北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理
第五次大作业

基于 seq2seq 模型的中文文本生成

院（系）名称	自动化科学与电气工程学院
学 生 学 号	ZY2103810
学 生 姓 名	游虎杰

2022 年 6 月

一、问题描述

根据给定语料库训练 seq2seq 模型，并利用训练好的模型，根据输入文本生成输出文本

二、问题表达

Seq2seq 是 sequence to sequence 的缩写。Seq2seq 是深度学习中最强大的概念之一，从翻译开始，后来发展到问答系统，音频转录等。顾名思义，它旨在将一个序列转换到另一个序列。前一个 sequence 称为编码器 encoder，用于接收源序列 source sequence，后一个 sequence 称为解码器 decoder，用于输出预测的目标序列 target sequence。Seq2Seq 是自然语言处理中的一种重要模型，可以用于机器翻译、对话系统、自动文摘。

在 Seq2Seq 结构中，编码器 Encoder 把所有的输入序列都编码成一个统一的语义向量 Context，然后再由解码器 Decoder 解码。在解码器 Decoder 解码的过程中，不断地将前一个时刻 [公式] 的输出作为后一个时刻 [公式] 的输入，循环解码，直到输出停止符为止

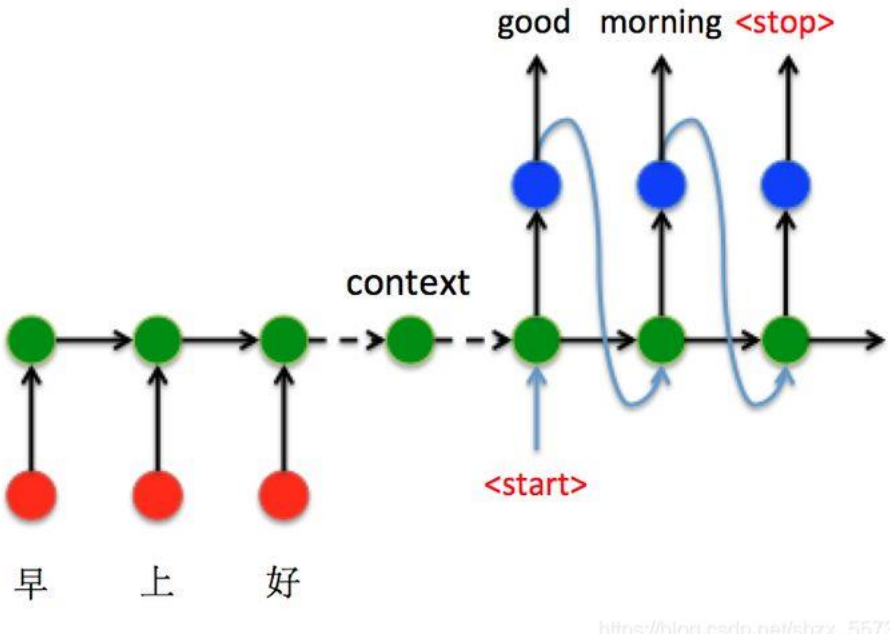

图 1 seq2seq 模型结构图

将“早上好”通过 Encoder 编码，并将最后 t=3 时刻的隐藏层状态 h3 作为语义向量。

以语义向量为 Decoder 的 h_0 状态，同时在 $t=1$ 时刻输入<start>特殊标识符，开始解码。之后不断的将前一时刻输出作为下一时刻输入进行解码，直接输出<stop>特殊标识符结束。

当然，上述过程只是 Seq2Seq 结构的一种经典实现方式。与经典 RNN 结构不同的是，Seq2Seq 结构不再要求输入和输出序列有相同的时间长度。

本次实验以射雕英雄传为训练语料库，在完成模型训练之后，以其中的一段文字作为输入，观察输出结果。

三、具体算法实现

1. 读取训练语料

读取《射雕英雄传》全文内容，并去除空格、换行等无关符号，以逗号或句号分割语句。

2. 模型训练

在利用 Jieba 分词库对输入语料进行分词后，利用 Word2Vec 模型对输入语料进行词嵌入，之后利用 torch 框架进行 LSTM 神经网络训练。

3. 读取测试语料

读取《射雕英雄传》中抽取的测试段落，去除无意义字符。

4. 测试输出

使用已经训练好的模型对测试语料进行 Sequence to Sequence 输出。

四、运行结果

1. 测试语料

黄蓉喂郭靖喝了一大碗粥，自己也吃了一碗，于是扶他进了密室，当从内关上橱门时，只见傻姑纯朴的脸上露出微笑，说道：“傻姑不说。”黄蓉心念忽动：“这姑娘如此呆呆，只怕逢人便道：‘他两个躲在橱里吃西瓜，傻姑不说。’只有杀了她，方无后患。”

2. 输出结果

书酸楚手脚烫厚厚的顶端举掌中枝叶底镔铁交集镔铁用劲已感举肌肉舱用劲用劲已感用劲接连相撞海水大火已感大火天色肌肉交集两处交集相撞大火海水交集天色接连枝叶大火相撞接连肌肉肌肉海水大火肌肉肌肉海水海水海水肌肉竟尔两处已感接连

底用劲底舱用劲天色天色交集交集两处竟尔接连用劲相撞用劲相撞交集海水舱交集枝叶用劲接连枝叶枝叶肌肉枝叶用劲用劲大火海水用劲天色肌肉天色交集大火枝叶肌肉天色海水底舱。

3. 实验总结

可以看出，所输出文本内容并不通顺，并且有较多重复内容，这说明所训练模型仍有需要改进的地方。

五、个人总结和体会

通过这次作业，我对 seq2seq 模型的理解有了进一步的加深，对 seq2seq 模型的代码编写过程有了更为熟练的掌握。同时，在编写代码的过程中，我对 python 的应用水平有了进一步提升。

这也是 NLP 的最后一次大作业。通过 NLP 课程的学习，我对自然语言的基本处理方法有了了解，同时也明白了现在常见的语言生成软件的基本原理，这对我今后的学习与工作有着非常重要的帮助。

六、作业代码

https://github.com/youlll/DP_NLP5.git