

基于预训练语言模型的中文古典诗词生成

李正阳

(福建师范大学 福建 福州 350000)

【摘要】中国古典诗词承载着厚重的中华民族传统文化，因此中文古典文艺作品的生成具有非常深刻的研究意义和文化意义。由于这类作品的创作难度较大，机器自动生成不仅在一定程度上降低诗词创作的门槛，更能作为一个辅助工具来帮助创作者进行相关文学创作。当今机器在创作中文文学作品时，使机器理解中文诗词以及典籍中繁杂的各种规则也是一个难点，控制以使得模型适应于不同主题的诗词生成的更为精确的方法也仍在探索之中。本文针对各类诗词，采用一种基于 GPT-2 模型的方法进行自动生成。根据输入的提示信息自动生成诗句。在收集整合的绝句与宋词数据集上进行了相关实验，经过训练后，最终 perplexity 达到了 8.81。

【关键词】深度学习；自然语言处理；中文古典诗词；机器生成

引言

中国古典诗词是中华民族独特的文化遗产，蕴含着丰富的文化内涵和艺术价值。机器自动生成不仅在一定程度上降低诗词创作的门槛，更能作为一个辅助工具来帮助创作者进行相关文学创作。古典诗词的评价标准是难以定义的，对于一首优秀的诗词，既有复杂的格律要求，又有作者的情感倾注，使得模型理解古典诗词的意义、情感、格律以及准确地生成目标内容仍然是困难的。

早期的古诗词自动生成技术主要基于规则生成和模板匹配，虽然取得了一定的成果，但诗歌质量和创新性尚有待提高。近年来，随着深度学习的发展，自然语言处理技术在中文古诗词生成方面取得了一定的突破，尤其是循环神经网络 (RNN) 和长短时记忆网络 (LSTM)^[1] 的引入，使得中国古典诗词生成的研究逐渐从基于规则转向基于神经网络生成，但随着技术的迭代，基于神经网络的中国古典诗词的生成方法也逐渐从循环神经网络转向 Transformer^[2]，并已在生成质量、创新性和多样性方面取得了显著的改进，但诗词生成任务中仍有许多问题待以解决。

本文提出一种基于 GPT-2 的古诗词自动生成方法。在公开预训练模型 gpt2-chinese-poem^[3] 上通过古诗词语料进行微调。考虑到生成多种不同类型诗词，我们采取一种特定序列的输入方法以使得模型生成目标内容。首先对数据集进行格式化预处理，模型根据输入的格式化序列生成并最终得到诗词内容。实验通过将收集和整合的绝句和宋词数据组成新的数据集并格式化成特定格式的提示信息，

在预训练的 GPT-2 模型上进行微调，最终得到的 perplexity 达到了 8.81 和相对较好的输出效果。

一、相关工作

在早期，机器生成诗歌的方法通常基于词汇、语法规则来生成诗歌，随着统计方法在自然语言处理中的应用，研究者们尝试使用基于统计方法的诗词生成，如 Pennington 等人提出的一种基于矩阵分解的词向量表示方法 Glove^[4]，利用全局的统计信息来学习每个单词的向量表示，通过对每个单词在语料库中的共现频率进行建模，来捕捉单词之间的语义关系。Jiang^[5] 等和 Fan^[6] 等人基于统计学方法，以统计机器翻译为基础，在其中设置主题模型和统计机器翻译方法，使得系统能够自动生成符合规定格律的中文诗歌和根据用户输入的主题进行内容创作。近年来随着深度学习的兴起与快速发展，基于神经网络的方法逐渐被人们所采纳，如 Wu 等人提出一种基于循环神经网络的中文诗歌生成模型^[7]，模型能够学习中文诗歌的韵律和语法等特征，Zhang 等提出的方法与之类似^[8]，但使用了独特的训练方法和基于前缀匹配的句子生成方法，能够在保证生成符合格律要求的诗句并增加诗句的多样性和创意性。随着注意力机制的提出，使得神经网络拥有更强的处理长序列任务的能力。如 Yang 等提出了一种基于注意力机制生成具有特定主题的中文诗歌方法^[9]，论文提出的方法能够在生成时对模型的注意力进行约束，控制诗歌的主题和风格等因素。然而，随着自注意力语言模型 Transformer 以及 BERT、GPT-2 等大模型的提出，预训练语言模型

