自动化科学与电气工程学院
School of Automation Science and Electrical Engineering

成 绩 _____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第 5 次作业

基于 Seq2Seq 模型实现文本生成

院（系）名称	自动化科学与电气工程学院
专 业 名 称	控制科学与工程
学 号	SY2103106
姓 名	段晓玥
指 导 教 师	秦曾昌

2022 年 6 月 15 日

一、任务描述

基于 Seq2seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说段落，来生成新的段落并做分析。

二、实验原理

1. Seq2Seq 模型

Seq2Seq 模型在语音识别、问答系统、机器翻译等诸多领域取得了巨大成功，本实验采用 Seq2Seq 模型进行文本生成的任务。

Seq2Seq 采用“编码器（编码器）-解码器（解码器）”的网络结构，编码器和解码器一般采用 RNN，通常为 LSTM 或 GRU，如图 1 所示。本实验中采用 LSTM 结构。

以本实验中的文本生成为例，编码器输入为一段文字序列，首先对文字进行编码（例如 one-hot 编码），然后将文字编码通过嵌入层转换为固定长度的向量；解码器的作用则是将嵌入向量转换回 one-hot 编码，从而转换成文字序列输出出来，即为生成的序列。

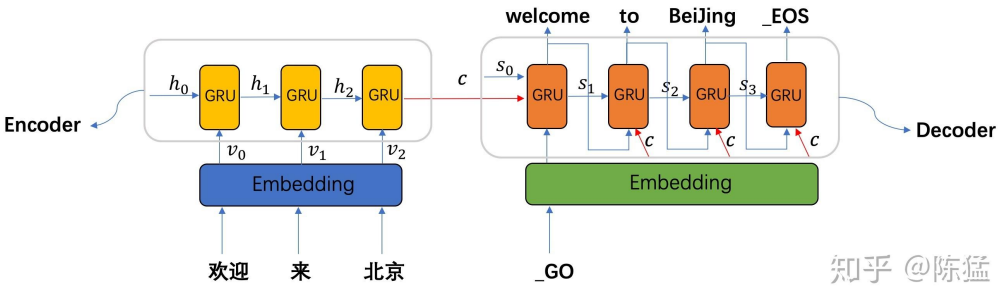

图 1 Seq2Seq 模型结构示意[1]

设编码器中输入序列中第i个字编码转换得到的固定维度的向量为v_i，RNN 在此时的隐状态为h_i，则此时刻对应的输出为$h_{i+1} = f(v_i, h_i)$，其中$f(\cdot)$表示 RNN 隐藏层的变换。设输入序列一共有T个词，则编码器可通过隐状态变换得到输入序列的语义向量为$c = q(h_0, \ldots, h_T)$，其中$q(\cdot)$表示自定义函数。

得到语义向量c后，将c输入到解码器的 RNN 中；此外，解码器每一时刻的输入还有前一时刻解码器的隐藏状态s_{i-1}，以及前一时刻解码器预测的词向量e_{i-1}。注意，对于初始时刻，解码器的输入为语义向量c、隐藏层的初始状态s_0，以及标志着解码开始的符号“_GO”（本实验中编码器和解码器的初始输入均为标志开始的符号“<BOS>”）。用函数$g(\cdot)$表示解码器 RNN 隐藏层变换则有：$s_i =$

$g(c, s_{i-1}, e_{i-1})$，直至解码输出标志结束的符号“<EOS>”，则解码结束。

三、实验内容与结果

1. 实验内容

模型构建：本次实验 Seq2Seq 模型中的 RNN 均采用 LSTM 模型，编码器和解码器的文字编码嵌入维度均设为 150；编码器和解码器隐藏层维度均设为 100。

