**实验九 神经机器翻译 作业报告**

1. 实验目的
2. 掌握循环神经网络及其变种的结构。
3. 了解Encoder-Decoder框架， word2vec技术，以及自然语言处理中的attention机制。
4. 进一步熟悉TensorFlow的使用，学会使用其自带的RNN框架、或LSTM框架，并结合收集到的相关数据，搭建Encoder-Decoder机器翻译模型。
5. 实验要求

利用Python语言和TensorFlow框架，构造简单的机器翻译模型，以实现英语和汉语的相互转换。

1. 实验原理

循环神经网络相关概念和原理参考《深度学习》课程讲授内容。

Encoder-Decoder模型是使用神经网络进行机器翻译的基本方法，一般也称作seq2seq模型。具体来说，它会先用一个Encoder将输入的序列编码为一个上下文向量c，再使用Decoder对c进行解码（将c作为初始状态的输入，或将c当作每一步的输入），将之变为输出序列，如图1、图2所示。在机器翻译任务中，输入序列为待翻译的英文，输出序列为其对应的汉语，Encoder和Decoder分别为两个RNN。



图1 c作为初始状态



图2 c作为每一步的输入

在传统的Encoder-Decoder结构中，c中必须包含原始序列中的全部信息，它的长度是限制模型性能的主要瓶颈。注意力机制（Attention）通过在每个时间输入不同的c来解决这个问题，每一个c会自动选取与当前所要输出的y最合适的上下文信息，即，用来衡量Encoder中第j阶段的和Decoder中第i阶段的相关性。是模型在训练时自动学出来的，且满足。如图3所示。



图3 带有注意力的Decoder

1. 实验所用工具及数据集

本实验主要针对中英机器翻译，使用的数据库来自NiuTrans提供的开源中英平行语料库，包含中、英文各10万条。如图4所示。

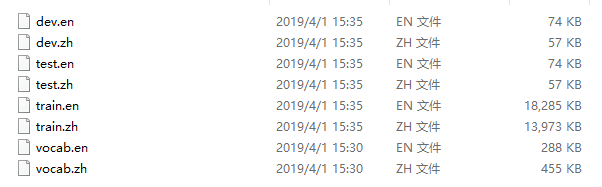


图4 NiuTrans语料库

数据集包含四部分：训练集（train.zh，train.en）、验证集（dev.zh，dev.en）、测试集（test.zh，test.en），以及词汇表（vocab.zh，vocab.en），其中以.en为后缀的文件为英语语料，.zh为后缀的为中文语料。值得注意的是，每个词汇表都会包含三个特殊的单词：<unk>、<s>、</s>。<unk>表示所有不常见的单词（对文本中使用的单词进行计数，取最常用的单词训练，剩下的单词都被替换为<unk>），<s>表示句子的开头，</s>表示句子的结尾。

本实验用到的数据集已经做好了中文分词，中文的数据样例如下：

北约 不少 飞机 不得不 携 弹 返航 , 降低 了 军事 能力 的 使用 效能 , 增加 了 战斗 成本 .

