迭代与发展,以及在更丰富的数据集上进行训练,模型在文本摘要上会拥有更优越的表现。除了提升模型的表现能力,在摘要任务上的研究重点还应当包括以下几方面:

(1) **强调互动** 本次的实验表明, 现有大模型未能按照提示词的长度限制进行总结。而未来的大模型应该更加注重与用户的交互, 允许用户在生成过程中提供实时反馈, 从特定角度捕捉关键信息, 或是不断调整摘要的内容以及关注重点, 实现更加定制化的文本摘要输出。

此外, 明明语义相近的提示词却导致生成的摘要得分存在差距, 对于不太了解最优提示词的用户, 模型如何清晰理解用户的意图以生成质量最好的摘要, 同样也值得探索。

(2) **评估方法** 现有对大模型摘要能力的评估多数都是基于 ROUGE 评分, 并且多数研究表明, 在 ROUGE 评分未能超过微调预训练模型的情况下, 大模型所生成的摘要还是获得了人类评估的一致好评。因此, 更全面和客观的评估指标也值得研究, 以更准确地衡量摘要的质量, 包括内容准确性、连贯性和创新性。

本文使用的事实一致性评估指标 FactCC 基于 BERT 模型训练, 其语义理解能力并不如实验模型。从结果上来看, 对 ChatGLM2 模型的评估可能会存在偏差。所以, 针对大模型所生成内容的事实进行一致性评估, 是应当选取更大的模型来进行更准确的评估, 还是选取采用算力资源略小的模型提升计算效率, 尚需要实验进行论证。

(3) **数据集** 本次评估所使用的中文新闻数据集的公布时间为 2015 年和 2017 年, 而今后大模型所具备的现实世界的知识也将会逐步更新, 在时间较旧的数据集上进行总结, 时间或事实不一致的概率也会大大增加。此外, 用户对摘要的需求可能并非仅仅是通过新闻生成“标题”, 可能还包括对其他类型文本的总结, 故构建符合用户需求的高质量中文摘要数据集也同样具有研究价值。

参考文献

- [1] QI S, LI L, LI Y, et al. SAPGraph: Structure-aware extractive summarization for scientific papers with heterogeneous graph [C]//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022: 575-586.
- [2] NIE P, ZHANG J, LI J J, et al. Impact of evaluation methodologies on code summarization[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland: Association for Computational Linguistics, 2022: 4936-4960.
- [3] OUYANG S, CHEN J, HAN J, et al. Compositional data augmentation for abstractive conversation summarization [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 1471-1488.
- [4] LIANG Y, MENG F, XU J, et al. Summary-oriented vision modeling for multimodal abstractive summarization[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 2934-2951.
- [5] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint arXiv: 2303.18223, 2023.
- [6] TAM D, MASCARENHAS A, ZHANG S, et al. Evaluating the factual consistency of large language models through news summarization[C]//Proceedings of the Association for Computational Linguistics, 2023: 5220-5255.
- [7] 车万翔, 窦志成, 冯岩松等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学: 信息科学, 2023, 53(09): 1645-1687.
- [8] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.
- [9] YANG J, JIN H, TANG R, et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and Beyond [J]. ACM Transactions on Knowledge Distillation from Data, 2024, 18(6): 1-32.
- [10] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [11] TAORI R, GULRAJANI I, ZHANG T, et al. Stanford alpaca: An instruction-following llama model [CP/OL]. https://github.com/tatsu-lab/stanford_alpaca[2024-01-06].
- [12] ZENG A, LIU X, DU Z, et al. GLM-130b: An open bilingual pre-trained model[C]//Proceedings of the 11th International Conference on Learning Representations, 2023.
- [13] LIU J, GONG R, WEI X, et al. QLLM: Accurate and

