基于BPE分词的中国古诗主题模型及主题可控的诗歌生成

张家瑞[1,2], **李文浩**[2,4], **孙茂松**[2,3,4]
[1]清华大学电子工程系
[2]北京信息科学与技术国家研究中心
[3]清华大学人工智能研究院
[4]清华大学计算机科学与技术系
{zhangjr18, wh-li20}@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn

摘要

中国古代诗歌是人类文化的瑰宝，其短小精悍的语言却能表达出极其丰富的含义和主题，从古至今吸引了无数的爱好者的欣赏。本文以超过80万首古诗为研究对象，基于BPE算法，按照共现频率对古诗集进行分词，以便于下游任务对古诗的语义进行更准确的理解，我们还将分词后的古诗语料利用隐狄利克雷分配(LDA)模型进行了主题分析。通过比较、调整主题的数量得到了准确度较高的主题模型。更进一步，我们还对语料中的绝句和律诗逐句套用了主题模型，得到了一首诗内部的主题转移矩阵，并进行了一些相关的分析。最后，我们利用了简单的控制码方法将主题模型嵌入到诗歌生成模型中，实现了主题可控的诗歌生成，同时检验了我们训练的主题模型的有效性。

关键词： 古诗分词；主题模型；诗歌生成

Topic model and topic-controlled poetry generation of Chinese ancient poem based on BPE

Jiarui Zhang[1,2] **Wenhao Li**[2,4], **Maosong Sun**[2,3,4]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology
[3]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[4]Department of Computer Science and Technology, Tsinghua University, Beijing, China
{zhangjr18, wh-li20}@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn

Abstract

Ancient Chinese poetry is a great treasure of human culture. The short and concise language can express extremely rich meanings and themes, which has attracted countless lovers since ancient times. We took over 800,000 ancient poems as the research dataset, and trained a word segmentation model on it based on Byte-Pair Encoding algorithm, which according to the co-occurrence frequency of the ancient poetry set for word segmentation. This model can further improve a more concise understanding of the semantics in ancient Chinese poems. Furthermore, we trained a topic model on the post-segmentation ancient poetry corpus based on the Latent Dirichlet Allocation algorithm. By comparing and adjusting the number of topics, we get the more accurate topic distribution of each ancient poem. We also calculated the theme transfer matrix within a poem after annotating the topic of Jueju and Lvshi sentence by sentence. Finally, we use a simple method-Control Code to embed the topic model into the poetry generation model, in order to control the theme of our generated poem and to examine the effectiveness of our topic model.

第二十届中国计算语言学大会论文集，第862页-第873页，呼和浩特，中国，2021年8月13日至15日。
(c) 2021 中国中文信息学会计算语言学专业委员会

Keywords: poetry word segmentation , topic model , poetry generation

1 引言

1.1 研究背景

中国古代诗词是中华古代艺术文化的宝贵遗产，是古人智慧的结晶，诗词的研究和普及对中国传统文化的传播有着重要意义和深远影响。其语言高度凝练，能够表达出十分丰富的语义和情感。近些年来，随着自然语言处理技术的兴起，古典诗歌的数字化存储、分析以及生成也吸引了很多研究者的注意力，早在1996年，刘岩斌and 孙钦善 (1995) 就建立了首个用计算机辅助研究古诗的系统，而后续胡俊峰and 俞士汶 (2001)，苏劲松 (2005) 等也对自动诗歌分析进行了研究

众所周知，古诗的语言具有凝练的特征，单字所代表的含义更为丰富，故而现在大多数的研究工作都是以单字为单位进行古诗的分析与生成。但是，基于字的分析方式忽略了很多常见的惯用语，如“小桥”，“年华”，“人间”等词语，而对古诗进行分词则可以有效解决这一问题。由于现在古诗词没有整体的标注好的分词语料，故而本文使用了无监督的BPE(Byte-Pair Encoding)算法，基于共现的频率对诗中出现的词语进行分割，以词为单位对古诗进行自动分析，以期能更好地捕捉诗歌的语义，对下游任务有所襄助。

诗歌具有极强的抒情性，诗人撰写诗歌时，往往有其想表达的特定的主题和情感。故而对诗歌的主题进行分析，即是对诗歌的语义进行整体把握，对诗歌分析、生成都有着重要的意义，也能帮助我们从计算的角度对诗人写诗的意图进行分析和理解。并且，诗词中经常出现借景抒情、虚实结合等表达技巧，诗歌不同句子之间的主题也存在一定的逻辑关系。故而，我们采用了自然语言处理中最常用的隐狄利克雷分配模型，对古诗进行了主题分析，并验证了BPE分词对主题模型的影响，并对诗内的主题转移进行了建模，并进行了简要分析。

同时，近年来，随着深度学习技术在自然语言处理领域的应用以及自然语言生成领域的迅速发展，诗歌生成也成为创造性文本生成的一个重要任务。故而，我们也想以诗歌生成为任务，测试我们的主题模型的有效性，并同时开发出主题可控的诗歌生成模型。