的方法因更好的效果逐渐成为主流,如Wang等提出的一种使用GPT-2模型来生成符合韵律和语义约束的中国古典诗歌的方法^[10]。上述诗歌生成方法主要与诗歌的生成效果和主题控制相关,但仍然存在一些问题,本文的方法将重点放在模型生成内容的控制当中,以更好地控制模型生成内容并得到更好的生成效果。

二、主要方法

(一) 模型

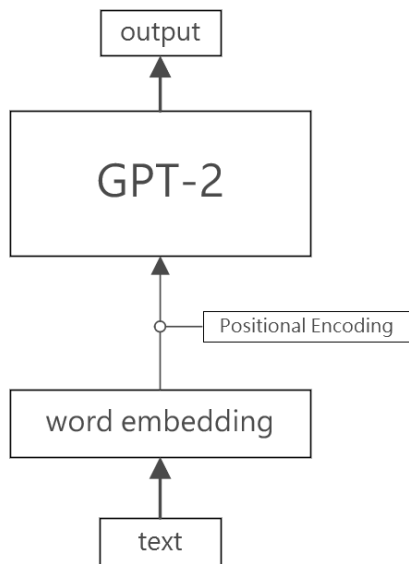


图1 模型微调流程

本文基于预训练的GPT-2模型进行古诗词生成。gpt2-chinese-poem模型基于800,000左右首中国古典诗词作为语料库进行预训练,这些数据由chinese-poetry和Poetry项目收集,预训练了200,000个步骤,在诗词语料库中出现次数大于或等于100的词将会被添加到词汇表中。图中text为经过处理后特定格式的文本数据,文本数据经过词嵌入和位置编码处理后输入到GPT-2模型中进行预测,模型输出得到预测结果,根据模型的预测内容计算损失并用于更新迭代。

(二) 数据处理

为了达到实验目的,本文方法首先对数据集中数据进行格式化,以类似于key-value的形式将这些信息进行格式化并组合成用于模型训练的提示数据。如图2所示,输入信息由两部分组成,第一部分为提示内容,第二部分则为预测内容。提示词信息的标识个数为自定义内容。这种输入格式非常灵活,可以根据不同情况设置相应的提示词。对于例如藏头诗这种特殊类型的生成,本文方法为将诗歌正文信息每句的开头第一个字提取出拼为首字

字符串组成其中一个提示词信息,可见该方法十分便捷。

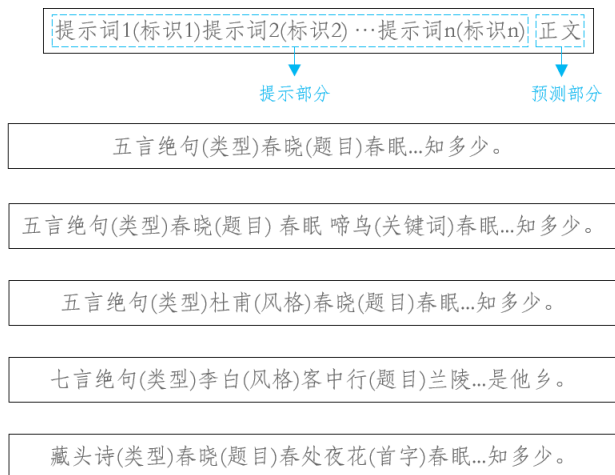


Figure 2 提示信息格式

(三) 模型训练

由于诗词的评判标准没有准确统一的定义,本文不对模型生成采取过多的限制,因此用的损失函数为交叉熵损失函数,将所有经过预处理后的数据输入到GPT-2中,模型根据输入信息输出一个概率分布,并选取概率较高的结果,我们的训练目的是最大化输入序列的概率。

$$P(X) = \sum_{i=0}^{1 \leq i \leq n} \log_p(x_i | x_1, \dots, x_{i-1})$$

需要注意的是,若迭代次数过多,模型可能会产生过拟合,此时更倾向于检索得到语料库中的原始数据。

(四) 诗词生成

当训练完成后,在给定的提示信息要求下,将数据以3.2中介绍的方法格式化后输入到模型中,模型会对剩余字符进行逐一解码,直到遇到“EOS”标记。对于符号的生成位置,我们没有使用硬约束的方法进行限定,此外,采用截断的top-k抽样策略来选择每个字,为了保证多样性和随机性,对这些概率较高的字进行随机采样,最终生成诗词。

