计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):28−36 [doi: 10.15888/j.cnki.csa.009779] [CSTR: 32024.14.csa.009779]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于大语言模型的文本摘要质量评估①

谭琛瀚[1,2], 贾克斌[1,2], 王浩宇[1,2]

[1](北京工业大学 信息学部, 北京 100124)
[2](先进信息网络北京实验室, 北京 100876)
通信作者: 贾克斌, E-mail: kebinj@bjut.edu.cn

摘　要: 自动文本摘要是自然语言处理 (NLP) 领域中的一个重要分支, 其主要难点之一是在于如何快速、客观且准确地评估生成摘要的质量. 针对现有文本摘要质量评估方法中评估准确度不高、需要参考文本以及计算资源消耗大的问题, 本文提出一种基于大语言模型的文本摘要质量评估方法, 设计基于思维链原理的提示词构建方法以提高大语言模型在文本摘要质量评估任务上的性能, 同时生成思维链数据集并以模型微调的方式对小型大语言模型进行训练, 显著降低了计算需求. 本文方法首先根据文本摘要的特点确定评估维度, 并基于思维链原理 (chain of thought, CoT) 构建提示词; 使用提示词对大型大语言模型进行引导, 使其根据摘要样本生成思维链过程与评估结果, 同时以此为基础生成思维链数据集; 使用生成的思维链数据集对小型大语言模型进行微调训练; 最后使用微调后的小型大语言模型完成文本摘要的质量评估任务. 本文在 Summeval 数据集上进行了对比实验与分析, 实验结果表明, 本评估方法显著提高了小型大语言模型在文本摘要质量评估任务上的评估准确度, 实现了一种无需参考文本、评估准确度高、计算需求低、便于部署的文本摘要质量评估方法.

关键词: 文本摘要; 质量评估; 大语言模型; 思维链; 微调训练

引用格式: 谭琛瀚,贾克斌,王浩宇.基于大语言模型的文本摘要质量评估.计算机系统应用,2025,34(2):28–36. http://www.c-s-a.org.cn/1003-3254/9779.html

Text Summarization Quality Evaluation Based on Large Language Model

TAN Chen-Han[1,2], JIA Ke-Bin[1,2], WANG Hao-Yu[1,2]

[1](Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)
[2](Beijing Laboratory of Advanced Information Network, Beijing 100876, China)

Abstract: Automatic text summarization is an important branch in the field of natural language processing (NLP), and one of its main difficulties lies in how to evaluate the quality of the generated summaries quickly, objectively, and accurately. Given the problems of low evaluation accuracy, the need for reference texts, and the large consumption of computing resources in the existing text summary quality evaluation methods, this study proposes an evaluation method for the quality of text summaries based on large language models. It designs a prompt construction method based on the principle of the chain of thought (CoT) to improve the performance of large language models in the evaluation of text summary quality. At the same time, a chain of thought data set is generated and a small large language model is trained in the way of model fine-tuning, significantly reducing the computing requirements. The proposed method first determines the evaluation dimension according to the characteristics of the text summary and constructs the prompt based on the principle of chain of thought. The prompt is utilized to guide the large language model to generate the chain of thought process and evaluation results based on the summary samples. Accordingly, a chain of thought data set is generated. The

① 基金项目: 北京市自然科学基金 (4212001)
收稿时间: 2024-07-15; 修改时间: 2024-08-13; 采用时间: 2024-09-19; csa 在线出版时间: 2024-12-19
CNKI 网络首发时间: 2024-12-20

generated chain of thought data set is used to fine-tune and train the small large language model. Finally, the study uses the fine-tuned small-scale large language model to complete the quality evaluation of the text summary. Comparative experiments and analyses on the Summeval dataset show that this evaluation method significantly improves the evaluation accuracy of the small-scale large language model in the task of text summary quality evaluation. The study provides a text summary quality evaluation method, which is a method with high evaluation accuracy, low computing requirements, and easy deployment without reference texts.

Key words: text summarization; quality evaluation; large language model (LLM); chain of thought (CoT); fine-tuning

随着人工智能的快速发展以及大量预训练语言模型的出现, 自然语言处理 (NLP) 得到了快速的发展, 而文本摘要生成则是其中一类重要的任务. 目前自动文本摘要的一大难点即是如何更加客观、准确地评估生成摘要的质量. 一个好的评估指标不仅能够高效地指导模型拟合数据分布, 还能够客观地评估文本摘要生成模型的性能, 从而进一步推动自动文本摘要的研究与发展.

目前的文本摘要评估方法可以分为主观评估和客观评估, 其中主观评估一般指人工评估. 客观评估则包含如 BLEU[1]、ROUGE[2]、greedy matching score[3]等基于词向量或词重叠的评估方法以及如 BERTScore[4]、BARTScore[5]等基于语言模型的评估方法, 这两类方法的基本原理多为以词为单位计算生成文本与参考文本间的相似度, 并以相似度为主要评估指标, 这也导致其难以对语法或篇章结构进行评估, 评估维度单一.

