小说文本生成

姓名 李祎柔

学号 20231181

一、实验要求

利用给定语料库（金庸语小说语料链接见作业三），用Seq2Seq与Transformer两种不同的模型来实现文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论两种方法的优缺点。

二、实验原理

2.1 seq2seq 模型

Seq2seq 模型就是一种能够根据给定的序列，通过特定的方法生成另一个序 列的方法。它在许多领域产生了一些运用。目前， 它主要的应用场景有：机器翻 译、聊天机器人、文本生成等。

Seq2seq 模型主要由编码器和解码器两部分构成，在这个结构中，输入一个 句子后，生成语义向量 c，编码过程比较简单；解码时，每个 c、上一时刻的 yi- 1，以及上一时刻的隐藏层状态 si-1 都会作用到 cell，然后生成解码向量。

编码器端往往采用序列模型，如 RNN，LSTM 等。在编码的每个时刻，模型 的输入除了上一时刻产生的隐层状态编码，还有当前时刻的输入字符，并将最后 模型最后一个时刻的隐层状态做为整个序列的编码表示，传递给解码器。

解码器端与编码器端近乎相同，不过解码器端需要保存模型的输出用于产生输出序列。在模型的训练阶段，模型的输入是文本内容以及上一刻的状态变量，模型输出为预测的下一个序列变量。在模型的预测阶段， 模型输入为上一个时刻的输出以及状态，来预测下一个时刻的输出。

整个编码-解码器的预测阶段的工作流程如下图所示

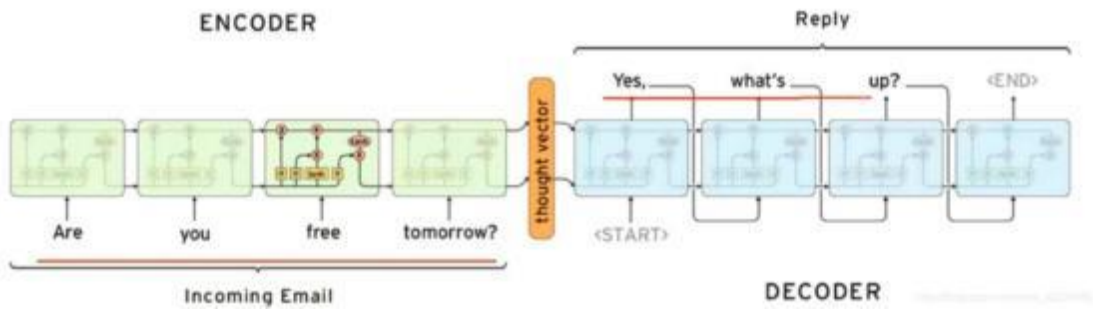


图 1 seq2seq 模型

2.2 LSTM模型

长短期记忆网络（Long Short-Term Memory，LSTM）是一种常用的循环神经网络（Recurrent Neural Network，RNN）架构，用于处理序列数据，特别是具有长期依赖关系的序列数据。

