基于模型融合和数据增强的条件大纲故事生成

作者姓名：廖泽明 莫林璋 刘永康
导师姓名：黄清宝
学　　校：广西大学
研究方向：多模态

2022.04.23

1 任务定义

任务定义

输入："筛子安慰", "筛子指着", "筛子换上来接替", "筛子问道", "不畏艰难不辞劳苦", "对着困难望而却步", "筛子失望", "任何人提供现成"

输出：一只淘金的筛子筛洗了一大堆一大堆的沙子，还没有得到黄金，筛子却已经破了。一只新的筛子换上来接替它的工作。“老兄，”新来的筛子问道：“你的成绩呢？”“那就是！”破旧的筛子指着那些沙堆回答。新来的筛子失望地说：“我不知道为什么要来接替你这种徒劳无益的工作！你留给我的就是这些失败的沙堆，我有什么希望获得有用的黄金呢！”“不，”老筛子安慰它说：“你要知道，大自然不会轻易地给任何人提供现成的果实。在对着困难望而却步的人的面前，确实只有失败的沙堆；而对于那些不畏艰难不辞劳苦的淘金者，却有可能从失败的沙堆里淘出成功的金子。

2 模型使用

模型使用与训练数据统计

使用模型：

LongLM (Chinese Long text pretraining Language Model)[1], LongLM is a Transformer-based model with an encoder-decoder architecture, which has three different versions: 60 million, 200 million and 1 billion parameters. We utilize the base version (200 million parameters).

训练数据分布：

	训练集	验证集	测试集
样本数	4824	661	1149
大纲的平均词数	20.78	20.55	20.23
故事的平均词数	256.74	241.28	191.67
大纲的最长和最短长度	最长：118 最短：5	最长：88 最短：8	最长：94 最短：7
故事的最长和最短长度	最长：7687 最短：42	最长：3574 最短：57	-

[1] Guan, Jian, et al. "LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation." Transactions of the Association for Computational Linguistics 10 (2022): 434-451.

3 训练和策略

策略一：数据增强

对输入的大纲进行乱序：

"速度高于房价上涨"，"写作过程"，"撕毁合同重签"，"未来买不起房子"，"句话"，"交往过程"，"心存疑虑"，"先是签"

"未来买不起房子"，"速度高于房价上涨"，"先是签"，"写作过程"，"心存疑虑"，"撕毁合同重签"，"句话"，"交往过程"

策略二：数据预处理

阿柔是我本科时的同学，她来自农村，身材高挑，长发披肩，笑起来眼睛像弯月一般，那样美好的笑容，犹如午后的太阳晒入心房，让人感觉暖暖的。从小在城市长大的我，起初和她并没有什么共同话题。和她聊明星八卦时，她总是在我话音落下之后，先是使劲地点点头表示认同，然后用手捂着嘴再偷偷地追问：“可是……他到底是谁啊？”和她聊未来环游世界的梦想时，她总是羡慕地注视着、支持着，仿佛一个小女孩隔着橱窗看到一件昂贵的嫁衣，喜欢，却清晰地知道那不会属于她。她每每和我说起她家乡的一些事情，我也不是很能体会，只是觉得她在黄河边游泳捉鱼是一件很有趣的事。……（省略583字）看到她的信息，我可以想象到她在打下这些字时脸上果敢坚毅的表情。那时我才发现，虽然名字里有个“柔”字，阿柔却从不曾是个柔弱的女子。……（省略153字）可是我太想做口译员了，如果放弃的话以后我可能会后悔，只能再尝试一次了。”最近一次和阿柔联系时，她已经获得高级口译证书，在上海一家外贸金融公司做口译员。……（省略112字）。（**共1251个字**）

