**团队学习汇报——2020.10.27**

**最近所做的工作：**

1. 语料的出处：

论文名称：《A Large-Scale Chinese Short-Text Conversation Dataset》

论文地址：*https://arxiv.org/abs/2008.03946*

项目地址：*https://github.com/thu-coai/CDial-GPT*

2.搭建了服务器训练环境、重新对语料进行了清洗工作

3.读了一些论文：

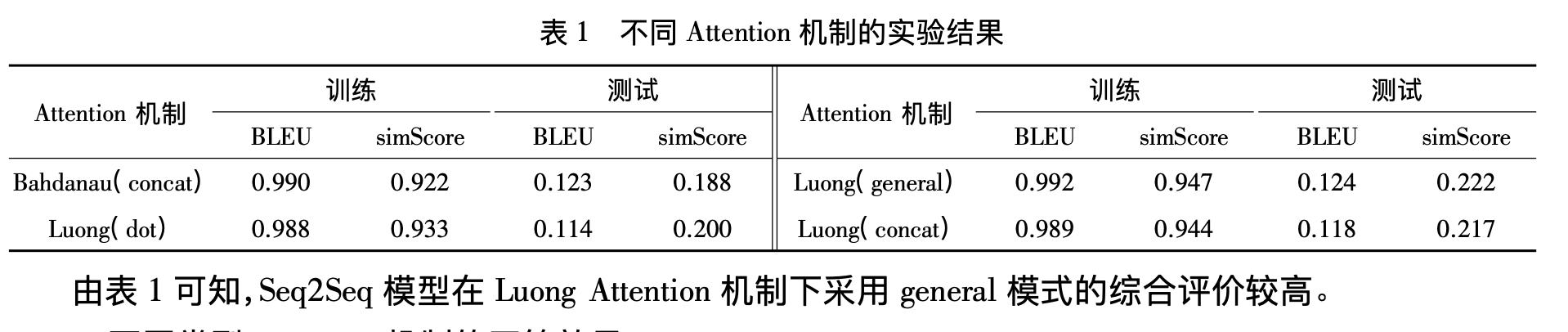
生成式的聊天机器人使用的方法包括（seq2seq-tensorflow、

seq2seq-pytorch、

seq2seqGAN、

BERT+seq2seq、

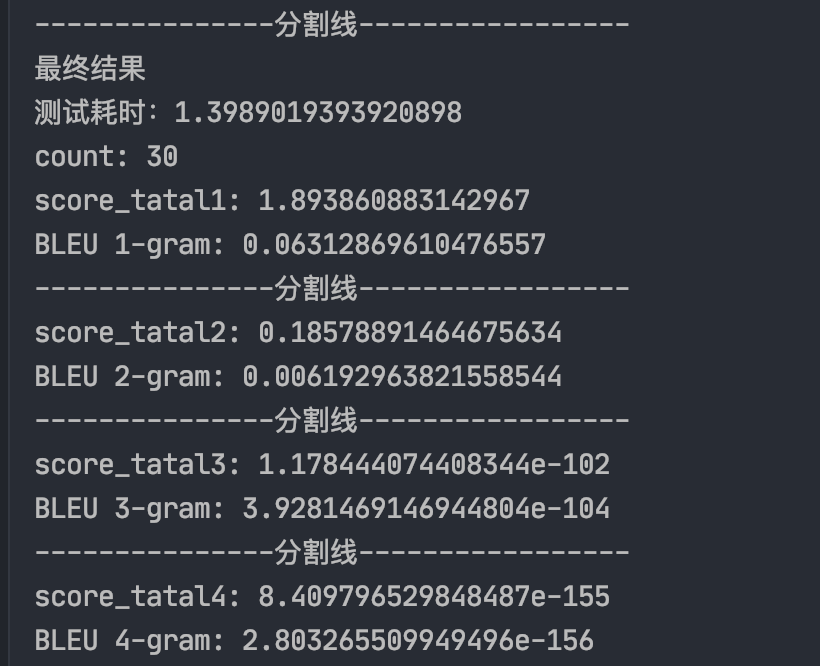
UNILM ......）

使用的评测方法有（BLEU、困惑度、平均问答长度、句子相似度）、

1. 对自己训练的模型进行了 BLEU 评测：

一般来说1-gram的结果代表了文中有多少个词被单独翻译出来了，因此它反映的是这篇译文的忠实度；而当我们计算2-gram以上时，更多时候结果反映的是译文的流畅度，值越高文章的可读性就越好。

