乐府

——预训练语言模型在诗词对联生成中的应用

华为诺亚方舟实验室

廖亿

Email: liaoyi9@huawei.com

Content

- GPT在乐府作诗、作对联的应用
- PMLM在乐府写宋词中的尝试
  - PMLM是我们发表于ACL2020的工作

基于GPT的乐府作诗、作对联

背景简介

- 中文传统诗歌对联生成：
  - 在格律、平仄、押韵方面具有严格的要求
  - 常规的诗歌对联生成模型加入规则对格式进行限制
- GPT：
  - GPT（包括GPT-2、GPT-3）是由OpenAI 推出的大规模预训练语言模型，其具备强大的文本生成能力

乐府作诗机训练流程

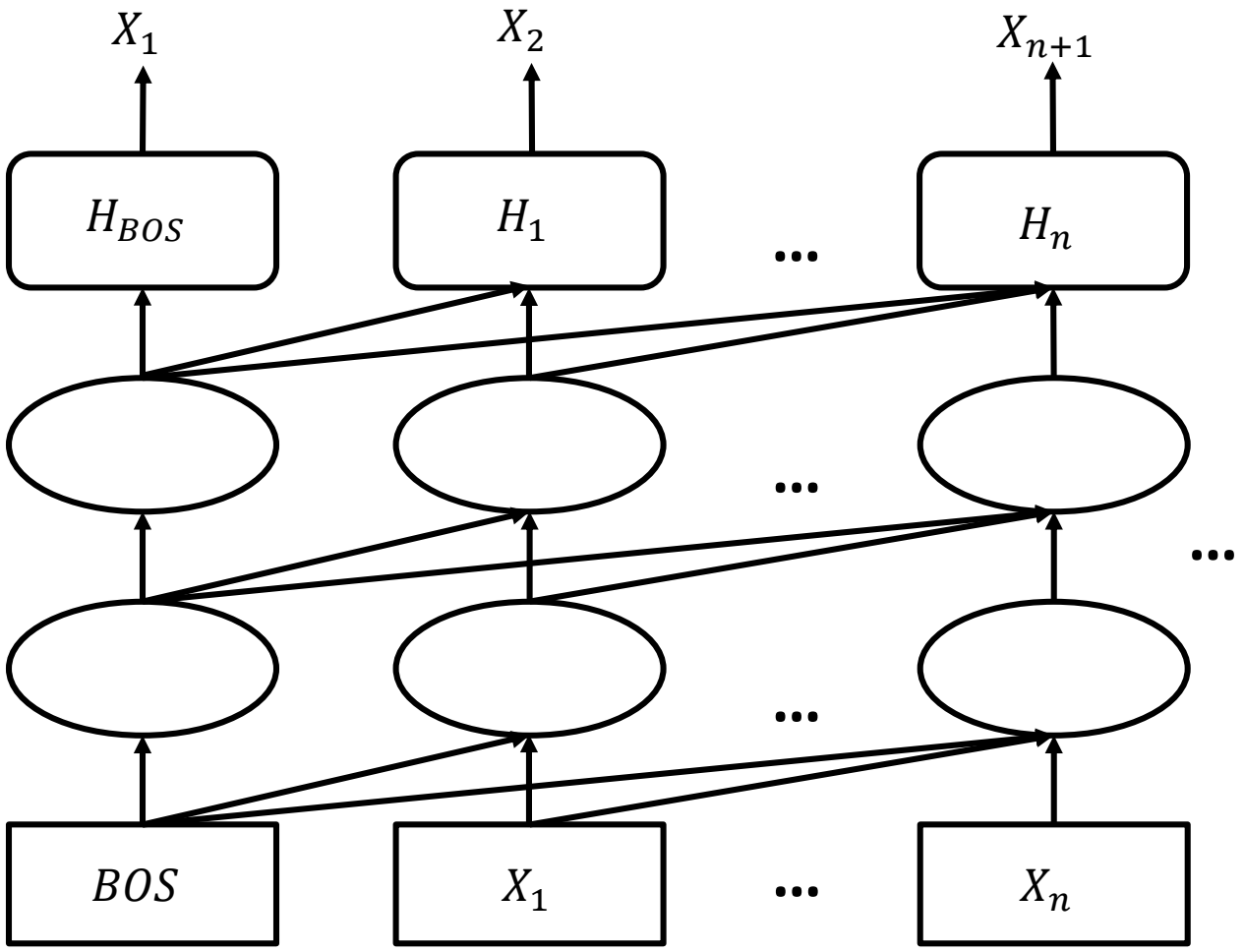
- 第一步：在大规模中文语料上预训练中文GPT模型
- 第二步：使用小规模的古诗词语料，在GPT模型上进行微调