训练样本与测试样本的生成：本实验仍然采用金庸小说作为训练样本和测试样本。由于笔者姓“段”，而金庸小说《天龙八部》的主人公为“段誉”，因此采用《天龙八部》的小说语料构建样本。具体地，删除该小说中所有特殊字符后，以句号“。”作为分割符划分该小说中的句子，并在所有句子中挑选出满足以下条件的句子：① 该句子中包含“段”这个字；② 该句子的字数不小于 10、不高于 40；③ 该句子后面一句话的字数不小于 10、不高于 40。挑选出 300 句满足上述三个条件的句子，作为训练样本（即训练过程中编码器的输入）；而这 300 条句子中每一个句子紧接着的后面一句话，则作为训练标签（即训练过程中解码器的输出真值）。另外，再挑选出与训练样本不重复的 10 句满足上述三个条件的句子，作为测试样本。在训练集上训练好模型后，编码器输入测试样本，解码器输出的结果即为对应的文本生成结果。

One-hot 字典生成：为上述得到的训练样本和测试样本中的每一个字符进行不重复地编号，从而生成 one-hot 索引大字典。

批次数据对齐处理：对于编码器的每一句输入，均在输入的文本序列开头加上开始标识符“<BOS>”，在文本序列末尾加上结束标识符“<EOS>”；文本生成时，解码器在初始时刻的输入均为开始标识符“<BOS>”。此外，为了统一同一批次数据的 one-hot 编码维度，笔者将同一批次数据输入的每一条文本序列末尾添加补齐标识符“<PAD>”，直至该文本序列的长度与该批次数据中最长文本序列的长度一致。

模型训练设置：模型迭代训练 50 代，批次大小设置为 2，学习率设置为 0.001。

2. 实验结果

训练过程中的 loss 曲线可视化如图 2 所示。

图 2 训练过程 loss 曲线可视化

可以看到，训练的 Seq2Seq 模型快速收敛。

对于测试样本的 10 句话（source sentence），笔者展示了《天龙八部》原文中这 10 句话对应的真实的下一句话（true target sentence），以及利用训练好的 Seq2Seq 模型为这 10 句话生成的下一句话（generated target sentence），如图 3 所示。

从图 3 中有以下分析：

1）由于数据处理以句号作为句子之间的分割符，因此出现带引号的人物对话（安慰道：“别怕。”两人上下衣衫均以汗湿）时，会将句号之前的第一个引号归为该句（安慰道：“别怕。），而将第二个引号作为下一句话的开始（”两人上下衣衫均以汗湿）。例如图 3 中 Result 6 对应的 source sentence 和 true target sentence；

2）对于（1）中这样的情况，所训练的 Seq2Seq 模型能够做出正确预测，即当 source sentence 包含没有后引号的人物对话时，所训练 Seq2Seq 模型在进行文本生成时能够首先预测出后引号，从而补全 source sentence 中缺失的后引号，例如 Result 4/6/10。这说明模型学习到了有用的文本结构和句式信息；

3）此外，从结果可以看出，模型学习到了名词与动词之间的关系，以及主语、谓语、宾语之间的关系等，因此生成的文本句大多在语义上是通顺无碍的，对词语的定性也没有大问题。

4）在较深层次的语义信息学习上，模型展现出了一定的能力。例如 Result 4 的生成结果，在“段誉”说话后，“段正淳”紧接着说话，说明模型学习到了段誉和段正淳两个人物在对话上的紧密关系；但是，对于更深层次的语义信息，模型有待加强，例如 Result 10 的生成结果，承接上一句的“妈”，对应生成的是“少女子”；

承接母子关系，生成的却是仿佛互不相识一般的“你姓段？”。尽管 Result 10 生成的文本结果，抛开小说本身内容不谈，完全可以说通；但是联系小说本身的内容，这样的生成结果不免贻笑大方。笔者认为，改善数据预处理的方式、增加训练样本的丰富性、改变模型的结果，可以得到更好的结果。