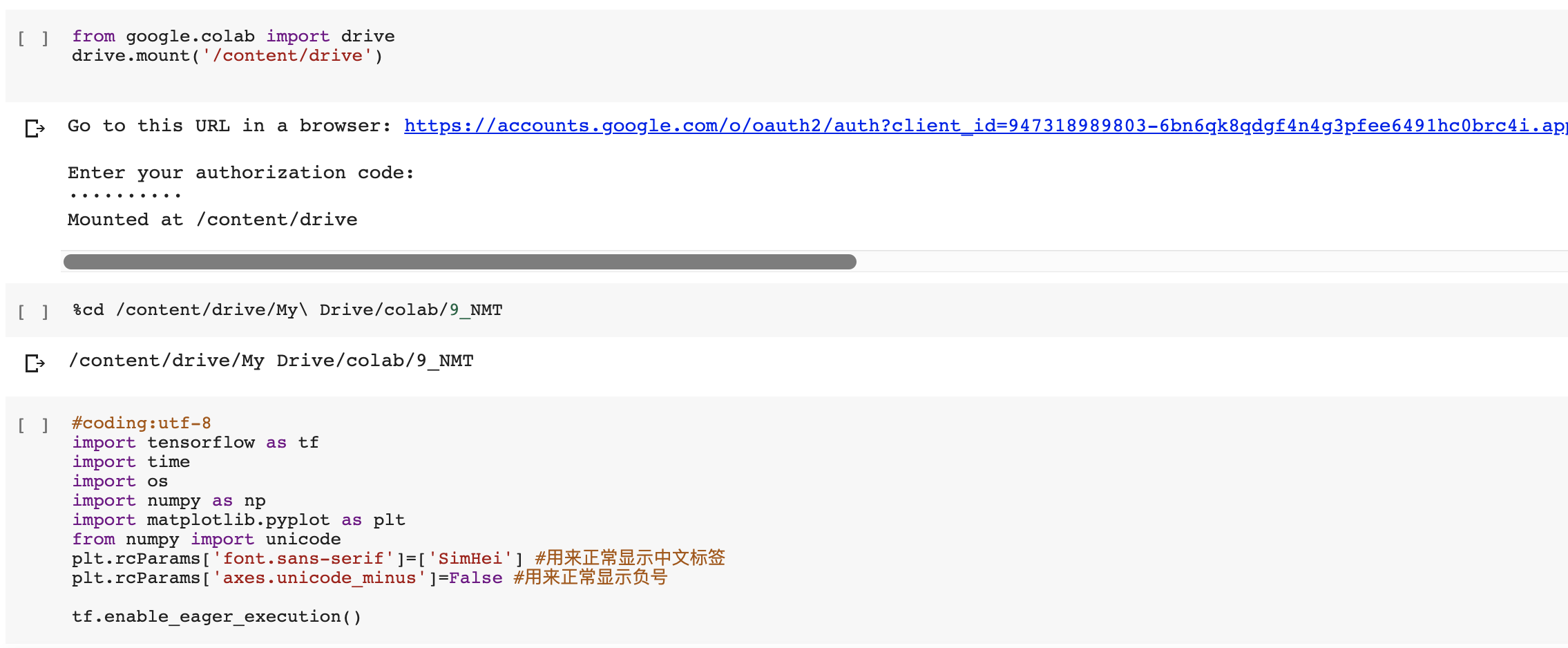
每个词之间用空格分隔，标点符号也算作一个单词。相应的英文样例如下：

many nato planes had to return to base laden with munitions , thus lowering the efficiency of use of military power and increasing the costs of fighting .