Training Phases	Corpus type	Corpus size
Pre-training	Chinese news	235M sentences
Fine-tuning	Jueju and Lüshi Cipai Couplet	250,000 Jueju and Lüshi 20,000 Cis 700,000 pairs of couplets

训练语料统计

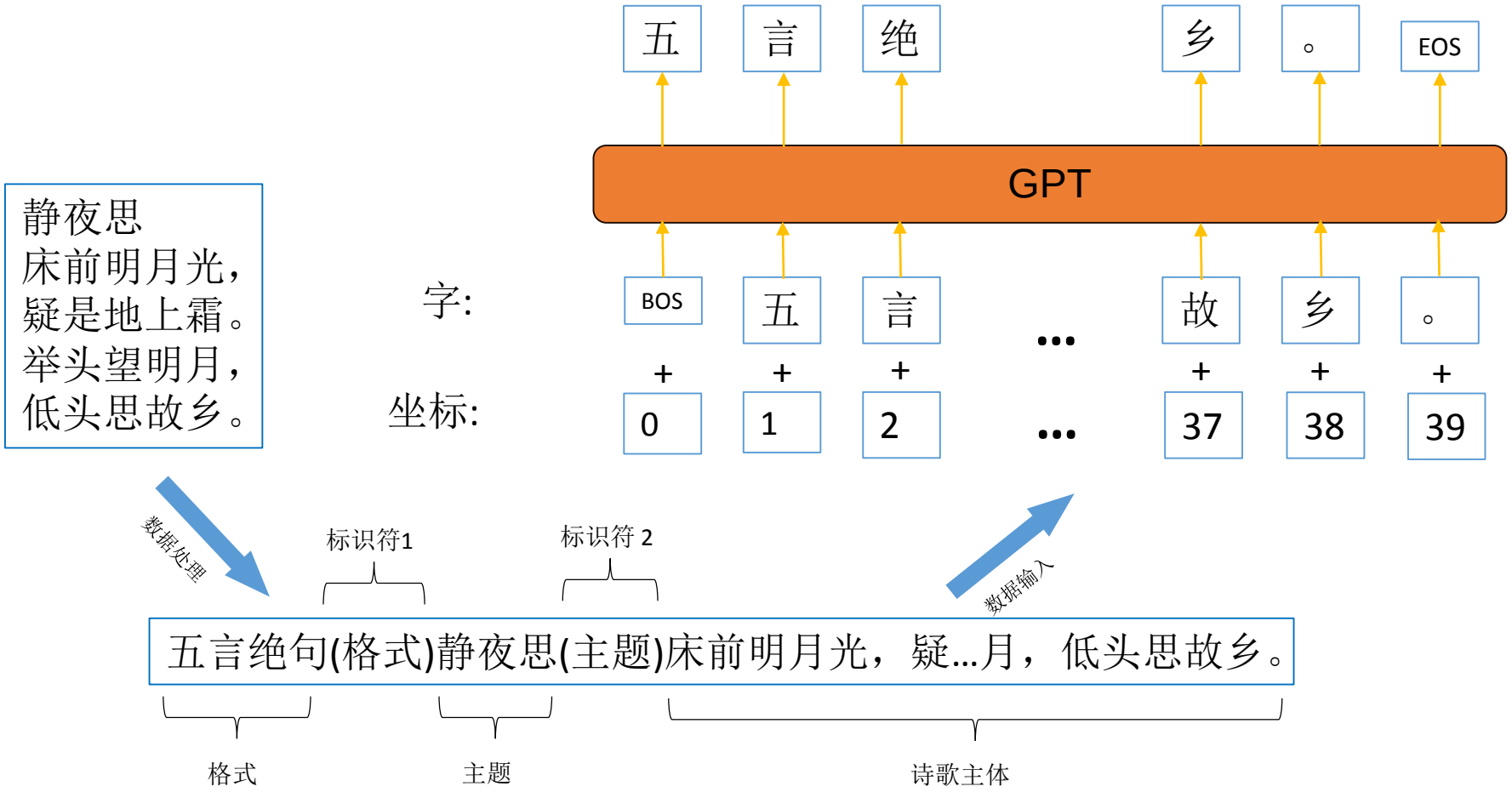
第一步：预训练中文GPT 模型

- 训练语料：30GB 中文文本
- 单向Transformer
- 模型大小：
  - Layer=12
  - Hidden size= 768
  - Intermediate size =3076
  - Attention head = 12
- 总参数量：1.1亿



