

文章编号: 1003-0077(2026)01-0070-15

CPQD: 用于提升大语言模型文学表现力的中文诗词引用数据集

胡亚豪¹, 徐磊¹, 潘志松¹, 陶蔚², 邹傲¹, 陈满¹

(1. 陆军工程大学 指挥控制工程学院, 江苏 南京 210001;

2. 军事科学院 战略评估咨询中心, 北京 100091)

摘要: 随着预训练技术的发展, 大语言模型的生成能力得到显著提升, 能够自动生成近似自然语言的文本, 如新闻、摘要、故事等, 为人们的日常生活和工作提供了便利。然而, 大语言模型在引用中国古典诗词方面的表现尚缺乏系统性研究, 而这个问题对于提升模型生成内容的文学表现力具有重要意义。为此, 该文构建了一个大规模诗词引用数据集 CPQD, 并从诗意、上下文连贯性和流畅性等多个角度建立了一套自动和人工评测指标。随后, 在零样本和少样本设定下, 利用国内开源(中文, 3 款)、国内商用(中文, 3 款)、国外开源(英文, 1 款)和国外商用大语言模型(英文, 1 款)共 8 款大语言模型进行了实验与比较分析。实验结果表明, 仅有 1 款国内商用大语言模型的性能超过了微调 LSTM 模型。而其余 7 款大语言模型在零样本学习下的 BLEU 值普遍偏低, 仅能达到微调 LSTM 模型的 15%~93%, 即便增加少量样本, 其性能提升也较为有限。为了提升大语言模型在诗词引用方面的能力, 该文进一步探究了开源大语言模型在 CPQD 数据集上进行监督微调的效果。结果显示, 微调后开源模型的 BLEU 值提升了近 10 倍, 并显著超越了商用大语言模型的性能。这表明, 该数据集为大语言模型的微调提供了有力支持, 有效地提升了模型在文学表现力, 特别是在诗词引用方面的能力。

关键词: 自然语言生成; 大语言模型; 中国古典诗词; 引用推荐

中图分类号: TP391

文献标识码: A

CPQD: A Chinese Poetry Quotation Dataset for Enhancing the Literary Expressiveness of Large Language Models

HU Yahao¹, XU Lei¹, PAN Zhisong¹, TAO Wei², ZOU Ao¹, CHEN Man¹

(1. Command and Control Engineering College, Army Engineering University of PLA,

Nanjing, Jiangsu 210001, China;

2. Strategic Assessments and Consultation Institute, Academy of Military Science, Beijing 100091, China)

Abstract: With the development of pre-training technology, the capability of generative artificial intelligence has been significantly improved, enabling the autonomous generation of texts that closely mimic natural language. However, their performance in quoting Chinese classical poetry has not been thoroughly explored, which is crucial for enhancing the literary expressiveness of the generated content. To address this, we construct a large-scale Chinese Poetry Quotation Dataset (CPQD), and propose a set of automatic and manual evaluation metrics from multiple perspectives including poeticness, coherence, and fluency. Subsequently, we analyze three open-sourced Chinese LLMs, three commercial Chinese LLMs, one open-sourced and one commercial English LLMs, in both zero-shot and few-shot setting. The results demonstrate that only one of the commercial Chinese language models outperforms the fine-tuned LSTM model. We fine-tune the open-source models using the proposed poetry quotation dataset, and the BLEU increases nearly tenfold, significantly surpassing that of the commercial LLMs.

Keywords: natural language generation; large language models; Chinese classical poetry; quote recommendation

收稿日期: 2024-06-28 定稿日期: 2024-09-03

基金项目: 国家自然科学基金(62076251, 62106281)

0 引言

随着 AI 技术的迅猛发展,智能文本生成已成为自然语言处理领域的前沿研究方向,特别是在 AI 生成内容^[1-2](AI-Generated Content, AIGC)方面取得了显著进展。然而,如何提升生成文本的文学表现力和文化内涵一直是该领域面临的挑战^[3]。

中国古典诗词^①源远流长,是一种高度文学化、艺术化和凝练化的语言形式,是中华优秀传统文化的重要代表之一。近年来,中国古典诗词热度持续提升,人们将古老的经典诗词和鲜活的当下生活创造性地结合起来,让古典诗词在当代充满活力。

在此背景下,将中国古典诗词与现代智能文本生成系统相结合,不仅可以为现代文本创作提供更为丰富和深邃的文化底蕴,同时也是对中华优秀传统文化的传承与创新。周等人^[4]提出了一种基于遗传算法的宋词自动生成技术,为古典诗词的现代化呈现开辟了新的路径。此外,Zhipeng 等人^[5]和 Liu 等人^[6]分别推出了“九歌”和“深诗”这两款古典诗词自动写作系统,进一步促进了古典诗词与现代科技的融合。最近发布的“荀子”古籍大语言模型^②,则为古籍研究和古典诗词生成提供了新的视角和方法。

本文主要聚焦于探索大语言模型在诗词引用任务中的表现和挑战。考虑真实的使用场景,本文研究在给定上下文中恰当地引用中国古典诗词的问题。如表 1 所示,将中国古典诗词中的经典名句“路漫漫其修远兮,吾将上下而求索”恰到好处地引入原文本中的“# mask #”位置,不仅丰富了文本的表达,还提升了整体的文学性。同时,这句诗词的引用也更为深刻地传达了原文本所要表达的百折不挠的“执着精神”,让读者能够更深入地理解和感受这种精神的内涵和力量。

诗词引用可视为引用推荐^[7-14]的一个全新子任务,相比已有工作,本文在以下三个方面存在显著不同:

首先,本文构建了大规模的诗词引用数据集 CPQD。先前的研究中,引用数据库的规模相对较小,例如,Zheng 等人^[15]构建的成语库只有 3 848 个成语,Qi 等人^[13]构建的古诗词有 4 438 条,Tan 等人^[7]构建的名言库只有 3 158 个句子。相比之下,本文构建的诗词库规模扩大了近 10 倍,包含了 35 561 诗词,为模型的训练提供更丰富的资源,并

且全部开源。

其次,本文聚焦于大语言模型的研究。先前的模型主要使用长短时记忆网络^[16](Long Short-Term Memory, LSTM)或预训练语言模型(如 BERT^[17])作为主干网络,其模型能力在处理复杂的引用任务时受到一定限制。而本文的模型主要研究大语言模型在生成诗词引用方面的能力,其丰富的预训练知识库为生成高质量的引用提供了有力支持。具体而言,本文选取了 8 款国内外开源和商用的大语言模型,并在零样本、少样本和监督微调三种不同的设定下进行了测试。

表 1 诗词引用任务示例

输入	坚持问题导向,以“# mask #”的执着精神,聚焦实践遇到的新问题、改革发展稳定存在的深层次问题、人民群众急难愁盼问题、国际变局中的重大问题、党的建设面临的突出问题,不断提出真正解决问题的新理念新思路新办法。
输出	坚持问题导向,以“路漫漫其修远兮,吾将上下而求索”的执着精神,聚焦实践遇到的新问题、改革发展稳定存在的深层次问题、人民群众急难愁盼问题、国际变局中的重大问题、党的建设面临的突出问题,不断提出真正解决问题的新理念新思路新办法。 ^[18]

最后,本文采用了序列到序列的方式来生成引用。大多数现有引用推荐方法^[7-14]将上下文和引用编码成向量,在此基础上进行引文-上下文匹配,并根据相似度得分返回候选库中与上下文内容相关的引用。相比之下,本文将诗词引用任务视为序列到序列的文本生成任务,这使得模型能够创造性地生成与上下文紧密相关的引用^③。这一转变不仅增加了任务的挑战性,同时也拓宽了引用的范围,使得生成的引用不再受限于候选库中的已有引用,而是能够根据上下文内容进行适当的创作。此外,这一转变导致推荐系统的评价指标^[19],如查全率和平均倒数排名不再适用。