由于英语中每个单词之间都有空格，故不需要分词，需要做的就是把首句大写字母还原为小写。

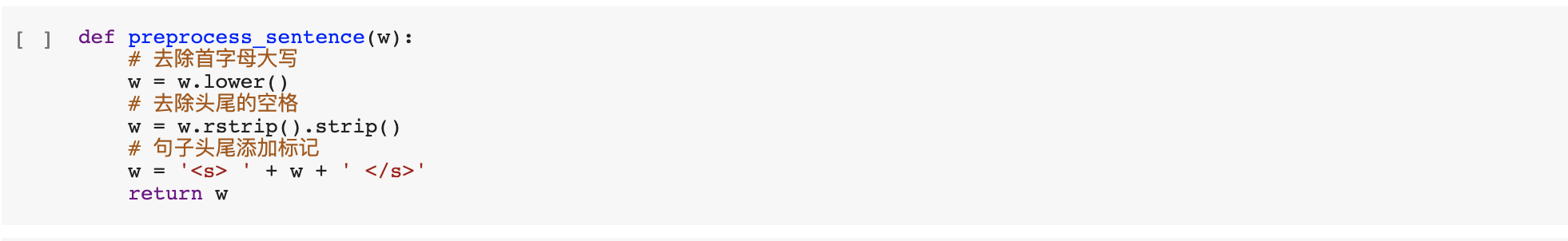
1. 实验步骤与方法
2. **代码依赖**

实验在Google Colab上完成， 包括设置Google Drive和路径

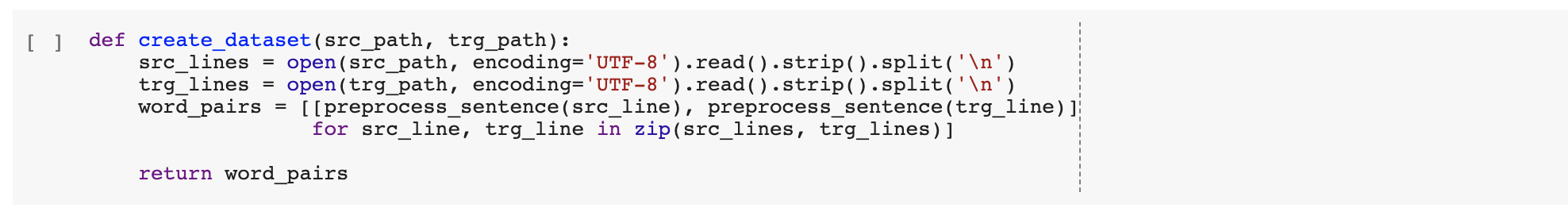
****

1. **数据处理**