近几年来, 大语言模型不断发展, 其通过大量数据和学习训练, 展现出了对语言和文本的优秀理解能力, 其在各类任务上都表现出色. 而近年来也不断涌现出各类优秀的开源大语言模型, 如 LLaMA 2[6]、Qwen[7]、Baichuan2[8]等. 由此, 使用大语言模型来完成文本摘要评估任务成为一种可行且优秀的方案. 相较于传统方法, 使用大语言模型进行文本摘要质量评估具有无需参考文本、评估维度多样、语义理解能力强等优点, 但同时也存在硬件需求苛刻、计算成本高昂的问题, 而模型参数量较小的小型大语言模型虽然更加灵活且便于部署, 但其模型的评估准确度却有待提升.

基于此现状, 本文提出一种基于大语言模型的文本摘要质量评估方法. 基于思维链原理对提示词进行设计, 并引导大型大语言模型生成思维链过程与评估结果, 以此为基础生成思维链数据集并对小型大语言模型进行微调训练, 最终使用小型大语言模型来更加准确且高效地完成文本摘要评估任务.

本文的主要贡献如下.

(1) 提出一种基于思维链原理的提示词构建方法并应用于文本摘要质量评估任务. 先显式地要求模型对评估任务进行分解, 之后要求模型按其给出的步骤进行逐步推理, 最终形成评估依据与评估结果, 使模型自动生成高质量的思维链推理过程, 在丰富评估结果的同时提升了评估准确度.

(2) 生成思维链数据集并以模型微调的方式对小型大语言模型进行训练. 由大型大语言模型生成评估样本并制作为思维链数据集, 基于该思维链数据集对小型大语言模型进行微调训练, 在保证较高评估准确度的同时大幅降低硬件需求, 节省计算资源.

1 相关研究现状

1.1 文本摘要质量评估

文本摘要质量评估是一个重要的研究领域, 其需要对生成文本的质量进行评估, 确保其准确性和有效性. 人工完成这类评估任务虽然可以得到较为准确的评估结果, 但往往耗时且成本较高, 为了满足实际应用需求, 研究者们对自动化的评估方法展开了探索. 近年来人工智能的发展与各类语言模型的出现为文本摘要质量评估带来了新的可能性, 使用语言模型来完成文本摘要质量评估成为研究的热点. 研究者们以各类预训练语言模型为基础, 结合其具有丰富知识储备的特点来准确、便捷、高效地完成文本摘要质量评估任务.

1.2 预训练语言模型

Transformer[9]的诞生极大地促进了自然语言处理领域的研究与发展, 在短时间内诞生出了许多优秀的预训练语言模型. BERT[10]就是一种预训练的自然语言处理模型, 其以 Encoder-only 的 Transformer 结构和自注意力机制为核心来捕捉文本上下文之间的信息关联,

并通过大规模文本数据进行无监督的预训练使模型对文本语义进行学习, 使其可以应用于各类不同的下游任务. 与 BERT[10]不同, GPT[11]采用 Decoder-only 的结构来根据输入的上文对下文进行预测, 经过大量语料的训练, 其在各类生成式任务上表现出色. BART[12]结合了自回归和自编码模型, 建立在标准的 Transformer 结构上, 其相较于 BERT[10]更适合文本生成的场景, 相较于 GPT[11]能更好理解上下文语境信息.

1.3 基于预训练语言模型的文本质量评估

预训练语言模型的出现, 为文本质量评估任务提供了许多新的方案. BERTScore[4]使用 BERT[10]模型的文本嵌入来衡量两个文本间的相似性, MoverScore[13]则在此基础之上以 N-gram 为计量单位来得到更加鲁棒的结果. CTC[14]为文本生成任务定义了 3 种特定的度量维度, 并根据具体维度使用不同模型进行评估. BARTScore[5]采用预训练的 BART[12]模型将生成文本的评估转化为文本生成问题, 以模型生成文本的概率来对文本进行评估. UniEval[15]使用预训练的 T5 模型[16], 并将评估任务、源文本和目标文本转化为问题-答案形式的数据, 以问答的形式对文本进行评估. 近些年 GPT[11]系列的模型发展迅速, 其展现出了出色的文本生成和语义理解能力, GPTScore[17]就利用了 GPT-3 等生成式预训练模型, 利用模型丰富的知识蕴含以及强大的语义理解能力完成对生成文本的评估.

2 基于大语言模型的文本摘要质量评估方法

本文提出的基于大语言模型的文本摘要质量评估方法首先根据思维链原理对提示词进行设计, 随后由大型大语言模型根据提示词与具体评估维度生成思维链过程与评分, 随后将模型给出的评分与标准评分进行对比, 筛选出评分相近的样本来生成思维链数据集. 之后使用该数据集对小型大语言模型进行微调训练, 最后由小型大语言模型完成文本摘要质量评估任务, 得到最终的评估结果. 流程包含下几个部分: (1) 文本摘要质量评估维度; (2) 基于思维链原理的提示词构建方法; (3) 思维链数据集生成; (4) LoRA 微调. 整体流程如图 1 所示.