综上,本文贡献主要有以下三点:

(1) 本文构建了一个包含 35 561 句诗词、总计 429 044 条上下文-诗词引用对的大规模数据集 CPQD。该数据集不仅为大模型的诗词引用能力评测提供了有力支持,还进一步支持对模型的微调。

① 在本文中,“中国古典诗词”泛指基于文言文和传统格律创作的古代诗文,包括诗、词、曲、赋等多种文学体裁。

② <https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM>

③ 大部分情况下是引用已有的中国古典诗词。

本文的数据集开源在 ModelScope 上^①。

(2) 为了全面评估大语言模型在诗词引用任务上的性能,本文在零样本、少样本和监督微调设定下,对 8 款国内外的大语言模型进行了实验与比较分析。实验结果显示当前大语言模型在诗词引用方面仍有很大的改进空间。

(3) 在模型评估方面,本文构建了针对诗词引用任务的自动评测指标,包括诗意、上下文连贯性和流畅性等方面。与人工评测的对比显示,这些指标在评估模型性能时具有一定有效性,能够反映出模型在不同方面的优势和不足。

1 相关工作

1.1 诗歌生成

诗词生成是指利用诗词生成模型,根据给定的一个或多个关键词、描述性语句等输入,自动地创作出一首完整的诗歌。基于深度学习的诗词生成模型通常在大量诗词语料上训练,以学习诗词的表达方式、主题风格等特征,并生成新诗词。梁等人^[20]使用循环神经网络来生成中国诗词,获得了比基于模板方法更好的效果。Yang 等人^[21]以无监督的方式实现了相同主题的多风格诗词生成,保证了一致性和连贯性,并提高了主题风格的表现力。Yi 等人^[22]将历史背景和生活经历等视为隐变量的影响因素,通过控制影响因素的设置生成不同的诗词,有效增加了生成诗词的多样性。此外,为提升诗词生成中的韵律规范,吴等人^[23]提出了基于 Transformer^[24]结合韵律特征的诗词生成模型。在将关键主题词融入模型的研究中,王等人^[25]提出了一种基于条件变分自编码器的生成模型,该模型能够在更加丰富的语义信息指导下,生成更符合关键字描述和用户满意度的诗词。

相比之下,诗词引用任务则要求模型根据一段输入文字,理解其上下文语境,选择恰当的诗词进行引用,以达到增强表达效果或传递特定信息的目的。

1.2 引用推荐

引用,是一系列他人所说或所写的文本序列,包括成语、谚语、格言和诗词等。通过恰当的引用,我们可以更清晰地阐明和强调想表达的意思,同时赋予写作或对话更加优雅的风格。为了帮助人们更有效地找到与其上下文相匹配的引用,引用推荐任务

旨在推荐与上下文相匹配的引用,进而提升写作质量和表达能力。

在模型方面,Tan 等人^[7]提出了一种基于手工特征的排序框架,包含了 16 个特征,如频率、投票、网络流行度。然而,由于手工特征提取工作非常耗时,且缺乏通用性。随后,Tan 等人^[8]尝试用神经网络处理引用推荐,通过学习上下文和引用的分布式表示,利用余弦相似度计算上下文与引用之间的相似度,实现基于上下文内容的推荐。Liu 等人^[9]利用 LSTM^[16,26]网络提出了一个基于神经机器翻译的引用生成框架,其中假设引用是目标语言,上下文信息是要翻译的源语言。然而,采用简单的 LSTM 网络不能很好地对上下文和引文进行建模,需要更先进的模型。Wang 等人^[10]引入预训练语言模型 BERT,利用其强大的语义表示能力来增强推荐系统,通过上下文和引用之间的表示相似性对候选引用进行排名。然而,处理引用的表示尤为挑战,因为它们简洁且富含深意,这使得学习过程更为复杂。Qi 等人^[13]进一步优化,通过在预训练语言模型的基础上融入通用词汇知识,以增强引用的表示,提高推荐的准确性。

在数据集方面,Tan 等人^[7]构建了一个包含 3 158 条引用和 64 323 个上下文的英文数据集,数据来自古腾堡计划^②的电子书。Ahn 等人^[12]也创建了一个类似的英文数据集,其中包含了来自古腾堡计划的电子书和博客的 400 条最常见引文及其上下文。另外,Zheng 等人^[15]构建了中文成语引用数据集,其中上下文中的成语被空白替换,需要从 3 848 个候选成语中选择正确答案。此外,Qi 等人^[13]构建了一个名言推荐数据集,包括英语、现代汉语和古诗词引用三部分,其中古诗词引用包含 4 438 条诗句和 116 53 个上下文,数据来源于中文互联网上的问答网站和电子书。

2 问题定义

在智能文本生成系统中,诗词引用任务可以被定义为序列到序列的文本生成任务。给定输入序列 $X = \{C_{\text{left}}, <\text{mask}>, C_{\text{right}}\}$,其中包含一个待引用的诗词槽位 $<\text{mask}>$ 以及该槽位的左右上下文信息 C_{left} 和 C_{right} 。诗词引用任务的目标是基于上下文

① https://modelscope.cn/datasets/huyahao/poetry_quotation

② <https://www.gutenberg.org/>

信息在<mask>槽位中引用恰当的诗词,从而提升输入序列的文学性和可读性。具体来说,给定输入序列 $X = \{C_{\text{left}}, \langle \text{mask} \rangle, C_{\text{right}}\}$,模型根据上下文信息来预测被屏蔽的诗词序列 $Y = \{y_1, y_2, \dots, y_m\}$ 的条件概率,如式(1)所示。

$$P(Y | X) = \prod_{i=1}^m P(y_i | y_{<i}, X) \quad (1)$$

其中, $y_{<i}$ 表示到目前为止生成的诗词内容。

3 数据集构建

3.1 引号段落收集

在日常写作中,人们经常通过引用中国古典诗词、模仿诗词风格或借鉴诗词的意境和表达方式,来为自己的文字增添文学气息和情感渲染力。例如下面这段文字:

社交平台上与诗相关的内容越来越多,有人分享诗意生活、朗诵优美诗句,有人解析《中国诗词大会》文案,为写作积累素材……绵亘千年的古典诗词已“飞入寻常百姓家”,与当代生活产生紧密联系。^[27]

正如上面例子所展示的,通过引入“飞入寻常百姓家”这句古诗,文字被赋予更浓厚的文学氛围。值得注意的是,在很多情况下,被引用的诗词往往被置于引号内,这一特点为构建诗词引用数据集提供了便利。

本文选取了大规模中文开源数据集 MNBVC^① 作为语料库,该语料涵盖通用文本、新闻、问答、小说、书籍、杂志、论文等各类中文数据。在清洗过的 430 GB 数据中,检索并抽取包含引号的段落,构建了一个基于 MNBVC 数据集的引号段落集合,其中包含了约 2 500 万条带引号的文本段落。

3.2 诗词-非诗词分类

在初步采集了包含引号的段落之后,发现其中掺杂了大量的非诗词数据,包括歌词(如“我用无悔,刻永世爱你的碑”)、标语(如“游西湖、品龙井、看小百花”)以及对话内容(如“禀告大帅,此乃在下所为”)等。由于引号段落集合规模庞大,人工过滤显然是不现实的。为了准确、高效地从中筛选出诗词数据,本文通过训练诗词-非诗词分类器来判断引号中的内容是否为诗词。

由于现有研究缺乏满足本文需求的分类数据,因此需要构建相关的分类数据集以便用于训练分类

器。具体而言,收集了 1 万多条古典诗词名句^②,作为诗词数据基准。与此同时,非诗词数据类型多样,包括歌词、标语、对话内容等,其质量的好坏直接影响分类器的性能。

由于缺乏全面覆盖的非诗词数据,采用人工-机器迭代的方法来构建数据集。首先,将一部分诗词的现代文翻译作为初始的非诗词数据,与诗词数据一起用于分类器的初始训练。随后,从引号段落集合随机选取少量数据用训练得到的分类器进行尝试分类。这些数据随后经过人工标注,以确定其真实类别,并用这些标注结果来检验分类器的准确性。如果分类准确率超过 90%,则认为分类器达到了满意的性能水平,迭代过程将终止。否则,将人工标注的数据整合到分类数据集中,重新训练分类器以提升其性能。

