**基于深度学习的自然语言处理**

**第6次实验报告**

**一、实验名称**

实现自动问答功能

**二、实验日期**

2020.12.05 – 2020.12.18

环境：python3.7，tensorflow1.13.1

**三、实验目的**

（1）熟悉自动问答的基本理论

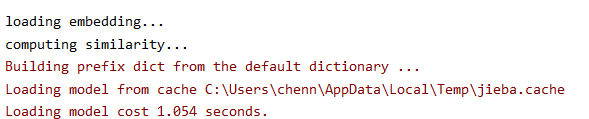
（2）掌握自动问答的代码实现

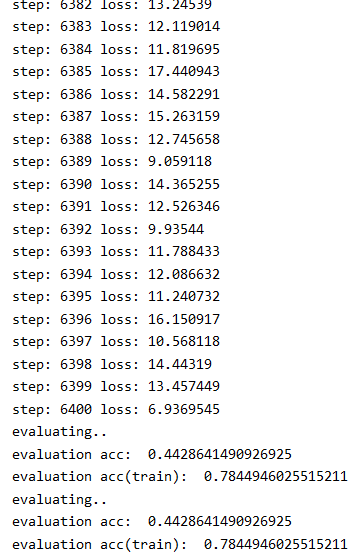
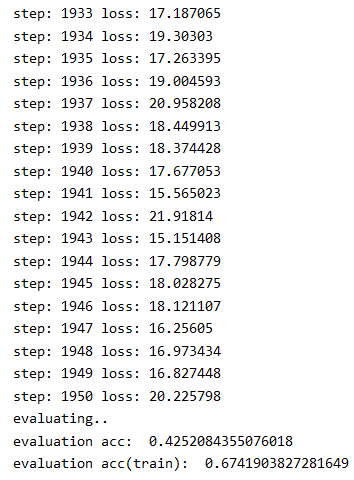
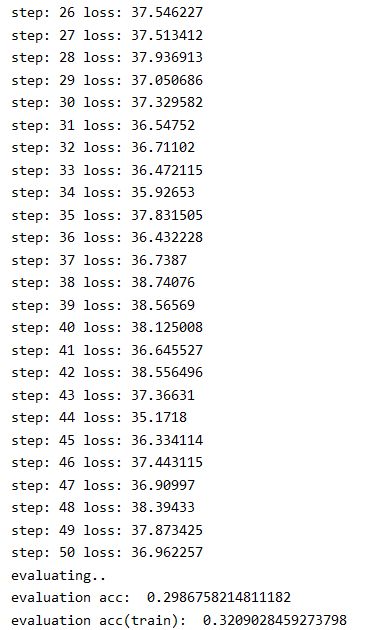
**四、实验数据**

