

# 基于预训练模型的古典诗歌风格判定方法

## 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

中国古典诗歌研究领域的重要课题之一，即是分析诗人的诗歌风格。而考察历代相关论著可知，研究者评判某位诗人的诗风，多凭其自身阅读和创作经验主导下的的主观感受，并参考已成为某种定论的前人评价。这种判断方式虽往往有效，但也可能存在一些误差。本文搭建了当下最完善的中国古典诗歌数据集，在此数据集上训练了一个BART-poem预训练模型，并在BART-poem模型基础上提出一种通用的诗歌风格判定方法，创新性地深度学习引入到计量风格学领域，为古典诗歌研究提供了一种新的研究方式。本文尝试使用此方法解决唐宋诗风格识别问题，并以明前后七子，江西诗派，同光诗派等被认为具有较为明确且一致的诗风的诗歌流派为研究对象，取其代表性诗人之诗作进行测试。实验表明，该模型对所测试的诗歌作品做出的判定结果与历代评论家给出的定论基本吻合，更验证了钱锺书先生一些较为前卫的判断，较好地解决了唐宋诗风格识别的任务。

**关键词：** 深度学习；预训练模型；计量风格学；唐诗；明七子；江西诗派；同光体

## A method to judge the style of classical poetry based on pre-training model

### First Author

Affiliation / Address line 1

Affiliation / Address line 2

Affiliation / Address line 3

email@domain

### Second Author

Affiliation / Address line 1

Affiliation / Address line 2

Affiliation / Address line 3

email@domain

## Abstract

One of the important topics in the research field of Chinese classical poetry is to analyze the poet's poetic style. By examining the relevant works of previous dynasties, it can be seen that researchers judge a poet's poetic style mostly by their subjective feelings dominated by their own reading and creative experience, and refer to the previous evaluations that have become a certain conclusion. Although this judgment method is often effective, there may be some errors. This paper builds the most perfect data set of Chinese classical poetry at present, trains a BART-poem pre-training model on this data set, and puts forward a generally applicable poetry style judgment method based on this BART-poem model, innovatively introduces in-depth learning into the field of computational stylistics, and provides a new research method for the study

of classical poetry. This paper attempts to use this method to solve the problem of poetry style identification in the Tang and Song Dynasties, and takes the poetry schools that are considered to have a relatively clear and consistent poetic style, such as the Hongzheng Qizi and Jiajing Qizi, Jiangxi poetic school and Tongguang poetic school, as the research object, and takes the poems of their representative poets for testing. Experiments show that the judgment results of the tested poetry works made by the model are basically consistent with the conclusions given by critics of previous dynasties, verify some avant-garde judgments of Mr. Qian Zhongshu, and better solve the task of poetry style recognition in the Tang and Song dynasties.

**Keywords:** Deep learning , pre-training , computational stylistics , Tang poetry , Qizi of Ming dynasty , Jiangxi poetic school , Tongguang poetic school

## 1 引言

由于我国诗学传统的悠久与连贯，诗人常有师法前代的行为。宋之江西诗派，取径杜甫，遥尊其为“一祖”（方回, 1986）；明之前后七子，诗法盛唐（张廷玉, 1974）；晚清同光体，常被视为宋诗派，诗多宋调（钱仲联, 1988）。这些诗人的作品中，明确展露出了他们对某些前人诗风的学习与继承。后人若研究这些诗人，必然无法绕开对他们诗学脉络的探索。如何判定一位诗人的作品是否与前人（如学杜，学韩）或前代（如学唐，学宋）的风格相似，在古典诗歌研究领域是个相当广泛且相当重要的课题。

过往学者的研究方法，主要凭借诗歌接受史已成定论的前人论断，并结合个人对诗歌的阅读体验。这种受限于主观感觉的方法有时会使不同的研究者产生截然不同的结论，且需要学养较为深厚的研究者花费大量时间进行文本细读。