经过多轮迭代,最终构建了一个诗词-非诗词分类数据集,其中包含 13 672 条诗词数据和 12 469 条非诗词数据。这种迭代的半自动构建方法可以帮助提高分类模型的准确性和鲁棒性,同时也减轻了人工标注的负担。表 2 展示了诗词-非诗词分类数据集的统计数据。

表 2 诗词-非诗词分类数据集统计

数据集	类别	训练集	验证集	测试集
诗词-非诗词	诗词	10 968	1 322	1 382
	非诗词	9 944	1 292	1 233

本文采用一种融合 BERT^[17] 和 CNN^[28] 的分类算法^③。具体而言,利用中文预训练语言模型 BERT 对文本进行句子层面的特征向量表示,并将所得到的特征向量输入 CNN 结构中进行分类与训练。表 3 展示了在诗词-非诗词测试集上的分类结果。

表 3 诗词与非诗词分类结果

(单位: %)

数据集	类别	准确率	召回率	F_1 值
诗词-非诗词	诗词	91.46	95.37	93.38
	非诗词	94.55	90.02	92.23

① <https://github.com/esbatmop/MNBVC>

② <https://github.com/caoxingyu/chinese-gushiwen>

③ <https://github.com/649453932/Bert-Chinese-Text-Classification-Pytorch>

综合来看,分类器能够满足识别诗词与非诗词识别的要求。

3.3 引号段落筛选

利用分类器对 3.1 节中收集到的引号段落集合进行分类,从中筛选出 150 万个包含诗词样本。为了构建上下文与诗词数据对 (X, Y) ,本文将这些样本段落中的诗词部分替换为掩码标记<mask>,作为模型输入 X ,同时将被屏蔽的诗词内容保留作为模型参考输出 Y 。

经过分类筛选后的数据还不能直接应用于诗词引用任务,仍需进一步地处理以确保其质量。具体而言,为了降低任务复杂度,仅保留诗词长度在 5-25 之间的数据。同时,为了控制段落的长度,本文提取诗词前后的句子(如果有的话)作为上下文信息,并限制每个样本最多包含 3 句话,总长度不超过 256 字。此外,为了确保有足够多的上下文信息来生成诗词,诗词的长度不超过段落总长度的 30%。最后,由于引用的诗词并非独立的,仅保留诗词引用的次数不少于 3 次的数据。这一筛选标准虽然可能导致数据集规模略有下降,但能够明显提升数据集的质量,确保所包含的诗词具有更高的研究价值。

3.4 数据统计分析

经过上述步骤后,构建了一个数据集 CPQD-MNBVC,其中包含 429 044 条上下文-诗词引用对。该数据集的诗词库共有 35 562 个不同的诗词,确保了引用的多样性。各个诗词对应的样本数量的分布情况如图 1 所示,其中横轴代表 CPQD-MNBVC 数据集中样本数量按从多到少排序的诗词索引,纵轴

代表每句诗词在数据集中的样本数量。

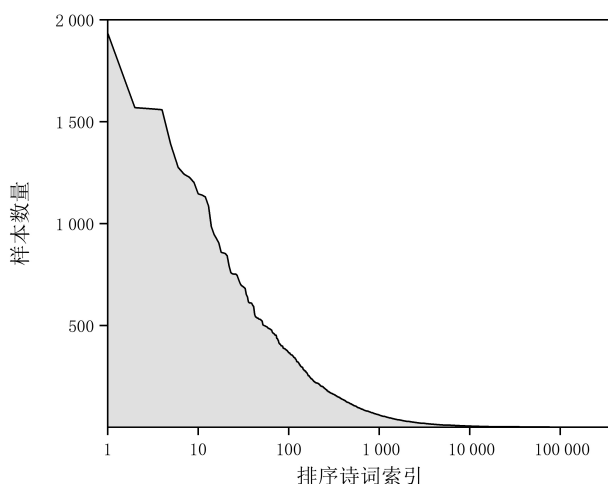


图 1 各诗词对应样本数量分布

从图 1 中可以看出,CPQD-MNBVC 数据集呈现出典型的长尾分布^[29]特征。为了进一步分析这一分布特征,将数据集中的诗词分为了三类:稀有诗词、常见诗词和频繁诗词。具体来说,样本数量在 $[3, 10)$ 之间的类别定义为稀有诗词;样本数量在 $[10, 50)$ 之间的类别定义为常见诗词;样本数量在 $[50, \infty)$ 之间的类别定义为频繁诗词。在 CPQD-MNBVC 数据集的诗词库中,稀有诗词有 28 679 个,占据了绝大多数;常见诗词和频繁诗词分别有 5 638 个和 1 245 个,数量相对较少。部分样本如表 4 所示。

为了公正地评估模型的性能,本文从稀有诗词、常见诗词和频繁诗词中分别抽取了 2 000 个样本作为验证集,再分别抽取 1 000 个样本作为测试集,其余的全部归入训练集。

表 4 诗词数据样本

频繁诗词	常见诗词	稀少诗词
工欲善其事,必先利其器	一缕香魂随风散,三更不曾入梦来	晚知书画真有益,却悔岁月来无多
先天下之忧而忧,后天下之乐而乐	出淤泥而不染,濯清莲而不妖	善歌声必先调其气
路漫漫其修远兮,吾将上下而求索	严以律人,宽以待己	经得一番寒彻骨
己所不欲,勿施于人	山回路转不见君,雪上空留马行处	重温往昔,雄关漫道真如铁
采菊东篱下,悠然见南山	无名天地之始,有名万物之母	舟必漏而水入,土必湿而生苔

3.5 域外测试

为了评估模型的泛化性,本文从互联网上爬取了 2 300 篇高考优秀作文,并从中筛选出 1 300 个符

合要求的样本,构成了一个域外测试集 CPQD-Gaokao。在筛选过程中,我们邀请了两位研究生从两个方面进行质量把关:一是确保引号内的内容确实是诗词;二是确保这些诗词与上下文内容相关。

表 5 展示了所有数据集的统计信息。其中,项 m/n 表示涉及 n 个引用的 m 个上下文-诗词引用对。

表 5 诗词引用数据集统计

数据集	训练	验证	测试	总计
CPQD-MNBVC	420 044/ 35 562	6 000/ 4 411	3 000/ 2 422	429 044/ 35 562
CPQD-Gaokao	—	—	1 300/935	1 300/935

4 基线模型

本文尝试了四大类方法以应对诗词引用任务: 基于检索推荐的方法、基于单任务学习的方法、基于提示学习的方法和基于模型微调的方法。对这些方法进行综合比较,全面地分析大语言模型在诗词引用任务上的表现。

4.1 基于检索推荐的方法

本文提出一种基于检索推荐的方法(Retrieval),即利用上下文信息来检索匹配的诗词。具体而言,首先利用中文词嵌入模型^①将收集的 35 562 句诗词转换成向量表示并存储在数据库中。当需要引用诗词时,将相应的上下文也转换为向量表示。然后,通过相似度度量方法计算上下文向量与数据库中的诗词向量之间的相似度^[30],并根据相似度结果对诗词进行排序。最终,返回与上下文最相似的诗词作为引用。

