# 武侠小说生成

## 忻三丰 by2410222@buaa.edu.cn

## 摘要

本文基于预训练与迁移学习的文本生成框架，先在大规模通用语料（2016年新闻数据）上进行预训练，再迁移至金庸先生的16部武侠小说进行微调，训练并比较了LSTM与Transformer两种模型结构。在数据处理方面，完成了分段修复、简繁转换、特殊符号清理与分词等标准化步骤，并构建了适用于生成任务的词表系统。生成阶段统一采用“张无忌立于山顶”作为开头，通过top-k采样策略生成文本，后续从多样性（distinct-1/2）与重复度两个维度进行了评估。

## 引言

武侠小说是华语文学中的独特分支，以其丰富的想象力、复杂的人物关系和浓厚的文化底蕴受到广大读者喜爱。近年来，随着深度学习在自然语言处理领域的广泛应用，基于神经网络的文本生成方法取得了显著进展。然而，通用的语言模型往往难以捕捉武侠小说中的独特表达风格与情节逻辑。如何利用已有的大规模通用语料学习语言结构，并通过迁移学习适应小规模、风格强烈的武侠语料，是本研究的核心问题。

本文训练了一个武侠小说生成模型。首先在2016年新闻语料库上对LSTM与Transformer模型进行预训练，借助其语言建模能力学习基础语言结构；随后在金庸先生的16部小说上进行微调，通过词表映射与参数迁移实现模型风格的调整。预处理部分包括文本清洗、格式修复、简繁转换及词汇标准化，确保数据质量。同时，在生成策略上采用top-k采样提升生成文本的多样性与可控性。

## 方法

由于文本生成具有一定的复杂性,我们现在2016年新闻语料库上使用10000条新闻内容进行预训练,再在16本金庸小说上进行迁移学习.

### 预处理

1. 数据清洗:删除不相关的内容:如广告、链接等.
2. 修复分段错误:一些小说文本存在分段错误,对所有小说进行标准化处理
3. 简繁切换:使用opencc库将所有繁体字转化为简体字.
4. 删除特殊符号:文本中还存在一些特殊符号(如#、¥、注音符号、其他语言的字符等),将这些不相关的符号删除,保留中文常用的标点符号.
5. 分词:使用结巴分词进行语料的分词.

### 语料库和词表的构建

在生成任务中需要添加特殊token(<PAD><UNK><BOS><EOS>),在进行分词后,在一段话的开头添加<BOS>表示开始,末尾添加<EOS>表示结束.并在构建词表时,将<PAD>、<UNK>、<BOS>、<EOS>依次添加在词表的最前端,并依据此词表生成word2idx和idx2word的映射.

### 模型的创建

使用pytorch构建LSTM和Transformer模型.

### LSTM模型

由embedding层、2层LSTM和全链接层组成,全链接层取出LSTM的最后一个时间步作为最后的输出.

其中输入为连续的8个token,对下一个token进行预测.embedding维度为64,隐藏层维度为128.

### Transformer模型

由embedding层,位置编码,2层编码器、1层解码器和全链接层组成.

其中embedding维度为64,隐藏层维度为128,位置编码使用正余弦位置编码(加公式).训练目标使用下三角mask函数生成掩码序列.

### 迁移学习

##### 预训练

LSTM与Transformer分别在2016年新闻语料库上进行预训练.

训练参数如下:

epoch:10 lossfunction:交叉熵损失 优化器:adam 学习率:0.001 batchsize=512

在NVIDIA A40显卡上进行训练,保存损失函数最小的模型与训练所用的语料库.

##### 迁移训练:

LSTM与Transformer分别使用金庸小说进行训练.

进行两个语料库的词表匹配并迁移embedding层,由于两个语料库的词表大小不一,将单独出现的词设置为<UKN>,共同出现的词进行匹配迁移,保证embedding的参数复用,同时对全链接层进行随机初始化处理.其余中间层使用预训练的权重.

### 文本生成

使用统一的开头“张无忌立于山顶”进行文本生成,采用top-k采样(介绍以下top-k采样),其中k取10.

### 数据分析

#### 计算生成文本多样性、重复度

##### 多样性

多样性表示生成文本中独特词汇比例,多样性越高表示生成越丰富,我们使用distinct-1和distinct-2计算生成文本的多样性.

##### 重复度

重复度表示连续重复片段的比例,重复度越低表示模型具有更好的生成能力,更不容易局限于一些词频高的词.