本文尝试用自然语言处理技术来解决这一问题：从计算机的角度看，诗歌风格判定的本质，是判断某首诗歌是否可以被分类到某类具有特定风格的诗歌之中去，一首诗被预测为该风格分类的概率，即可被视作它与此风格诗歌的相似度。据此，本文首先搜集了大量语料，训练出适用于诗歌文本相关任务的BART模型，而后将诗风判定任务转化为一个基于预训练模型的分类任务，选取合适训练集对模型进行微调，最终得到某种特定风格的诗风判定模型。

为了验证该方法的有效性，本文使用此方法训练出了唐宋诗风格判定模型，该模型可以判定一首七言律诗更贴近唐诗的风格还是宋诗的风格。该模型在测试过程中表现优异，学唐的明前后七子代表人物之诗作与唐诗相似度极高，而宋诗的典型代表江西诗派之诗以及学宋的同光体诗歌与唐诗相似度极低，与宋诗相似度极高。模型的这些结果与过往古典诗歌研究领域的定论是一致的，这不仅证明了本文解决了唐诗风格判定的问题，也说明本文提出的诗歌风格判定的方法是有效的。

本文的主要贡献是：

- 1.通过书本、网络等渠道搜集了大量中国古典诗歌的语料。在本文作者进行语料收集工作前，最完备的公开诗歌语料数据集约收录了80万首诗歌作品（Werneror, 2018），本文建立的语料库收录诗歌约120万首，且提取出了每首诗歌的主题词和关键字，为研究者提供了更为丰富的数据集。

- 2.基于完善后的诗歌语料，本文训练了一个适用于中国古典诗歌相关任务的BART-poem预训练模型，研究者可在此模型的基础上进行微调来实现诗句补全，诗歌生成等任务。

- 3.本文提出了一种摒弃主观因素的较为客观且有效的诗歌风格判定的方法，为古典诗歌领域研究者提供了更好的研究工具。

- 4.通过使用这一方法，本文训练出了一个唐宋诗风格识别模型，大体解决了唐宋诗风格判定问题，并验证了钱锺书先生“王世贞晚年学宋”的观点（钱锺书, 2012）。

## 2 相关研究

诗歌风格计算任务属于计算风格学领域的研究范畴，早在1995年，刘岩斌et al. (1997)在建立首个用计算机辅助研究古诗的系统时，便提出可以通过该系统提供的字频统计，体裁统计，平仄规律统计等功能来研究陆游的诗歌风格。而后李良炎et al. (2005)将诗歌分为豪放与婉约

二类，开发了基于词联接的诗词风格评价技术，在四百首宋词的测试集上，专家组给出的结果与程序结果的相关系数为0.495。苏劲松 (2007)在构建全宋词语料库时，将宋词风格评判问题转化为文本分类问题，使用KNN算法来识别豪放词与婉约词，分类准确率达到90%。Mu et al. (2015)等人工作与苏劲松相似，以诗歌关键词为特征，使用SVM算法对豪放词与婉约词进行分类，在六百首宋词的测试集上准确率达到了95%。

过往的研究多抽取字频，词频，主题词，关键字等特征来构建向量空间模型用以表示诗歌文本，随着自然语言处理技术的进步，文本可以通过预训练模型获得更好的表示。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。清华大学自然语言处理与社会人文计算研究中心 (2021)以约92万首诗歌数据训练了BERT-CCPoem来完成智能诗歌检索、推荐和情感分析在内的各种下游应用。本文所选用的BART模型由Lewis et al. (2019)提出的一种使用Transformer结构的预训练语言模型，该模型相比于BERT等预训练模型在自然语言理解任务上的表现没有下降，并且在自然语言生成任务上较之有明显的提高。

### 3 数据集构造

#### 3.1 数据收集

- GitHub用户Werneror (2018)所开源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。

#### 3.2 数据处理

- 首先将诗歌文本用孙茂松et al. (2016)开发的THULAC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，主题词的数量是诗歌文本长度的1/12，将得到的主题词存储到csv中。
- 本文选择使用Li et al. (2018)在四库全书语料上使用word2vec算法训练得到的字向量，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。根据篇幅长短，关键字的数量是诗歌文本长度的1/10。