把目标文本的长度从1251压缩到768，压缩后的文本包含全部大纲词汇。

阿柔是我本科时的同学，她来自农村，身材高挑，长发披肩，笑起来眼睛像弯月一般，那样美好的笑容，犹如午后的太阳晒入心房，让人感觉暖暖的。从小在城市长大的我，起初和她并没有什么共同话题。和她聊明星八卦时，她总是在我话音落下之后，先是使劲地点点头表示认同，然后用手捂着嘴再偷偷地追问：“可是，他到底是谁啊？”和她聊未来环游世界的梦想时，她总是羡慕地注视着、支持着，仿佛一个小女孩隔着橱窗看到一件昂贵的嫁衣，喜欢，却清晰地知道那不会属于她。她每每和我说起她家乡的一些事情，我也不是很能体会，只是觉得她在黄河边游泳捉鱼是一件很有趣的事……（省略397字）。看到她的信息，我可以想象到她在打下这些字时脸上果敢坚毅的表情。可是我太想做口译员了，如果放弃的话以后我可能会后悔，只能再尝试一次了。”最近一次和阿柔联系时，她已经获得高级口译证书，在上海一家外贸金融公司做口译员。（**共766个字**）

策略三：模型融合

Input：木头吃梨#老中医告诉#囫囵吞下去#集市上买#嘴里咀嚼#眼泪直流#伤脾胃#伤牙齿

Base answer（from best model）：有一个伤脾胃的人,在集市上买了一块木头吃梨,囫囵吞下去。过了一会儿,他的肚子开始疼了起来。他大声喊着:“救命啊!救命!”一个集市上的人听到这个消息,赶紧来救他。一个老中医告诉他,这个人的胃里塞满了梨子,不能吃，吃多了可能伤牙齿。这个人听了,眼泪直流。他流着泪水把梨子给了那个伤心人,还告诉他:“我知道错了,我以后再也不吃梨了。”

Supplementary answer（from other model）：有一个伤脾胃的人,在集市上买了一块木头吃梨,囫囵吞下去,半个小时过去了,他嘴里还叼着一根木头。老中医告诉他,这个人的脾气非常暴躁,一旦发脾气,就有可能伤牙齿,甚至连牙齿都碎了。那个人听后,吓得面如土色,嘴里咀嚼着木头也不敢吃了。

Final answer：有一个伤脾胃的人,在集市上买了一块木头吃梨,囫囵吞下去。过了一会儿,他的肚子开始疼了起来。他大声喊着:“救命啊!救命!”一个集市上的人听到这个消息,赶紧来救他。一个老中医告诉他,这个人的胃里塞满了梨子,不能吃，吃多了可能伤牙齿。这个人听了,眼泪直流。他流着泪水把梨子给了那个伤心人,还告诉他:“我知道错了,我以后再也不吃梨了。”嘴里咀嚼着木头也不敢吃了。

4 实验结果

评价指标

① **BLEU-1/2：**衡量输出的故事和标准答案之间的相似性。

② **Distinct-3/4：**衡量输出的故事的多样性。

③ **Coverage：**计算生成的故事和给定的情节序列之间的 Rouge-L recall 分数。 衡量输出的故事包含了多少给定大纲中的情节。

④ **Order：**用 1 减去逆序短语对的数目除以总短语对的数目，这里情节的出现位置通过比对最长公共序列来确定。 衡量短语出现顺序是否得当。

$$Order = 1 - \frac{N_{inv\_pairs}}{N_{all\_pairs}}$$

⑤最终评价指标：$Score = \sum_{i=1}^{M} \frac{w_i}{\sum_{j=1}^{M} w_j} S_i$，其中 $M = 4$，w_i 表示权重，其等于Ground-Truth的得分与 LongLM-small 的得分的比值，即如果LongLM-small 的得分越小，其对应的权重越大。

实验结果

Model	Score
Baseline	0.312
Baseline+数据增强	0.326
Baseline+数据增强+数据预处理	0.330
Baseline+数据增强+数据预处理+模型融合	**0.357**

表2 实验结果

融合三种策略的模型和baseline在Coverage和Order评测指标上的结果比较：

Coverage：0.8656 --> 1

Order：0.6653 --> 0.6924

5 总结

总结

基于条件大纲的故事生成任务，分别采取了数据增强、数据预处理和模型融合的策略，在测试集上总分从0.312提升到了0.357。

感谢聆听~