- efficient low-bitwidth quantization for large language models [C]//Proceedings of the 12th International Conference on Learning Representations, 2024.
- [14] GU Y, DONG L, WEI F, et al. MiniLLM: Knowledge distillation of large language models [C]//Proceedings of ICC, 2024.
- [15] VALICENTI T, VIDAL J, PATNAIK R. Mini-GPTs: Efficient large language models through contextual pruning [J]. arXiv preprint arXiv:2312.12682, 2023.
- [16] LEWIS M, LIU Y, 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, 2020: 7871-7880.
- [17] ZHANG J, ZHAO Y, SALEH M, et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization [C]//Proceedings of the 37th International Conference on Machine Learning, 2020: 11328-11339.
- [18] HE P, PENG B, LU L, et al. Z-Code++: A pre-trained language model optimized for abstractive summarization [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 5095-5112.
- [19] LIU Y, LIU P. SimCLS: A simple framework for contrastive learning of abstractive summarization [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1065-1072.
- [20] RAVAUT M, JOTY S, CHEN N F. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland: Association for Computational Linguistics, 2022: 4504-4524.
- [21] LIU Y, LIU P, RADEV D, et al. BRIO: Bringing order to abstractive summarization [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland: Association for Computational Linguistics, 2022: 2890-2903.
- [22] LIN C Y. ROUGE: A Package for automatic evaluation of summaries [C]//Proceedings of the Workshop on Text Summarization Branches Out, 2004: 74-81.
- [23] CHEN Y C, BANSAL M. Fast abstractive summarization with reinforce-selected sentence rewriting [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia: Association for Computational Linguistics, 2018: 675-686.
- [24] WANG Y, ZHANG Z, WANG R. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 8640-8665.
- [25] ZHANG H, LIU X, ZHANG J. Extractive summarization via chatgpt for faithful summary generation [C]//Proceedings of the Association for Computational Linguistics, Singapore, 2023: 3270-3278.
- [26] KRYSCINSKI W, MCCANN B, et al. Evaluating the factual consistency of abstractive text summarization [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020: 9332-9346.
- [27] LIU Y, FABBRI A R, LIU P, et al. On learning to summarize with large language models as references [C]//Proceedings of the NAACL, 2024: 8647-8664.
- [28] GOYAL T, LI J J, DURRETT G. News summarization and evaluation in the era of GPT-3 [J]. arXiv preprint arXiv:2209.12356, 2022.
- [29] ZHANG T, LAHDAK F, DURMUS E, et al. Benchmarking large language models for news summarization [J]. arXiv preprint arXiv:2301.13848, 2023.
- [30] PU X, GAO M, WAN X. Summarization is (Almost) dead [J]. arXiv preprint arXiv:2309.09558, 2023.
- [31] FABBRI A R, LI I, SHE T, et al. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 1074-1084.
- [32] WANG J, LIANG Y, MENG F, et al. Zero-Shot cross-lingual summarization via large language models [C]//Proceedings of the 4th New Frontiers in Summarization Workshop, 2023: 12-23.
- [33] WANG J, MENG F, LU Z, et al. ClidSum: A benchmark dataset for cross-lingual dialogue summarization [J]. arXiv preprint arXiv:2202.05599, 2022.
- [34] 武俊宏, 赵阳, 宗成庆. ChatGPT 能力分析与未来展望 [J]. 中国科学基金, 2023, 37(05): 735-742.

- [35] HU B, CHEN Q, ZHU F. LCSTS: A large scale Chinese short text summarization dataset [C]//

Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1967-1972.



王俊超(1988—), 博士, 副教授, 主要研究领域为先进计算、网络空间安全。
E-mail: wangjunchao11@126.com



樊可汗(1994—), 通信作者, 硕士研究生, 主要研究领域为自然语言处理。
E-mail: khan@mail.ustc.edu.cn



霍智恒(1996—), 硕士研究生, 主要研究领域为自然语言处理。
E-mail: huozhiheng@foxmail.com

欢迎订阅《中文信息学报》

《中文信息学报》(Journal of Chinese Information Processing)是全国一级学会—社团法人中国中文信息学会和中国科学院软件研究所联合主办的学术性刊物,创刊于1986年10月,现为月刊。

《中文信息学报》是我国计算机、计算技术类中文核心期刊。主要刊登中文信息处理基础理论与应用技术方面的高水平学术论文,内容涵盖计算语言学(包括语音与音位、词法、句法、语义、语用等各个层面上的计算),语言资源建设(包括计算词汇学、术语学、电子词典、语料库、知识本体等),机器翻译或机器辅助翻译,汉语和少数民族语言文字输入输出及其智能处理,中文语音识别及文语转换,信息检索,信息抽取与过滤,文本分类、中文搜索引擎,以自然语言为枢纽的多模态检索,与语言处理相关的数据挖掘、机器学习、知识获取、知识工程、人工智能研究,与语言计算相关的语言学研究等。也刊登相关综述、研究报告、成果简介、书刊评论、专题讨论、国内外学术动态等稿件。

读者对象主要为从事中文信息处理的研究人员、工程技术人员和大专院校师生等。

《中文信息学报》(国内统一刊号: CN11-2325/N; 国际统一刊号: ISSN 1003-0077)国内外公开发行人,国内定价每期50元,全年600元。

国内发行处:《中文信息学报》编辑部

国外发行处: 中国图书进出口总公司 100020 北京 88-E 信箱

1. 支付宝转账: (请注明期刊征订)

账号: cips_pay@163.com

姓名: 中国中文信息学会

2. 银行转账

开户银行: 工商银行北京市分行海淀西区支行

户名: 中国中文信息学会

账号: 0200004509014415619

《中文信息学报》编辑部

地址: 北京海淀区中关村南四街4号7号楼201房间

电话: 010-62562916

电子信箱: jcip@iscas.ac.cn

《中文信息学报》编辑部