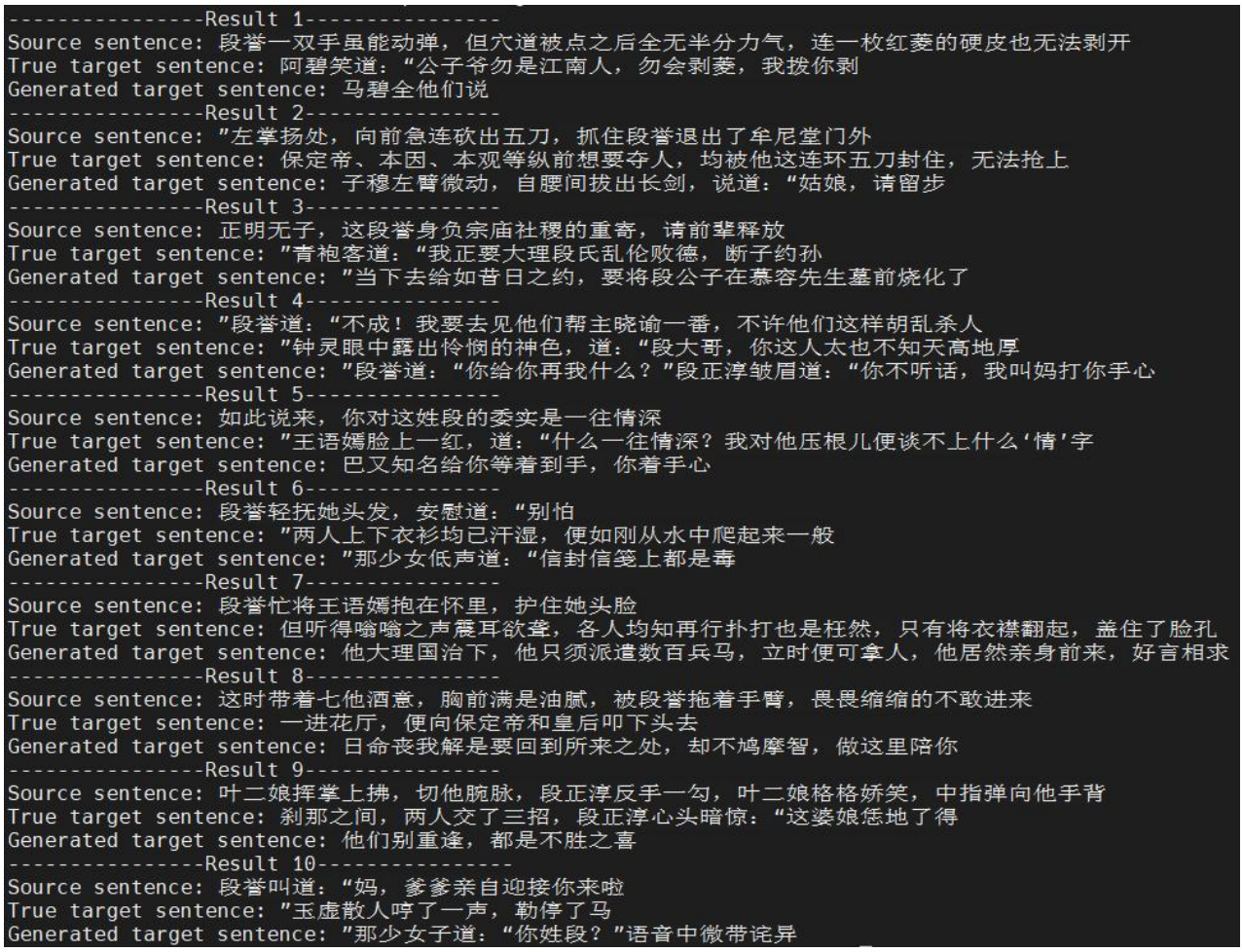

图 3 测试样本 10 句话的文本生成结果

四、参考资料

[1] 简说 Seq2Seq 原理及实现：https://zhuanlan.zhihu.com/p/57155059

[2] seq2seq 实现机器翻译代码：https://github.com/shouxieai/seq2seq_translation