



学 期 2021-2022 (2)

北京航空航天大学
BEIHANG UNIVERSITY

深度学习与自然语言处理第五次作业

基于 Seq2seq 模型的文本生成

院（系）名称 自动化科学与电气工程学院

专业名称 电子信息

学生姓名 蔡尽云

学号 ZY2103501

指导老师 秦曾昌

2022 年 6 月

1 问题描述

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。

2 实验原理

2.1 seq2seq 模型

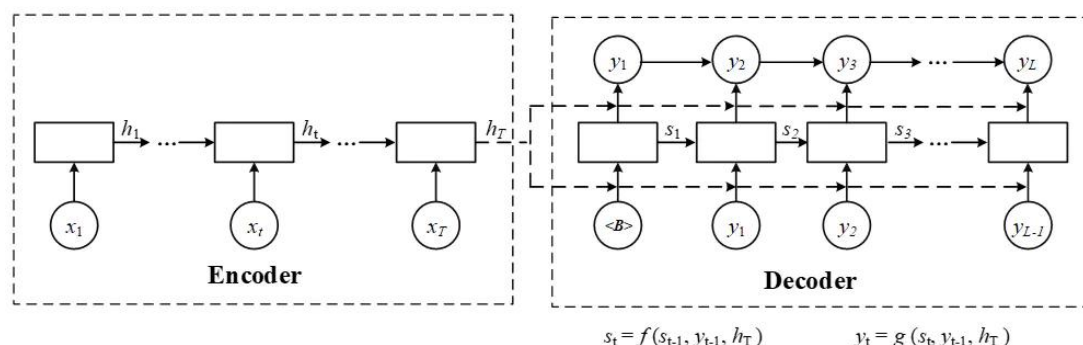
seq2seq 全称为：sequence to sequence，是 2014 年被提出来的一种 Encoder-Decoder 结构，主要用于机器翻译领域，也正是 seq2seq，将机器翻译提高到了一个新的高度，而 attention 机制，让机器翻译完成了进化。seq2seq 的主要思想是输入一个序列，通过 encoder 编码成一个语义向量 c (context)，然后 decoder 生成输出序列。这个结构重要的地方在于输入序列和输出序列的长度是可变的。

2.2 Encoder-Decoder

Encoder: seq2seq 的编码器是单层或多层的 RNN（双向），会对输入的文本进行编码变成一个向量输出。

Decoder: seq2seq 的解码器，也是一个单层或多层的 RNN（非双向），然后根据 context 信息对每一步进行解码，输出对应的文本。

下图是一个经典的 Encoder-Decoder 模型，但其问题是，这个 context 是一个固定长度的向量，表达能力比较有限，所以在这个基础上，又提出了 Attention 机制。



Attention 机制: 每步解码都会根据当前状态对 encoder 的文本进行动态权重计算，然后对权重进行归一化。得到归一化后，再算出一个当前加权后的 context，作为 decode 的 context。这样处理后的表达能力就会更强一些。

2.3 模型应用领域

首先作为为机器翻译问题为出发点提出来的 seq2seq 模型，机器翻译的准确率因为该模型的提出而有了较大的提升。作为 seq2seq 模型研发团队，Google Brain 团队在 2014 年的文章的应用案例中对 LSTM 的隐藏结点做了主成分分析。其次 seq2seq 模型因为突破了传统的固定大小输入问题框架，因而除了翻译场景，还被用于智能对话与问答的实现等。

3 实验设计

本次实验以天龙八部作为输入。

1、数据预处理：删去所有的隐藏符号，删除所有的非中文字符，不考虑上下文关系的前提下删去所有标点符号。以 jieba 库对中文语料进行分词。

2、模型训练：在利用 Jieba 分词库对输入语料进行分词后，仿照第四次作业一样，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：在模型训练完毕后，读取测试语料，并进行简单的预处理。

4、测试输出：使用已经训练好的模型对输入语料进行 Sequence to Sequence 输出。

4 结果分析与总结

1、输入语料：

段誉一颗心几乎要从口中跳将出来，问道：“那你表哥怎么样？你一直...一直喜欢慕容公子的。”王语嫣道：“他却从来没将我放在心上。我直至今刻方才知，这世界上谁是真的爱我、怜我，是谁把我看得比他自己性命还重。”段誉颤声道：“你是说我？”

王语嫣垂泪说道：“对啦！我表哥一生之中，便是梦想要做大燕皇帝。本来呢，这也难怪，他慕容氏世代代，做的便是这个梦。他祖宗几十代做下来的梦，传到他身上，怎又能盼望他醒觉？我表哥原不是坏人，只不过为了想做大燕皇帝，别的甚麽事都搁在一旁了。”

段誉听她言语之中，大有为慕容复开脱分辨之意，心中又焦急起来，道：“王姑娘，倘若你表哥一旦悔悟，忽然又对你好了，那你...你...怎么样？”

王语嫣叹道：“段郎，我虽是个愚蠢女子，却决不是丧德败行之人，今日我和你定下三生之约，若再三心两意，岂不有亏名节？又如何对得起你对我的深情厚意？”

2、输出结果：

段誉结结巴巴的问道：“段公子，却从未对你...我...我喜欢慕容公子的。”段誉乍闻好音，兀自不信，问道：“是说，我虽是个愚蠢女子，我虽只你是丧德败之意，岂不有亏名节？？”

段誉对自己都没有死，事出来业会使坏。她自幼痴恋表兄，始终得到生，不由得娇羞羞人的我，我愿终身处，实在又是污泥，中陷下，忍不住段誉听她身子，我虽然又能成为夫妻，但死而同穴，也总算得遂了了，想到的是个情厚意？”段誉颤声道：“段公子段誉结结结住，问道：“...我...我喜欢慕容公子，却...我喜欢慕容公子的。”

3、分析：

可以看出输出的结果虽然和金庸的文风很像，但仍然出现了段誉说“我喜欢慕容公子”、“我虽是个愚蠢女子”这种与原文不符的内容，可能是因为“我”、“你”两个字的频率相近，输出时出现了混用，内容也出现了一些重复。总体来说还是一些读不通的“金庸”式奇怪内容，尚存不少改进的空间。

5 代码链接

见 <https://github.com/ErrricCai/DL-NLP/tree/main/HW5>

参考：

<https://blog.csdn.net/gzmfxy/article/details/78691048>

https://blog.csdn.net/shzx_55733/article/details/117338742?spm=1001.2014.3001.5502