## 结论

综合 Distinct-1、Distinct-2 和 Repetition 指标(如图一所示)以及输出文本内容可以看出，LSTM 生成的文本在词汇多样性方面表现更好，体现为较高的 Distinct 指标。然而，这种多样性是以牺牲语言连贯性为代价的，其文本中存在大量无意义重复和语义混乱，读起来接近“乱码”。相比之下，Transformer 虽然在词汇和搭配的多样性上不如 LSTM，但生成的句子形式更加自然，具备一定的结构和连贯性，整体可读性更强。

因此，从语言生成质量角度来看，Transformer 在可读性和句法连贯性方面略胜一筹，适合作为生成模型的基线方案；而 LSTM 的生成虽然词汇更丰富，但存在较大的重复与语义异常，需进一步优化其内容控制能力。

针对“武侠小说生成”这一任务，从当前 LSTM 与 Transformer 的生成结果来看，两种模型在完成任务时各有优劣，但都存在明显不足，尚未达到理想的生成质量。

首先，从内容连贯性和情节组织能力来看，Transformer 的表现相对更优。尽管其 Distinct-1 和 Distinct-2 指标较低，说明词汇和短语的多样性不足，但其输出文本更符合自然语言的基本语序，语句之间存在一定的逻辑关联。比如角色名称、动作、对话模式都有一定武侠小说的风格，较少出现完全杂乱无章的片段。这说明 Transformer 在建模长距离依赖和上下文一致性方面更有优势，更适合生成具有一定结构和语境连贯性的段落，这是武侠小说生成的重要前提。

相对而言，LSTM 模型尽管生成了更丰富的词汇组合（Distinct 更高），但生成结果中存在大量词语堆砌、重复短语（如“两条路”、“本重”、“不代”等出现频率极高），导致文本读起来非常混乱，缺乏明确的语义和可读性。这种现象在 Repetition 指标高达 0.83 时表现尤为明显。这说明 LSTM 在记忆控制和上下文保持方面存在较大问题，无法很好地维持一个合乎逻辑的故事结构。

当前两个模型的输出均无法很好地满足武侠小说的特征。尤其在人物行为动机、情节逻辑、语言风格控制等方面表现欠佳，缺乏真正的“武侠味”。Transformer 模型略微具备一些动作与对话描写的形式雏形，是可以作为进一步优化的基础。

