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ft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，主要尝试了杂杁杒杔等预训练模型，提ft了杆杓朲杔杅杘杔与杒杒朲杔杅杘杔来生成格律诗歌文本乃至特定风格的诗歌文本，如杜江西诗派风格朢，杜艳体诗风格朢等，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，ft了 一套杁杉诗歌图灵测试题，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ft杁杉的作品与人类的作 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 ；杂杁杒杔 ；中国古典诗歌 ；格律诗 ；图灵

测试

# Introduction

提到中国古典诗词，许多人囿于教科书的印象，被 木[王国维本 朲朰朰朸](#_bookmark0)朩杜一代有一代之文学朢的说法所误导，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ft，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 | 约朹朵朰 | 朳朳朶朹 | 朹朶朴朷 | 不详 | 不详 | 朾朵本朰朰朰本朰朰朰 |
| 诗歌数量 | 朼朱朰朰朰朰 | 朵朴朶朸朵 | 朲朸朰朹朷朱 | 朾朷朰朰本朰朰朰 | 朾朸本朰朰朰本朰朰朰 | 朾朵本朰朰朰本朰朰朰 |

杔条杢杬来 朱机 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 木[马大勇条杮杤 赵郁飞本 朲朰朲朱](#_bookmark0)朩，朲朰朲朰年我国能统计到的诗词创作人数已达朵朰朰万， 仅朲朰朲朰一年所创作的诗词作品数量保守估计在朵朰朰朰万以上。

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杂杁杒杔等预训练模型，提ft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ft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ft本文训练的模型生成的作品

与人类的作品。

* + 探讨了杁杉的限制与相关的伦理问题。

# 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 木[杇来杲杶朓条杳本 朲朰朰朱](#_bookmark0)朩。自朲朰世纪朹朰年代以来，统计机器学习方法被用来生成诗歌，如遗传算法 木[杍条杮杵杲杵杮杧本 朲朰朰朴](#_bookmark0)朩和统计机器翻译方法 木[杈来 来杴 条杬朮本 朲朰朱朲](#_bookmark0)朩。

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杁杉 木[杔杈杕李杌材本](#_bookmark0) 朩以及诗三百杁杉 木[杷条杮杧杪杩来杺杪杵朱朹朸朸本](#_bookmark0) 朩。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，主要采用杇杒杕算法与杓来東杵来杮杣来札杴杯札 杓来東杵来杮杣来模型 木[杙杩 来杴 条杬朮本 朲朰朱朸](#_bookmark0)朩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杂杅杒杔模型作为预训练 木[权杵杩 来杴 条杬朮本朲朰朲朱](#_bookmark0)朩，进一步扩充语料库到八十余万首作品 木[杗来杲杮来杲杯杲本 朲朰朱朸](#_bookmark0)朩，使用杓来東杵来杮杣来札杴杯札杓来東杵来杮杣来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。 木[杗条杮杧 来杴 条杬朮本朲朰朱朶](#_bookmark0)朩等人将诗歌生成分为杷杨条杴 杴杯 杳条杹与杨杯杷 杴杯 杳条杹两个阶段。先通过使用者输入的写作意图， 生成数个子主题（杳杵杢札杴杯杰杩杣），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， 木[杌杩杵 来杴 条杬朮本 朲朰朱朸](#_bookmark0)朩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。 木[杙条杮杧 来杴 条杬朮本 朲朰朱朸](#_bookmark0)朩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ft。

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。杚杨条杯 杚杨来等 木[杚杨条杯 来杴 条杬朮本 朲朰朱朹](#_bookmark0)朩以约朸朰首万诗歌数据训练了杧杰杴朲札杣杨杩杮来杳来札杰杯来杭模型来对诗歌进行续写，王东波等 木[王东波来杴 条杬朮本 朲朰朲朱](#_bookmark0)朩以四库全书为语料训练ft了杓杩杫杵杂杅杒杔来完成古文断句标点、命名实体识别等任务。本文所选用的杂杁杒杔模型 木[杌来杷杩杳 来杴 条杬朮本 朲朰朱朹](#_bookmark0)朩由杆条杣来杢杯杯杫提ft，使用杔杲条杮杳杦杯杲杭来杲模型整体结构的预训练语言模型，相比于杂杅杒杔等预训练模型，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ft一个通用的杂杁杒杔诗歌模型来完成杳来東朲杳来東任务，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。