2.1 文本摘要质量评估维度

文本摘要指的是将长文本中包含的重要信息进行提取并使用简短的若干句话来完成概括, 是一类经典自然语言处理任务. 随着深度学习的发展和硬件算力的提升, 这些摘要生成方法的性能逐渐提高, 其生成摘要的质量也不断提升, 而人们也开始意识到对这些摘要的质量进行评估时, 仅依靠词重叠或语义重叠度已经很难区分哪个模型更加优秀. 于是要实现更加规范且准确的评估, 一套规范且细粒度的评估维度便是不可或缺的.

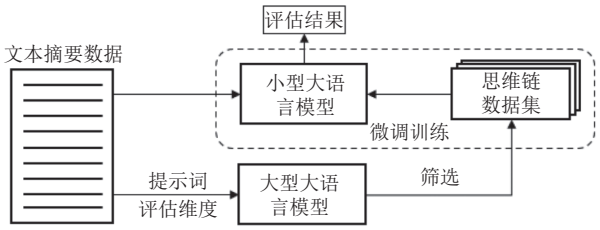

图 1 整体流程

Kryściński 等人[18]为文本摘要定义了 4 个重要的评估维度: 连贯性 (coherence)、一致性 (consistency)、流畅性 (fluency) 和相关性 (relevance). 这 4 个维度共同构成了文本摘要的综合评估标准, 其确保了生成的摘要不仅要在内容上与原文保持一致, 同时也应在表达和逻辑上清晰且准确. Summeval 数据集[19]即是在这 4 个评估维度的基础上建立的, 其后续也被应用于诸多文本摘要质量评估的相关工作中. 本文方法沿用此标准, 以连贯性、一致性、流畅性和相关性这 4 个维度作为文本摘要质量的评估维度.

表 1 给出了各评估维度及每个维度对应的评估标准, 其中的评估标准包含了对评估维度的进一步解释及具体打分的相关说明, 其也是提示词构建环节中的重要组成部分.

2.2 基于思维链原理的提示词构建方法

虽然大语言模型具有强大的上下文理解能力, 在各类下游任务上均表现出了卓越的性能, 但其也很难处理需要多个推理步骤的复杂任务, 于是构建合适的提示词来激发模型的推理能力便成为大模型处理复杂任务过程中的重要的一环. Wei 等人[20]提出了使用思维链 (chain of thought, CoT) 来帮助大语言模型进行推理和思考, 即使用 few-shot-CoT 的方式来人为地为模型提供逐步思考的示例, 引导模型通过逐步思考的方式来得到更加准确的答案. 与 few-shot-CoT 不同, zero-shot-CoT[21]省去了人为提供示例的过程, 其通过添加特殊提示词 (“let’s think step by step”) 来实现让模型先生成推理过程再生成答案的效果. 上述两种方法共同成

为思维链提示 (CoT prompting) 的基础, 后续也衍生出例如 self-consistency[22]、DiVeRSe[23]、auto-CoT[24]及 least-to-most prompting[25]等多种思维链提示的优化方式.

对大语言模型而言, 进行文本摘要质量评估需要完成包括文本语义理解、比对、判断、打分等一系列步骤, 因此构建合适的思维链提示词对模型在这一任务上性能的提升至关重要, 而由于人工为模型构建文本摘要质量评估的提示示例耗时长且工作量大, 所以本文提出一种基于思维链原理的提示词构建方法. 本方法通过提示词使模型根据任务描述完成对复杂任务的分解, 同时使模型给出完成任务所需要的步骤, 随后要求模型按其给出的步骤进行逐步推理并形成最终答案, 由此可以使模型自动生成高质量的思维链推理过程. 本文将此提示词构建方法应用于文本摘要质量评估任务, 具体过程如图 2 所示.

表 1 评估维度与评估标准

评估维度	评估标准
连贯性 coherence (1–5)	关注摘要中所有句子的整体质量, 其是否结构良好、组织良好, 是否是从句子到主题的连贯信息体. 评估时应对信息连贯、信息主题统一的高质量摘要打出较高分数; 对信息杂乱、信息相关性差的低质量摘要打出较低分数.
一致性 consistency (1–5)	关注摘要与源文本间的主题、事实的一致性. 评估时应对和源文本中的主题、事实高度一致的高质量摘要打出较高分数; 对和源文本中主题、事实有偏差或不准确的低质量摘要打出较低分数.
流畅性 fluency (1–5)	关注摘要在语法、拼写、标点、用词和句子结构方面的质量. 评估时应对流畅、优美的高质量摘要打出较高分数; 对含有语法或拼写错误, 阅读体验不通畅的低质量摘要打出较低分数.
相关性 relevance (1–5)	关注摘要是否主要包含源文本中的重要信息, 而非无用信息. 评估时应对包含了源文本中各个重要信息的高质量摘要打出较高分数; 对源文本中重要信息包含不全或包含较多冗余信息的低质量摘要打出较低分数.

