基于预训练模型的中国古典诗歌生成

第一作者
工作关系/ 地址一
工作关系/ 地址二
工作关系/ 地址三
email@domain

第二作者
工作关系/ 地址一
工作关系/ 地址二
工作关系/ 地址三
email@domain

摘要

中国古典诗歌作为中国最主流的文学形式的历史持续了数千年，时至今日，仍有超过五百万创作者坚持古典诗歌创作。人工智能是否可以像人类一样创作出合格的古典诗歌？这个课题是对人工智能尝试人类艺术创作的一种重要探索，也可以帮助数量众多的创作者更好地进行古典诗歌创作。

中国古典诗歌以唐初定型的格律诗（即近体诗）为主，且现有的算力及算法只支持显式规则较多且文本长度较短的格律诗文本生成，所以本文以格律诗文本的生成为研究对象，主要尝试了BART等预训练模型，提出了FS2TEXT与RR2TEXT来生成格律诗歌文本乃至特定风格的诗歌文本，如“江西诗派风格”，“艳体诗风格”等，并解决了使用者的写作意图与生成的诗歌文本后文相关性逐渐降低的问题。

为了测试模型效果，本文选取了一些古代诗人的作品，与模型生成的作品混合，出了一套AI诗歌图灵测试题，联合一些诗词创作者与诗歌领域研究者进行了评测。参与人数超过六百，最终结果显示水平较高的诗词爱好者都无法分辨出AI的作品与人类的作品，这表明本文的模型作品与人类的作品已无明显差别。为了惠及数量庞大的诗歌爱好者群体，本文的模型已与相关诗词网站达成合作以供大家使用，希望此模型的作品能给人带来创作上的启发。

关键词： 深度学习；文本生成；预训练；BART；中国古典诗歌；格律诗；图灵测试

1 Introduction

提到中国古典诗词，许多人囿于教科书的印象，被 (王国维, 2008)“一代有一代之文学”的说法所误导，认为古典诗歌兴盛于唐宋而衰亡于明清，现在更是早已死亡。然而与平常人的认知不同，如果以诗歌数量，诗人数量来作为判断诗词是否兴盛的标准，那么可以明显看出，古典诗歌是随着时间的发展而愈发兴盛，从未中断过。以下是各个时代留存下来的诗人数量与作品数量：

时代	先唐	唐	宋	明	清	仅2020
诗人数量	约950	3369	9647	不详	不详	¿5,000,000
诗歌数量	¡10000	54685	280971	¿700,000	¿8,000,000	¿5,000,000

Table 1: 各时代诗人数量与诗歌作品数量

到了当代社会，尽管格律诗这种文学体裁已经不再有古代的地位，但由于受教育人口急剧增长，获取信息难度降低，发表渠道多样，诗词作品数量多到无法估计。据中华诗词研究院发布的报告 (马大勇and 赵郁飞, 2021)，2020年我国能统计到的诗词创作人数已达500万，仅2020一年所创作的诗词作品数量保守估计在5000万以上。

从诗人数量与诗歌作品数量来看，当今社会的诗词创作与古代相比，处于一个极度繁盛的状态，甚至还处于上升趋势，是一种非常有生命力的文体。然而由于古典诗歌所使用的语言与现代汉语存在较大差异，大多数创作者经常会在进行诗歌创作时遇到语言上的困难，空有情感

却找不到合适的字句来抒发。为了接续数千年来的诗歌传统，也为了帮助当今数百万诗词创作者进行更好地创作，探索如何使用深度学习模型生成近体诗是一个很有意义的研究。

本文的主要贡献是：

- 构造了目前最完善的中国古典诗歌数据集。
- 在此数据集上使用了BART等预训练模型，提出了两种不同的诗歌生成模型以适用于不同的场景。
- 古典诗歌生成会遇到生成的作品后面的句子与使用者的写作意图相关性逐渐降低的问题，本文较好地解决了这个问题。
- 提出相关算法，尝试生成不同风格的诗歌。
- 进行了类图灵测试，发现水平较高的诗词爱好者都无法分辨出本文训练的模型生成的作品与人类的作品。
- 探讨了AI的限制与相关的伦理问题。

2 Related research

作为人工智能的一个长期关注点，诗歌自动生成的研究可以追溯到几十年前。这一领域的第一步是基于规则和模板 (Gervás, 2001)。自20世纪90年代以来，统计机器学习方法被用来生成诗歌，如遗传算法 (Manurung, 2004)和统计机器翻译方法 (He et al., 2012)。