#### 3.3 数据存储

经过以上的步骤，本文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，每篇作品对应文件中一行，分为“题目”、“朝代”、“作者”、“内容”、“关键字”、“主题词”六个字段进行储存。多个关键字与多个主题词之间使用空格隔开。以下是一首诗在csv文件中的存储示例：

题目	朝代	作者	内容	关键字	主题词
估客行	唐	李白	海客乘天风，将船远行役。 譬如云中鸟，一去无踪迹。	云中	天风 行役

Table 1: 数据集存储样例

由于本文构造的数据集不仅记录了诗歌的标题内容作者等基础信息，还提取了关键字和主题词以方便后来的研究者，因此相关领域的学者可以基于此数据集做一些方便的分析，如图1与图2所示，使可用词云来观察唐诗与宋诗中哪些主题词较为常见。



Figure 1: 唐诗主题词

Figure 2: 宋诗主题词

4 BART-poem模型

虽然现在已经有了数据量极其庞大的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判然有别，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。而现有的基于诗歌语料的预训练模型，数据量都要比本文搜集的语料更少，因此基于本文搜集的诗歌语料重新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

4.1 模型训练

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的汉字加到谷歌发布的中文词表中，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。

4.2 模型应用

BART-poem模型可以帮助古典诗歌研究者做许多工作。比如执行诗歌生成，诗句补全等下游任务。某些诗歌由于年代久远或遭到禁毁，流传至今的版本中，部分字句湮灭不可辨识。比如明末遗民诗人的作品，很多因为带有反清倾向而流传不易，阎尔梅的《白耄山人诗集》与陈子龙的《丙戌遗草》等作品中存在广泛的缺字现象，文本的不全给相关研究者造成了困扰。如果能补全这些缺少的字句，必然可以给研究者新的启发。陈子龙是晚明云间派的代表人物，由于清廷的禁毁，《陈子龙诗集》(陈子龙, 2006)中存在大量的缺字，这无疑影响了后人对陈子龙诗歌的研究。本文使用BART模型尝试去补全陈子龙作品中缺漏的字句，取得了不错的效果。以下是两首经模型补全前后的陈子龙作品对比，原作中缺字处用“？”代替：

第一首七绝模型将“中丞今夜？？还”补为“中丞今夜梦魂还”，不仅符合上下文逻辑，也符合这首诗的平仄规律。第二首五律模型将“前军临马邑，？？祀龙城。”补为“前军临马邑，故地祀龙城。”，不仅符合平仄规律，而且与前一句保持了对仗关系。由此可见模型的补全效果不错，可以为相关研究者提供帮助，更好地理解缺字的诗文作品。



原诗	模型补全结果
海风吹断古榆关，一线斜通万叠山。 数道角声西塞口，中丞今夜??还。	海风吹断古榆关，一线斜通万叠山。 数道角声西塞口，中丞今夜梦魂还。
六郡旧知名，三河出射声。 前军临马邑，??祀龙城。 万灶寒烟薄，千山堠火明。 汉家飞将在，绝漠好横行。	六郡旧知名，三河出射声。 前军临马邑，故地祀龙城。 万灶寒烟薄，千山堠火明。 汉家飞将在，绝漠好横行。

Table 3: 陈子龙原诗与模型补全结果对比

5 诗风判定方法

5.1 问题转化

诗歌的风格判定任务可以转化为一个文本分类问题。将某种风格的诗歌标注为类型A，第二种风格的诗歌为类型B，以此类推，将这些标注为不同类型的诗歌混合，在BART-poem预训练模型上进行迁移学习，以完成文本分类任务。迁移学习完毕后，得到包含数种风格的诗风判定模型，输入一首诗歌，该模型会返回这首诗歌被分类为每种类型的概率，如被分类为类型A的概率。此概率是一个介于0到1之间的实数，我们可以将其视作输入的诗歌与类型A风格的相似度。