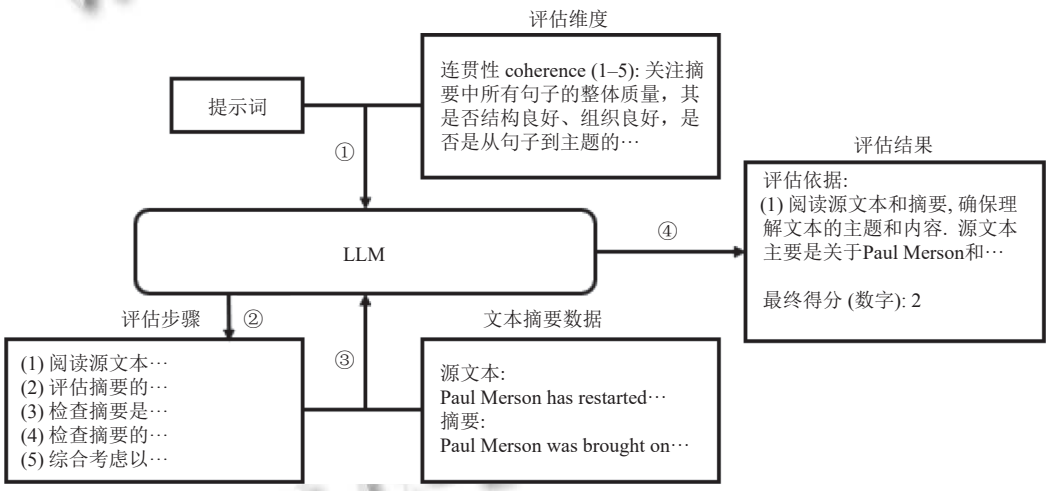

图 2 思维链推理过程

表 2 中给出了本文方法中所设计的提示词与对应模型回答的相关示例. 其中提示词 1 包含任务描述与任务分解引导, 提示词 2 包含具体任务文本与思维链推理引导. 首先由提示词 1 来引导模型根据评估维度给出具体评估步骤 (模型回答 1), 随后由提示词 1 和模型回答 1 组成对话历史, 并由提示词 2 来要求模型根据具体评估任务给出思维链推理过程以及最终得分 (模型回答 2), 由此来完成无参考的文本摘要质量评估任务.

表 2 中模型回答的示例由 GPT-4 生成, 可以看到在经过提示词的引导后, 模型可以做到通过思维链的方式逐步进行思考并给出最终答案, 实现了对复杂问题的分解拆分和连续推理的过程. 这种结合思维链原理的提示词构建方法可以有效提升模型在文本摘要评估任务上的评估准确度, 相关对比实验结果在本文第 3.4 节进行说明.

2.3 思维链数据集生成

使用思维链提示虽然可以让模型解决复杂任务的能力大幅提升, 但这种方法对于模型参数量有很高的要求, 通常需要拥有数百亿参数规模的大型大语言模型才能够正确完成复杂的推理任务. 由于这些模型计算要求和推理成本过于庞大, 导致其难以大规模部署和应用到实际任务中.

为解决这一问题, 本文利用大型大语言模型生成的思维链推理过程来生成思维链数据集, 并以此为基础对小型大语言模型进行微调训练, 使小型大语言模型得到解决复杂任务的能力.

表 2 提示词与模型回答示例

类型	内容示例
提示词1	作为一名文本生成质量评估专家, 你将收到由“源文本”和其“摘要”组成的一组摘要文本. 你的任务是在单一维度上对该“摘要”进行质量评估打分. 请确保你仔细阅读并理解以下“评估标准”并根据需要进行参考. “评估标准”: {评估标准} (相关性 relevance (1–5)…) 请严格遵循“评估标准”, 给出不多于5条评估的“具体步骤”. “具体步骤”:
模型回答1	(1) 阅读和理解源文本: 在开始评估之前, 彻底阅读源文本, 以便深入理解其主要论点、事实和详情. 注意各种信息点的重要性层级, 以便能够辨识哪些是关键信息. (2) 阅读摘要并识别关键信息: 细致阅读摘要文本, 标记出摘要中出现的关键信息点. 这些信息点应该是源文本中最重要的内容摘要. (3) 比较摘要和源文本的关键信息: 将摘要中标记的关键信息与源文本中的相应信息进行比较. 检查摘要是否捕捉到了所有关键信息, 以及这些信息是否已经被准确且恰当地传达. (4) 评估信息的相关性: 对于摘要中的每个关键信息点, 评估其与源文本的相关性. 如果摘要包含了所有关键信息且没有无关紧要的信息, 考虑给出高分. 如果重要信息缺失或摘要包含了不相关的信息, 给出较低的分数. (5) 打分和提供反馈: 根据摘要中关键信息的完整性和相关性, 使用1–5的等级打分. 如果摘要在相关性上做得很好, 即包含了所有重要信息且未包含多余信息, 给出5分. 如果摘要遗漏了重要信息或包含了过多无关信息, 根据其严重程度适当降低分数.
提示词2	请严格遵循“具体步骤”, 并将其应用于下面给出的“源文本”和“摘要”. “源文本”: {源文本} (This is the embarrassing moment…) “摘要”: {摘要文本} (buckingham palace guard slipped…) 严格遵循“具体步骤”中的每一个步骤, 并结合“源文本”和“摘要”的内容, 给出“评估依据”. “评估依据”: 根据该“评估依据”, 给出“最终得分”. “最终得分” (数字):
模型回答2	评估依据: (1) 阅读和理解源文本: 源文本详细描述了一名守卫在白金汉宫换岗仪式中不慎滑倒的情形. 文中包含了事件发生的时间、地点、目击者的描述以及守卫的反应等详细信息. (2) 阅读摘要并识别关键信息: 摘要文本简要地提到了守卫滑倒的事件, 包括滑倒的动作和结果 (摔倒、撞击熊皮帽、掉落步枪). (3) 比较摘要和源文本的关键信息: 摘要中未提及的关键信息包括目击者的身份、守卫的具体反应以及军事部的官方回应. 摘要正确地传达了事件的主要部分, 但缺少了一些细节. (4) 评估信息的相关性: 摘要与源文本的关键信息相比较, 包含了事件的主要核心, 即守卫滑倒的情形. 然而, 它并没有包含目击者的描述、守卫的尴尬情绪以及摔倒之后的反应, 这些信息虽非核心却丰富了故事. (5) 打分和提供反馈: 摘要捕捉到了事件的主要情节, 但遗漏了使故事更完整的一些细节. 因此, 我会给摘要打4分. 这反映了它在相关性方面做得好, 但不是完美无缺. 有趣的细节, 比如目击者的反应和守卫的尴尬, 虽然不是必需的, 但如果包含其中, 将使摘要更加生动和完整. 最终得分: 4