本文的主要贡献如下：

1.基于BPE算法对数据集中的古诗进行分词切割，得到了古诗上的分词模型；

2.首次以词为单位对古诗进行主题分析，经过调优得到了一个准确度较高的主题模型；

3.创新地从计算角度分析了古诗的主题转移的问题，得到了有价值的结论；

4.通过嵌入主题改进了诗歌的生成模型，有效增强了模型的主题控制能力。

1.2 古诗的计算研究相关工作

在我国，对古诗的计算分析起源于约30年前，研究的内容主要是对语料库的构建、诗歌的自动分析、诗歌生成等。在古诗语料库的构建方面，刘岩斌and 孙钦善 (1995)建立了首个用计算机辅助研究古诗的系统，并提供了词汇、格律、风格等研究功能；穗志方et al. (1998)实现了宋诗的自动注音系统；胡俊峰and 俞士汶 (2001)深入研究了唐宋诗词的计算机辅助系统；而苏劲松et al. (2007)在全宋词语料库上进行了分词，建立了全宋词的切分语料库。在诗歌的自动分析方面，Yu) and Hu) (2003)提出了基于共现频率、互信息的唐宋诗词汇自动分析技术；李良炎et al. (2005)开发了基于词联接的诗词风格评价技术；苏劲松 (2005)利用多重松弛迭代方法，对宋词情感进行了分析。

特别地，很多诗歌自动分析和生成的研究工作也将诗歌的主题作为主要的研究对象。胡韧奋and 诸雨辰 (2015)是第一个提出对古诗进行自动分类的工作，利用向量空间模型对唐诗进行了无监督地主题聚类；刘昱彤et al. (2020)虽然主要研究的是古诗的知识图谱构建，但也将主题分类作为知识图谱的下游任务之一，并基于特征自动构建了一个主题分类的数据集。这两篇工作中的主题数分别为7和10，是较为粗粒度的主题分类，而我们的工作的主题数定为100，实现了更加细粒度的主题分析。这也是基于我们利用BPE分词更好地捕捉了诗歌的语义所完成的。

在诗歌生成方面，第一篇研究工作为He et al. (2012)于2012年发表，其利用统计机器翻译模型，对诗歌生成进行建模。而Zhang and Lapata (2014)第一次将深度神经网络引入诗歌生成中，取得了很好的效果，后续还有很多研究者(Yi et al. (2017), Li et al. (2018), Yi et al. (2018), Wang et al. (2016))等都利用不同模型对古诗生成进行了研究。而在系统方面，Zhipeng et al.

(2019)所研发的九歌系统也取得了很大的社会影响力。这其中，与我们的工作最相关的是力图生成不同风格诗歌的两篇工作，Cheng et al. (2018)通过最大化风格分布与输出分布之间的互信息，用无监督的方式生成不同风格的诗歌；而Yi et al. (2020)将诗歌风格解释为多种因素的结合，采用半监督的VAE框架，生成了语义更加丰富的诗歌。这两篇工作是主要聚焦于设计复杂机制将风格设计成因变量并进行建模。而我们在设计主体可控生成模型的同时，也意图以此对我们所训练的主题模型进行测试。故而我们采取了一种简单而灵活的方式——控制码法，进行主题控制，我们会在第5节中详述我们的方法。

2 基于BPE的古诗分词

2.1 概述

本文利用BPE算法对80余万首古诗语料库进行了基于共现频率的词汇切分，将数千个汉字的字符集拓展到了数万个词汇的词汇集。并根据分割的词表大小以及分割出的词汇含义设定了合理的词频阈值，得到了较为理想的分割结果。

2.2 BPE分词

由于古代汉语并没有公开的分词数据，我们使用了无监督的BPE算法，对古诗进行词汇切分。BPE(Byte-Pair Encoding)算法由Gage (1994)提出，是一种无监督的分词算法，最初提出是被用于语料库的压缩。近年来由Sennrich et al. (2015)用于解决机器翻译中的开放词汇(OOV)问题。近些年，很多预训练语言模型，如 [Devlin et al. (2018] 与 [Radford (2019]等，均使用BPE作为其确定词表的算法，也从侧面证明了BPE算法的有效性。

该算法统计语料库中前后出现频率最高的单词对，将其作为一个新的未出现过的单词按照原来的位置加入语料库，同时扩大词表，进行反复迭代，直到最高频率的单词对低于某一人为设定的阈值时结束迭代。我们按照词频由高到低的顺序截取了BPE算法切分出的部分词表部分词如表1所示。

词汇	出现频率	词汇	出现频率	词汇	出现频率
何处	23912	万事	5015	送归	300
万里	23904	此地	4012	金甲	200
不知	20119	银河	3001	道合	150
千里	20030	鸡犬	2010	多风	130
春风	20005	不易	1901	飞星	110
人间	18659	新月	1801	扁舟一叶	90
不见	18559	长空	1699	鞠躬尽瘁	89
不可	17534	使者	1600	观世音	80
十年	16349	谪仙	1501	思共	70
白云	16066	驱车	1400	欲逃	60

表 1: BPE分词结果及其出现频率

2.3 参数调节

为了平衡词汇数量，该算法需要设定是否进行分词的频率阈值。为了得到更好的阈值，我们分别将最低词频设定为50，100，200个进行实验，其中，最低词频为200时的词表大小为33152个，最低词频为100时的词表大小为49518个，最低词频为50时的词表大小为78490。后两者共相差28972个词，通过观察，我们发现这28972个词中的两字词更多的是因为前后两个字的出现频率都比较高，如表1中的“欲逃”、“思共”，直观上看，这些也并非常识所认为的古诗词汇。