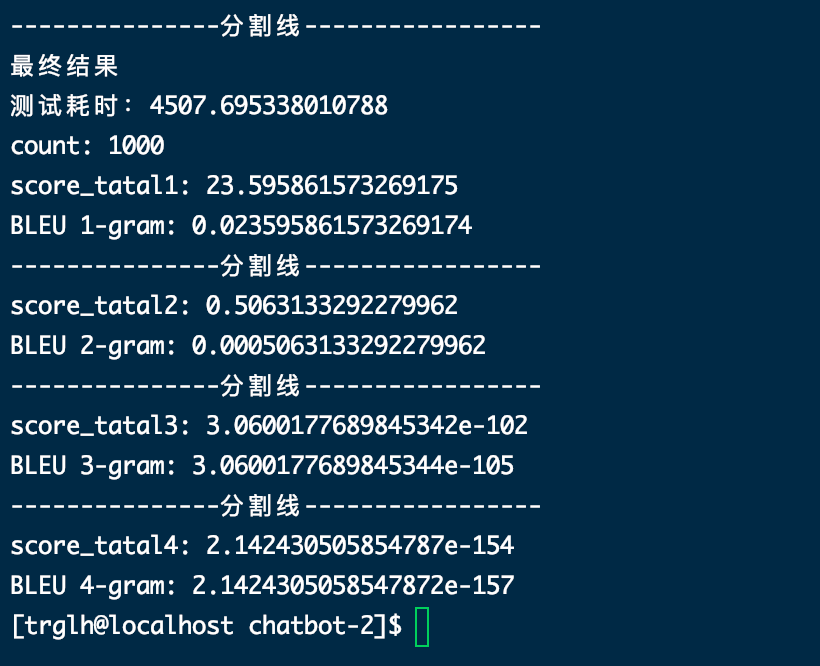
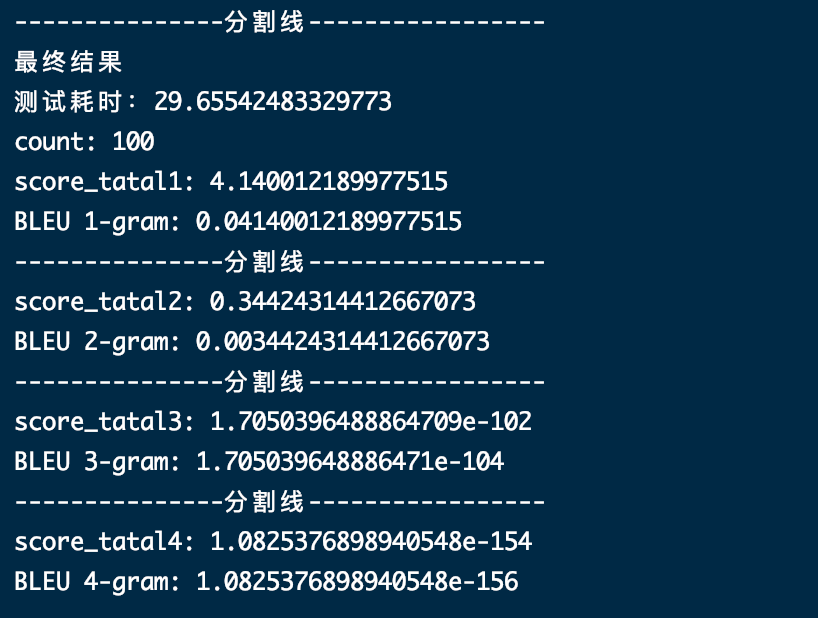
1万对分字语料(100epoch 训练的模型）：



10万对分词语料、字典大小为1000（1500epoch 训练的模型）：



40万对分字语料、字典大小为4000,（300epoch 训练的模型）：

****

**后期准备做的工作：**

1. 学习预训练模型—— BERT 、UNILM、OPENAI（GPT）
2. 使用别的评测指标

3.使用TF-IDF的方法实现检索式聊天机器人。