4.2 基于单任务学习的方法

在古典诗词中,经常出现省略关键信息、调整语法顺序以及运用典故的现象,这对于诗词引用带来了挑战。尤其在丰富世界知识的情况下,仅依赖上下文信息来生成引用诗词可能会变得困难。为了探究这一问题,本文设计了一个序列到序列^[31]的模型,该模型不依赖预训练语言模型的知识。具体而言,本文采用了一个带注意力机制的 LSTM 作为编码器-解码器模型的骨干网络。此外,模型的词向量初始化采用了基于人民日报数据训练的 Word2Vec^[32]模型。在此基础上,使用 CPQD-MNBVC 数据集来训练模型参数。

4.3 基于提示学习的方法

大语言模型^[33-35]是一个庞大的知识库,为各种

文本生成任务提供了关键的支持。本文采用零样本(Zero-Shot)提示学习和少样本(Few-Shot)示例学习两种提示学习^[36-37]方法,以评估八种国内外开源与商用大语言模型在诗词引用任务上的性能。商用模型包括 OpenAI 的 GPT4(所使用的版本为 GPT-4,调用时间为 2024 年 5 月 8 日),百度的 ERNIE-4.0(所使用的版本为 ERNIE-4.0-8K)以及智谱 AI 的 GLM3(所使用的版本为 GLM-3-Turbo)和 GLM4(所使用的版本为 GLM-4);国内开源模型涵盖智谱 AI 开放的 ChatGLM 系列,包括 ChatGLM-6B、ChatGLM2-6B、ChatGLM3-6B;国外开源模型为 Llama2-7B。在零样本提示学习中,通过在输入前添加任务相关的提示,引导模型生成符合要求的诗词。在少样本示例学习中,首先检索与输入上下文相关的样本,并将其作为示例添加到输入前,然后让模型根据这些示例和输入上下文来预测诗词。

4.4 基于模型微调的方法

尽管大语言模型在提示学习中表现出强大的性能,但在处理复杂任务时,微调模型仍然具有不可替代的优势。本文以 ChatGLM 系列开源模型为基础,采用了两种不同的微调方法:全参数微调和部分参数高效微调^[38]。具体来说,全参数微调(Fine-Tuning, FT)调整了模型的所有参数,使其更符合 CPQD-MNBVC 数据集的特性。此外,本文采用 P-Tuning^[38]作为高效微调方法,其只优化小部分模型参数,其余参数保持不变,这大大降低了计算和存储成本。

在实验环境方面,LSTM 模型的训练在 1 张 Tesla A100 GPU 上进行,P-Tuning 和 Fine-Tuning 方法在 8 张 Tesla A100 GPU 上进行。

5 评测方法

为了定量评价不同模型的性能,需要对模型所生成的诗词质量进行评测,本文采用的性能评测主要分为人工评测和自动评测两类方法。

5.1 人工评测方法

基于先前对诗词质量评估的相关研究^[5],结合本文的研究重点,总结了以下人工评测指标:

诗意(Poeticness): 评估生成的诗词是否展现出某种完整、优美的意境,是否具备艺术性和感染力。

① <https://huggingface.co/GanymedeNil/text2vec-large-chinese>

上下文连贯性(Context Coherence): 评估诗词在主题、意境、逻辑结构上是否与上下文关联紧密, 连贯一致。

流畅性(Fluency): 评估生成的诗句是否具备语法正确性、语义合理性, 并且是否足够流畅自然。

整体质量(Overall Quality): 除了细粒度的评测指标之外, 还收集了评测者对所生成诗词的整体质量评价。

具体来说, 将每个模型生成的 200 句诗词及其对应的上下文一同提交给三名评测者(均是研究生)。评分范围设定为 1~5 分, 其中 1 分表示最低质量, 5 分表示最高质量。使用下面的公式计算模型的在各个评测指标上的最终评分:

$$\text{Score}(\text{model}) = \frac{1}{|P|} \sum_{p \in P} \frac{1}{|J|} \sum_{j \in J} \text{score}(p_{\text{model}}, j) \quad (2)$$

其中, model 是基线模型之一, P 是一个包含 200 个测试集样本的子集, J 是 3 个评测者的集合, p_{model} 是由该模型生成的诗词及其上下文信息, $\text{score}(p_{\text{model}}, j)$ 表示 p_{model} 在评测者 j 下的得分。

5.2 自动评测方法

人工评测存在主观性, 且易受评测者个体差异的影响, 因此需要采用客观、稳定、可复现的自动评测指标来反映模型的基础性能。本文针对上述四个测评角度分别设计了自动评测指标:

诗意 假设现有的诗词具有优美的意境。为了评估诗句的诗意, 本文利用在 3.2 节中训练的分类器, 该分类器对诗词-非诗词数据的判断准确率为 93%。在评估过程中, 将生成的诗词输入到该分类器中, 利用其判断这些诗句是否具有诗意。通过分类器的处理, 测试集中被分类为诗词的样本比例 (Accuracy, ACC) 作为诗意指标。

上下文连贯性 假设参考诗词与上下文高度相关。对于生成的诗词, 本文采用 BLEU^[39] 指标来评估其与参考诗词之间的 n -gram 重合程度。同时, 计算两者之间的余弦距离来进行度量语义相似度 (BERTScore^[40])。较高的词汇重合度或者语义相似度表明模型生成的内容更接近人类诗词, 从而更有可能具备良好的上下文连贯性。

流畅性 假设现有的诗词语法正确并且足够流畅自然。为了评估诗句的流畅性, 本文利用在 3.2 节中收集的诗词数据训练了一个 3-gram 的 KenLM^[41] 语言模型。在评估过程中, 将生成的诗

词输入到语言模型中, 计算其在模型上的困惑度 (Perplexity, PPL), 以此作为衡量流畅性的指标。较低的困惑度值意味着模型能够更好地预测文本中的单词序列, 因此生成的诗词更加通顺自然。

整体质量 假定上述各项指标的重要性相等。本文对 ACC、BLEU、BERTScore 和 $(\log \text{PPL})^{-1}$ 指标取几何平均 (Geometric Mean, GM), 作为衡量整体质量的指标。

6 实验

本节将首先探讨基线模型在自动评测 (6.1 节) 和人工评测 (6.2 节) 中的性能, 并分析它们各自的优势与不足。接着, 分析自动评测指标和人工评测指标之间的相关性, 以衡量本文提出的自动评测指标的有效性 (6.3 节)。同时, 对人工评测进行了信度分析 (6.4 节)。由于本文所构建的诗词引用数据集具有长尾分布特点, 我们还分析了这种数据分布对实验结果可能产生的影响 (6.5 节)。此外, 本文还探讨了预训练语言模型版本对诗词引用任务性能的影响 (6.6 节)。最后, 通过实例分析 (6.7 节) 进一步阐明这些模型在实际应用中的表现。

6.1 自动评测分析

自动评测结果如表 6 所示, 加粗显示的为每项指标得分最高的前三项。Retrieval 模型根据输入上下文信息, 从诗词库中检索出一句相关的诗词作为引用。尽管该模型在诗意指标上达到了完美的 ACC=100.00%, 但在评估上下文连贯性的 BLEU 指标上却表现不佳, 两个数据集上的评分分别为 6.09 和 3.49。这可能是由于现代文本与古典诗词之间在语义、语法和表达习惯上存在差异, 使得嵌入模型在尝试将两者表征时难以形成有效的映射, 从而导致检索到的诗词与上下文信息不连贯。

相较于 Retrieval 模型, LSTM 模型在上下文连贯性上有显著性能提升, BLEU 指标分别提高了 8.05 和 9.64。这主要得益于 LSTM 模型基于 Seq2Seq 的有监督训练, 使其能够更好地理解和生成与上下文相匹配的诗句。然而, 由于诗词的深邃意境和抽象表达, 仅仅依赖单任务训练, 缺乏对广泛知识的整合, 导致它往往以较高的概率输出频繁诗词, 而对于常见和稀少诗词的场景, 其应对能力相对较弱, 从而导致在诗意方面的 ACC 指标上表现不佳, 分别仅为 58.00% 和 74.38%。值得注意的是, LSTM 模型在流畅性 PPL 指标上明显优于其他方法和参考 Reference。然

