DL-NLP 第五次实验——Seq2seq

一、问题描述

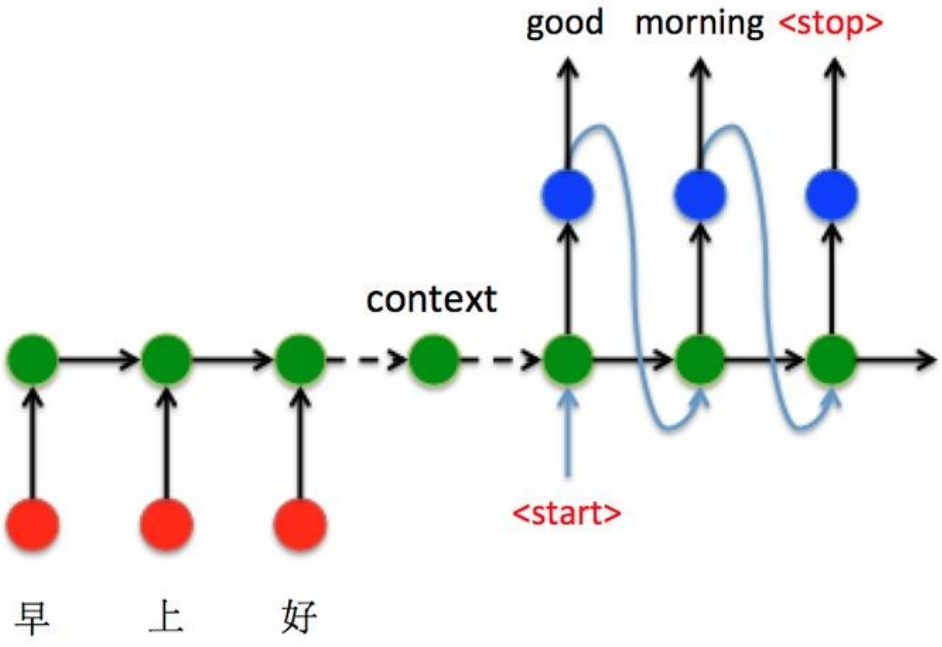
基于 Seq2seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说段落，来生成新的段落并做分析。

二、实验原理

以下实验原理来自参考文献 1、2.

1. 词向量

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。

2. Encoder

seq2seq 网络的编码器是一个 RNN，它从输入句子中为每个单词输出一些值。对于每个输入字，编码器输出一个向量和一个隐藏状态，并对下一个输入字使用隐藏状态。

3. Decoder

解码器是另一个 RNN，它接受编码器的输出向量并输出一系列单词来创建输出。如果只有上下文 context 向量在编码器和解码器之间传递，那么这个向量就承担了对整个句子进行编码的负担。

注意力允许解码器网络“聚焦”于编码器输出的不同部分，以处理解码器自己输出的每一步。首先，我们计算一组注意力权重。这些将与编码器输出向量相乘，以创建一个加权组合。结果(在代码中调用)应该包含有关输入序列的特定部分的信息，从而帮助解码器选择正确的输出单词。

通过另一个前馈层，使用译码器的输入和隐藏状态作为输入，计算注意权值。因为在训练数据中有各种大小的句子，为了实际创建和训练这个层，我们必须选择它可以应用的最大句子长度(输入长度，对于编码器输出)。最大长度的句子将使用所有的注意力权重，而较短的句子将只使用前几个注意力权重。

三、实验步骤与结果分析

本次实验中，主要分为：文本处理与训练数据准备、Word2vec 训练、Seq2seq 模型训练、测试与分析四部分。

1. 文本处理与训练数据准备

同实验一，删除压缩包中的无关文档和文件，仅留下 16 个文档文件，将每个文档中的“本书来自 www.cr173.com 免费 txt 小说下载站 更多更新免费电子书请关注 www.cr173.com”删除。本步骤为手工完成。

jieba 的词汇表中并没有收录很多金庸的武侠小说这种特定环境下的很多专有名词，包括一些重要人物的名称，一些重要的武功等。这里整理了三份 txt 文本，分别记录了武侠小说中的人物名称、武功名称和门派名称，并把这些词汇添加到词汇表中。与之前不同的是，为了生成语句，停用词不要删去。

总体步骤：替换无用字符、jieba 添加词汇、jieba 分词、保存为 txt 以便下一步训练。见 sentences.py

2. Word2vec 训练

同上个实验，训练 CBOW 模型，保存模型。见 word2vec.py

3. Seq2seq 模型训练

搭建 Seq2seq 模型，其中代码参考 pytorch 官网提供的 seq2seq 代码，encoder 中为 GRU，decoder 为带 attention 的 GRU。将语句分词后每个词送入 word2vec 转换为向量，依次送入 encoder，将输出和隐状态接着送入 decoder，输出向量，再用 word2vec 模型找出最接近该向量的词汇。见 seq2seq.py

4. 结果分析

见 testmodel.py，实验结果如下：

输入语句：

郭襄怒道：“这位大师是忠厚老实的好人，你们欺他仁善，便这般折磨于他，哼哼，天鸣禅师呢？

无色和尚、无相和尚在哪里？

你去叫他们出来，我倒要问问这个道理。”

两个僧人听了都是一惊。

天鸣禅师是少林寺方丈，无色禅师是本寺罗汉堂首座，无相禅师是达摩堂首座，三人位望尊崇，寺中僧侣向来只称“老方丈”、“罗汉堂座师”、“达摩堂座师”，从来不敢提及法名，岂知一个年轻女子竟敢上山来大呼小叫，直斥其名。

那两名僧人都是戒律堂首座的弟子，奉了座师之命，监视觉远，这时听郭襄言语莽撞，那瘦长僧人喝道：“女施主再在佛门清净之地滋扰，莫怪小僧无礼。”

输出语句：

阿绣公主道杨大哥，表哥喀丝丽我。”我你人家成亲你将来将来你以后别人以后别人以后将来别人别的歹事谢大侠怨仇只不过人家，我我将来侄儿爹娘别人人家。”侄儿，别人将来别人罚别人怨仇，别人爹娘我。”侄儿你侄儿。人家，将来我现下人家侄儿人家，爹娘别人爹娘人家将来人家人家别人将来别人罚他下毒的怨仇怨仇。。”了它我罚。谢大侠侄儿。她们，谢大侠爹娘我。”她们怨仇谢大侠怨仇谢大侠，义父爹娘怨仇了她们。

可以发现，模型输出效果很差，由于本实验难度太大，加上时间有限，seq2seq 模型训练次数较少，仅选取了 20000 句进行训练，并且模型参数，结构，以及程序的构建均充满了主观性，并没有很好地预测上下文的输出。

四、参考文献

1. https://blog.csdn.net/shzx_55733/article/details/117338742
2. https://blog.csdn.net/qq_38163755/article/details/106813809