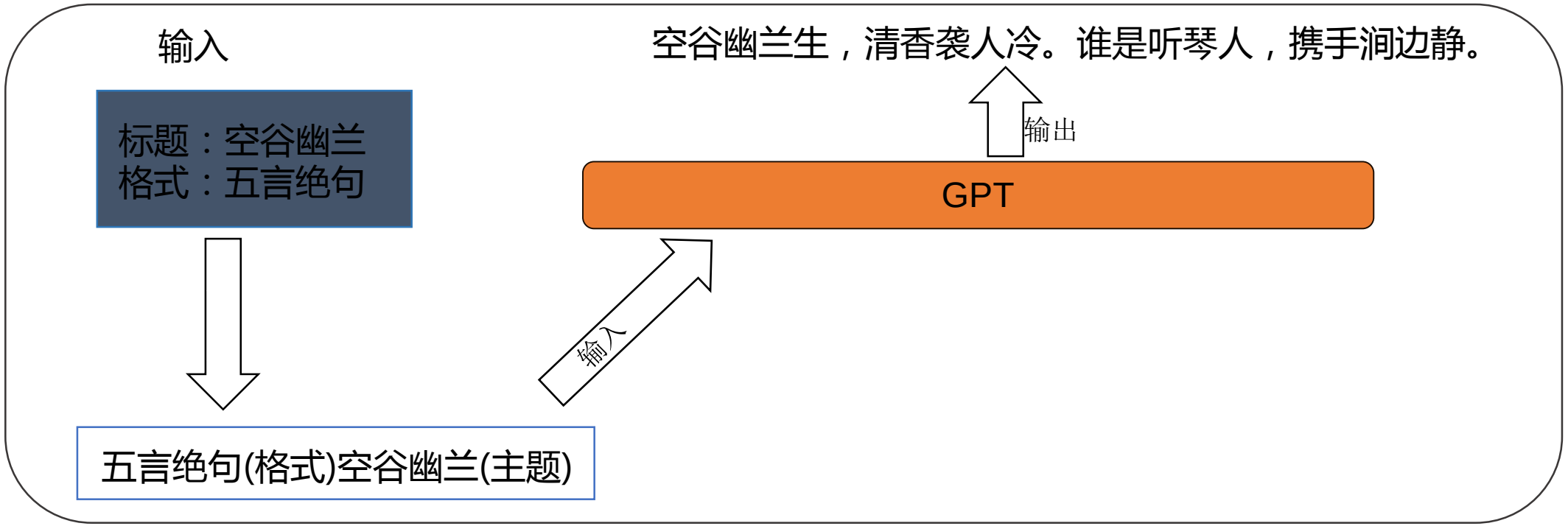
GPT模型结构

第二步：使用古诗词数据进行模型微调



诗歌生成流程

- 假设输入：
  - 主题：空谷幽兰
  - 格式：五言绝句
- 生成流程：



乐府诗歌生成的优点

- 格律、押韵、平仄等等格式层面表现非常好。
- 生成诗歌的多样性很高。
- 在深层次的内容、意境、氛围、引经据典等艺术层面，也经常有出人意料的表现。
- 生成高质量诗歌的概率很高，不需要从大量生成的例子中挑选少数好例子。
- 经常可以生成普通人难以辨别真假的诗歌。

江上田家

村南喧鸟雀，江北梦悠悠。
桑熟蚕三眠，人家半依楼。
一身千万里，何处得穷愁。
日暮歌明月，长河满斛秋。

江上田家

江边田舍好，茅屋远相迎。
竹里开门入，芦中引水行。
犬来沙上吠，鸥去岸间鸣。
不是无吟兴，谁知乐此生。

江上田家

近海川原薄，人家本自稀。
黍苗期腊酒，霜叶是寒衣。
市井谁相识，渔樵夜始归。
不须骑马问，恐畏狎鸥飞。

江上田家

野水通渔路，江村带夕阳。
数家深竹里，一树隔芦塘。
牧去牛羊下，人行果橘旁。
相逢皆贺岁，还有醉眠乡。

以上四首古诗，只有一首是由古代诗人原创，其余三首均为乐府作诗机生成的。

乐府诗歌生成的主要不足之处

- 诗歌质量对输入标题较为敏感
  - 如标题中出现古诗中常见的意象，如秋、竹、月、莲花、送别等等，通常能创作出意境贴切的诗歌。
- 有时在语义的连贯性会出现前后不一致，例如输入中秋为主题，尾句描写的景色并非属于秋天。

更多例子