表 6 自动评测结果

基座模型	方法	CPQD-MNBVC					CPQD-Gaokao				
		ACC ↑ /%	BLEU ↑	BERTScore ↑	PPL ↓	GM ↑	ACC ↑ /%	BLEU ↑	BERTScore ↑	PPL ↓	GM ↑
—	Reference	100.00	100.00	100.00	250.50	20.63	100.00	100.00	100.00	184.15	20.92
—	Retrieval	100.00	6.09	60.36	155.30	9.24	100.0	3.49	59.22	114.09	8.13
LSTM	Fine-Tuning	58.00	14.14	62.60	128.66	10.14	74.38	13.13	62.21	66.06	10.97
ChatGLM-6B	Zero-Shot	73.97	3.84	58.05	360.87	7.27	75.62	3.28	58.08	237.72	7.16
	Few-Shot	73.17	4.55	58.81	580.18	7.45	77.78	4.70	59.05	570.59	7.64
	P-Tuning	79.57	10.53	60.38	336.21	10.17	91.56	12.60	61.34	184.35	10.79
	Fine-Tuning	96.03	41.75	77.01	262.17	15.34	98.84	49.62	79.93	190.01	16.53
ChatGLM2-6B	Zero-Shot	66.17	4.40	58.64	461.12	7.26	75.23	4.78	59.04	317.70	7.79
	Few-Shot	69.67	5.00	58.80	437.65	7.62	79.63	5.14	59.07	227.99	8.17
	P-Tuning	79.93	11.63	64.29	275.51	10.16	90.39	13.48	64.17	183.59	11.07
	Fine-Tuning	97.57	41.07	76.71	240.00	15.39	98.84	50.07	80.14	211.75	16.50
ChatGLM3-6B	Zero-Shot	53.37	2.15	50.99	1431.62	5.35	60.16	1.86	51.22	1330.71	5.34
	Few-Shot	63.60	5.55	59.36	790.26	7.49	78.40	5.13	59.40	586.80	7.82
	P-Tuning	81.23	12.54	61.47	350.96	10.17	89.32	13.86	66.27	200.85	11.15
	Fine-Tuning	97.17	39.88	76.25	257.86	15.19	99.00	48.09	79.30	212.43	16.29
GLM3	Zero-Shot	65.10	11.65	62.70	1259.66	9.03	73.77	13.62	63.86	836.85	9.88
	Few-Shot	62.03	12.97	63.60	1611.31	9.12	72.85	15.94	64.92	990.60	10.22
GLM4	Zero-Shot	71.43	13.17	63.98	764.82	9.78	78.00	16.67	65.43	701.05	10.67
	Few-Shot	75.67	16.08	65.46	609.22	10.56	86.85	19.52	66.89	340.84	11.81
ERNIE-4.0	Zero-Shot	77.07	23.53	67.77	826.06	11.63	87.54	27.60	70.40	260.83	13.22
	Few-Shot	75.57	25.67	68.12	818.77	11.85	87.46	32.45	72.78	348.34	13.71
Llama-2-7B	Zero-Shot	61.54	1.64	50.71	965.67	5.22	68.59	2.06	57.64	1058.37	5.85
	Few-Shot	67.82	3.19	52.43	840.31	6.41	70.32	4.63	58.01	856.41	7.27
GPT4	Zero-Shot	76.10	5.73	59.53	516.85	8.03	76.62	5.41	59.72	433.87	7.99
	Few-Shot	76.70	6.77	60.42	606.75	8.36	82.31	7.05	60.62	507.14	8.67

而,实例分析显示,在稀少诗词场景下,LSTM 模型倾向于生成带有常用词的非自然句子。例如,当参考句子为“劳神苦思,代百司之职役”时,LSTM 模型的输出为“不而不,,而,后”。尽管从 PPL 值来看,生成诗词的 1.05 远低于参考诗词的 163.32,但从内容层面来看,生成诗词的流畅性并不理想。这表明自动评测指标如 PPL 在评估流畅性时的局限性,进一步强调了人工评测在评估生成诗词质量方面的重要性。

此外,为了评估大语言模型在处理该任务时的性能表现,本文分别进行了零样本提示学习和少样

本示例学习的实验。在开源的 ChatGLM 系列及 Llama-2-7B 等大语言模型上,相较于缺乏大语言模型基础能力的 LSTM 模型,BLEU 和 BERTScore 指标均显著下降,表明当前主流的开源模型对诗词引用任务的适应性有限。然而,尽管国外商用的 GPT4 在其他自然语言处理任务上超越了开源的 ChatGLM 系列,但其在诗词引用任务上的性能与 ChatGLM 系列相近。相较于 GPT4 和 ChatGLM 系列,国内商用的 ERNIE-4.0 和 GLM4 在诗词引用任务上的性能显著提升。这表明,相比商用模型,开源模型的性能较弱,可能是因为其参数规模不足。

GPT4 相对 ERNIE-4.0 和 GLM4 的性能较弱,可能是由于 GPT4 对中文的学习不足。另一方面,少样本示例学习方法的性能要略优于零样本提示学习方法。

除了提示学习,本文还在开源的 ChatGLM 系列模型上进行了两种微调实验,以验证本文所构建数据集的有效性。相较于无训练的 Zero-Shot 和 Few-Shot 方法,参数高效微调 P-Tuning 在 ChatGLM 系列模型的所有三个版本中均显著提升了性能,其表现甚至达到了与同系列最先进的商用模型 GLM4 相近的水平。此外,全量微调 Fine-Tuning 方法的性能远超商用模型 GPT4、ERNIE-4.0 和 GLM4。这些结果表明,虽然开源模型在参数规模和语言学习能力上可能比商用模型稍逊一筹,但是通过有效的微调策略,如 P-Tuning 和 Fine-Tuning,它们仍然能够在特定任务上达到甚至超越商用模型的性能。更重要的是,这些结果证明了本文构建的数据集对于提升模型在诗词引用任务上的性能起到了关键作用。

6.2 人工评测分析

如表 7 所示,人工评测结果与自动评测结果基

本一致,这在一定程度上验证了自动评测指标的有效性。在 CPQD-MNBVC 和 CPQD-Gaokao 两个任务上,ChatGLM2-6b & Fine-Tuning 在各项评测指标上均取得了最高分,这表明在 ChatGLM2-6b 模型上进行全参数微调能够充分利用训练数据,使模型更好地适应诗词引用任务。虽然 P-Tuning 在某些指标上接近 Fine-Tuning,但整体仍略逊一筹,这表明参数高效微调可以取得不错的性能提升,但全参数微调通常更优。相比之下,商用模型上的 Zero-Shot 和 Few-Shot 两种基于提示学习的方法在各项指标上的表现相对较差,尤其是在上下文连贯性和整体质量方面。这表明在诗词引用任务中,当前大语言模型仅凭预训练知识或少量样本不足以解决该任务。尽管如此,Few-Shot 在大多数指标上稍胜于 Zero-Shot,这表明提供模型一些示例数据可以提高其性能。此外,LSTM 模型在所有指标上均表现最差,这可能与诗词引用任务的复杂性以及模型本身缺少世界知识有关。

表 7 人工评测结果

基座模型	方法	CPQD-MNBVC				CPQD-Gaokao			
		诗意	上下文连贯性	流畅性	整体质量	诗意	上下文连贯性	流畅性	整体质量
—	Retrieval	3.05	2.49	3.29	2.53	3.21	2.16	3.02	2.31
LSTM	Fine-Tuning	2.09	1.87	1.90	1.74	2.43	1.98	2.15	1.94
GPT4	Few-Shot	3.66	2.13	3.86	2.05	3.99	2.62	3.91	2.62
ERNIE-4.0	Few-Shot	3.80	3.28	4.11	3.11	4.09	3.39	4.13	3.46
GLM4	Few-Shot	3.84	2.73	4.03	2.63	4.06	3.33	4.19	3.36
ChatGLM2-6B	Zero-Shot	3.36	2.13	3.44	2.34	3.52	2.10	3.36	2.32
	Few-Shot	3.46	2.65	3.54	2.75	3.87	2.56	3.68	2.77
	P-Tuning	3.43	2.73	3.55	2.84	3.97	2.84	3.80	3.05
	Fine-Tuning	4.05	3.71	4.17	3.72	4.30	3.87	4.26	3.96