考虑到词频较低的词汇中无意义的词汇占比过多，并且，阈值为50时会大幅扩张词表大小，进而显著提升程序运行时间，因此我们最终将阈值设定到了100次。分词之后，词汇的字数分布如表2所示。

单字词	二字词	三字词	四字词	五字及以上
15140	31747	2328	141	3

表 2: BPE分词后的词汇分布

原诗句	分词后的诗句	原诗句	分词后的诗句
小桥残月芦花	小桥 残月 芦花	百年不得逢	百年 不得 逢
秋风不解年华	秋风 不解 年华	无家不得终	无家 不得 终
酒冷心头怕醉	酒 冷 心头 怕 醉	迄今雄飞雌随十载矣	迄今 雄飞 雌 随 十载 矣
不如一盏清茶	不如 一盏 清 茶	家祭无忘告乃翁	家 祭 无忘 告 乃翁

表 3: 诗句分词结果

2.4 结果分析

分词后诗句的结果如表3所示，第一句例句中，原诗多为6个字一句，BPE较为准确的把“小桥”，“秋风”等词分割出来，而酒、冷，怕、醉等本就具有独立性的字，我们的模型没有对其进行分割，可见我们的分词模型具有较高的准确度。

但这种分词方式也有一定的局限性，例如，第二首中的雄飞在这里并不是一个词，但是“雄飞”一词也有着比喻奋发有为的含义。这是古诗本身的歧义性所导致的，解决方式只能是引入附加的排歧模块，这一点我们留待以后的工作。

3 基于LDA的古诗主题模型

3.1 概述

本章基于前一章的BPE分词结果，对古诗进行LDA主题分析，我们基于古代汉语的虚词表建立了主题分析时的停用词表，提高了模型的性能。同时，我们还对不同的主题数量进行了对比和分析，也通过对比试验证实了BPE分词对主题分析的帮助作用，最终得到了较为准确的古诗主题分析模型。

3.2 LDA主题模型

隐狄利克雷分配(Latent Dirichlet Allocation, LDA) 模型由(Blei et al. (2001))提出，是一种简单、高效的文档主题生成模型，它可以在大规模数据集上基于三层贝叶斯概率模型提取出其中隐含的主题分布。近年来，此模型广泛应用于各种自然语言处理任务，比如文本分割([石晶et al. (2008])、相似文档查找([王振振et al. (2013])、文本摘要([Xu et al. (2015])等等。

对于长度较短、语言精炼的古诗，其主题更加鲜明，判断出古诗的主题对人而言是十分简单的任务。前人将古诗的主题大致分为送别诗、边塞诗、山水田园诗、怀古诗、悼亡诗、咏物诗、军旅诗等。但一首古诗实际的主题十分复杂，可能处于两个或多个主题的交融区域，也可能不属于任一人为划分的主题范围内，且对于数十万甚至上百万的诗词逐个进行主题分析对人而言是几乎不可能完成的任务。因此，我们引入了LDA模型对诗词集进行自动的主题分析。

LDA是无监督的学习模型，学习时仅需要指定我们所需要的主题数，结合诗词的主题规模，我们尝试了将主题数设置为100，200个，对数量的分析和研究将在3.3.2中展开。

单字词						多字词					
与	止	直	于	动	止	何当	假令	倏尔	而已	尽皆	随而
元	了	至	在	加	自	毕竟	两两	往往	从此	每每	向来
再	于	总	则	互	最	常常	偶尔	依然	从而	平生	亦复

表 4: 部分停用词表(单字和多字词)

主题号	代表词
主题24	也 是 了 得 好 儿 便 来 他 人
主题30	兮 之 而 其 以 于 我 乎 有 彼
主题39	中 同 通 空 天地 有 无 自 一 外
主题43	怀 既 以 亦 与 有 志 何 所 无
主题46	长 翔 何 飞 鸣 伤 我 有 复 悲
主题49	春 人 新 尘 身 亲 频 真 客 津
主题61	老 平生 已 愧 君 我 久 犹 尚 少
主题64	之 以 与 者 其 如 为 我 此 有
主题68	为 不 民 以 其 死 之 事 人 言

表 5: 不去除停用词的主题代表字

主题号	代表词
主题15	泪 悲 魂 哭 哀 空 恨 死
主题25	兵 将军 军 边 战 马 城 将
主题29	空 英雄 地 江山 中原 兴亡 山河
主题38	梅花 梅 雪 香 寒 春 花 一枝
主题44	声 琴 弦 弹 听 音 调 曲
主题45	舞 歌 花 香 曲 玉 红 娇
主题47	别 千里 故人 远 寄 梦 忆 雁
主题59	仙 仙人 鹤 蓬莱 丹 神仙 玉 人间
主题84	船 江 岸 舟 水 风 帆 扁舟

