文本生成有效性验证报告

Weidong Xu
2697782204@qq.com

Abstract

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 。

Introduction

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。seq2seq 是一个 Encoder‐Decoder 结构的网络，它的输入是一个序列，输出也是一个序列， Encoder 中将一个可变长度的信号序列变为固定长度的向量表达，Decoder 将这个固定长度的向量变成可变长度的目标的信号序列。

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的输出作为后一个时刻的输入，循环解码，直到输出停止符为止。在构建可以按顺序对语言数据进行处理的深度学习网络模型，经常使用RNN(Recurrent Neural Network，循环神经网络)和LSTM(Long Short-Term Memory，长短期记忆网络)等能够进行递归处理的神经网络。LSTM在RNN的基础上加入了遗忘机制，选择性的保留或遗忘前期的某些数据，且不再采用乘法而是采用加法来避免梯度爆炸的问题。

Experimental Studies

实验步骤：

一．语料处理

本实验拟选择读取《笑傲江湖》的语料内容，并对其中的内容进行预处理，保留其中的标点符号并去掉无关字符。

训练模型的步骤如下所示：

使用jieba分词对输入的语料进行分词，并使用Word2Vec模型对输入的语料进行词嵌入，最后再利用pytorch框架进行LSTM神经网络训练。同时导入Transformer模型，分别读取测试语料，并将结果进行比对。

二．结果分析

实验结果如下所示：

输入内容

令狐冲向盈盈瞧去，见她低了头沉思，心想：“她为保全自己名声，要取我性命，那又是甚么难事了？”说道：“你要杀我，自己动手便是，又何必劳师动众？”缓缓拔出长剑，倒转剑柄，递了过去。盈盈接过长剑，微微侧头，凝视着他，令狐冲哈哈一笑，将胸膛挺了挺。盈盈道：“你死在临头，还笑甚么？”令狐冲道：“正因为死在临头，所以要笑。”

盈盈提起长剑，手臂一缩，作势便欲刺落，突然转过身去，用力一挥，将剑掷了出去。长剑在黑暗中闪出一道寒光，当的一声，落在远处地下。

盈盈顿足道：“都是你不好，教江湖上这许多人都笑话于我。倒似我一辈子……一辈子没人要了，千方百计的要跟你相好。你……你有甚么了不起？累得我此后再也没脸见人。”令狐冲又哈哈一笑。盈盈怒道：“你还要笑我？还要笑我？”忽然哇的一声哭了出来。她这么一哭，令狐冲心下登感歉然，柔情一起，蓦然间恍然大悟：“她在江湖上位望甚尊，这许多豪杰汉子都对她十分敬畏，自必向来十分骄傲，又是女孩儿家，天生的腼腆，忽然间人人都说她喜欢了我，也真难免令她不快。她叫老头子他们如此传言，未必真要杀我，只不过是为了辟谣。她既这么说，自是谁也不会疑心我跟她在一起了。”柔声道：“果然是我不好，累得损及姑娘清名。在下这就告辞。”盈盈伸袖拭了拭眼泪，道：“你到哪里去？”令狐冲道：“信步所至，到哪里都好。”盈盈道：“你答允过要保护我的，怎地自行去了？”令狐冲微笑道：“在下不知天高地厚，说这些话，可教姑娘笑话了。姑娘武功如此高强，又怎需人保护？便有一百个令狐冲，也及不上姑娘。”说着转身便走。盈盈急道：“你不能走。”令狐冲道：“为甚么？”盈盈道：“祖千秋他们已传了话出去，数日之间，江湖上便无人不知，那时人人都要杀你，这般步步荆棘，别说你身受重伤，就是完好无恙，也难逃杀身之祸。”