5.2 特定风格数据集构造

若想要判定的风格只有一种，例如研究一首诗是否有艳体诗的风格，这种情况下我们可以收集一些艳体诗作品，标注为类型1，同时从诗歌数据集中随机抽取同等数量的作品，作为无特定风格的类型，标注为类型2作为对照，以便分类。

如果想要判定的风格有多种，比如清朝诗坛有“格调派”、“神韵派”、“性灵派”三个影响比较大的流派。我们如果想要研究一个清朝诗人受哪一个流派影响较大，可以取“格调派”的作品标注为类型1，“神韵派”为2，“性灵派”为3。

5.3 迁移学习

在BART-poem模型编码器上接两层前向神经网络，然后在构建好的数据集上进行文本分类任务。根据数据量大小，运行数个epoch即可停止。训练后的模型会返回输入的诗歌与每种风格类型的相似度，模型结构如图 3所示：

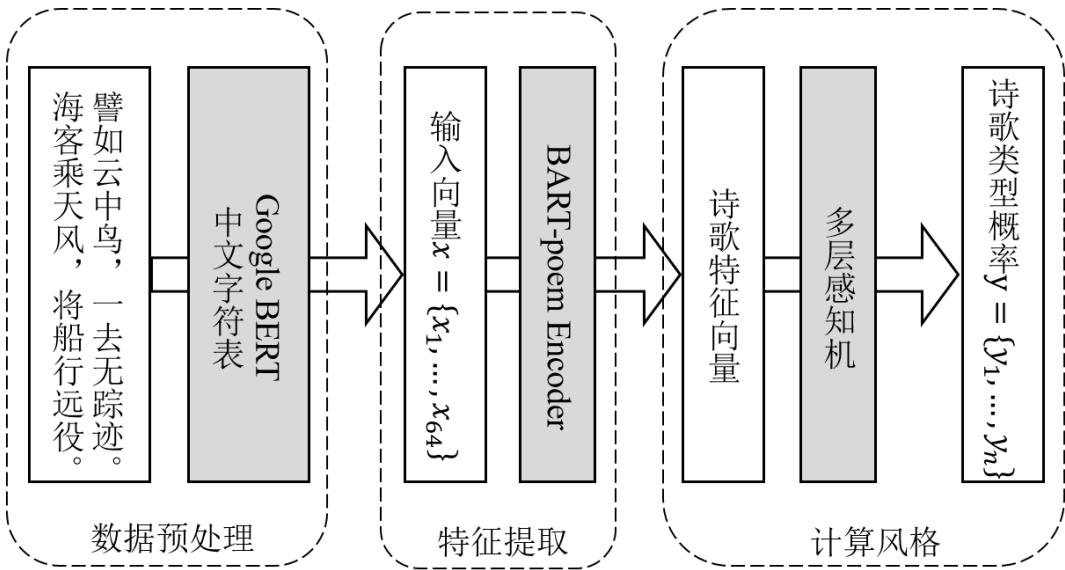


Figure 3: 风格判定模型结构

我们可以抽取某位诗人全部的作品进行判定，取全部作品的相似度平均值作为这位诗人与某种风格的相似度。

## 6 唐宋诗风格判定

### 6.1 问题背景

作为中国古代诗歌史上两个重要的时代，唐、宋皆多有能诗者。钱锺书先生认为唐宋诗之争并非时代之争，而是风格之争。由此可见，唐诗与宋诗各有鲜明的风格特点。一般认为，唐诗重意兴，宋诗重思理。钱锺书先生即指出“唐诗多以丰神情韵擅长，宋诗多以筋骨思理见胜。”(钱锺书, 2012)对于这种风格特点的不同及其论述，古代文学研究领域已从主题、意境生成、接受史等多种角度展开了论述。