表 6: 去除停用词后部分主题及其代表字

3.3 LDA影响因素分析

3.3.1 停用词

诗词中会有一些虚字或者虚词没有特定的主题含义，如“之”，“也”，“兮”，“哉”等词，如果将这些词引入主题模型的凝练和推理中，对其分析的准确性可能会存在一定的干扰。通过文言文的虚词表以及根据词频表的人为经验判断，我们总结得出了一个停用词表，部分停用词如表4所示，在进行主题分析时，在停用词表的词将不会放入主题模型中参与计算。为了验证停用词对主题分析的干扰情况，我们基于去除停用词和不去除停用词的两种语料，分开训练了两个主题模型，并对结果进行了对比。表5展示了不去除停用词时，部分主题概率最高的代表字，其中红色为停用词，可以发现，如“在，为，自，无”等高频词在大量主题中反复出现，对主题的分析起到了一定的干扰作用。

去除停用词后，在100个主题上选取了10个主题，表6展示了它们的代表词情况，直观上很容易发现，主题15与悲伤相关，主题25与怀古相关，主题29与边塞、战争相关，可以说他们的鲜明程度都较高。值得一提的是主题44、45都包含“曲”一字，但是他们所描述的一个是宫廷的歌舞，另一个则是高山流水，可见100个主题的模型下，主题的分割十分细腻。

3.3.2 主题数量

本章以《石灰吟》和《黄鹤楼》为例，分别利用主题数为100、200的主题模型对这两首诗进行主题分析，得出了其主题概率分布，以及置信概率最高主题的代表词，如表7所示。

石灰吟

千锤万凿出深山，烈火焚烧若等闲。
粉骨碎身浑不怕，要留清白在人间。

黄鹤楼

昔人已乘黄鹤去，此地空余黄鹤楼。
黄鹤一去不复返，白云千载空悠悠。
晴川历历汉阳树，芳草萋萋鹦鹉洲。
日暮乡关何处是？烟波江上使人愁。

我们知道，《石灰吟》本身所描述焚烧、烈火的主题在古文中并不常见，因此主题数为100时这个主题并没有被总结出来，或者说这个主题被蕴含在了概率最高的大主题中，没有被单独体现。而主题数为200时这个主题则出现了，可见主题数增加后会使主题的分割粒度更细，会生成出一些更加罕见的主题。但从《黄鹤楼》一例中可以发现，主题数量增加时，常见的主题会被切分，也可能会导致主题不准确的情况，因此需要选定一个合适的主题大小，基于主题准确度和多样性的综合考量，本文选取了100作为最后的主题数这一超参数取值。

3.3.3 BPE对主题分析的影响

为了更好地探究BPE分词对主题模型的影响，我们也同时训练了基于单字的LDA主题模型，并进行了案例研究以期进行比较，我们以《悯农》和《黄鹤楼》两首脍炙人口的诗歌为例进行对比，表格的格式和3.3.2相同。

《石灰吟》	100个主题的结果		
	排名	代表词	概率
	1	玉 石 金 珠 成 宝 磨 出 光 珍	0.5007028
	2	事 看 人间 人 好 作 不知 去 老 著	0.25199613
	3	花 香 红 艳 色 芳 春 开 露 染	0.10168867
	200个主题的结果		
	排名	代表词	概率
	1	火 热 烧 气 寒 冰 炎 炉 焚	0.3687777
	2	剑 龙 铁 光 鬼 神 飞 血 惊	0.1389110
	3	间 还 山 闲 关 颜 斑 攀 湾	0.0930812
《黄鹤楼》	100个主题的结果		
	排名	代表词	概率
	1	楚 江 湘 洞庭 秋 水 愁 楼 潇湘 州	0.793512
	2	路 去 问 语 记 旧 梦 住 愁 天涯	0.104517
	3	吾 事 乐 足 老 人 求 心 我 志	0.001831
	200个主题的结果		
	排名	代表词	概率
	1	秋 落日 晚 远 暮 望 树 愁 空 入	0.2174512
	2	江 双 降 窗 邦 缸 庞 襄阳 长江 武昌	0.2167459
	3	空 荒 古 旧 碑 废 当年 遗 千年 当时	0.1814344
	4	秋 游 流 愁 留 休 忧 收 求 浮	0.0729394

表 7: 两首诗在100个、200个主题数下的主题分析

悯农

锄禾日当午，汗滴禾下土。
谁知盘中餐，粒粒皆辛苦。

不难看出，《悯农》一诗在概率最高的主题上两种模式的表现基本一致，但在第二、第三的主题把控上分词的结果分布更加集中，但因为这首诗本身较短，分词之后这首诗包含的词汇更加有限，从一定程度上限制了它的主题分析能力。而长度较长的《黄鹤楼》一诗，基于BPE分词后语料的主题模型，前两个主题和诗歌所描述的长江有关的意象和表达的怀古幽情有很好的对应，而基于单字的主题模型所给出的概率第二高的关于春天景色的主题，和原诗主题关联程度并不大，但模型却给出了0.19的置信度，说明基于BPE的主题模型对这类诗歌主题的分析能力比起基于单字的主题模型对这类诗歌主题的分析能力比起基于BPE的主题模型还是有一定的提升。由于主题模型本身缺少标注数据，我们无法对其进行评测，在5中的结果能够印证BPE分词在主题分析上的有效性。