LSTM通过引入一种称为"门"的机制来解决传统RNN中的梯度消失和梯度爆炸问题，使其能够有效地捕捉和利用长期依赖关系。下面是LSTM的主要组成部分及其工作原理：

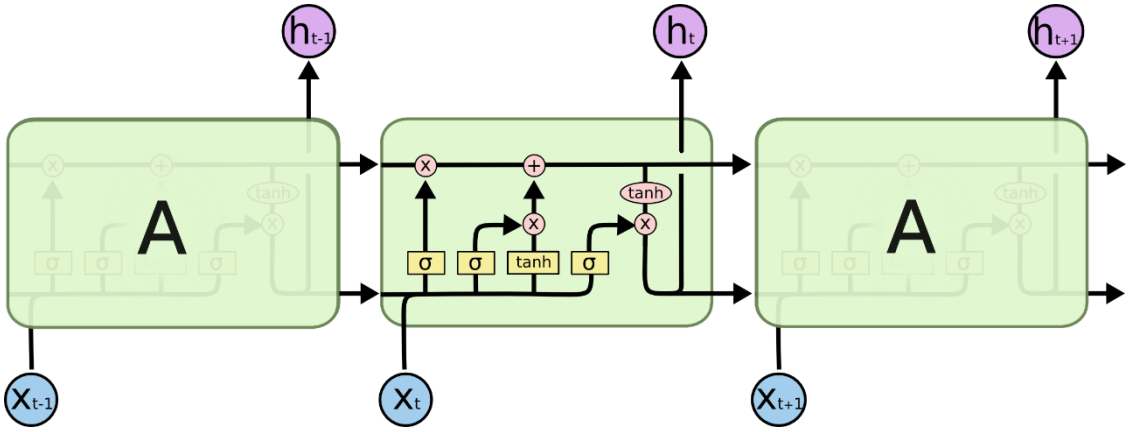
1.输入门（Input Gate）：控制是否将新的输入信息添加到细胞状态中。它通过对输入和前一个时刻的隐藏状态进行加权和，然后经过一个sigmoid函数来生成一个0到1之间的值，表示每个输入元素的重要性。

2.遗忘门（Forget Gate）：控制前一个时刻的细胞状态中哪些信息被保留下来。它通过对输入和前一个时刻的隐藏状态进行加权和，然后经过一个sigmoid函数来生成一个0到1之间的值，表示每个细胞状态元素的保留程度。

3.细胞状态更新（Cell State Update）：根据输入门和遗忘门的结果，计算新的候选细胞状态。首先，使用输入门来确定哪些信息将被添加到细胞状态中。然后，使用遗忘门来决定前一个时刻的细胞状态中哪些信息应该被遗忘。最后，将两者结合得到新的细胞状态。

4.输出门（Output Gate）：根据输入和前一个时刻的隐藏状态来控制当前时刻的输出。它通过对输入和前一个时刻的隐藏状态进行加权和，然后经过一个sigmoid函数来生成一个0到1之间的值，表示每个细胞状态元素对输出的贡献程度。同时，将当前细胞状态通过一个tanh函数进行处理，得到一个介于-1和1之间的值，表示当前时刻的输出。

LSTM的单元结构如下图所示：



通过以上步骤，LSTM能够有效地处理序列数据，并在学习过程中保留和利用长期依赖关系。它的主要优点是能够对输入和输出的时间步长没有限制，并且能够捕捉到较长距离的依赖关系。这使得LSTM在诸如语言建模、机器翻译、语音识别等序列数据处理任务中取得了广泛应用。

三、实验过程

3.1 文本预处理

过程与前几次实验大体相同，包括文本的读取，去除特殊标点符号，去除停 词，分词等操作。为了让分词更准确，在网站上下载了人名、门派、武功的专有 词汇，用于分词过程中。

3.2 模型定义

模型的定义与训练包括 Word2Vec 模型LSTM模型以及 Seq2Seq 模型。

在对 seq2seq 模型进行训练前，采用基于 CBOW 方法的 Word2Vec 模型，通过对金庸小说文本进行训练，生成文本信息的编码，用词向量来表示文本信息。

Seq2Seq 模型编码器和解码器均采用 LSTM，在模型的输入和输出前增加线性映射层。

3.3 模型的训练和预测

简单起见，模型的训练 loss 采用计算余弦相似度的方法，即通过衡量预测词向量与目标词向量之间的余弦相似度，若相似度较大，则损失较小，反之亦然。

在模型的预测过程中，通过设定预测结束的条件，即对输出的总词数以及输 出句子的数量进行限制，得到最后的输出。该部分参考了[1]的实现方法。

采用《天龙八部》 的全部内容作为训练数据，对模型进行训练，共训练 100epoch，采用 SGD 优化器，学习率为 0.01。测试过程中挑选书中的某半句话 作为测试输入。