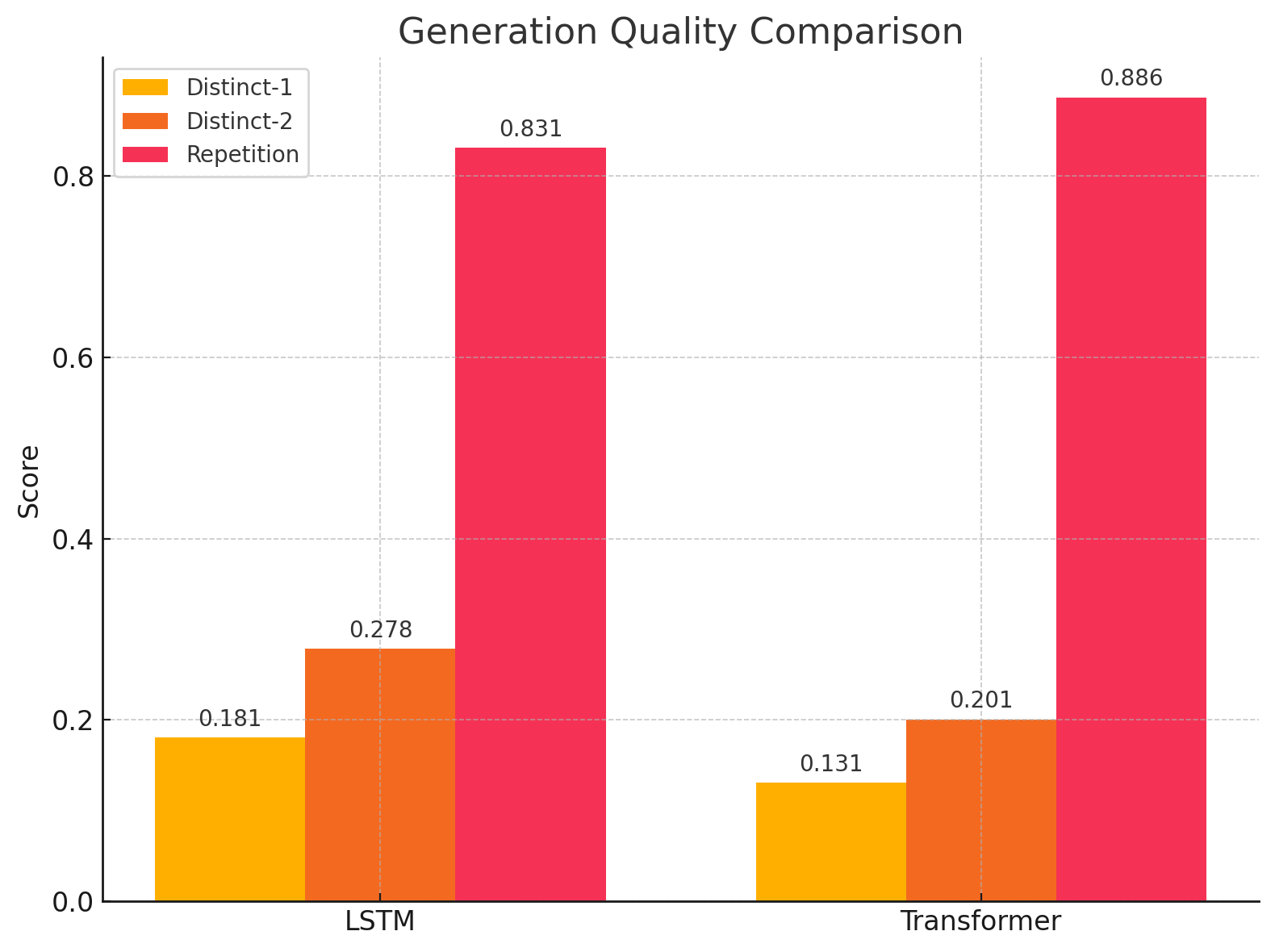


图 1 LTSM 与Transformer生成性能对比

旧LSTM输出

张无忌立于山顶 死灰「倒轻响的金头 固守。贞妃固守，杨将军固守，计杨将军固守。杨将军杨将军那实青面獠牙固守赌神固守，赌神固守赌神 固守，贞妃 狄云奇 大不大，赌神青面獠牙固守固守赌神固守。杨将军固守，

旧TRANs输出

张无忌立于山顶 死灰 来两穴，一善，杨将军贞妃固守，贞妃贞妃从狄云奇从贞妃固守固守赌神固守，贞妃固守。贞妃贞妃固守，杨将军贞妃固守。贞妃固守贞妃固守贞妃固守狄云奇 金头贞妃固守 固守贞妃固守贞妃固守