4 诗歌的主题转移分析

4.1 概述

诗歌不仅有一致的表达意图，更有丰富的内部结构，在整体表达诗人整体意图的同时，同一首诗内部在描述的具体主题上也可能有细微的差别，这种差别才构成了诗歌起承转合的内部结构。故而，本节拟利用上一节训练得到的主题模型，对这种主题转移在计算上进行简单的建模与分析。我们挑选了我们数据集中的50余万首律诗和绝句，绝句分为上下两联，律诗分为首颔颈尾四联逐联对其进行主题分析，并统计每首诗内部的主题转移，归一化后建立了古诗主题转移的概率矩阵，利用其建模诗歌主题转移的相关特征。我们还结合古代诗歌结构、主题分布以及因果关系等因素对其进行了分析，以期能对古诗相关的数字人文研究有一定的参考价值。

4.2 结果分析

由于古诗篇目较多，结构和内容不尽相同，我们选择了句式结构较为规整的律诗和绝句作为分析对象。我们以一联两句为一个单元，对每首诗内部的转移频度进行统计，构建一个

	基于分词的LDA结果		
	排名	代表词	概率
《悯农》	1	田 耕 村 麦 牛 雨 农 稻 熟 种	0.648160
	2	泪 悲 魂 哭 哀 空 恨 死 泣 痛	0.178415
	3	吾 事 乐 足 老 人 求 心 我 志	0.003122
	基于单字的LDA结果		
	排名	代表词	概率
	1	田 耕 农 桑 雨 种 野 麦 禾 亩	0.422501
	2	味 盘 酒 香 玉 鱼 甘 食 金 羹	0.281855
	3	我 君 相 今 见 日 时 子 知 为	0.188891
	基于分词的LDA结果		
	排名	代表词	概率
	1	楚 江 湘 洞庭 秋 水 愁 楼 潇湘 州	0.793512
	2	路 去 问 语 记 旧 梦 住 愁 天涯	0.104517
	3	吾 事 乐 足 老 人 求 心 我 志	0.001831
《黄鹤楼》	基于单字的LDA结果		
	排名	代表词	概率
	1	悠 水 怅 云 流 然 日 去 山 惆	0.305959
	2	春 花 燕 飞 莺 风 草 啼 蝶 柳	0.191542
	3	仙 蓬 云 海 莱 天 鹤 山 神 游	0.153331
	4	十 年 三 二 五 月 日 四 今 百	0.115322

表 8: 基于BPE的主题模型和基于单字的主题模型的案例对比

了100*100的矩阵A，并对其进行归一化，矩阵A定义如下：

$$A_{i,j} = \frac{N_{topic[i] \to topic[j]}}{\sum_{k=0}^{99} N_{topic[i] \to topic[k]}}$$

该矩阵的热力图如图1所示。通过观察可以发现，诗歌不同句之间有很大一部分保持了原有的主题，即图中的对角线部分，符合我们的直观感受。

表9列出了5个转移到自身概率最高和最低的主题，可以发现，转移概率最高的五个主题，即分别和战争、宫廷、花草、佛法、丰收相关的主题，鲜明度明显高于转移概率最低的五个主题，最低的5个主题中则模糊的情感成分较多。而我们也希望更好地观察不同主题之间的转换，为此，我们将矩阵的对角线置零，重新进行归一化，该矩阵的热力图如图2所示。

同时，我们还排序得到了得到了5对最容易相互转化的主题，如表10所示。我们可以发现这些概率最高的主题转移对不仅仅表现在主题相似度高，除了第一对主题的相似程度高之外，其他四对主题都能看出较为明显的因果性。比如宫廷的主题转移到报恩谢恩的主题，送别的主题转移到相思，战争的主题转移到忠诚等。值得注意的是，47号主题最容易被其他主题所转化，而47号描述的离别和思念的主题，是古代诗歌中出现最多的情感，这也与我们的认知一致。

5 融入主题模型的古诗生成

5.1 概述

在此章节中，我们基于（Radford (2019)）提出的GPT-2模型训练了一个古诗生成模型，并利用一种简单的方法-控制码(Control Code)法将我们的主题模型嵌入了诗歌生成模型中，实现了简单的主题可控的诗歌生成。我们还利用此诗歌生成模型对基于单字和BPE分词后的主题模型进行了对比。实验显示，基于主题的生成模型在困惑度(PPL)等语言建模能力指标及JS、Dist两个多样性评价指标上优于基础模型，而且生成模型也能较好的控制诗歌的主题。