闻秋虫有感

西风黄叶堕阶前，秋客愁思正可怜。
夜静子规啼滴滴，天寒乌鹊影翩翩。
一声塞雁江南去，几处家书海北连。
莫道征鸿无泪落，年年辛苦到燕然。

——乐府 2019.08.30

伊妹赞

二八佳人质，娇羞不肯妆。
临风笑相语，问我意何长。
独立鸟关外，含颦花甸傍。
世间儿女多，无奈自迷茫。

——乐府 2019.08.31

荔枝

绛纱囊擘水晶团，
风味何人似此看。
一骑红尘二千里，
不知谁是素馨冠。

——乐府 2019.09.03

端午

彩丝百缕纫为佩，
角黍千钟食有鱼。
安得此身无一事，
江湖随分乐樵渔。

——乐府 2019.09.03

乐游西湖

西湖万顷清秋月，
酒客重来独自游。
好事长宜勤载酒，
青山何必待人愁。

——乐府 2019.08.31

中秋望月

一年此夜最宜晴，况是新秋两夕生。
万里无云天更净，四溟收水月犹明。
吟怀耿耿还孤酌，静语悠悠忽老成。
坐想玉娥应有意，试临华席问寰瀛。

——乐府 2019.09.03

模型开源

- 开源链接： https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/NEZHA-Gen-TensorFlow
- 开源内容包括：
  - 预训练的中文GPT模型
  - 乐府作诗机
    - 注：我们开源的乐府作诗机可用于学术用途，如果您有商业上的需求，请使用华为云提供的接口 https://support.huaweicloud.com/api-nlp/nlp_03_0070.html