三、实验

模型在3,0000左右首五言绝句、10,0000左右首七言绝句、13000左右首宋词上进行训练, batch_size 设定为64,总共训练10个epoch。对于五言绝句和七言绝句,序列为[类型(类型)题目(题目)关键词(关键词)正文]由于诗词的题目对正文内容信息有一定的概括作用,因此这里也

将题目作为一个提示词作为输入,对于关键词,指需要在正文中原词复现的词;对于宋词,序列为[类型(类型)题目(题目)正文]。

(一) 评价标准和实验结果

使用 perplexity 来评价所生成诗词正文的流畅性和准确性。经过实验,观察得知模型生成诗歌能力较强,但生成宋词的效果则相对较差。在训练 10 个 epoch 后,模型生成诗词得到的 perplexity 约为 8.81,而分别基于统计模型^[10]、循环神经网络、Transformer 的诗词生成模型得到的结果分别为 26.62, 23.5, 27.23, 相比之下,该模型生成的诗词有着最好的效果。

(二) 生成案例

春晓 (五言绝句)

东风昨夜至,先入客衣寒。

故园芳草绿,惆怅倚阑干。

静夜思 (七言绝句)

花影重帘映酒光,一轮明月照清漳。

当年只有如今梦,不信流年夜夜长。

一剪梅 (宋词)

(部分)

小雨乍晴寒,人在霜天晚。一声何处,

玉笛新翻,梅花初绽。春梦到江南。

模型生成的诗词在平仄、格律等表现效果较好,且生成文本质量更佳,内容有着良好的意境和美感。但由于没有进行严格的格式限制,特别是对于宋词,因此模型生成的宋词在格式上的表现欠佳,在本身有着严格格式要求的绝句诗歌数据集上表现较好。

四、总结

本文提出一种基于 GPT-2 的古诗词生成方法。

创新的将多种提示词方法组合在一起对模型进行提示,使其输出相应诗词内容,并且这种方法灵活多变,可以应用于多种场景,在定义各种类型后准确生成目标诗词文本。此外由于没有对模型的生成做出过多限制,相对其它研究进一步增强了生成文本的随机性和多样性,生成内容和特点更取决于所使用数据集本身富含的文本特征,由模型自身完成各种关系的映射和学习。实验结果表明,最终的生成结果不仅更为精确可控,也富有良好的流畅性和准确性,以及意境和美感。

参考文献:

- [1]Hochreiter,S.,&Schmidhuber,J.(1997).Longshort-term memory.Neural computation,9(8),1735-1780.
- [2]Vaswani,A.,Shazeer,N.,Parmar,N.,Uszkoreit,J.,Jones,L.,Gomez,A.N.,&Polosukhin,I.(2017).Attentionisallyouneed.InAdvancesinneuralinformationprocessing systems(pp.5998-6008).
- [3]Huggingface.(2023).Uer/gpt2-chinese-poem.[Online].Available:https://huggingface.co/uer/gpt2-chinese-poem [Accessed:24 May2023].
- [4]Pennington],SocherR,Manning C.Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1532-1543.
- [5]蒋锐滢,崔磊,何晶,等.基于主题模型和统计机器翻译方法的中文格律诗自动生成[N].计算机学报, 2015,38(12):2426-2436.
- [6]Zhang J, Gan Z, Fan K, et al. Generating Chinese classical poems with statistical machine translation models. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 465-476.
- [7]Wu, F., Li, H., Wang, J., et al. "Towards neural network-based reasoning for generating Chinese classical poems." EMNLP 2016, 2071-2076.
- [8]张新华,刘艳华,李鹏飞.基于循环神经网络的中文诗歌自动生成[N].中文信息学报,2014,28(6):12-20.
- [9]杨欣,林涛,李宇阳,肖纯.基于约束注意力机制的主题中文诗歌生成[N].中文信息学报,2018,32(6):95-102.
- [10] Wang, Yilin, et al. "Generating Chinese Classical Poems with Rhythmic and Semantic Constraints using GPT-2." arXiv preprint arXiv:2011.11491 (2020).