唐诗宋诗正如太极之两仪，两者的风格对后世诗人产生了巨大的影响。明清诗人不学唐则学宋，余者寥寥无几。纵有王运等宗汉魏六朝者，也无法影响大势。明清两代影响力较大的诗派，几乎都把唐诗或宋诗作为学习的目标。明朝的前七子与后七子学唐，清朝的神韵派学唐，格调派学唐，秀水派学宋，同光派学宋。因此在明清诗歌研究领域，判断某位诗人的风格是接近唐诗还是接近宋诗，是一个重要且广泛的问题。

本文训练了一个模型来判断一首诗更接近唐诗的风格还是更接近宋诗的风格。由于唐宋诗体裁多样，歌行，五古，乐府，七绝，七律，五绝，五律，不同的体裁纵使是一个诗人所写，风格也不一样。于是本文选择了风格较为鲜明的七言律诗进行研究。

### 6.2 模型训练

首先将所有搜集到的唐代七言律诗抽取出来，计7763首，标注为1。又在与唐诗风格差异较大的宋代七律中抽取相同数量的作品作为对照，标注为2。将两者混合，随机抽取10%作为验证集，剩余90%作为训练集。数据集中的作品不包含之后用来测试的诗人，如黄庭坚，陈师道等人的作品。

在BART-poem模型编码器上接两层前向神经网络，设batch size 为64，在训练集上训练3个epoch后，验证集0.85，loss 0.18，停止训练并得到唐宋诗风格判定模型。该模型会对输入的诗歌文本返回其与唐诗和宋诗的相似度，二者相加为1。

### 6.3 数据分类合理性

唐宋诗的差异更多地被认为是一种风格上的差异，而非单纯的时代的差异。“唐诗、宋诗，亦非仅朝代之别，乃体格性分之殊，天下有两种人，斯分两种诗。……曰唐曰宋，特举大概而言，为称谓之便。非曰唐诗必出唐人，宋诗必出宋人也。故唐之少陵（杜甫）、昌黎（韩愈）、香山、东野，实唐人之开宋调者；宋之柯山、白石、九僧、四灵，则宋人之有唐音者。”(钱锺书, 2012)

但鉴别某首诗属于唐诗风格亦或宋诗风格是一个较为主观性的问题。由于唐代、宋代传世诗歌的数量较多，本文认为可以尊重唐代诗、宋代诗内部的相似性，规避人为判断的主观性。因此，在数据集的实际处理过程中，本文并没有剔除掉所谓“唐人的宋诗”和“宋人的唐诗”，而是直接将唐代诗与宋代诗视作“唐诗风格作品”与“宋诗风格作品”。

这种处置方式是否忽略了唐代诗或宋代诗中与“唐诗风格”或“宋诗风格”不同的作品，从而会造成较大误差？本文在训练时抽取了杜甫、韩愈等被认为“唐人之开宋调者”的诗人的部分七言律诗，未将这部分作品放入模型中进行训练。现将这部分作品使用训练后的模型进行判定。

诗人	杜甫与韩愈部分作品	其他唐代诗人未经训练作品
唐诗相似度	0.65	0.85
宋诗相似度	0.35	0.15

Table 4: 杜甫与韩愈部分作品与唐宋诗相似度

可以看出相对于其他唐人未经训练的作品，杜甫韩愈等人的作品与唐诗的相似度低了一些，但总的来看还是与整体唐代的诗歌较为相似。因此，虽然唐代或宋代有些不太像“唐诗”或“宋诗”的作品，但对整体数据集造成的影响在可接受范围内。本模型直接将唐代或宋代的诗视作“唐诗风格作品”或“宋诗风格作品”有其合理性。

### 6.4 明七子验证

宋后学唐的诗人如过江之鲫，而影响较大的有明朝的前后七子，虽然前七子是否可被视为一个一气同声的诗人群体有待商榷，但其主要代表人物李梦阳、何景明等人取法唐人是学界共识，而后七子的李攀龙等人更在诗歌创作上对唐人诗作字拟句模。由于前后七子“文必秦汉，诗必盛唐”的文学主张广为人知(张廷玉, 1974)，且在诗歌创作实践中对唐人诗作进行了较多的模仿，故本文选出前后七子中较有代表性的几位诗人的七律进行测试。