PMLM在乐府写宋词中的尝试

GPT模型能完成宋词生成吗？

- 难度非常大
  - 宋词格式种类多。
  - 训练数据少。历史上宋词总量很少，平均每个词牌的训练数据只有十几二十首（一个近体诗格式四五万首），GPT模型无法从少量的数据中学到丰富的格式变化（ 6首钗头凤词牌）
- 我们发表于ACL2020的模型可以完成宋词生成的问题

Probabilistically Masked Language Model (PMLM) Capable of Autoregressive Generation in Arbitrary Word Order

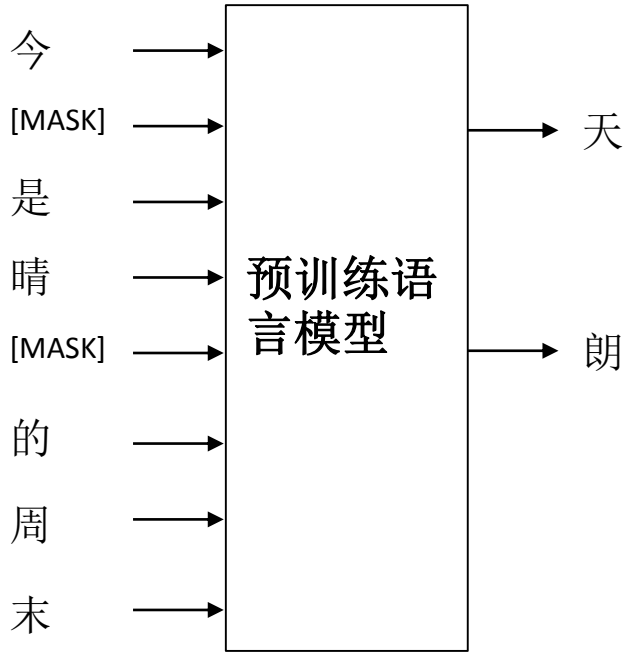
Yi Liao, Xin Jiang, Qun Liu

Huawei Noah’s Ark Lab

{liaoyi9, jiang.xin, qun.liu}@Huawei.com

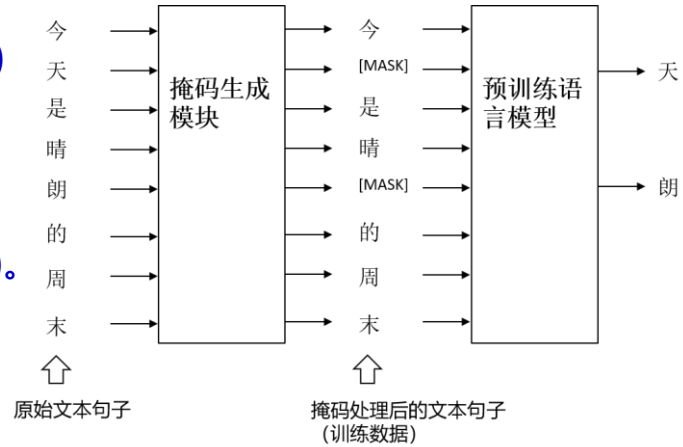
BERT的训练方式——Masked Language Model （MLM）

- MLM模型定义：部分词被替换成掩码[MASK]，模型预测被掩码处理部分的原词。
- 传统的MLM模型（如BERT）假设每个词被掩盖的概率是固定的值r=0.15



PMLM:

- Probabilistically Masked Language Model (PMLM)
  - 注：PMLM已收录至ACL2020
- PMLM基本思想：对MLM进行泛化，假设每个词被掩盖的概率值r独立地服从一个连续的先验概率分布，即$r \sim p(r)$。
- 掩码生成过程：
  - 掩码生成模块内置了先验概率p(r)
  - 对于每个输入的句子
    - 从p(r)采样每个词被掩盖的概率值$r \in [0,1]$
    - 每个词按照概率r被掩码处理，即被掩码处理的概率为r
- PMLM的特例：
  - BERT可看做PMLM在p(r)为狄拉克函数时候的特例
  - p(r)为0~1之间的uniform distribution时，uniform-PMLM数学上等效于autoregressive permutated language model (APLM)（论文中已证明）
  - GPT是一个autoregressive language model，但是不具备”permutated”的特性
- 注：胶片接下来的内容，PMLM均代指uniform-PMLM