句子首位添加<s>和</s>标记，去除大写和头尾空格



将原和目标句子组成一对放入数据集



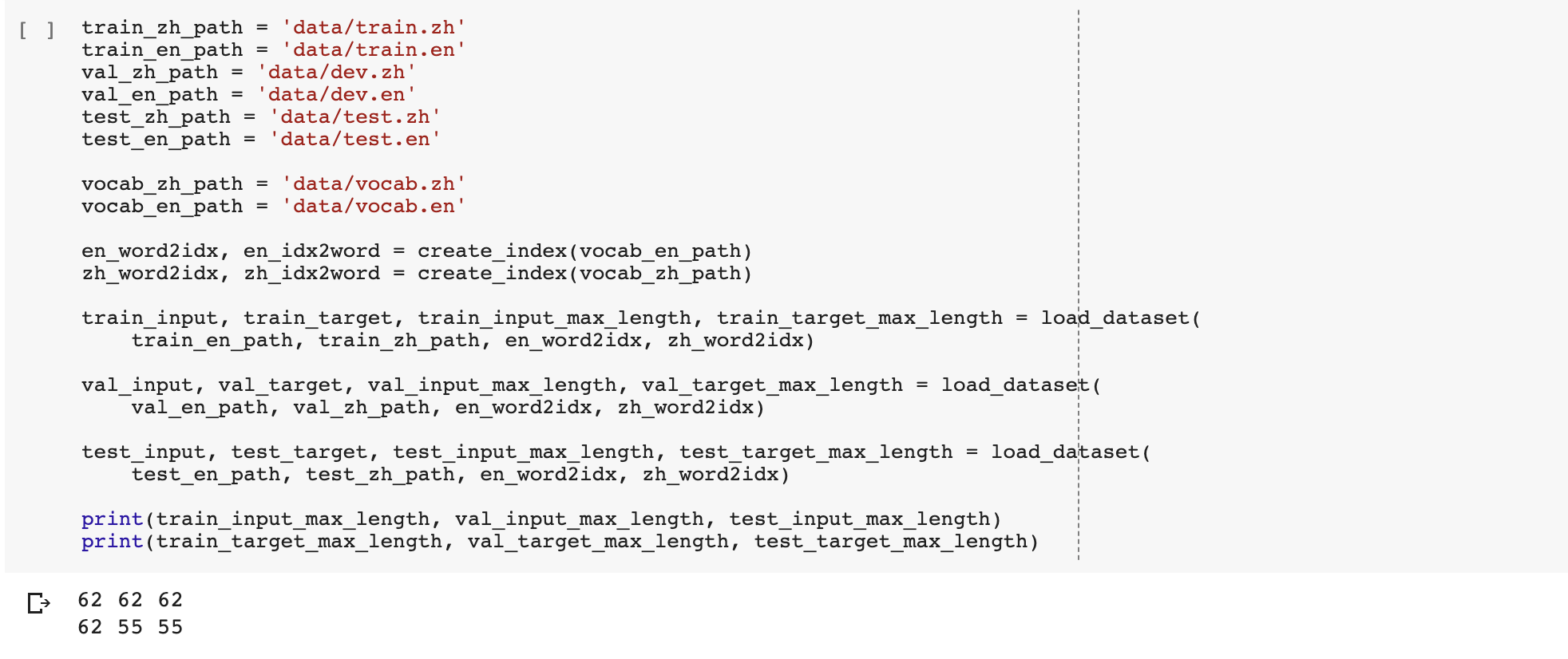
创建word2id和id2word转换的词表

****

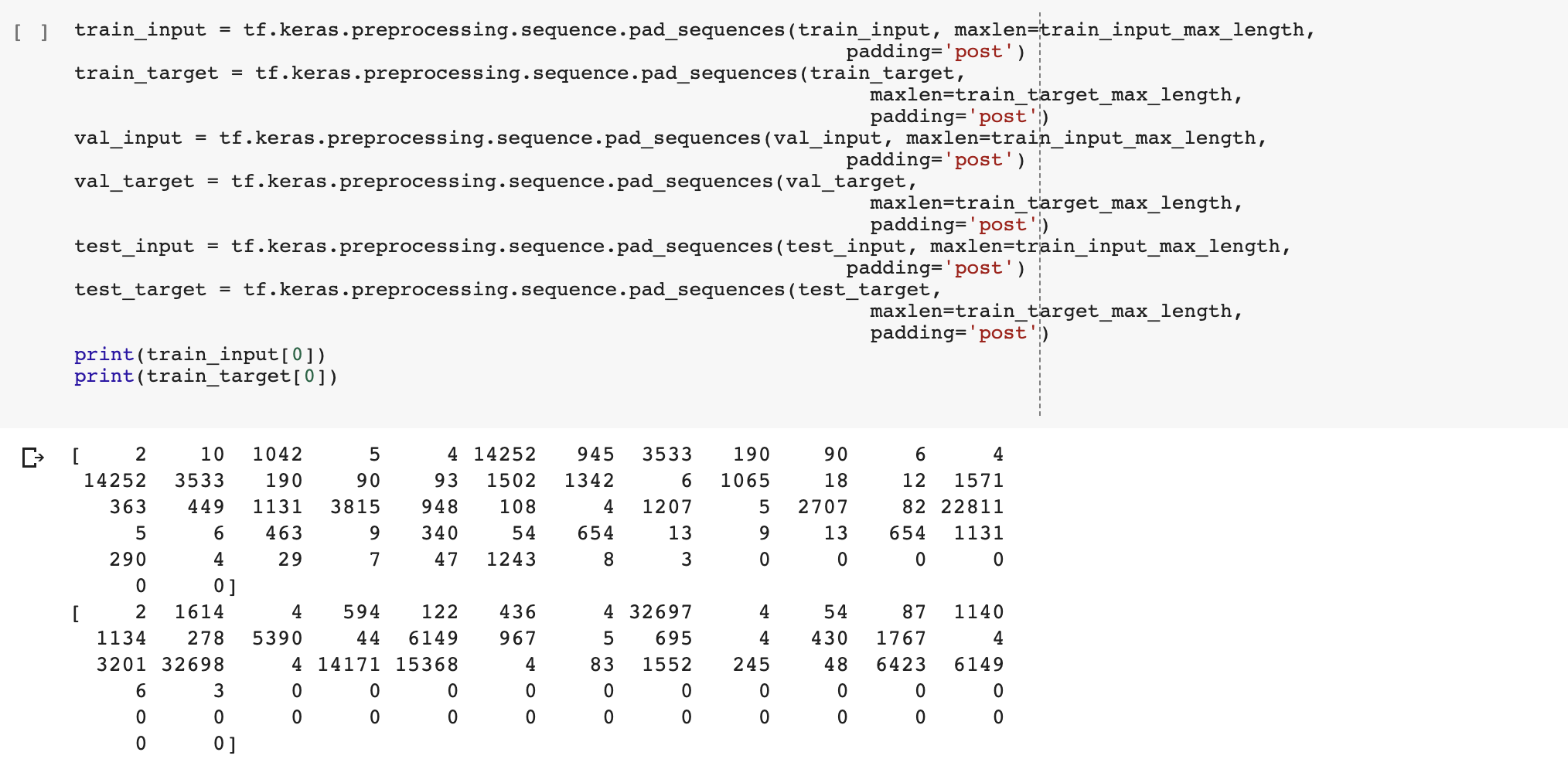
读取数据集，将原句和目标句子的单词分别转换为id，未知则为<unk>