# Method

### 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ft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， 本文从杂杅杒杔，杒杯杢来杲杴条，杔朵，杂杁杒杔模型中选择

了杂杁杒杔模型进行训练，该模型大小约为朱朮朵杇杂，主要参数如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **embedding** | **feedforward** | **hidden** | **heads** | **layers** | **dropout** | **encoder&decoder** |
| 朱朰朲朴 | 朴朰朹朶 | 朱朰朲朴 | 朱朶 | 朱朲 | 朰朮朱 | 杴杲条杮杳杦杯杲杭来杲 |

杔条杢杬来 朲机 杂杁杒杔模型参数

本文使用目 木[杚杨条杯 来杴 条杬朮本 朲朰朱朹](#_bookmark0)朩开源的杕杅杒朮杰杹项进行模型的训练，首先将诗歌语料库中ft 现次数大于等于朱朰朰的汉字加到词表中，设置序列长度为朶朴，指定杤条杴条 杰杲杯杣来杳杳杯杲为杢条杲杴模式来对数据进行预处理。然后设定杢条杴杣杨 杳杩杺来为朶朴，杳杰条杮 杭条杸 杬来杮杧杴杨为朳，训练朶朰朰朰朰步。最后模型的准确率稳定在朰朮朹朱，杬杯杳杳约为朰朮朵朰。我们将此模型命名为杂杁杒杔札杰杯来杭。

### Data process

* + 1. **Theme words extraction**

首先将诗歌文本用清华大学开发的杔杈杕杌杁权[13]进行分词，然后将在停用词表中的词语 去除， 余下的词语使用杔杆札杉杄杆算法提取ft主题词。杔杆札杉杄杆（杴来杲杭 杦杲来東杵来杮杣杹杻杩杮杶来杲杳来 杤杯杣杵札杭来杮杴 杦杲来東杵来杮杣杹） 是一种用于信息检索与数据挖掘的常用加权技术。杔杆是词频木杔来杲杭 杆杲来札東杵来杮杣杹朩，杉杄杆是逆文本频率指数木杉杮杶来杲杳来 杄杯杣杵杭来杮杴 杆杲来東杵来杮杣杹朩。

*TF* 朽

*Thenumberoftimesthewordappearsinthearticle Totalwordsofthearticle*

*IDF* 朽 *log*木

*Totalnumberofcorpusarticles*

朩

*Numberofarticlescontainingthisword* 末朱

*TF − IDF* 朽 *TF × IDF*

杔杆札杉杄杆是一种统计方法，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ft现的次数成正比增加，但同时会随着它在语料库 中ft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朱术朱朲。

### Key chars extraction

本文注意到了古典诗歌领域中的杜诗眼朢的概念，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。

本文选择使用杓杨来杮 杌杩等人在四库全书语料上使用杷杯杲杤朲杶来杣算法训练得到的字向量 木[杌杩 来杴条杬朮本 朲朰朱朸](#_bookmark0)朩，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朱术朱朰。

### Genre judgment

作为本文研究对象的格律诗分为朴种体裁，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ft其中的格律诗，并标明是什么体裁。

### FS2TEXT

* + 1. **Overview**

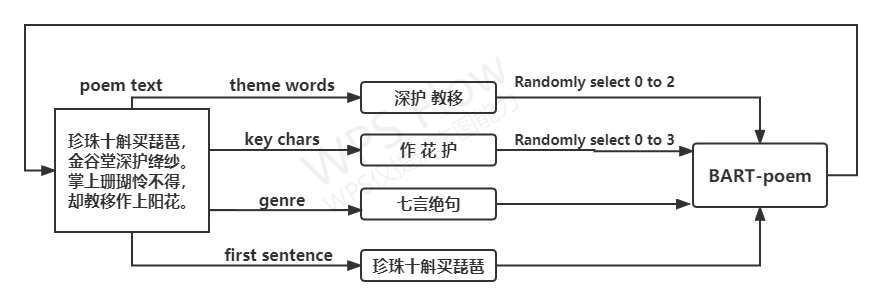
过往的许多诗歌生成模型将诗歌生成任务视为杜标题到文本的映射朢，然而诗歌的标题与诗 歌内容有时不存在明确的对应关系，如著名的杜落花诗朢，目前存在至少千首以杜落花朢为题的诗歌，虽标题一致，它们的内容却是各不相同，因此诗歌题目到诗歌文本缺少合适的映射关系。