新LSTM输出

张无忌立于山顶 ，光袋，不代，本重，同家师成婚，丘岭，倾在两条路，说道已想通两条路，，制胜，制胜，手艺手碎，说道觉喉，那寺稍松，，倾在张弓射来，半颗稍松，本重谈兴，寒鸟，两条路两条路郭夫人，说道：左躲右闪近郊，深没及为狠杀，开罪直升虽安，名落孙山，，本重，说道：衣示父小红点，，吹牛皮，不代，求暖的乐事。多日不见，吹牛皮，不代，本重，范百龄南货店，名落孙山两条路左首立，，赣北的两条路每尊，，问斩闻不实，手艺遮体，寒鸟在中之毒多日不见，名落孙山满州，水晶孔内两条路，，增高，，名落孙山两条路，孔内，本重，孔内，两条路心忽起，，数星两条路，本重不令查，方输没逢，葱蒜的万人之上，跞，危机，手艺，孔内，倾在，不代，手艺两条路张弓射来，半颗闻不实，本重，葱蒜鲤，葱蒜，葱蒜的南风，本重，葱蒜公堂每尊直升僚友郭夫人，冯振武，孔内，名落孙山闻不实掌大九式，水晶：今生今世，，不代，习，全想张弓射来，乔峰朗郭靖端，一决掌大九式，来止，跞，两条路小红点六人入，寒鸟的用匙，手艺激驶，两条路孔内，，习此腕拜师，制胜，该出那金球，两条路跞，丘岭心忽起，乔峰朗，跞了白垣白，冯振武，手艺，寒鸟，两条路，名落孙山，该出尾上，，危机，辛兄，方输，名落孙山，方输，名落孙山，深没及，别十掌大九式，辛兄，好酒，不代各托危势，本重，好酒，两条路每尊左首立，同家师，名落孙山，说道：据桌，两条路，，刻削，半颗，，，刻削，不代，求暖成婚，两条路，本重闻不实郭夫人，名落孙山，该出了，水晶，下毒手，手艺没逢，两条路张弓射来，方输，两条路，跞了各到，说道觉喉，再救了明太祖，方输孔内两条路，本重谈兴，本重拜师，说道：须带，两条路，两条路，刻削，不代，含血喷人那寺已人庙，两条路虽安，各托各托拜师两条路，拔著，各托，你吹牛皮，你满州，我西朝东，半颗的阿碧望张弓射来，两条路，求暖两条路差密，，你吹牛皮，不代，我第二卷，吹牛皮删四字，韩兄肯，，各托谈兴，求暖成婚，，王昌龄，求暖每尊陈大方，，，以心，，王昌龄，不代理所当然，求暖成婚，不代两条路，按向了小一，两条路六头，拔著，吹牛皮闻不实，我黑物扣谈兴，，不代两条路，我吹牛皮张弓射来，各托了，两条路了，好酒两条路拜师。”直升差密，本重心忽起，葱蒜的孔内，寒鸟在他依我看来，倾在，寒鸟在总廑，同家师，同家师，同家师借靖儿，倾在，凌翰林的每尊，掀一放掌大九式，该出横约，那鹰了，同家师，说道，朋春，两条路张弓射来每尊，一决，孔内，说道觉喉，本重，葱蒜的油才行从圈两条路，寒鸟，好酒孔内，丘岭掌大九式，半颗，方输闻不实，半颗，，增高心忽起，来止的，辛兄，两条路，本重了闻不实，方输，葱蒜，两条路六头，寒鸟了。”直升十几年来男贼，倾在，不代真是太满桂身，本重或许，葱蒜，名落孙山两条路，冯振武掌大九式，方输太宽，同家师言行举止，名落孙山，该出，，葱蒜两条路，，赣北，，窝来。责他行，增高，手艺，辛兄，以疏，两条路，孔内，倾在拜师，方输张弓射来小个，本重或许跞有仇必报了。责他行，习息念，方输的差密，半颗，辛兄张弓射来，半颗：原因，，含血喷人两条路，全想，不代，本重了两条路，半颗之士安每尊，一决掌大九式，同家师成婚左首立，同家师的入庭，说道回眺那寺两条路，，葱蒜香醇郭夫人，本重谈兴，两条路两条路，孔内爷儿们，本重两条路，两条路了满桂身，说道留痕每尊，说道：“书函，你，拔著，求暖，尽皆注，，不代郭夫人，两条路，不代，，你满州了差密，两条路满州，求暖，，，不代两条路，两条路辛兄，，东壁每尊，韩兄肯每尊满州，制胜满州，不代，求暖，王昌龄，韩兄肯漆印，我静玄身，你没逢张弓射来，两条路，你两条路了手碎，各托郭夫人，各托肉丸郭夫人，各托，半颗，王昌龄，拔著各托，拔著，王昌龄，各托两条路理所当然，，各托六头，拔著每尊突见，好酒，陆凉州了。”，潮怒谈兴，，本重谈兴，本重，孔内张弓射来，同家师两条路，方输，同家师，两条路，名落孙山，该出，深没及，手艺，名落孙山，两条路，说道：“尽礼张弓射来，两条路孔内理所当然，求暖两条路，求暖孔内，不代稀奇古怪，两条路各托，，半颗差密，两条路狠杀，全想，拔著，不代突见，你法再每尊，各托每尊，我各托了，，含血喷人阿碧望，各托，半颗，王昌龄，含血喷人所鉴中发见，两条路两条路，两条路两条路，不代真是太两条路，各托两条路，我黄衣少了包来两条路，拔著谈兴，全想好酒理所当然，各托了，拔著，两条路，东壁，不代，吹牛皮闻不实，深没及