任意字序的文本生成

- GPT只能按照从左往右的常规顺序进行生成。
  - 给定的context只能置于开头，模型接着开头生成篇章。

Tom is a cat and Jerry is a mouse .“ It ’ s very sad ! ” . The writers had wanted Tom to have “ something big to tell it . . . and a fun place to get excited ” . The writers believed that the “ little animal ” and the “ little black dog ” at the end of the episode would have attracted more attention from viewers , but it never took place . Tom ’ s first television role was that of the boy scout “ Mr . Krabs ” in the 1978 NBC Western comedy pilot , The Search for Mr . Krabs .

- PMLM可以进行任意顺序的填词式的文本生成。
  - 给定的context可以放置在篇章的任意位置，模型接着填充给定context词中间的空白，完成篇章生成。
  - 下面的例子中，即使限制任意两个context词之间都包含14个词，但是PMLM依然可以生成流程通顺的篇章

The wolf has an extraordinary speed , and it can often jump from a spot **quick** enough to escape a spot already occupied by an adult wolf . Unlike the **brown** and black bear , where it is easily distracted by wolves , the gray **fox** does not run over a wolf , and is often driven mad . Having **jumps** with high speed that breaks the wolf ’ s legs before it is run **over** , a grey wolf could defend itself against an adult of other species as **the** best predator at any time . The black bear may kill packs of four **lazy** , though the gray fox can inflict significant wounds on a **dog** .

Arbitrarily Ordered Text Generation

Step	Prediction Index	State of the sequence							
0	n/a	_	_	_	_	_	_	_	_
1	3	_	_	a	_	_	_	_	_
2	7	_	_	a	_	_	_	random	_
3	1	This	_	a	_	_	_	random	_
4	2	This	is	a	_	_	_	random	_
5	4	This	is	a	sentence	_	_	random	_
6	6	This	is	a	sentence	_	in	random	_
7	5	This	is	a	sentence	generated	in	random	_
8	8	This	is	a	sentence	generated	in	random	order
Generation Order: 3→7→1→2→4→6→5→8									
Output: This is a sentence generated in random order									

Table 1: An example of how u-PMLM generates a sequence in random order. The special token `[MASK]` is simplified as the symbol "_".

Generation Algorithm:

1. Initialize with a sequence of blank([MASK]) tokens
2. In each iteration
    1. specify a position/index of the token to be predict
    2. replace the [MASK] at this position with the predicted token to update the generated sequence.

任意字序的文本生成方式的应用

- 例：PMLM可用于格式要求严格的文本生成场景
  - 以基于GPT和基于PMLM的宋词生成模型对比为例
  - 我们基于GPT的诗歌生成模型“乐府1.0” 可以生成格式正确的**近体诗**。但“乐府1.0”无法实现**宋词**的生成。通过引入PMLM技术，将乐府升级为2.0版本，支持宋词的生成。
  - 宋词生成是比近体诗生成难度大很多的任务。原因在于历史上宋词总量很少，平均每个词牌的训练数据只有十几二十首（一个近体诗格式四五万首），GPT模型无法从少量的数据中学到丰富的格式变化（ 6首钗头凤词牌）。
  - PMLM通过先填入标点符号框定格式，随后填充中间文本的方式，避开让模型学习格式的问题，可以实现格式完全正确的宋词生成。
  - 一首PMLM以“赤壁怀古”为标题生成的词：

    钗头凤·赤壁怀古

    江水流，云霭收，一声渔笛隔汀洲。

    风飕飕，月悠悠。

    千载英雄，如何不休。

    愁、愁、愁。

Evaluation of Language Generation

- Perplexity
  - Finetuned the models using the training set and evaluate them using the test set.
  - Results:
    - u-PMLM generates better sentence than BERT and can even outperform GPT in wikitext_103 dataset
    - The generated text in random order are also in good quality

Model	PPL(sequential)	PPL(random)
BERT	23.12	25.54
GPT	21.23	N/A
u-PMLM-R	19.58	21.51
u-PMLM-A	19.32	21.30

Table 2: Perplexity on Wikitext103.