本文方法使用现有的文本摘要数据集, 利用设计好的思维链提示词来引导大型大语言模型根据摘要与源文本生成质量评估的推理过程与最终得分, 随后将其给出的答案分数与数据集中的人工打分进行比对, 筛选出得分相近的样本进行保留, 最后将筛选完成的数据制作成多轮对话形式的思维链数据集.

表 3 中给出了思维链数据集的结构, 数据包括 instruction、output 和 history 这 3 部分, 其分别对应输入指令、模型输出和对话历史, 其内容由第 2.2 节中提到的提示词与模型回答组成. 后续使用该数据集对小型大语言模型进行微调训练, 可以让模型对基于思维链原理的推理过程进行学习, 从而使小型大语言模型能够以思维链推理的方式完成文本摘要质量评估任务.

2.4 LoRA 微调

大语言模型有着参数量巨大这一特点, 对其重新预训练或进行全参数微调都会消耗极高的计算资源, 成本十分高昂. 为避免这一点, 文本方法采用 LoRA[26] 的方式对模型进行微调训练以减少训练成本, 提高计算效率. LoRA 微调的具体原理由图 3 所示.

表 3 思维链数据集结构

类型	内容
Instruction	{提示词2} (请严格遵循“具体步骤”, 并将其应用于下面给出的“源文本”和“摘要”…)
Output	{模型回答2} (模型给出的评估依据与评估结果)
History	[{提示词1} (作为一名文本生成质量评估专家, 你将收到由“源文本”和其“摘要”组成的一组摘要文本…)] [{模型回答1} (模型给出的评估步骤)]

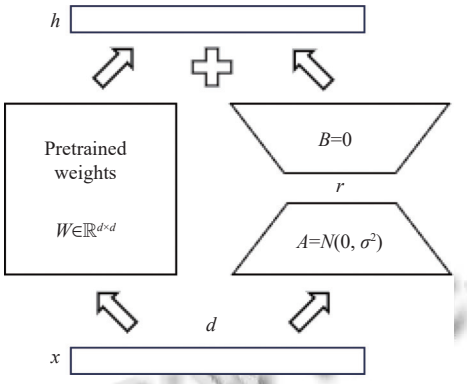

图 3 LoRA 微调原理

与普通的训练方式不同, LoRA 微调采用添加旁路的方式, 在模型 Transformer 结构的权重矩阵旁添加了 2 个低秩矩阵 $A \in \mathbb{R}^{r\times d}$ 和 $B \in \mathbb{R}^{d\times r}$. 其中矩阵 A 采用随机高斯初始化, B 采用 0 初始化. 最终的输出 h 表示如下:

$$h = Wx + BAx \tag{1}$$

在微调训练过程中, 只对矩阵 A 和 B 进行权重更新. 由于 r 远小于 d, 所以该旁路的参数量和计算成本都远小于原网络矩阵 W. 通过这种方式, 实现了极大降低训练开销的同时又使模型完成了针对下游特定任务的微调.

3 实验结果与分析

3.1 数据集

本文使用 Summeval 文本摘要数据集[19], 其包含 16 个不同模型在 100 篇 CNN/DailyMail 新闻文章上生成的摘要和其在 4 个维度上的人工评分, 共计 1 600 组由源文本、摘要、评分所组成的数据. 实验中按 3:1 的比例将数据集进行划分, 对于大型大语言模型, 使用全部 1 600 组数据进行评估以生成数据样本并得到评估结果; 对于小型大语言模型, 使用 1 200 组大型大语言模型生成的数据样本, 以模型评分与标准评分的相关性准确度为基准对数据样本进行筛选, 使用筛选后的数据对基座模型进行微调训练, 之后对其余 400 组数据进行评估并得到评估结果.