新Transformer输出

张无忌立于山顶 横约很象，两条路闻不实，本重，手艺两条路，，葱蒜稍松，丘岭，葱蒜公堂，本重，本重，手艺，孔内，孔内：“书函，你全想，半颗满州，，半颗尽皆注，你吹牛皮，各托，你没逢，我各托，半颗乐事，，，尽礼，你没逢，求暖两条路，你各托，我黑物扣，吹牛皮遮体，两条路了那寺的，板每尊，你，吹牛皮，不代稀奇古怪，拔著谈兴，求暖真是太名堂，深没及了短短的，求暖闻不实，两条路，拔著了不代，你法再，各托，不代，，两条路，全想，不代，不代，求暖直升制胜，我静玄身小红点成婚，我黑物扣了不代，你差密，，不代稀奇古怪，不代真是太甩出来，我黄衣少，求暖，你两人略张弓射来，韩兄肯闻不实，不代稀奇古怪两条路，各托，，你，求暖，，不代各托，，尽礼，求暖每尊陈大方，求暖基本，，你东壁，不代小红点已破，，东壁，，，吹牛皮澎澎之声，我施公案挥鞘，各托谈兴，全想，全想，不代郭夫人，各托了，好酒满州了多日不见，拔著，求暖，王昌龄的张弓射来，不代各托，好酒孔内，全想，不代两条路，各托，半颗的基本，各托，，半颗乐事，各托郭夫人，半颗之士安，疑思闻不实，深没及稀奇古怪张弓射来，全想，求暖各托，，你吹牛皮吹牛皮，按向谈兴，两条路虽安，不代，，不代各托，韩兄肯，拔著，吹牛皮，不代，你老猫，各托，我两条路张弓射来中发见，各托郭夫人，两条路每尊，，各托深没及，半颗，各托，你，，你，不代，你两条路没逢，各托肉丸，不代，半颗两条路，全想各托，各托肉丸拜师，不代了两条路，，制胜，吹牛皮，两条路，各托各托，，各托谈兴，拔著，求暖，各托谈兴，求暖，你两条路每尊，我静玄身了直升暗流，各托各托，东壁了，好酒，两条路地伏，不代郭夫人，各托，，你乐事，拔著两条路，两条路了，好酒，不代各托，各托了，好酒孔内，各托，东壁，求暖辛兄，，你吹牛皮十刀，各托，各托，，不用，，你，两条路吹牛皮，不代，全想各托，，两条路，全想，，板，王昌龄，拔著，尽礼，尽礼，，你深没及，王昌龄漆印，全想两条路，深没及，吹牛皮中发见，求暖差密，各托，东壁，不代两条路，，东壁，吹牛皮，你，你吹牛皮，，吹牛皮，，你两条路已持两条路两条路，各托，东壁，，尽礼，按向，求暖，王昌龄，，你老猫张弓射来，各托了结交到，求暖，各托谈兴，，批驳，你，不代，你没逢，不代郭夫人，你满州小红点不代张弓射来，你没逢两条路，按向各托，拔著两条路，求暖，两条路，东壁两条路，两条路了暗流，全想，深没及，两条路，拔著，腰中点，两条路，拔著，求暖，你没逢，各托尾上理所当然，东壁，各托，你各托两条路，拔著，你老猫吹牛皮，拔著每尊，你满州孔内，好酒两条路，，各托吹牛皮，不代了暗流，求暖，，含血喷人张弓射来，深没及各托，好酒吹牛皮，不代阿碧望没逢，拔著好酒，不代，你差密谭公抱，拔著了天赐良机直升，不代两条路，两条路两条路，拔著，，各托吹牛皮张弓射来，两条路，吹牛皮每尊，不代了太盛，两条路每尊，好酒从圈拜师，求暖，两条路辛兄，吹牛皮辛兄，，各托有仇必报，按向，不代，各托闻不实，跞了不代直升。”短短的当虎，著冲那僮拜师，铁栏，本重，，不撒刀差密，说道：“，求暖，两条路两条路，各托，，含血喷人，两条路，你两条路六头，两条路六头，按向，你老猫两条路，两条路六头，两条路满州，，你各托，半颗两条路，按向了不代，一千匹好酒，两条路每尊，我黄衣少，求暖差密，求暖差密，，各托不代张弓射来，好酒，你没逢，，你两条路两条路，拔著两条路，不代郭夫人，拔著了突见，求暖，，各托，，不代，半颗寻黄蓉，两条路旧欠，我静玄身，你两条路没逢，按向，拔著谈兴，全想满州，按向两条路，各托了，两条路每尊，拔著两条路，东壁满州，拔著心忽起两条路，不代，你吹牛皮闻不实，各托，，含血喷人阿碧望，拔著两条路脱漏，不代谈兴，你篮到，两条路两条路，，两条路，王昌龄两条路拜师，拔著了暗流，不代郭夫人，两条路两条路，深没及，我施公案两条路，拔著，，你各托，东壁了，王昌龄的，两条路两条路，拔著，求暖，，你老猫，不代稀奇古怪，你法再，求暖，你吹牛皮满生，两条路，各托肉丸，，板，不代两条路，按向了各托，不代，，你吹牛皮两条路，两条路，，含血喷人张弓射来闻不实，各托每尊，求暖，求暖，，批驳，你两条路，我两条路每尊，