6.3 评测指标相关性分析

为了分析人工评测指标与自动评测指标之间的相关性,本文采用斯皮尔曼秩相关系数(Spearman Rank Order Correlation Coefficient, SROCC)对评测指标做相关性分析,并得到如图 2 所示的评测指标相关性热力图。SROCC 的取值范围是 $[-1, 1]$,数值的绝对值越大,代表评测指标之间的相关性越高。在图 2 中,人工评测指标和自动评测指标分别用中文和英文缩写进行表示。

如图 2 所示,诗意与 ACC 之间的相关性较大,

表明在评测诗意时,本文训练得到的自动评测诗意的分类器是有效的,即分类准确率 ACC 指标具有可靠性。同时注意到,诗意与 $(\log \text{PPL})^{-1}$,诗意与流畅性、ACC 与 $(\log \text{PPL})^{-1}$ 之间的相关性都很高,这意味在自动评测和人工评测中,诗意和流畅度都是紧密相关的,这与古诗词的阅读赏析习惯是相符的。此外,上下文连贯性与 BLEU 和 BERTScore 之间的相关性较高,这表明在评测上下文连贯性时,本文采用的词汇重合度 BLEU 和语义相似度 BERTScore 指标与人类评判标准较为一致。整体质量与诗意、上下文连贯性和流畅性都具有较强的相关性。其中,

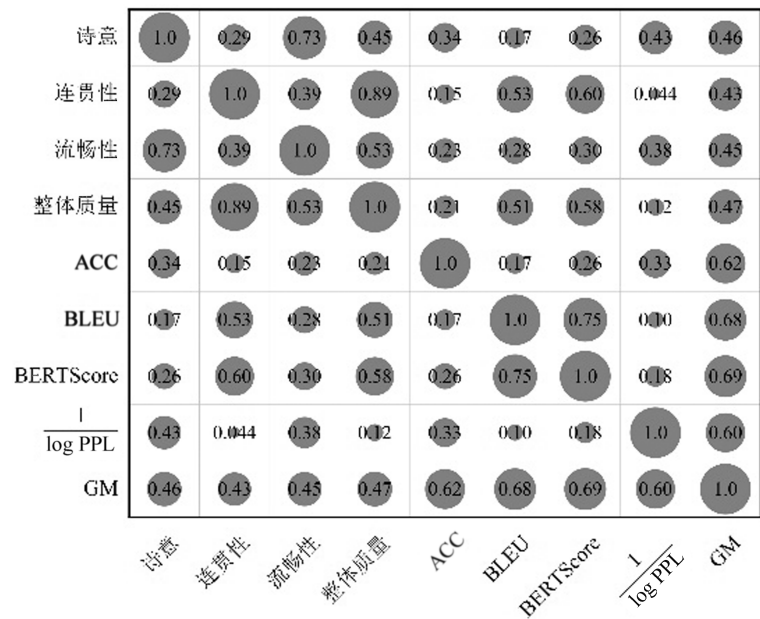


图 2 评测指标相关性热力图

与连贯性的相关性最大,其相关系数为 0.89,远大于与诗意和流畅性的相关系数 0.45 和 0.53,这表明在评估古诗词的整体质量时,连贯性是一个尤为重要的因素。评估整体质量的自动评测指标 GM 与所有指标的相关性均大于 0.4,这表明该指标在评测过程中综合考虑了诗意、上下文连贯性以及流畅性这三个方面,具有较高的综合性和全面性。

6.4 人工评测信度分析

由于人工评测的主观性及其易受评测者个体差异的影响,数据难以避免地存在评测误差,为此本文进行了人工评测的信度分析。通过运用斯皮尔曼秩相关系数(SROCC)对不同评测者之间的相关性进行探究,同时采用组内相关系数(Intraclass Correlation Coefficient, ICC)检验,来衡量相同研究对象多次观测数据的一致性。如表 8 所示,各项指标的 SROCC 和 ICC 值均呈现出较高的相关性和一致性,且显著性水平均小于 0.001,表明人工评测结果具有较高的信度。

表 8 人工评测信度分析结果

指标	相关性分析		一致性分析	
	SROCC	显著性	ICC	显著性
诗意	0.600	<0.001	0.509	<0.001
上下文连贯性	0.635	<0.001	0.639	<0.001
流畅性	0.540	<0.001	0.433	<0.001
整体质量	0.632	<0.001	0.643	<0.001

6.5 数据长尾分布的影响

本文分析了数据长尾分布可能对实验结果产生的影响。表 9 展示了在 ChatGLM2-6B 模型上进行全参数微调后,针对不同频次的诗词类别所得到的实验结果。从表 9 可知,频繁诗词在各项指标上都取得了较好的表现,而稀有诗词的表现则相对较差。这一现象可能源于数据长尾分布,即频繁诗词在训练数据中出现频率较高,使模型能更好地学习和掌握其特征;而稀有诗词则因出现次数较少,模型难以充分学习其特征。因此,在未来的研究中,对于如何平衡数据分布,提升模型对稀有类别的学习能力,是一个值得关注的方向。

表 9 长尾分布的影响

类别	CPQD-MNBVC				
	ACC /% ↑	BLEU ↑	BERTScore ↑	PPL ↓	GM ↑
频繁诗词	99.31	55.95	82.19	345.88	16.72
常见诗词	98.17	47.65	79.53	246.70	16.12
稀少诗词	93.70	21.26	69.24	208.19	12.68

6.6 基座模型版本的影响

本节对比了开源的 ChatGLM 系列模型与商用的 GLM3 和 GLM4 模型的性能表现,如图 3 所示。在图 3 中,少样本示例学习方法 Few-Shot 和全参数微调方法 Fine-Tuning 分别用 FS 和 FT 缩

写代替。首先,在少样本示例学习方法下,开源的 ChatGLM 系列模型在诗词引用任务上的性能并未随着模型版本的升级而带来明显提升,这反映了该任务的复杂性以及当前大语言模型在处理该任务时的局限性。此外, GLM3 相较于开源版本的

ChatGLM 系列在 BLEU 指标上有显著的性能提升,表明了开源与非开源模型之间的性能差距。同时,与 GLM3 相比, GLM4 的性能进一步提升,表明其在理解和生成诗词方面取得了显著的进步。