诗人	何景明	李攀龙	谢榛	王世贞
唐诗相似度	0.70	0.71	0.70	0.52
宋诗相似度	0.30	0.29	0.30	0.48

Table 5: 明七子代表人物与唐宋诗相似度

从这几位明七子的代表人物的测试结果可以看出，他们的七言律诗有很重的唐人风味，相似度平均在0.7左右。王世贞略低，但也超过0.5，根据钱锺书先生的理论：“（王世贞）於嘉靖七子，实为冠冕；言文必西汉，言诗必盛唐。《四部稿》中，莫非实大声弘之体。……《续稿》一变矜气高腔，几乎缀言之癖，刮法之痕，平直切至。屡和东坡诗韵……虽词气尚负固矜高，不肯遽示相下，而乃心则已悦服。”(钱锺书, 2012)，可以看出王世贞晚年诗染宋调，这可能是他的七律与唐诗相似度略低的原因。

### 6.5 江西诗派验证

江西诗派虽然主张学习杜甫，韩愈等人，却是典型宋代诗风的代表。根据前文测试，杜甫韩愈等人虽是唐人，作品却已经与唐诗风格有了一些差别，而学习杜甫韩愈等人的江西诗派，远袭杜甫，韩愈，近承梅尧臣，欧阳修等人，形成了与唐诗迥异的诗歌风格。我们选取了江西诗派“一祖三宗”中的“三宗”进行验证

诗人	黄庭坚	陈与义	陈师道
唐诗相似度	0.32	0.24	0.25
宋诗相似度	0.68	0.76	0.75

Table 6: 江西诗派“三宗”与唐宋诗相似度

实验结果符合现有文学史论断，作为宋诗代表人物，黄庭坚等人的七律作品与唐诗存在较大差异，与宋诗相似度高，是典型的宋代诗风。

### 6.6 同光诗派验证

同光体是近代诗派之一。主要代表人物有陈三立，郑孝胥，陈曾寿，范当世，俞明震等人。陈衍宣称此派诗人特点为“同、光以来诗人不墨守盛唐者”(陈衍, 2001)，但根据钱仲联先生的说法，同光体“远承宋代江西派而来”(钱仲联, 1988)。《光宣诗坛点将录》评价陈三立“双井（黄庭坚）风流谁得似，西江一脉（江西诗派）此传薪。”(汪国垣, 2019)，明确指出了陈三立对江西诗派尤其是黄庭坚的宗尚。郑孝胥则被称赞为“真后山（陈师道）复生也”(李晨冉, 2017)，范当世自评其诗“无意于学人，出手类苏黄（黄庭坚）”(范当世, 2006)，陈曾寿也写诗自道“深吸西江（江西诗派）得我师”(陈曾寿, 2012)。这些同光体代表诗人无一不表现出了他们对江西诗派的学习与诗歌风格的继承。也就是说，同光体诗人以学宋为主，本文选几位同光巨擘进行验证——

诗人	陈三立	郑孝胥	陈宝琛	范当世	俞明震	陈曾寿
唐诗相似度	0.27	0.31	0.25	0.27	0.36	0.29
宋诗相似度	0.73	0.69	0.75	0.73	0.64	0.71

Table 7: 同光诗派代表人物与唐宋诗相似度

实验结果与江西诗派三宗相似,由此可以看出同光体诗人学宋的主张在他们的七律创作中确实有所体现。

## 7 结论

虽然大多数诗人的诗歌作品都有自己的特点与原创性,但不可否认的是,我们如今对某位诗人进行研究仍需找出其诗学脉络,然后讨论这位诗人诗之所宗,如学杜韩、学苏黄、学江西派、学西昆体,乃至学唐,学宋。本文的意义在于寻找到了一种通用且消除了主观误差的方法,来判断一篇作品与某类风格的诗歌是否相似。为验证该方法的效果,本文使用该方法实现了唐宋诗风格判定任务,该任务为古典诗歌领域影响广泛、讨论较多的重要问题。