3.2 实验环境

本文实验使用的基座模型为 Baichuan2-13B 的 chat 模型和 base 模型, 大型大语言模型为 GPT 系列的 GPT-4 和 GPT-3.5-turbo. 实验使用的计算机系统环境为 Ubuntu 20.04 系统, GPU 配置为 A40, CPU 配置为 Intel Xeon Platinum 8358P, 微调方式采用 LoRA 微调. 实验使用大型大语言模型生成的训练数据对基座模型进行训练, 根据其训练损失在两轮训练后基本不再下降的情况, 实验将 epoch 设定为 3, 最后以 *temperature*=0.3, *top-P*=0.85 的设置使模型输出最终评估结果.

3.3 摘要质量评估结果对比

本文实验以数据集中的人工评分作为基准, 将模型输出的评分与其进行对比, 并计算 Spearman (ρ) 和 Kendall-Tau (τ) 相关性分数作为评估指标. 具体计算公式如下:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

$$\tau = \frac{4P}{n(n-1)} - 1 \tag{3}$$

其中, Spearman (ρ) 计算公式中 n 为样本大小, d_i 为两组数据排序后秩次的差值. Kendall-Tau (τ) 计算公式中的 n 为样本大小, P 为两组数据中排名都在给定数据后的数据的数量之和. 实验通过计算 Spearman (ρ) 和 Kendall-Tau (τ) 相关性分数来得到模型输出的评分和数据集基准评分之间的单调关系和排序一致性度量, 并以此来衡量模型对文本摘要质量评估的准确度.

本文实验分别以 GPT-3.5-turbo 和 GPT-4 作为大型大语言模型, 以 Baichuan2-13B-base 和 Baichuan2-13B-chat 作为小型大语言模型, 在 Summeval 数据集[19]上进行实验并最终共得到 7 组模型的实验结果, 各组设置如下.

(1) Ori: 未经微调的 Baichuan2-13B-chat 模型.

(2) G-3.5: GPT-3.5-turbo 模型.

(3) Base-3.5: 以 GPT-3.5-turbo 作为大型大语言模型, 对 Baichuan2-13B-base 模型进行微调得到的模型.

(4) Chat-3.5: 以 GPT-3.5-turbo 作为大型大语言模型, 对 Baichuan2-13B-chat 模型进行微调得到的模型.

(5) G-4: GPT-4 模型.

(6) Base-4: 以 GPT-4 作为大型大语言模型, 对 Baichuan2-13B-base 模型进行微调得到的模型.

(7) Chat-4: 以 GPT-4 作为大型大语言模型, 对 Baichuan2-13B-chat 模型进行微调得到的模型.

表 4 中展示了本文的 7 组实验结果与现有方法对文本摘要质量评估准确度的对比. 对于大型大语言模型, G-4 组的结果优于其他所有结果, 这说明了在文本摘要评估任务的表现上, GPT-4 的性能优于其他大语言模型, 也说明本文的使用大语言模型来实现文本摘要质量评估的方法是准确且有效的. 对于小型大语言模型 Ori 组, 即未经微调的 Baichuan2-13B-chat 在文本摘要质量评估任务上的准确度很低, 而经过微调后的模型在评估准确度上有了极大的提升. 由于实验的打分过程符合人类对话逻辑, 所以 Chat 模型的表现会略优于 base 模型, 而又因 GPT-4 模型生成的数据优于 GPT-3.5-turbo 生成的数据, 所以实验中小型大语言模型的最优结果为 Chat-4 组, 其在文本摘要的质量评估的能力可以达到与 GPT-3.5-turbo 相当的水平.

表 4 不同模型的文本摘要质量评估准确度对比

Approaches	Coherence		Consistency		Fluency		Relevance		Avg	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
CTC (consistency)	0.223	0.172	0.415	0.345	0.335	0.276	0.166	0.124	0.285	0.229
CTC (relevance)	0.402	0.310	0.366	0.301	0.299	0.245	0.428	0.336	0.374	0.298
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
GPTScore	0.434	—	0.449	—	0.403	—	0.381	—	0.417	—
Ori	0.075	0.063	0.040	0.038	0.022	0.020	0.045	0.040	0.046	0.040
G-3.5	0.397	0.327	0.345	0.306	0.358	0.313	0.356	0.295	0.364	0.310
Base-3.5	0.221	0.182	0.185	0.162	0.198	0.170	0.176	0.147	0.195	0.165
Chat-3.5	0.291	0.241	0.248	0.219	0.253	0.221	0.249	0.205	0.260	0.222
G-4	0.543	0.453	0.465	0.416	0.441	0.388	0.492	0.419	0.485	0.419
Base-4	0.380	0.310	0.301	0.269	0.293	0.255	0.337	0.281	0.328	0.279
Chat-4	0.437	0.363	0.364	0.323	0.351	0.305	0.395	0.324	0.387	0.329

本文提出的文本摘要质量评估方法通过微调训练的方式极大地提升了小型大语言模型的评估准确度. 与传统方法相比, 本文方法在具有更加优秀评估准确度的同时无需参考文本, 且可以在给出得分的同时给出评估依据, 评估结果更加丰富多样; 与现有的其他采用大语言模型的评估方法相比, 本文方法在保证较高评估准确度的同时将使用模型的参数量降低至 13 B, 极大减小了模型的硬件需求和计算成本, 更加灵活且便于部署.