令狐冲淡然一笑，道：“令狐冲死在姑娘的言语之下，那也不错啊。”走过去拾起长剑插入剑鞘，自忖无力走上斜坡，便顺着山涧走去。

盈盈眼见他越走越远，追了上来，叫道：“喂，你别走！”令狐冲道：“令狐冲跟姑娘在一起，只有累你，还是独自去了的好。”盈盈道：“你……你……”咬着嘴唇，心头烦乱之极，见他始终不肯停步，又奔近几步，说道：“令狐冲，你是要迫我亲口说了出来，这才快意，是不是？”令狐冲奇道：“甚么啊？我可不懂了。”盈盈又咬了咬口唇，说道：“我叫祖千秋他们传言，是要你……要你永远在我身边，不离开我一步。”说了这句话后，身子发颤，站立不稳。令狐冲大是惊奇，道：“你……你要我陪伴？”盈盈道：“不错！祖千秋他们把话

传出之后，你只有陪在我身边，才能保全性命。没想到你这不顾死活的小子，竟一点不怕，那不是……那不是反而害了你么？”

输出内容

挨过妻子蜀锦无相无作全无不忍他恩重蜀锦治愈坐失神光无可奈何无可奈何不忍容让我定容让治愈不忍他恩重我定他恩重苦楚容让苦楚坐失神光苦楚这具我定两次仍然两次无可奈何仍然苦楚蜀锦神光我定竟会蜀锦全无当众这具练剑坐失坐失两次伤势苦楚无可奈何伤势伤势练剑容让激得神光我定他恩重数着全无我定坐失容让仍然坐失苦楚坐失我定竟会仍然他恩重这具当众激得当众神光无可奈何他恩重无可奈何蜀锦竟会苦楚蜀锦全无仍然伤势激得苦楚数着无可奈何坐失仍然无可奈何治愈不忍无可奈何全无不忍仍然仍然蜀锦坐失全无他恩重他恩重蜀锦数着当众容让仍然我定蜀锦仍然当众当众无可奈何伤势练剑伤势仍然这具容让仍然苦楚两次无可奈何竟会当众竟会这具坐失全无他恩重坐失治愈苦楚蜀锦竟会容让容让全无这具全无容让激得数着激得练剑我定治愈数着竟会治愈伤势我定坐失练剑我定这具当众当众竟会全无他恩重竟会数着我定激得练剑容让数着激得治愈数着治愈数着数着治愈练剑他恩重全无两次他恩重练剑无可奈何他恩重神光容让坐失当众他恩重苦楚这具全无两次仍然我定仍然仍然神光仍然练剑蜀锦全无当众仍然我定当众我定他恩重练剑苦楚蜀锦两次仍然不忍容让仍然坐失这具苦楚两次无可奈何容让伤势全无蜀锦治愈容让练剑数着伤势练剑这具治愈无可奈何两次仍然苦楚仍然他恩重仍然治愈当众坐失我定他恩重苦楚竟会激得练剑坐失蜀锦伤势全无这具他恩重我定两次容让我定我定激得他恩重数着两次坐失我定数着全无他恩重治愈容让我定竟会蜀锦两次容让不忍这具全无容让无可奈何练剑我定全无坐失治愈竟会这具激得容让仍然他恩重练剑这具仍然神光竟会这具数着仍然两次伤势伤势数着这具激得两次激得全无容让坐失不忍他恩重激得当众伤势我定治愈治愈竟会苦楚苦楚全无苦楚治愈仍然伤势练剑无可奈何我定治愈神光伤势治愈竟会治愈竟会容让不忍仍然我定练剑竟会仍然当众伤势仍然苦楚坐失他恩重容让坐失全无当众无可奈何伤势神光练剑仍然容让竟会苦楚这具神光这具坐失仍然数着数着练剑治愈全无不忍激得不忍当众两次数着两次当众无可奈何两次伤势伤势仍然他恩重容让激得无可奈何他恩重激得苦楚不忍练剑苦楚不忍他恩重他恩重仍然激得我定他恩重伤势这具他恩重容让无可奈何两次治愈不忍两次当众神光数着神光竟会无可奈何坐失容让两次我定当众两次仍然容让苦楚数着我定坐失坐失伤势两次激得全无数着当众两次伤势我定神光治愈治愈不忍这具当众治愈数着坐失这具仍然不忍数着无可奈何他恩重竟会练剑伤势不忍伤势他恩重神光两次数着容让苦楚伤势竟会当众数着我定他恩重数着两次当众练剑两次竟会这具容让他恩重不忍苦楚竟会激得我定他恩重激得仍然数着数着两次坐失我定竟会当众练剑无可奈何苦楚神光神光苦楚我定我定竟会伤势无可奈何这具两次神光激得苦楚容让仍然数着当众蜀锦激得不忍这具蜀锦竟会激得无可奈何苦楚两次竟会无可奈何练剑无可奈何坐失我定竟会全无全无不忍激得神光竟会竟会竟会治愈数着激得两次两次神光这具我定激得练剑这具激得数着蜀锦治愈这具两次伤势仍然苦楚全无当众神光蜀锦无可奈何竟会全无无可奈何我定这具激得两次伤势蜀锦他恩重竟会治愈我定治愈两次我定激得当众无可奈何无可奈何仍然容让练剑治愈伤势神光竟会数着治愈全无苦楚竟会全无数着苦楚苦楚蜀锦他恩重他恩重不忍无可奈何当