****

分别读取训练集，验证集和测试集，输出最大句子长度

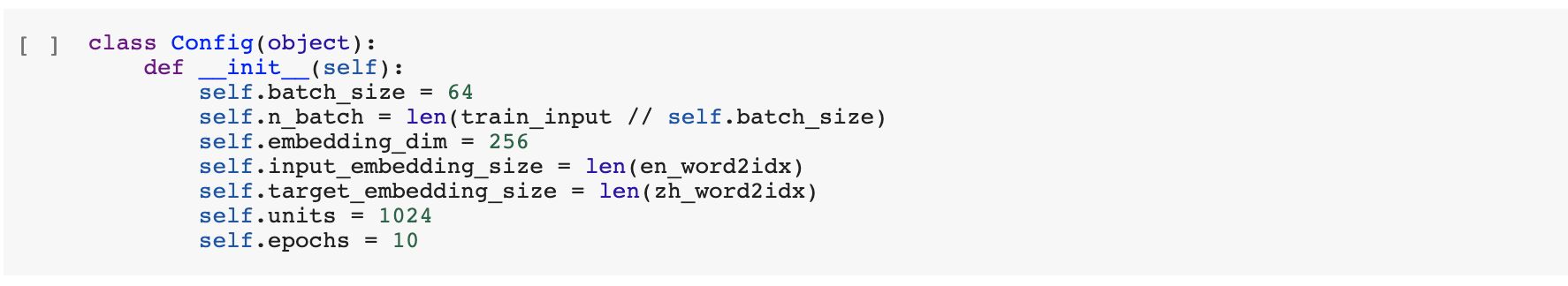
****

将所有句子按照最大句子长度进行填充，填充<pad>id为0

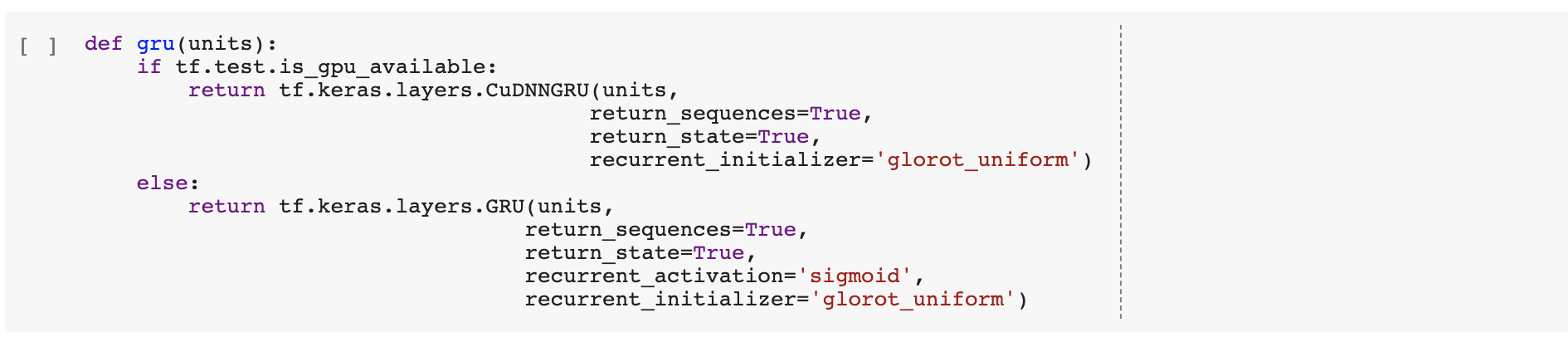
****

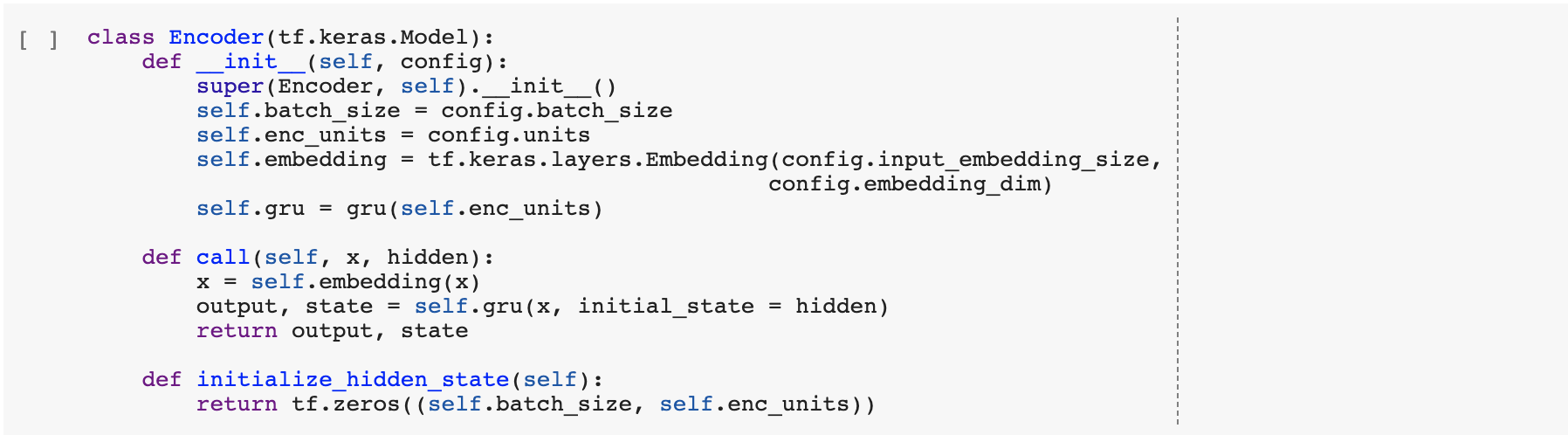
1. **模型构建**