表 5 给出 Ori 与 Chat-4 两组模型分别在 Summeval 英文数据集[19]和 LCSTS[27]中文数据集, 在“相关性 relevance”这一维度上的评估结果. 由表 5 中示例可以看出未经微调的模型给出的评估依据逻辑性较差且缺乏细节, 最终得分也与标准得分有较大出入, 而经过微调的 Chat-4 组的模型在评估过程中具有更强的逻辑性且可以给出详细且完整的评估依据, 且其在中文和英文数据上的评估效果都有极大的提升. 这表明本文提出的方法可以有效提升模型在文本摘要质量评估任务上的评估能力, 并同样能够在其他的不同类型或语种的数据上完成高质量的评估, 具有良好的泛化性.

3.4 思维链提示词有效性分析

本文实验采用结合思维链原理的提示词构建方法来提升大语言模型在文本摘要质量评估任务上的准确度. 为验证其有效性, 本文实验在 Summeval 数据集[19]中随机抽选了 320 条数据进行对比实验, 分别使用含思维链引导 (本文实验使用的提示词) 与不含思维链引导 (要求模型直接进行打分) 的提示词来引导模型完成打分评估.

实验分别使用了未经微调的 Baichuan2-13B-chat 模型 (Ori)、GPT-3.5-turbo (G-3.5) 和 GPT-4 (G-4) 进行了实验, 其中 Ori、G-3.5、G-4 为使用思维链的组, Ori (no CoT)、G-3.5 (no CoT)、G-4 (no CoT) 为不使用思维链的组. 表 6 给出了不同组模型在连贯性维度上的评分准确度. 由实验结果可以看出, 含思维链引导的组的实验结果明显优于不含思维链引导的组的实验结果. 实验结果表明本文采用的基于思维链的提示词构建方法可以有效提升模型在文本摘要评估任务上的评估准确度.

表 5 评估结果示例

源文本与摘要以及标准得分	Ori	Chat-4
源文本: Paul Merson has restarted his row with… 摘要: Paul Merson was brought… 标准得分: relevance: 1.667	评估依据: (1) 相关性: 摘要主要包含了源文本中的重要信息, 而非无用信息. 摘要中包含源文本中的各个重要信息, 如Paul Merson与Andros Townsend的争执、Merson对Townsend的批评、Townsend对Merson的回应等. (2) 完整性: 摘要包含了源文本中的大部分重要信息, 如Merson对Townsend的批评、Townsend对Merson的回应等. (3) 冗余信息: 摘要中未出现过多的冗余信息. 最终得分: 4	评估依据: (1) 理解源文本: 源文本提供了关于Paul Merson和Andros Townsend之间争执的背景信息. Merson在社交媒体上对Townsend的评论, 以及Townsend为英格兰队进球后对Merson的回应, 都被提及. (2) 阅读摘要: 摘要提到了Merson在比赛最后时刻被换上场, Townsend在比赛中的表现, 以及Merson对Townsend的评论. (3) 评估信息覆盖: 摘要没有提及Merson承认自己错误的部分. 摘要也没有提到Townsend对Merson的回应, 以及Merson在社交媒体上的评论. (4) 识别冗余信息: 摘要中提到了“click here for all the latest arsenal news”, 这是一个不相关的信息, 因为摘要应该专注于源文本的主要事件. (5) 打分: 摘要虽然提到了一些关键事件, 但遗漏了源文本中的一些重要信息, 并且包含了冗余信息. 因此, 根据摘要的相关性, 我会给出较低的分数. 最终得分 (数字): 2
源文本: 昨天, 杭城最高气温又达到了28.4℃, 今天最高气温可能冲到30℃. 不过明天到后天有一股强冷空气来袭, 杭州明天小雨转阴, 气温19–26℃. 秋雨寒, 出门最好还是备把伞. 随着冷空气的渗透, 后天会是本周最凉的一天, 气温16–21℃. 摘要: 明后天有冷空气南下后天最高气温降到21℃. 标准得分: relevance: 暂无	评估依据: (1) 相关性: 摘要包含了源文本中的关键信息, 如最高气温、冷空气和气温范围. 摘要主要包含了源文本中的重要信息, 而非无用信息, 因此得分较高. (2) 准确性: 摘要准确地传达了源文本中的关键信息, 如最高气温和气温范围. 摘要没有错误或误导性的信息, 因此得分较高. (3) 简洁性: 摘要简洁明了, 能够用简短的文字概括源文本的主要观点. 摘要简洁且明了, 因此得分较高. (4) 语言和风格: 摘要使用清晰、简洁的语言, 并遵循一致的格式和风格. 摘要的语言和风格恰当, 因此得分较高. 最终得分: 5	评估依据: (1) 阅读源文本: 源文本提供了关于杭州未来几天的天气情况的详细信息, 包括最高气温的变化、天气状况以及气温范围. (2) 阅读摘要文本: 摘要文本提到了“冷空气南下”和“最高气温降到21℃”的信息. (3) 比较关键信息: 源文本中的关键信息包括最高气温的变化、天气状况、气温范围以及建议的预防措施 (如带伞). 摘要中提到了最高气温的变化, 但没有提及天气状况和气温范围, 也没有提到建议的预防措施. (4) 评估信息的相关性: 摘要捕捉了源文本中关于气温变化的关键信息, 但遗漏了天气状况、气温范围以及建议的预防措施等其他信息. (5) 打分并提供反馈: 摘要虽然提到了冷空气南下和气温下降, 但遗漏了其他信息, 如天气状况和气温范围, 因此, 摘要的相关性评分不能是满分. 最终得分 (数字): 3.5