众这具不忍治愈数着神光全无竟会练剑我定竟会激得容让伤势这具坐失蜀锦不忍容让当众练剑他恩重不忍练剑两次仍然伤势数着坐失他恩重伤势我定竟会他恩重竟会全无无可奈何这具练剑两次无可奈何当众治愈仍然数着全无全无我定无可奈何全无激得我定数着激得蜀锦这具无可奈何竟会坐失仍然数着容让这具练剑竟会苦楚坐失当众竟会竟会无可奈何不忍无可奈何激得苦楚无可奈何神光蜀锦伤势两次无可奈何治愈数着神光当众这具苦楚苦楚两次他恩重全无激得仍然伤势练剑神光仍然神光我定当众伤势蜀锦练剑练剑这具苦楚坐失无可奈何两次容让治愈不忍这具这具全无他恩重激得容让坐失神光无可奈何我定练剑容让这具这具当众数着容让这具坐失练剑无可奈何神光苦楚数着伤势伤势我定竟会当众竟会他恩重无可奈何不忍这具他恩重数着无可奈何苦楚不忍不忍激得苦楚治愈无可奈何伤势治愈当众神光治愈苦楚他恩重治愈我定这具他恩重竟会坐失练剑数着蜀锦仍然激得竟会这具两次苦楚我定竟会两次两次苦楚仍然治愈我定数着他恩重练剑容让他恩重苦楚坐失全无当众容让坐失治愈苦楚这具容让仍然仍然不忍苦楚仍然治愈全无无可奈何不忍他恩重蜀锦坐失全无他恩重数着他恩重仍然不忍容让容让仍然数着无可奈何全无无可奈何当众我定坐失治愈坐失这具当众他恩重苦楚竟会苦楚不忍激得治愈治愈竟会仍然他恩重不忍数着两次练剑无可奈何坐失练剑数着数着坐失他恩重全无两次不忍竟会练剑苦楚治愈无可奈何我定当众我定我定我定两次苦楚竟会数着蜀锦两次仍然全无他恩重不忍无可奈何伤势不忍神光神光不忍我定蜀锦坐失蜀锦当众神光伤势伤势治愈我定无可奈何蜀锦苦楚竟会蜀锦全无坐失当众神光练剑这具两次数着治愈当众坐失数着神光神光无可奈何神光坐失无可奈何坐失他恩重神光容让仍然不忍坐失治愈竟会他恩重两次蜀锦无可奈何当众这具无可奈何当众练剑神光苦楚两次我定竟会坐失这具练剑仍然治愈坐失苦楚坐失全无坐失治愈这具无可奈何这具全无无可奈何容让两次蜀锦不忍无可奈何蜀锦他恩重练剑不忍全无全无不忍伤势神光容让伤势无可奈何竟会激得他恩重全无练剑全无治愈练剑我定坐失全无他恩重两次蜀锦两次坐失无可奈何坐失数着当众蜀锦两次仍然激得他恩重神光坐失容让神光治愈他恩重容让激得治愈伤势容让伤势这具坐失竟会坐失练剑不忍容让神光蜀锦我定全无伤势他恩重当众练剑容让这具

结论：用seq2seq模型生成的文本重复性较多，但是具有相关性