5.2 模型设置

我们的基础生成模型是基于（Radford (2019)）提出的GPT-2模型，该模型基于多

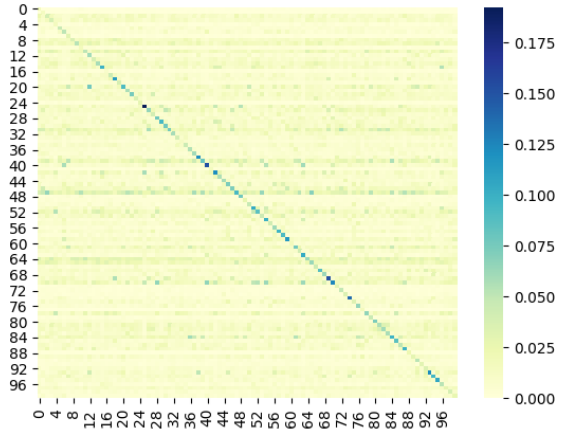

图 1: 主题转移矩阵热力图

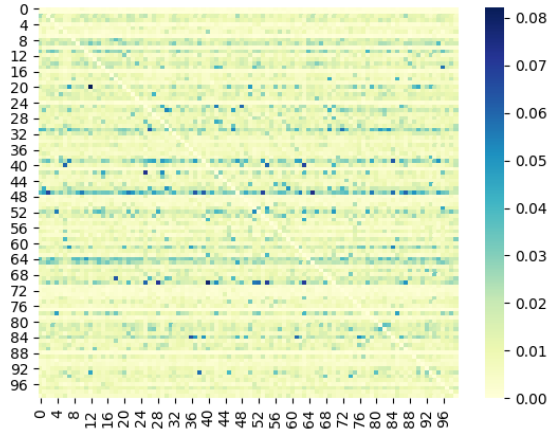

图 2: 主题转移矩阵热力图(去对角线)

主题号	代表字
主题25	兵 将军 军 边 战 马 城 将
主题40	御 殿 宫 赐 出 开 朝 下
主题93	花 香 红 艳 色 芳 春 开
主题69	僧 禅 佛 寺 法 师 经 梵
主题74	雨 望 喜 润 泽 欣 云 农
主题41	长 翔 伤 黄 光 香 乡 我
主题4	书 言 子 学 君 事 我 文
主题34	我 出 昏 行 气 不可 欲 见
主题10	君 君不见 歌 长安 黄金
主题16	悲 期 我 思 知 迟 驰 辞

表 9: 转移到自身概率最高和最低的5个主题

转移对	主题号	代表字
12 到20	主题12	春 花 东风 红 雨 春风
	主题20	春 东风 柳 绿 花 燕
40 到70	主题40	御 殿 宫 赐 出 开 朝
	主题70	重 恩 旧 归 新 望 出
65 到47	主题65	君 我 别 相逢 故人
	主题47	别 千里 故人 远 寄
25 到42	主题25	兵 将军 军 边 战 马
	主题42	国 死 臣 王 忠 大 贼
28 到70	主题28	吏 民 官 郡 政 邑 州
	主题70	重 恩 旧 归 新 望 出

表 10: 转移概率最高的5个主题对

层Transformer解码器的框架，是当前自然语言生成领域最为常用的模型。我们的模型大小设定与GPT2-base一致，在我们的80万首古代诗词语料库上进行了训练，得到了模型。此模型是在条件生成的设置下训练和测试的，即给定诗歌类型和题目，生成整首诗歌。

而在控制诗歌的主题方面，我们采用了一种简单而有效的控制方法——控制码法，即将所需要的控制指令，以字符的方式输入模型中，作为生成的条件，该方法广泛应用于Lewis et al. (2019)所提出的BART，Brown et al. (2020)所提出的GPT-3等预训练模型中。不同于他们用自然语言描述的控制码，我们对主题模型中的每个主题定义了新的字符，作为生成的条件融入模型中。经过测试，这种控制方法控制的生成模型生成的诗句有57%经过LDA模型分析后，置信度最高的主题与我们想控制的主题一致，因为共有100个主题，故而我们可以说这一方法已经达到了较为满意的控制效果。

具体的做法为，先用上一节所得到的的主题模型，以句子（格律诗以两联为一句，其他格式不规则的诗歌以除了逗号，顿号之外的其他标点符号分隔）为单位对诗歌进行切分，对每句进行主题分析得到其概率最大的一个主题，然后将该主题利用控制码的方式注入到每句最前的位置，具体的输入格式如图3所示。

5.3 具体实验

首先，我们对不带主题控制的基础生成模型(Basic)、基于单字和基于BPE的主题模型进行主体控制的生成模型(TopicGen-Char与TopicGen-BPE)进行了诗歌语言建模能力和生成诗歌多样性两方面的评测。在诗歌语言建模能力方面，我们采用了困惑度(PPL)和BLEU(Papineni et

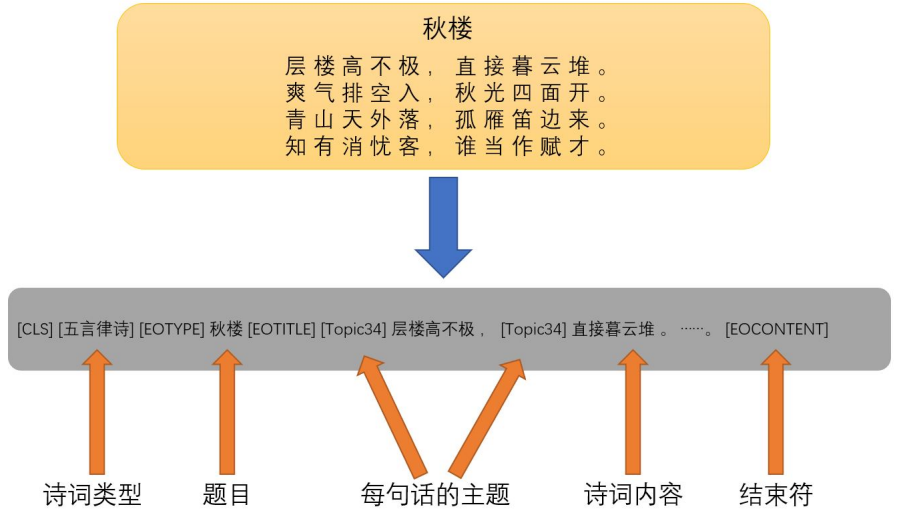

图 3: 主题可控的诗歌生成模型的输入格式

al. (2002))两个指标进行评测。困惑度指标是自然语言生成的常见测试指标，越低越说明模型更好地对训练集的语言进行了建模，其定义如下:

$$PPL(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

我们在全体测试集和测试集的格律诗上均进行了测试，得到了PPL-ALL和PPL-RP两个指标。而BLEU则是机器翻译中的常用指标，刻画了生成文本和参考文本中n-gram的重合程度。我们这里采用1-4gram的结果。而在多样性指标方面，我们采用了JS(Wang and Wan (2018))和Dist(Li et al. (2015))两个指标，JS是生成文本中任二句子Jaccard相似度的平均值，越低说明多样性越强。Dist则为生成句子中不重复的n-gram所占比例，对这两个指标，我们采用了最有区分度的2-gram的结果。这些指标的结果如表 11所示。

模型	PPL-ALL↓	PPL-RP↓	BLEU↑	BLEU-CTRL↑	JS↓	Dist↑
Basic	36.80	32.68	**0.301**	-	0.47	28.63
TopicGen-Char	32.50	29.06	0.230	**0.372**	0.37	31.03
TopicGen-BPE	**31.83**	**28.46**	0.233	0.332	**0.36**	**31.12**

表 11: 不同诗歌生成模型在测试集上的各指标结果

从困惑度指标我们可以看出，带有主题控制的模型显著的降低了生成模型的困惑度，而基于BPE主题模型的生成模型比起基于单字的困惑度又有降低。这说明基于BPE的主题模型与诗歌的潜在主题更加一致，更好地说明了BPE分词对于诗歌潜在语义捕捉等下游任务上的有效性。而在BLEU指标上，我们的模型差于基础模型，这主要是因为BLEU不仅测试了模型的语言建模能力，更测试了生成诗歌和原诗的相似程度。而我们的模型由于生成诗歌多样性更强，所以可能与原诗相似程度有所欠缺。因此我们还测试了在每句指定原诗主题情况下生成诗歌的BLEU值，即表中的BLEU-CTRL一列，可以看出，这样情况下生成的诗歌BLEU值高于原模型，展现了我们模型对生成过程中的强控制性。而在JS和Dist两个多样性指标上，主题控制的生成模型都优于基础模型，且基于BPE主题模型的生成结果均为最优，展现了其优越性。

同时，我们还想观察基于BPE主题模型的生成模型，自动生成的诗歌主题控制码，与测试集中古人的诗歌主题是否一致，因为这可以反映我们的生成模型是否很好地拟合了古人诗歌的主题转移和起承转合。故而我们将测试集中的主人诗歌输入模型，观察其预测的下一句的控制码对应的主题是否与古人的下一句诗歌一致。经过统计，其准确率为0.174，F-1分数为0.168，在100分类下准确率较为可观，可以说我们的模型也捕捉到了一定的古人主题转换，起承转合的特征。每个类别的具体F-1分数如图4所示。

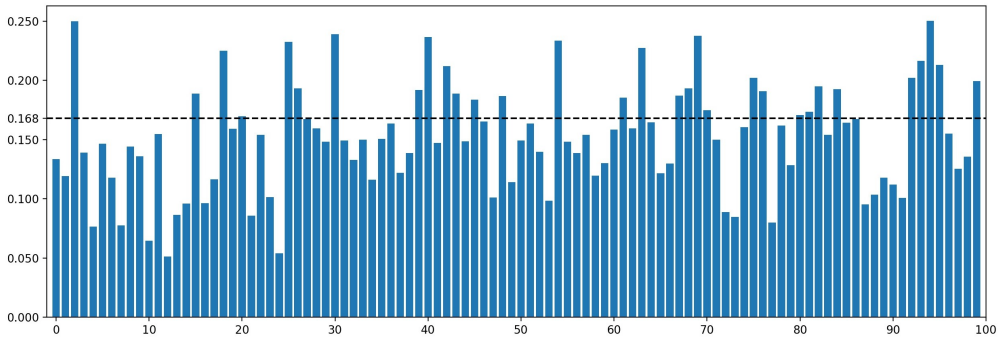

图 4: 不同主题控制码预测的F-1分数

5.4 示例展示

端州别高六戬·五言律诗			秋夜即事·七言律诗		
基础模型	主题控制	人类诗歌	基础模型	主题控制	人类诗歌
送君湘水上	相看愁万里	异壤同羁窜	风叶无声下砌墀	霜风瑟瑟雁声寒	庭前宿雨渍苍苔
惆怅惜芳时	况复异乡身	途中喜共过	独吟孤坐欲何为	一曲离歌独倚阑	闲倚阑干看斗回
春雨千峰暗	去路云随马	愁多时举酒	寒砧一倍伤乡思	明月不随乡信渺	野岸风微黄叶堕
斜阳五马驰	离情月照人	劳罢或长歌	明月双双照鬓丝	故人空作客愁看	银河星动白云开
客程猿穴近	蛮花迷瘴雨	南海风潮壮	老子此时真不觉	青毡旧物归心切	花明曲涧金铃护
归思鹭鸥迟	江树暗孤秦	西江瘴疠多	壮年他日定谁欺	白发新年往梦残	月下斜廊玉漏催
明日看山色	此夕思君泪	于焉复分手	人间万感从心起	无奈萧萧蓬户底	披草别寻松菊径
应题寄远辞	长悬粤水滨	此别伤如何	莫遣愁来入梦迟	几回惆怅拥炉叹	踏残寒露带珠来