表 6 不同模型是否含思维链提示的评估准确度对比

Approaches	Coherence	
	ρ	τ
Ori	0.0696	0.0561
Ori (no CoT)	0.0349	0.0330
G-3.5	0.3816	0.3192
G-3.5 (no CoT)	0.2038	0.1635
G-4	0.5207	0.4303
G-4 (no CoT)	0.4281	0.3643

4 结论与展望

本文提出了一种基于大语言模型的文本摘要质量的评估方法, 设计了一种基于思维链原理的提示词构建方法, 提升了大语言模型在文本摘要质量评估任务上的性能, 同时生成思维链数据集对小型大语言模型进行微调训练, 显著提高了小型大语言模型在文本摘要质量评估任务上的评估准确度. 与传统方法相比, 本文方法的评估准确度更高, 无需参考文本且能够给出评估依据, 评估结果更加丰富多样. 与其他基于大语言模型的评估方法相比, 本文方法在保证较高准确度的同时计算成本更低, 更加灵活且便于部署. 后续研究将围绕更多样的提示词设计和更加丰富的数据集构建来展开, 以继续提升模型在文本摘要质量评估任务上的评估准确度.

参考文献

1 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318.

2 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the 2004 Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.

3 Rus V, Lintean M. A Comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. Proceedings of the 17th Workshop

on Building Educational Applications Using NLP, Stroudsburg: ACL, 2012. 157-162.
4 Zhang T, Kishore V, Wu F, *et al*. BERTScore: Evaluating text generation with BERT. arXiv:1904.09675v3, 2020.
5 Lewis M, Liu YH, Goyal N, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Washington: ACL, 2020. 7871–7880.
6 Touvron H, Martin L, Stone K, *et al*. LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.
7 Bai JZ, Bai S, Chu YF, *et al*. Qwen technical report. arXiv:2309.16609, 2023.
8 Yang AY, Xiao B, Wang BN, *et al*. Baichuan 2: Open large-scale language models. arXiv:2309.10305, 2023.
9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
10 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–186.
11 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver: Curran Associates Inc., 2020. 1877–1901.
12 Lewis M, Liu YH, Goyal N, *et al*. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461, 2019.
13 Zhao W, Peyrard M, Liu F, *et al*. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 563–578.
14 Deng MK, Tan BW, Liu ZZ, *et al*. Compression, transduction, and creation: A unified framework for evaluating natural language generation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 7580–7605.
15 Zhong M, Liu Y, Yin D, *et al*. Towards a unified multi-dimensional evaluator for text generation. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 2023–2038.
16 Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text Transformer. The Journal of Machine Learning Research, 2020, 21(1): 140.
17 Fu JL, Ng SK, Jiang ZB, *et al*. GPTScore: Evaluate as you desire. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City: ACL, 2024. 6556–6576.
18 Kryściński W, Keskar NS, McCann B, *et al*. Neural text summarization: A critical evaluation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 540–551.
19 Fabbri AR, Kryściński W, McCann B, *et al*. SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 2021, 9: 391–409. [doi: 10.1162/tacl_a_00373]
20 Wei J, Wang XZ, Schuurmans D, *et al*. Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24824–24837.
21 Kojima T, Gu SS, Reid M, *et al*. Large language models are zero-shot reasoners. arXiv:2205.11916, 2022.
22 Wang XZ, Wei J, Schuurmans D, *et al*. Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171v4, 2023.
23 Li YF, Lin ZQ, Zhang SZ, *et al*. Making language models better reasoners with step-aware verifier. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 5315–5333.
24 Zhang ZS, Zhang A, Li M, *et al*. Automatic chain of thought prompting in large language models. arXiv:2210.03493, 2022.
25 Zhou D, Schärli N, Hou L, *et al*. Least-to-most prompting enables complex reasoning in large language models. arXiv:2205.10625, 2022.
26 Hu EJ, Shen YL, Wallis P, *et al*. LoRA: Low-rank adaptation of large language models. arXiv:2106.09685, 2021.
27 Hu BT, Chen QC, Zhu FZ. LCSTS: A large scale Chinese short text summarization dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 1967–1972.

(校对责编: 王欣欣)