Model	PPL(sequential)	PPL(random)
BERT	140.67	56.97
GPT	24.25	N/A
u-PMLM-R	35.24	38.45
u-PMLM-A	49.32	42.46

Table 3: Perplexity on One-Billion Words.

Evaluation of Language Generation

- Latency Analysis:
  - In theory, the computational complexity of the proposed generation method is N times of traditional autoregressive language model, where N is the length of the generated sequence.
  - However, since Transformer updates hidden states parallelly, the proposed generation method is only 20.1% slower.

Models	Cost Time
GPT	105.6 s
u-PMLM-A	126.8 s

Table 4: Latency for generating 100 128-length sequences.

PMLM生成的宋词示例（单次生成，未经人工筛选）

浣溪沙·溪流背坡村
溪流背北沙洲好，竹外茅茨三两家。人来未觉暮天涯。寒鸥独立云中雪，野艇闲行水带沙。风林淅沥雨吹花。

念奴娇·溪流背坡村
溪头深岸，石桥小，水色如蓝凝绿。竹里横塘，人正睡，数点荷花新雨。钓舟过后，棹歌声远，不觉村烟暝。

雨霖铃·溪流背坡村
溪流背郭上坡村，人寂寂，柴门向水开。山童携妇担来。炊新秫，煮以菊魁。秋阳满野，黄叶乱、红凋欲脱。吾其所乐惟田父，笑语同一杯。

沁园春·溪流背坡村
曲曲溪流，竹外茅檐，中有人家。但鸡犬声闻，烟火青荧。松下行莎，谷口吞霞。一叶扁舟，数竿修杖，来往无穷日未斜。归亦好，更清风快雨，乱放新衙。

鹊桥仙·溪流背坡村
石门深邃，水木苍凉，竹暗路迷回冈。人家鸡犬野烟外，一声鸣橹隔溪篁。风生苹末，露湿花裳，山禽对语清簧。不须更待白莲出，满耳但闻啼杜鹃。

踏莎行·溪流背坡村
野渡横舟，水桥断脚，数家茅屋临溪阁。白苹翠竹秋满篱，日暮鸡犬无人喧。渔歌声远，山云自碧，一片孤帆带斜色。夕阳西去又何之，芦花汀上萧萧泣。

菩萨蛮·溪流背坡村
溪流浅浅沙铺雪，低枝倚岸梅花发。日暮竹林村，水云香暗门。小桥横短彷，茅舍连鸡犬。烟际一行斜，牧童牛背家。

贺新郎·溪流背坡村
溪流带雨清且浑。岸边村，人家几处，水面柴门。山下有田三十亩，竹树阴森相倚。正东西，日光摇紫。一笛横飞归去尽，不管渔翁船过浦。更何许，钓竿把。

乐府体验

- 微信小程序：EI体验空间
- 运行在华为自研芯片昇腾910、310的版本：
  - https://www.huaweicloud.com/ascend/apps/poem

广告

- 华为诺亚方舟语音语义实验室长期招收对预训练语言模型研究感兴趣的实习生，实习地点：深圳、北京。
- 欢迎投简历到 liaoyi9@huawei.com

今日分享材料获取

- 通过以下二维码可以获取

入群步骤：

①请扫码关注公众号

②回复“知识图谱”后将自动弹出小助手二维码

③识别小助手二维码并添加为好友

④进群获取分享材料

谢谢，欢迎大家提问