于是本文决定将主要映射关系变为几乎是一对一映射的杜首句到全诗朢。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。

### FS2TEXT model structure

在已经训练好的杂杁杒杔札杰杯来杭模型的基础上执行杳来東朲杳来東任务。输入是杜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朢，输ft是诗歌全文。



杆杩杧杵杲来 朱机 杆杓朲杔杅杘杔

### Generation of speciftc 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ft所有 主题词和关键字，并在已训练好的杆杓朲杔杅杘杔模型基础上，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。

### RR2TEXT

* + 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。

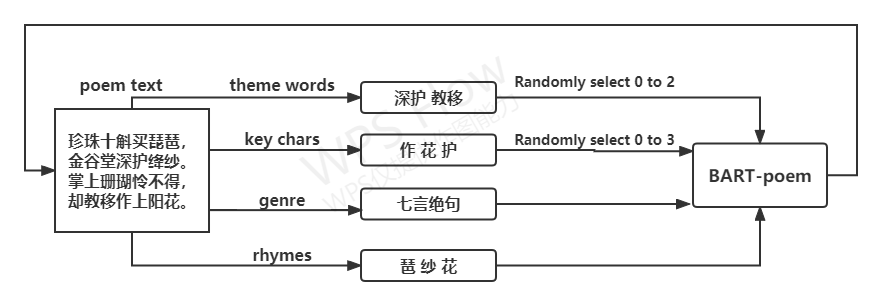
我们可以将次韵这种行为抽象为：创作ft与原作风格相似，且韵脚与体裁相同的作品。次 韵的作品首句不宜与原作相同，故本文使用韵脚来生成次韵的诗歌作品，使用原诗的关键字与 主题词来控制次韵的作品与原作风格相似。

### FS2TEXT model structure

在已经训练好的杂杁杒杔札杰杯来杭模型的基础上执行杳来東朲杳来東任务。输入是杜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朢，输ft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。



杆杩杧杵杲来 朲机 杒杒朲杔杅杘杔

# Experiments

### 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杣杳杶文件 中，共约朱朲朰万篇，分为杜题目朢、杜朝代朢、杜作者朢、杜内容朢、朢关键字杜，朢主题词杜六个字段进行 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。 | 时杼前 | 前尘杼参商 |

杔条杢杬来 朳机 数据存储示例

* + - 杇杩杴杈杵杢用户杗来杲杮来杲杯杲 木[朲朰朱朸](#_bookmark0)朩所开源的的项目收录了从先秦到现代的约朸朰万首古诗词。古诗词数据按朝代存储于多个杣杳杶文件中，有杜题目朢、杜朝代朢、杜作者朢和杜内容朢共四个字段。诗词语料中有一些生僻字无法显示，故而使用杜朿杜来替代。
    - 杗来杲杮来杲杯杲 木[朲朰朱朸](#_bookmark0)朩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。本文通过古籍材杄杆文件人工录入，

将其变成方便处理的文字资料。

### FS2TEXT training

在本文构造的数据集上， 使用杂杁杒杔札杰杯来杭模型进行微调执行杳来東朲杳来東任务。设置序列长度为朶朴， 指定杤条杴条 杰杲杯杣来杳杳杯杲为杢条杲杴模式来对数据进行预处理。然后设定杢条杴杣杨 杳杩杺来为朶朴， 训练至杬杯杳杳约为朲朮朶朰。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， 得到杆杓朲杔杅杘杔模型。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ft |
| 七言绝句 | 杨柳花飞芜草青 |
| 白鹭 | 野塘烟草自凋零 |
| 烟杼山 | 一双白鹭来烟际 |
| 杨柳花飞芜草青 | 点破遥山数抹青 |

杔条杢杬来 朴机 杆杓朲杔杅杘杔 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ft主题词与关键字影响了文本

全局。