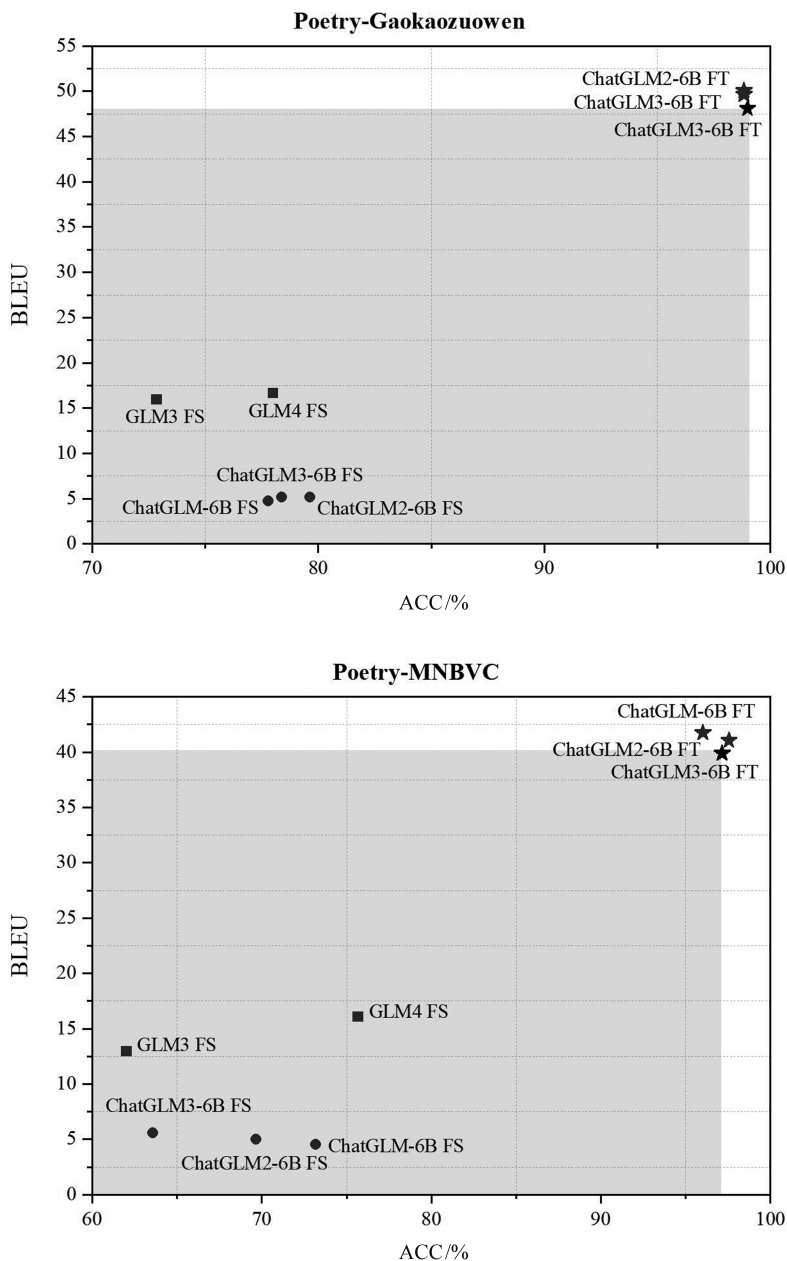


图3 不同基座模型版本对比分析

值得注意的是,对开源版本 ChatGLM 系列上进行全参数微调后的性能均显著超越了在非开源 GLM3 和 GLM4 的少样本示例学习性能。这一结果表明,针对诗词引用任务进行模型微调是提高性能的有效方法,并再次凸显了本文所构建数据的有效性以及当前大语言模型在处理该任务时的局

限性。

6.7 实例分析

表 10 展示了不同模型生成的诗词实例,并提供了每一输入对应的人类引用的诗词作为对比参考。可以清楚观察到, ChatGLM2-6B & Fine-Tuning 生

成的诗词更贴近参考诗词。例如,在实例(1)中, ChatGLM2-6B & Fine-Tuning 生成的“虽千万人吾往矣”语句通顺、语义合理,并且成功地传递了诗词的意境。相比之下, Retrieval 方法过于侧重于“庄子”这一上下文信息中,虽然通过相似度检索到了与“此训诂之学也,读庄子岂宜作如是解”相关的内容,但却忽略了“真的闯出自己的一片天地来”的这一关键信息。LSTM & Fine-Tuning 输出的“不之不,,而,后”无法形成通顺的句子,这表明其在处理稀少诗词时的生成能力不足。在少样本示例学习方法下, GPT-4 & Few-Shot 的输出内容质量明显低于

ERNIE-4.0 & Few-Shot 和 GLM4 & Few-Shot,这表明国内的大语言模型在处理中国古典诗词任务上可能更有优势。此外,在频繁诗词的场景下, LSTM 能正确理解上下文信息并生成合适的诗词,如实例(2)和实例(3)所示。然而,在稀少诗词的场景下, LSTM 倾向于生成包含常用字但非自然的句子,如实例(1)和实例(4)所示。相较而言,基于大语言模型的方法在生成诗词的流畅度方面表现出更好的性能,能够生成更自然的诗词,这表明了大语言模型所学习的通用知识在处理诗词引用任务上的关键作用。

表 10 不同模型生成的实例对比

模型 & 方法	(1)阿就读庄子啊!“得意时是儒家,失意时是道家”,除非你刚好是那万分之一的“人生胜利组”,你或许可以“知其不可为而为之”、“#mask#”,真的闯出自己的一片天地来。但大多数时候你不是,那时候你就需要庄子了	(2)桃花盛放,姹紫嫣红,一个春的运用与陪衬,更体现出李白对友人情谊的珍惜与留恋。“#mask#”在春天里,离愁渐渐消渐远渐无形,换来的是沉甸甸的友情
Retrieval	此训诂之学也,读庄子岂宜作如是解	樱桃花下送君时,一寸春心弃折枝
LSTM & Fine-Tuning	不之不,,而,后	桃花潭水深千尺,不及汪伦送我情
GPT-4 & Few-Shot	骋望舒耆邨	白头搔更短,浑欲不胜簪
ERNIE-4.0 & Few-Shot	天下无难事,只怕有心人	桃花潭水深千尺,不及汪伦送我情
GLM4 & Few-Shot	舍生而取义者也	桃花潭水深千尺,不及汪伦送我情
ChatGLM2-6B & Few-Shot	逆境时的道家,顺境时的儒家,才是真正的智慧	桃花潭水深千尺,不及汪伦送我情
ChatGLM2-6B & Fine-Tuning	虽千万人吾往矣	桃花潭水深千尺,不及汪伦送我情
参考诗词	虽千万人吾往矣	桃花潭水深千尺,不及汪伦送我情
模型 & 方法	(3) 古语说“无规矩不成方圆”,国人自古重视规则,重视公共意识的教育,从孩子学习文字开始,就反复教育他“#mask#”,让孩子知道人和人之间“礼”之重要性	(4) 如今开始重建,无论是文化还是思想上,回是回不去的,但重建是必需的。《论语》说“#mask#”,儒家文化就是以德服人。慢慢重建吧
Retrieval	以礼规行,以德树人	福自天申凭正直,德由义集要公平
LSTM & Fine-Tuning	不学礼,无以立	君子以德,,如之德
GPT-4 & Few-Shot	礼之用,和为贵	己所不欲,勿施于人
ERNIE-4.0 & Few-Shot	不学礼,无以立	为政以德,譬如北辰,居其所而众星共之
GLM4 & Few-Shot	不知礼,无以立也	礼之用,和为贵
ChatGLM2-6B & Few-Shot	千里之行,始于足下	士不可以不弘毅,任重而道远
ChatGLM2-6B & Fine-Tuning	不学礼,无以立	远人不服,则修文德以来之
参考诗词	发乎情,止乎礼	远人不服,则修文德以来之
模型 & 方法	(5)谋事在人,成事在天,天人合一,谁奈我何呢?如是,轻装上阵,带着我们的理想,向着目标进发,离“#mask#”的美妙境地,其实就已经不再遥远	(6)可是,今天迅猛发达的信息时代,文学天空中的明月却并不皎洁。“#mask#”审视我们今天的文化生活,我们心中的宁静便注定要被打破
Retrieval	谋事在人,成事在天	人有悲欢离合,月有阴晴圆缺

续表

LSTM & Fine-Tuning	大霞与孤鹜齐飞,秋水共长天一色	海上不明月,天,云上天
GPT-4 & Few-Shot	桃花潭水深千尺,不及汪伦送我情	人生如梦,一尊还酹江月
ERNIE-4.0 & Few-Shot	乘风破浪会有时,直挂云帆济沧海	月出皎兮,佼人僚兮
GLM4 & Few-Shot	海阔凭鱼跃,天高任鸟飞	雾失楼台,月迷津渡
ChatGLM2-6B & Few-Shot	谋事在人,成事在天,人间正道是沧桑	落霞与孤鹜齐飞,秋水共长天一色
ChatGLM2-6B & Fine-Tuning	采菊东篱下,悠然见南山	明月几时有,把酒问青天
参考诗词	众里寻他千百度,那人却在灯火阑珊处	古人今人若流水,共看明月皆如此