深度学习兴起之后，在诗歌文本生成问题上取得了巨大的优势，现有的对中文格律诗文本生成的尝试，较为成功的是清华的九歌AI (THUNLP,)以及诗三百AI (wangjiezju1988,)。九歌由清华大学自然语言处理与社会人文计算实验室开发，主要采用GRU算法与Sequence-to-Sequence模型 (Yi et al., 2018)该模型在三十余万首格律诗文本的语料库上进行训练，取得了较好的效果。

诗三百则更进一步，使用了哈尔滨工业大学开源的中文BERT模型作为预训练 (Cui et al., 2021)，进一步扩充语料库到八十余万首作品 (Werneror, 2018)，使用Sequence-to-Sequence模型进行题目到文本的生成。诗三百是目前较为知名的诗歌生成网站中效果最好的。

为了加强生成的作品后面的句子与使用者的写作意图相关性，九歌在生成每句诗歌时，保留最显著的部分，而后通过主题和保存的之前句子的信息来生成下一句诗。 (Wang et al., 2016)等人将诗歌生成分为what to say与how to say两个阶段。先通过使用者输入的写作意图，生成数个子主题（sub-topic），再通过子主题来生成每一句诗。这些研究都取得了一定成效。

至于特定风格诗歌生成方面， (Liu et al., 2018)通过计算生成的诗歌与某种特定风格诗歌作品的余弦距离，设计了一个风格匹配的奖励函数，可以生成三种特定风格的高质量诗歌。 (Yang et al., 2018)等人通过互信息该将不同风格的诗歌分离开来，并根据手动选择的风格输入生成特定风格的输出。

随着自然语言处理技术的进步，预训练模型在中国古代文学方面得到了更多的应用。Zhao Zhe等 (Zhao et al., 2019)以约80首万诗歌数据训练了gpt2-chinese-poem模型来对诗歌进行续写，王东波等 (王东波et al., 2021)以四库全书为语料训练出了SikuBERT来完成古文断句标点、命名实体识别等任务。本文所选用的BART模型 (Lewis et al., 2019)由Facebook提出，使用Transformer模型整体结构的预训练语言模型，相比于BERT等预训练模型，其在自然语言理解任务上表现没有下降，并且在自然语言生成任务上有明显的提高。

本文的研究在这些研究的基础上，先训练出一个通用的BART诗歌模型来完成seq2seq任务，通过指定诗歌中的主题词和关键词来加强整篇诗歌与使用者的写作意图相关性，并通过控制主题词和关键字来生成特定风格的诗歌，取得了不错的效果。

3 Method

3.1 BART-poem

虽然现在已经有了适用于现代汉语或古代汉语的预训练模型，但是古典诗歌的语言与现代汉语有不小的差别，而其语词的组合及词句的连缀逻辑，也与古代汉语在语用和语法上也有一些差异。因此训练出一个适用于诗歌任务的预训练模型是很有必要的。

通过对不同模型效果进行评估，本文从BERT，Roberta，T5，BART模型中选择了BART模型进行训练，该模型大小约为1.5GB，主要参数如下：

embedding	feedforward	hidden	heads	layers	dropout	encoder&decoder
1024	4096	1024	16	12	0.1	transformer

Table 2: BART模型参数

本文使用目 (Zhao et al., 2019)开源的UER.py项进行模型的训练，首先将诗歌语料库中出现次数大于等于100的汉字加到词表中，设置序列长度为64，指定data processor为bart模式来对数据进行预处理。然后设定batch size为64，span max length为3，训练60000步。最后模型的准确率稳定在0.91，loss约为0.50。我们将此模型命名为BART-poem。

3.2 Data process

3.2.1 Theme words extraction

首先将诗歌文本用清华大学开发的THULAC[13]进行分词，然后将在停用词表中的词语去除，余下的词语使用TF-IDF算法提取出主题词。TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF是词频(Term Frequency)，IDF是逆文本频率指数(Inverse Document Frequency)。

$$TF = \frac{The number of times the word appears in the article}{Total words of the article}$$

$$IDF = log(\frac{Total number of corpus articles}{Number of articles containing this isword + 1})$$

$$TF - IDF = TF \times IDF$$

TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

本文利用该算法提取主题词，每篇诗歌提取的文本主题词的数量是诗歌文本长度的1/12。

3.2.2 Key chars extraction

本文注意到了古典诗歌领域中的“诗眼”的概念，如果诗歌中某个字是上下文的中心，其它字的意思围绕着这个字展开，那么这个字就是诗眼，本文视作关键字。

本文选择使用Shen Li等人在四库全书语料上使用word2vec算法训练得到的字向量 (Li et al., 2018)，将诗歌正文去除停用词后，余下的字转化为字向量进行表示，通过计算确定这些字向量的中心点，找到离中心点欧氏距离最近的数个向量，这些向量对应的字符即为诗歌的关键字。每篇诗歌提取的关键字的数量是诗歌文本长度的1/10。