表 12: 生成诗歌示例

两个不同模型生成和人类诗人所写的示例如表12所示。对于“端州别高六戬”这一题目，基于主题控制的生成模型生成的这首诗，在首联直接抒发异乡的客愁，颔联借云、月等意象侧面衬托离别之苦，颈联描写作者对于离乡悲凉场景的想象，尾联直抒胸臆，表达思念的惆怅，起承转合布局合理，而且对于别离之情的抒发，比起基础模型生成诗歌更浓烈，体现出了更强的扣题性。而对于“秋夜即事”这一题目，主题生成模型生成的诗歌首联借景物引入主题，颔联直接与间接描写结合，点出“客愁”这一主题，颈联更进一步，利用“青毡”“白发”等意象对比衬托出作者年华老去而归乡不得的悲凉，尾联利用连续的直接情绪抒发，将悲伤情绪烘托到顶点。而这一题目基础模型生成的诗歌，颈联直接转移到了与上下文不太相关的话题上，十分跑题，而基于主题控制的生成模型则没有此问题，显示出了对诗歌整体话题的控制能力。

6 总结与展望

本文首先对古诗文进行了基于频率的BPE分词，然后基于分词的结果进行LDA主题分析，通过观察和对比，调节主题数、停用词得到了较好的结果。将主题模型在每一首诗上逐句话进行分析，统计出了在格式相对规整的律诗和绝句上的主题转移矩阵。我们从转移到自身的概率角度对该主题的鲜明度进行了衡量，也能够通过不同主题间的转移概率大小得到主题的关系，包含相似度和因果性。最后我们将主题模型融入诗歌生成模型，得到了较为理想的生成结果。

同时，基于频率的BPE分词难免会有错漏，今后我们可能会利用自然语言生成的模型，并观察模型在每一步输出的概率分布，在输出概率明显升高时对词汇进行切分。结合这个判据和BPE算法，可以得到更加准确的分词结果。同时，我们未来还会尝试更多的古诗词主题控制方法，例如将主题模型得到的向量嵌入其中，或者利用强化学习进行更强的主题控制等等。

7 致谢

本文受科技创新2030—“新一代人工智能”重大项目（2020AAA0106502）资助，且为国家

社科基金重大项目“基于大数据技术的古代文学经典文本分析与研究（18ZDA238）”阶段成果。

参考文献

D. M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Y. Cheng, M. Sun, X Yi, and W. Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

P. Gage. 1994. A new algorithm for data compression. *The C Users Journal*.

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*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3890–3900.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

A. Radford. 2019. Language models are unsupervised multitask learners.

R. Sennrich, B. Haddow, and A. Birch. 2015. Neural machine translation of rare words with subword units. *Computer Science*.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

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*.

J. A. Xu, J. M. Liu, and K. Araki. 2015. A hybrid topic model for multi-document summarization. *IEICE Transactions on Information and Systems*, E98.D(5).

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 211–223. Springer.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153.

X. Yi, R. Li, C. Yang, W. Li, and M. Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):9450–9457.

俞士汶(Shi-Wen Yu) and 胡俊峰(Jun-Feng Hu). 2003. 唐宋诗之词汇自动分析及应用. 语言暨语言学, 4(3):631–647.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative Chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30, Florence, Italy, July. Association for Computational Linguistics.

刘岩斌and 孙钦善. 1995. 古诗研究的计算机支持环境的实现. In 中国古籍整理研究出版现代化国际会议.

刘昱彤, 吴斌, and 白婷. 2020. 古诗词图谱的构建及分析研究. 计算机研究与发展, v.57(06):132–148.

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.

王振振, 何明, and 杜永萍. 2013. 基于lda主题模型的文本相似度计算. 计算机科学, 40(012):229–232.

石晶, 胡明, 石鑫, and 戴国忠. 2008. 基于lda模型的文本分割. 计算机学报, (10):1865–1873.

穗志方, 俞士汶, and 罗凤珠. 1998. 宋代名家诗自动注音研究及系统实现. 中文信息学报, 12(2):45–54.

胡俊峰and 俞士汶. 2001. 唐宋诗之计算机辅助深层研究. 北京大学学报*(*自然科学版*)*, 37(5):727–733.

胡韧奋and 诸雨辰. 2015. 唐诗题材自动分类研究. pages 262–268.

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.

苏劲松. 2005. 全宋词语料库建设及其风格与情感分析的计算方法研究. Ph.D. thesis, 厦门大学.