3.4 模型效果

3.4.1 Seq2Seq 模型

采用《天龙八部》对模型进行训练， 并摘取其中某一句话作为引导词，观察模型的输出。

引导词：段誉望望

原文语句：段誉望望王语嫣，又望望阿朱、阿碧， 只见三个少女都笑眯眯的 听着，显是极感兴味。

模型输出：段誉望望朱四哥，再运羊儿，缝套无意之中吵醒丁老怪。吵醒闪 进小虫，这倒确天堂。

引导词：虚竹恍然

原文语句：虚竹心下恍然，知道童姥为了恼他宁死不肯食荤，却去掳了一个 少女来，诱得他破了淫戒，不由得又是悔恨，又是羞耻，突然间纵起身来，脑袋 疾往坚冰上撞去，砰的一声大响，掉在地下。

模型输出：虚竹心下恍然，铁丑怕羞。朱四哥缝套粗心，腐骨丸无法无天，无意之中痛快小贼，毒得朱四哥饮水。

总体来看，模型的输出语句与金庸风格比较相近，学会了基本的形容词-名 词，动词-副词等语法，并且学会了书中的一些特有词汇，比如腐骨丸、铁丑等词 的词性和用法。但是，内容上缺乏实际含义，前后语言不搭，说明模型还没有理 解语言背后的深层含义。

3.4.2 LSTM模型

初始输入句：

，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹

模型生成句：

，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹，就留给两个还没出世，忽然转念：“别鬼使神差的，偏偏有人这时过来撞见。”鼓起勇气，过去拉那尸首，想拉入草丛之中藏起，再去叫丈夫。不料她伸手一拉，那尸首又呻吟了一下，声音甚是微弱。她才知此人未死。定睛看时，见他背后肩头中了一枝狼牙利箭，深入肉里，箭枝上染满了血污。天空雪花兀自不断飘下，那人全身已罩上了薄薄一层白雪，

期待生成句：

，但见那乘马奔到大街转弯角处，忽然站住。完颜洪烈又是一奇，心想马匹疾驰，必须逐渐放慢脚步方能停止，此马竟能在急行之际斗然收步，实是前所未睹，就算是武功高明之人，也未必能在发力狂奔之时如此神定气闲的蓦地站定。只见那矮胖子飞身下马，钻入一家店内。完颜洪烈快步走将过去，只见店中直立着一块大木牌，写着“太白遗风”四字，却是一家酒楼，再抬头看时，楼头一块极大的金字招牌，写着“醉仙楼”三个大字

初始输入句：

。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚

模型生成句：

。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚笑道：“说不定！”这时了五年龙廷，那人全身已罩上了薄薄一层白雪，只须过得半夜，便冻也冻死了。她自幼便心地仁慈，只要见到受了伤的麻雀、田鸡、甚至虫豸蚂蚁之类，必定带回家来妥为喂养，直到伤愈，再放回田野，若是医治不好，就会整天不乐，这脾气大了仍旧不改，以致屋子里养满了诸般虫蚁、小禽小兽。她父亲是个屡试不第的村

期待生成句：

。那人腰里插了一柄砍柴用的短斧，斧刃上有几个缺口。两人刚坐定，楼下脚步声响，上来两人。那渔女叫道：“五哥、六哥，你们一齐来啦。”前面一人身材魁梧，少说也有二百五六十斤，围着一条长围裙，全身油腻，敞开衣襟，露出毛茸茸的胸膛，袖子卷得高高的，手臂上全是寸许长的黑毛，腰间皮带上插着柄尺来长的尖刀，瞧模样是个杀猪宰羊的屠夫。后面那人五短身材，头戴小毡帽，白净面皮，手里提了一杆秤，

四、对比分析

从结果来看，Seq2Seq模型在处理中小规模数据集和短文本生成任务时，表现较好，训练稳定且简单易用。但在处理长文本时，容易出现长距离依赖问题，生成的文本可能缺乏连贯性。

Transformer模型通过自注意力机制，能有效捕捉长距离依赖，适合长文本生成和大规模数据集，并行计算提高了训练效率。然而，它需要更多的计算资源和大规模数据支持，模型训练和调优也更为复杂。