提供的开源的实验数据

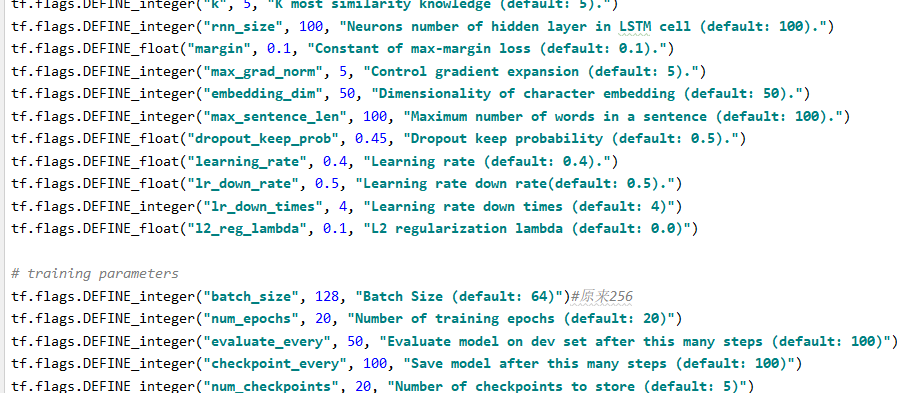
**五、实验步骤及结果分析**

**Run train.py结果如下：**

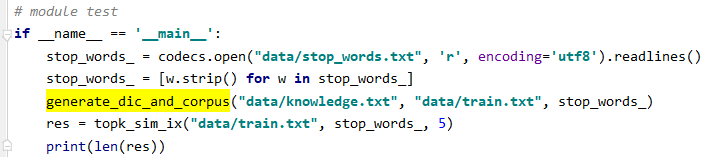


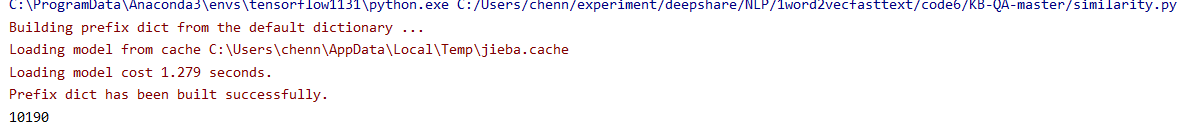


**其中batch\_size改成128**



**运行similarity.py**





5.1 基于知识库的中文问答系统. 整体流程如下:

1.根据Background和Question寻找到最相关的K个Knowledge，K Knowledge + Background + Question构成一个大问题

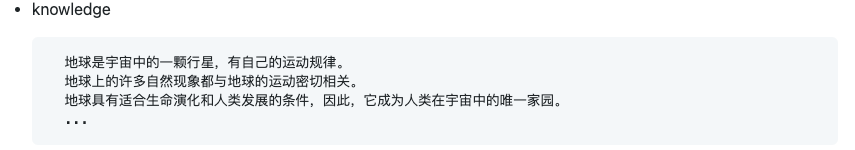
2.正确选项分别与该问题中所有错误选项组合，构成3个答案组合，分别与大问题组合构成3个样例，采用余弦距离计算大问题与正确选项和错误选项的相似度

正确选项相似度为t\_sim, 错误选项相似度为f\_sim,损失函数为：

loss = max(0, margin - t\_sim + f\_sim)

3.Model: 寻找相关Knowledge(LSI) 、训练（biLSTM）

4.



5.



6.参考链接：<https://github.com/DouYishun/KB-QA>

小的能跑通

<https://github.com/zhihao-chen/QASystemOnMedicalGraph>

大的 很慢 有兴趣可以跑 batchsize小一点

构建知识图谱

**六、实验感想**

**Knowledge-Based Question Answering**

基于知识库的中文问答系统. 整体流程如下: 1.根据Background和Question寻找到最相关的K个Knowledge，K Knowledge+Background+Question构成一个大问题.

2.正确选项分别与该问题中所有错误选项组合，构成3个答案组合，分别与大问题组合构成3个样例，采用余弦距离计算大问题与正确选项和错误选项的相似度.

正确选项相似度为t\_sim, 错误选项相似度为f\_sim,损失函数为

loss = max(0, margin - t\_sim + f\_sim)

**Model**

寻找相关Knowledge: LSI

训练: biLSTM

**Requirement**

python3, tensorflow

stop\_words, 中文word2vec(a2u6)

**Data Format**

**knowledge**

地球是宇宙中的一颗行星，有自己的运动规律。

地球上的许多自然现象都与地球的运动密切相关。

地球具有适合生命演化和人类发展的条件，因此，它成为人类在宇宙中的唯一家园。

...

**train&test**

**问题为选择题，每个问题的格式为 Background, Question, Right, Wrong, Wrong, Wrong.**

B:近年来，我国有些农村出现了“有院无人住，有地无人种”的空心化现象。

Q:“有院无人住，有地无人种”带来 R:土地资源浪费 W:农业发展水平提高 W:城乡协调发展 W:农村老龄化程度降低

B:广东省佛山市三水区被称为“中国饮料之都”。除青岛啤酒、伊利等国内著名饮料企业抢先布局外，百威、红牛、可口可乐、杨协成等国际巨头也先后落户于此，作为其在中国布局中的重要一环。

Q:众多国际饮料企业选址三水的主导区位因素是 R:市场 W:技术 W:劳动力 W:原料

B:凡是大气中因悬浮的水汽凝结，能见度低于1千米时，气象学称这种天气现象为雾 Q:深秋到第二年初春，晴朗的夜晚容易形成雾，这主要是因为 R:晴朗的夜晚大气逆辐射弱，近地面降温快 W:晴天大气中的水汽含量多 W:晴朗的夜晚地面水汽蒸发强 W:晴天大气中的凝结核物质较少 ...

**Usage**

python3 train.py 该数据集下最佳参数为 dropout:0.45 k:0.5