此方法不仅可以作为古典诗歌研究者用于寻找诗人取径的一个便捷工具,也可以与现有的研究成果互相参照。若袁枚之类标榜其诗学独树一帜,与古人无关,乃至扬言道“若问随园诗学某,三唐两宋有谁应。”(袁枚,1988),竭力掩饰自己诗学取径的诗人,研究者也可以用此工具来分析他的诗风与既往的某位诗人较为相似,然而,由于本文未能获取袁枚诗歌的全部数据,故并未进行此项研究。

当然,本文提出的方法不仅可以用来判断一位诗人的诗学取径,还可以对一个诗派,甚至一个时期的诗学创作情况进行研判。如同光诸家虽皆以宋诗为纲,但各有所本,故后人提出过“同光无体”的说法。同光体各位诗家之差异究竟如何,诸家能否算作同一个诗派,有待学者使用此方法判断出他们之间的相似性,用来辅助研究。再如竟陵派承继了公安派的“性灵说”,并对其进行了一定的修正(张啸,2019),那么竟陵派与公安派的诗歌创作之差别何在、诗风何似,借助本文的方法也可一窥究竟。

此外,为了应用这种方法进行研究,本文贡献了一个更为全面的诗歌数据集与一个适用于诗歌研究任务的BART模型,或可助益诗歌研究者在此基础上完成其工作。

## 参考文献

- 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. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.
- Yong Mu. 2015. Using keyword features to automatically classify genre of song ci poem. In *Workshop on Chinese Lexical Semantics*, pages 478–485. Springer.
- Werneror. 2018. Poetry. [EB/OL]. <https://github.com/Werneror/Poetry> Accessed May 30, 2022.
- 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.
- 刘岩斌, 俞士汶, and 孙钦善. 1997. 古诗研究的计算机支持环境的实现. *中文信息学报*, 11(1):27–36.
- 孙茂松, 陈新雄, 张开旭, 郭志, and 刘知远. 2016. Thulac: 一个高效的中文词法分析工具包.
- 孙茂松, 郭志, and 胡锦毅. 2021. Bert-ccpoem. [EB/OL]. <https://github.com/THUNLP-AIPoet/BERT-CCPoem> Accessed May 30, 2022.
- 张啸. 2019. 俚俗与幽深,同归于真情——公安派与竟陵派比较. *湖北社会科学*, (6):5.
- 张廷玉. 1974. 明史·李梦阳传. 北京: 中华书局.
- 方回. 1986. 瀛奎律髓汇评. 上海: 上海古籍出版社.
- 李晨冉. 2017. 近百年郑孝胥诗歌研究述评. *安庆师范学院学报(社会科学版)*, 36(004):18–22.
- 李良炎, 何中市, and 易勇. 2005. 基于词联接的诗词风格评价技术. *中文信息学报*, 19(6):100–106.



- 汪国垣. 2019. 汪辟疆诗学论集. 南京:南京大学出版社.
- 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.
- 苏劲松, 周昌乐, and 李翼鸿. 2007. 基于统计抽词和格律的全宋词切分语料库建立. 中文信息学报, 21(2):52–57.
- 范当世. 2006. 范伯子诗文选集. 浙江: 浙江古籍出版社.
- 袁枚. 1988. 小仓山房诗文集. 3册. 上海: 上海古籍出版社.
- 钱仲联. 1988. 近代诗坛鸟瞰. 社会科学战线, (1):267–276.
- 钱锺书. 2012. 谈艺录. 北京: 北京三联出版社.
- 陈子龙. 2006. 陈子龙诗集. 上海: 上海古籍出版社.
- 陈曾寿. 2012. 苍虬阁诗集. 上海: 上海古籍出版社.
- 陈衍. 2001. 陈石遗集. 福州: 福建人民出版社.