3.2.3 Genre judgment

作为本文研究对象的格律诗分为4种体裁，它们各自的句子长度，句子数量都不相同。

- 五言绝句：每句五字，共四句。
- 七言绝句：每句七字，共四句。
- 五言律诗：每句五字，共八句。
- 七言律诗：每句七字，共八句。

这四种体裁都是偶数句押韵，第一句可押韵也可不押韵。由于原始数据集中存在各种体裁的格律诗与非格律诗的诗歌，本文根据字数与押韵情况提取出其中的格律诗，并标明是什么体裁。

3.3 FS2TEXT

3.3.1 Overview

过往的许多诗歌生成模型将诗歌生成任务视为“标题到文本的映射”，然而诗歌的标题与诗歌内容有时不存在明确的对应关系，如著名的“落花诗”，目前存在至少千首以“落花”为题的诗歌，虽标题一致，它们的内容却是各不相同，因此诗歌题目到诗歌文本缺少合适的映射关系。

于是本文决定将主要映射关系变为几乎是一对一映射的“首句到全诗”。为了解决作品后面的句子与使用者的写作意图相关性逐渐降低的问题，本文使用主题词和关键字来控制诗歌的生成过程与特定风格作品的生成。

3.3.2 FS2TEXT model structure

在已经训练好的BART-poem模型的基础上执行seq2seq任务。输入是“每篇诗歌的首句这篇诗歌随机数量的主题词这篇诗歌随机数量的关键字”，输出是诗歌全文。

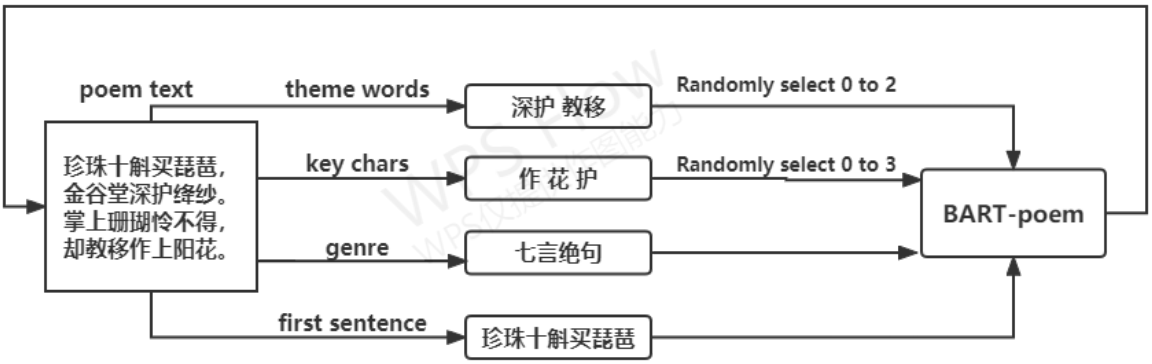

Figure 1: FS2TEXT

3.3.3 Generation of specific style poetry

如果想要生成特定风格的诗歌，那么需要构造只包含特定风格诗歌的数据集，提取出所有主题词和关键字，并在已训练好的FS2TEXT模型基础上，使用只包含特定风格诗歌的数据集中的主题词与关键字，与体裁和首句一起输入到模型中，得到一首特定风格的诗歌。

3.4 RR2TEXT

3.4.1 Overview

次韵是一种使用与目标作品相同的体裁与韵脚进行创作的行为，在诗词创作中非常流行。杜甫的《秋兴八首》，王士的《秋柳四章》，黄景仁的《绮怀十六首》等诗都被后人连篇累牍地次韵过。

我们可以将次韵这种行为抽象为：创作出与原作风格相似，且韵脚与体裁相同的作品。次韵的作品首句不宜与原作相同，故本文使用韵脚来生成次韵的诗歌作品，使用原诗的关键字与主题词来控制次韵的作品与原作风格相似。