相对而言,ChatGLM2-6B 结合 Fine-Tuning 技术,在所有生成诗词引用的方法中展现出了较优的性能。然而,即便是这一方法,在捕捉上下文深层意境方面也存在局限性。具体而言,在实例(5)中,ChatGLM2-6B & Fine-Tuning 生成的“采菊东篱下,悠然见南山”,虽然表达了一种超脱与宁静,但与上下文中的积极向前、追求目标的意境不完全吻合。而其他模型与方法,在生成与上下文意境相契合的诗词方面,同样未能达到理想效果。在实例(6)中,ChatGLM2-6B & Fine-Tuning 输出“明月几时有,把酒问青天”。虽然涉及明月元素,但更多是在表达思念,未能充分反映不同时代的人对同一事物共有的感悟,与上下文语境略显脱节。

7 结论

本文构建了大规模的中国古典诗词引用数据集,并从诗意、上下文连贯性和流畅性等多个角度,设计了自动评测指标和人工评测指标,在此基础上对大语言模型在诗词引用任务中的表现进行了全面研究。本文分别在零样本提示学习和少样本示例学习设定下,对 8 款国内外的大语言模型进行了实验与比较分析。实验发现,大语言模型在生成诗词引用方面的表现并不理想,尤其是有 7 款大模型在性能上甚至与微调后的 LSTM 模型相比还有一定的差距。为了提升大语言模型在诗词引用方面的能力,本文在所提出的诗词引用数据集 CPQD 上对开源大语言模型进行了监督微调。结果显示,微调后开源大模型在 BLEU 值上提升了近 10 倍,并显著超越了国内商用大语言模型的性能。这表明,该数据集为大语言模型的微调提供了有力支持,有效地提升了模型在文学表现力,特别是在诗词引用方面的能力。

在未来的研究中,可以沿两个方向进一步研究

诗词引用任务。一个方向是在预训练或微调阶段,注重并提升模型对中国古典诗词理解的能力;另一个方向是构建更高质量或风格特定的诗词引用数据集。目前的数据集虽然已经初步满足了研究需求,但为了进一步提升模型的性能,可以考虑构建更高质量、更为精细化的数据集,以涵盖更广泛的诗词风格和主题。

参考文献

- [1] CAO Y, LI S, LIU Y, et al. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT [J]. arXiv preprint arXiv: 2303.04226, 2023.
- [2] 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(9): 1645-87.
- [3] 万小军. 智能文本生成: 进展与挑战[J]. 大数据, 2023, 9(2): 99-109.
- [4] 周昌乐, 游维, 丁晓君. 一种宋词自动生成的遗传算法及其机器实现[J]. 软件学报, 2010, 21(3): 427-37.
- [5] ZHIPENG G, YI X, SUN M, et al. Jiuge: A human-machine collaborative chinese classical poetry generation system[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
- [6] LIU Y, LIU D, LV J. Deep poetry: A Chinese classical poetry generation system[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020.
- [7] TAN J, WAN X, XIAO J. Learning to recommend quotes for writing[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015: 2453-2459.
- [8] TAN J, WAN X, XIAO J. A neural network approach to quote recommendation in writings[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA; Association for Computing Machinery, 2016: 65-74.
- [9] LIU Y, PANG B, LIU B. Neural-based Chinese idiom

- recommendation for enhancing elegance in essay writing [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5522-5526.
- [10] WANG L, ZENG X, WONG K-F. Quotation recommendation and interpretation based on transformation from queries to quotations [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics 2021: 754-758.
- [11] WANG L, ZENG X, WONG K-F. Quotation recommendation for multi-party online conversations based on semantic and topic fusion[J]. ACM Trans Inf Syst, 2023, 41(4): Article 94.
- [12] AHN Y, LEE H, JEON H, et al. Quote recommendation for dialogs and writings [C]//Proceedings of the 3rd Workshop on New Trands on Content-Bnsed Recommender Systems, 2016: 39-42.
- [13] QI F, YANG Y, YI J, et al. QuoteR: A benchmark of quote recommendation for writing [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 336-348.
- [14] LIAO J, CHENG S, TAN M. Text polishing with chinese idiom: Task, datasets and pre-trained baselines[J]. ACM Trans Asian Low-Resour Lang Inf Process, 2023, 22(6): Article 177.
- [15] ZHENG C, HUANG M, SUN A. ChID: A large-scale Chinese idiom dataset for cloze test [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 770-787.
- [16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Comput, 1997, 9(8): 1735-1780.
- [17] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [18] 张政文. 深刻把握中华优秀传统文化同科学社会主义价值观主张的契合性[J]. 走向世界, 2023, (16): 8-9.
- [19] 袁晓丰, 邱锡鹏, 吴立德, 等. 基于短语检索和答案排序的列表问题回答方法[J]. 中文信息学报, 2008, 22(5): 74-83.
- [20] 梁健楠, 孙茂松, 吴晓沅, 等. 基于神经网络的集句诗自动生成[J]. 中文信息学报, 2019, 33(3): 126-135.
- [21] YANG C, SUN M, YI X, et al. Stylistic Chinese poetry generation via unsupervised style disentanglement [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 3960-3969.
- [22] YI X, LI R, YANG C, et al. MixPoet: Diverse poetry generation via learning controllable mixed latent space [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9450-9457.
- [23] 吴林东, 何向真, 万福成. 融合韵律特征的诗歌生成模型[J]. 计算机工程与应用, 2024, 60(13): 162-170.
- [24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
- [25] 王勇超, 周灵智, 赵亚萍, 等. 自动扩充关键词语义信息的诗歌生成算法[J]. 计算机科学与探索, 2023, 17(6): 1387-1394.
- [26] 张雄伟, 葛晓义, 孙蒙, 等. 音频隐写方法综述: 从传统到深度学习[J]. 数据采集与处理, 2023, 38(5): 995-1016.
- [27] 李宇明. 语言与人类文明 [N]. 中国社会科学报, 2021-02-09.
- [28] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [29] ZHANG Y, KANG B, HOOI B, et al. Deep long-tailed learning: A survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10795-10816.
- [30] 吴立德, 张奇, 黄萱菁. 一种新的句子相似度度量及其在文本自动摘要中的应用 [J]. 中文信息学报, 2005, 19(2): 94-100.
- [31] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems, 2014: 3104-3112.
- [32] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 138-143.
- [33] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models [J]. arXiv preprint arXiv: 2302.13971, 2023.
- [34] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 320-335.

- [35] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 1877-1901.
- [36] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [J]. arXiv preprint arXiv: 2107.13586, 2021.
- [37] WEI J, WANG X, SCHUURMANS D, et al. Chain of thought prompting elicits reasoning in large language models [J]. arXiv preprint arXiv: 2201.11903, 2022.
- [38] LIU X, JI K, FU Y, et al. P-Tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:211007602, 2021.
- [39] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
- [40] ZHANG T, KISHORE V, WU F, et al. BERTScore: Evaluating text generation with BERT[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
- [41] HEAFIELD K. KenLM: Faster and smaller language model queries[C]//Proceedings of the 6th Workshop on Statistical Machine Translation. Edinburgh, Scotland; Association for Computational Linguistics, 2011: 187-97.



胡亚豪(1995—), 博士, 主要研究领域为自然语言处理。

E-mail: huyahao@aeu.edu.cn



潘志松(1973—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理。

E-mail: panzs@nuaa.edu.cn



徐磊(2000—), 硕士, 主要研究领域为自然语言处理。

E-mail: 18751392792@163.com