3.4.2 FS2TEXT model structure

在已经训练好的BART-poem模型的基础上执行seq2seq任务。输入是“每篇诗歌的韵脚这篇诗歌随机数量的主题词这篇诗歌随机数量的关键字”，输出是诗歌全文。

由于次韵的作品一般与原作的风格和内容是相似的，所以我们可以通过输入与原作相同的主题词和关键字来实现对生成作品的控制，而大多数次韵的作品与原作的内容不会相似度特别高，所以本文通过在执行生成任务时也会只输入部分原作的主题词和关键字来确保生成的作品不会与原作高度相似。

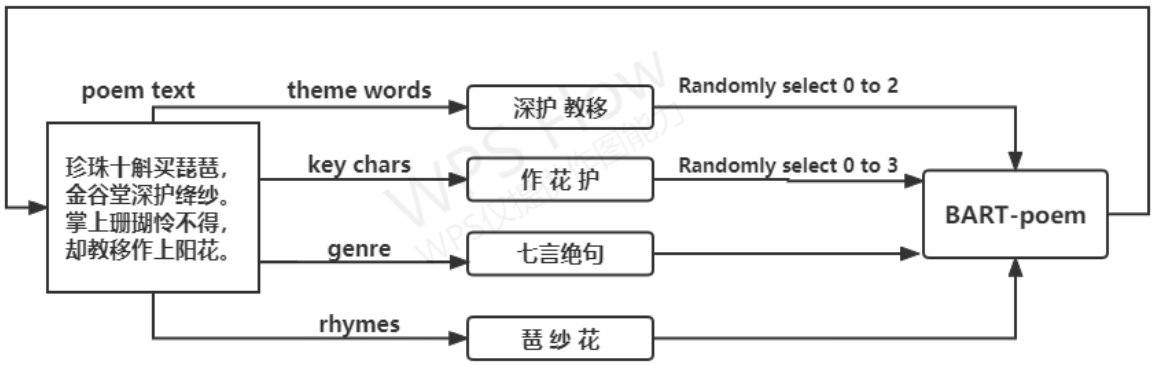

Figure 2: RR2TEXT

4 Experiments

4.1 Dataset

本项目构建了目前公开的最完善的诗歌数据集，将各个时代的诗歌作品都放在一个csv文件中，共约120万篇，分为“题目”、“朝代”、“作者”、“内容”、”关键字“，”主题词“六个字段进行储存。

题目	时代	作者	诗歌文本	关键字	主题词
失题	当代	杜随	后会何须约，前尘自可忘。一时同梦寐，万古各参商。	时—前	前尘—参商

Table 3: 数据存储示例

- GitHub用户Werneror (2018)所开源的的项目收录了从先秦到现代的约80万首古诗词。古诗词数据按朝代存储于多个csv文件中，有“题目”、“朝代”、“作者”和“内容”共四个字段。诗词语料中有一些生僻字无法显示，故而使用“?“来替代。
- Werneror (2018)项目缺少许多明清诗人的作品。本文通过网络爬虫，从各种诗词网站上搜集了许多金元明清诗人的作品。
- 由于某些少见的诗词资料，网上没有公开的数字资源。本文通过古籍PDF文件人工录入，将其变成方便处理的文字资料。

4.2 FS2TEXT training

在本文构造的数据集上，使用BART-poem模型进行微调执行seq2seq任务。设置序列长度为64，指定data processor为bart模式来对数据进行预处理。然后设定batch size为64，训练至loss约为2.60。此时经人工判断模型已经能生成质量尚可的诗歌，停止训练，得到FS2TEXT模型。此时指定体裁，输入首句与一定数量的关键字与主题词，即可生成对应的诗歌。使用者的意图可以通过关键字与主题词来影响生成的文本。

输入	输出
七言绝句 白鹭 烟—山 杨柳花飞芜草青	杨柳花飞芜草青 野塘烟草自凋零 一双白鹭来烟际 点破遥山数抹青

Table 4: FS2TEXT 结果

模型生成的结果符合指定体裁的平仄规律与格式，且可以看出主题词与关键字影响了文本全局。

4.3 FS2TEXT-amorous

本文选取王彦泓与孙原湘的部分作品作为艳体诗的数据集，提取出艳体诗数据集中的主题词与关键字后，在已训练好的FS2TEXT模型中指定只能使用此数据集中的主题词与关键字，即可生成艳体诗风格的作品。

输入	输出
七言律诗	相见时难别亦难，临歧无奈暂盘桓。
—	舟沿碧草同千里，人隔青天共一峦。
—	梦去不妨风浩荡，酒来犹喜月团圆。
相见时难别亦难	从今珍重琼瑶字，莫作鸳鸯万缕看。

Table 5: FS2TEXT-amorous 结果

4.4 RR2TEXT training

在BART-poem模型进行微调执行seq2seq任务，设置序列长度为64，指定data processor为bart模式来对数据进行预处理。然后设定batch size为64，训练至loss约为2.80。此时经人工判断模型已经能生成质量尚可的诗歌，停止训练，得到RR2TEXT模型。

使用者输入想要次韵的作品到程序中，程序经过处理后分析出体裁，韵脚，主题词，关键字等，输入到模型中。得到次韵作品。

输入	输出
独起凭栏对晓风	日没荒墟生晓风
满溪春水小桥东	满溪流水碧山东
始知昨夜红楼梦	不知渔父相扶醉
身在桃花万树中	独立苍茫烟雨中

Table 6: RR2TEXT 结果

4.5 test

4.5.1 test rules

为了验证模型生成的诗歌作品质量如何，本文模仿图灵测试出了一套测试题。流程如下：

- 选取文学史上重要诗人（代表人类较高水平）
- 从这些诗人的作品中选取一些不出名的作品（避免有人读过而影响实验结果）
- 根据体裁，每种体裁选4首，共16首。
- 选定后，输入首句令FS2TEXT模型生成作品，将FS2TEXT模型生成的作品与原作放在一起二选一，让人判断哪首是AI创作。

测试在小范围产生了爆炸式传播，并被“少年国故微刊”等公众号转载。最后收到616份有效回答，参与测试人员有北大复旦等名校的专业研究者，也有各大诗社社员等格律诗创作者，总体来看受测人员在诗词方面造诣明显高于常人。

4.5.2 Result analysis

根据问卷情况来看，大部分题目正确率都在50%左右波动，低于40%或高于60%的只有三题。总的来看，测试人员共做了9856次选择，其中正确了4960次，约占整体比例的50.32%，极其接近50%，因此我们得出结论，即使对于比较专业，诗词方面造诣较高的人，也很难分辨出AI的作品。

5 Conclusion and discussion

在进行类图灵测试时，有些诗词创作者看到了AI生成诗歌的水平，对有人可能会拿AI创作的诗词去参加诗赛表示忧虑。本文对此也没有太好的解决方法，唯一能做的就是呼吁九歌，诗三百这种诗歌生成网站将其模型生成的历史作品开放出来，这样就可以通过查询来判断一首诗词作品是本人写的还是这些网站的诗歌生成模型生成的。

自然语言处理领域的技术日新月异，AI生成的诗歌质量也水涨船高。本文所研究的诗歌生成模型，已经能写出高水平研究者都无法分辨的作品。当代我国诗词创作人数已达500万，但囿于从小接受的教育问题，相当一部分当代的创作者缺少语感，在字句，技巧上欠缺良多。本文的模型可以给他们以帮助，在他们遣词造句时，可以参考本模型根据他们现有的诗句生成作品，来创作自己的诗歌。诗歌的意义在于作者的思想情感，AI的诗写的再好也是没有意义的——它唯一的意义是让人看到且给人以启发。

参考文献

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Pablo Gervás. 2001. An expert system for the composition of formal spanish poetry. In *Applications and innovations in intelligent systems VIII*, pages 19–32. Springer.

Jing He, Ming Zhou, and Long Jiang. 2012. Generating chinese classical poems with statistical machine translation models. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Dayiheng Liu, Jiancheng Lv, and Yunxia Li. 2018. Generating style-specific chinese tang poetry with a simple actor-critic model. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):313–321.

Hisar Manurung. 2004. An evolutionary algorithm approach to poetry generation.

THUNLP. 九歌——人工智能诗歌写作系统. http://jiuge.thunlp.org/. Accessed May 16, 2022.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*.

wangjiezju1988. 诗三百. https://www.aichpoem.net/#/shisanbai/poem. Accessed May 16, 2022.

Werneror. 2018. Poetry.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3960–3969.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Zonghan Yang. 2018. Chinese poetry generation with a working memory model. *arXiv preprint arXiv:1809.04306*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

王东波, 刘畅, 朱子赫, 刘江峰, 胡昊天, 沈思, and 李斌. 2021. Sikubert与sikuroberta:面向数字人文的《四库全书》预训练模型构建及应用研究. 图书馆论坛, pages 1–14.

王国维. 2008. 宋元戏曲考. 上海:上海世纪出版集团.

马大勇and 赵郁飞. 2021. 中华诗词发展报告2020—诗词创作（年度创作生态概观）. 中中华诗词研究院.