模型参数

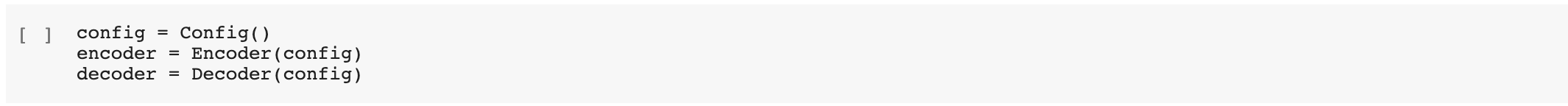
****

Attention + Seq2Seq模型

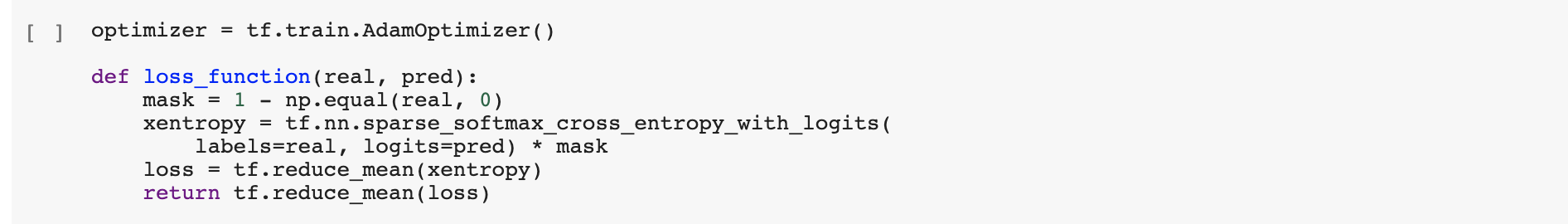
****

****

****

****

优化器和损失函数

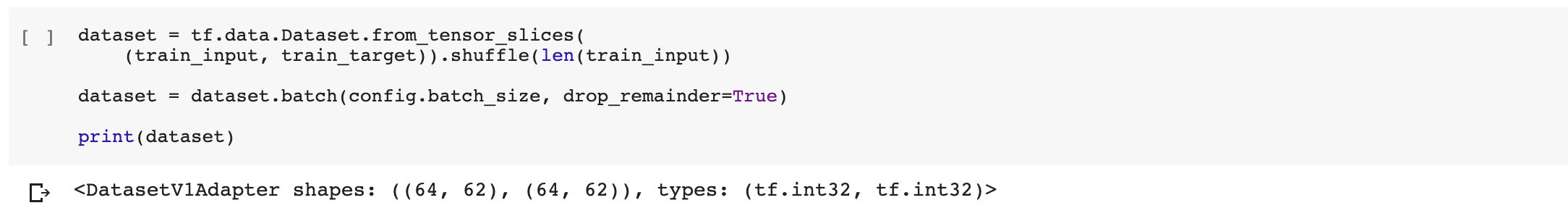
****

1. **训练**

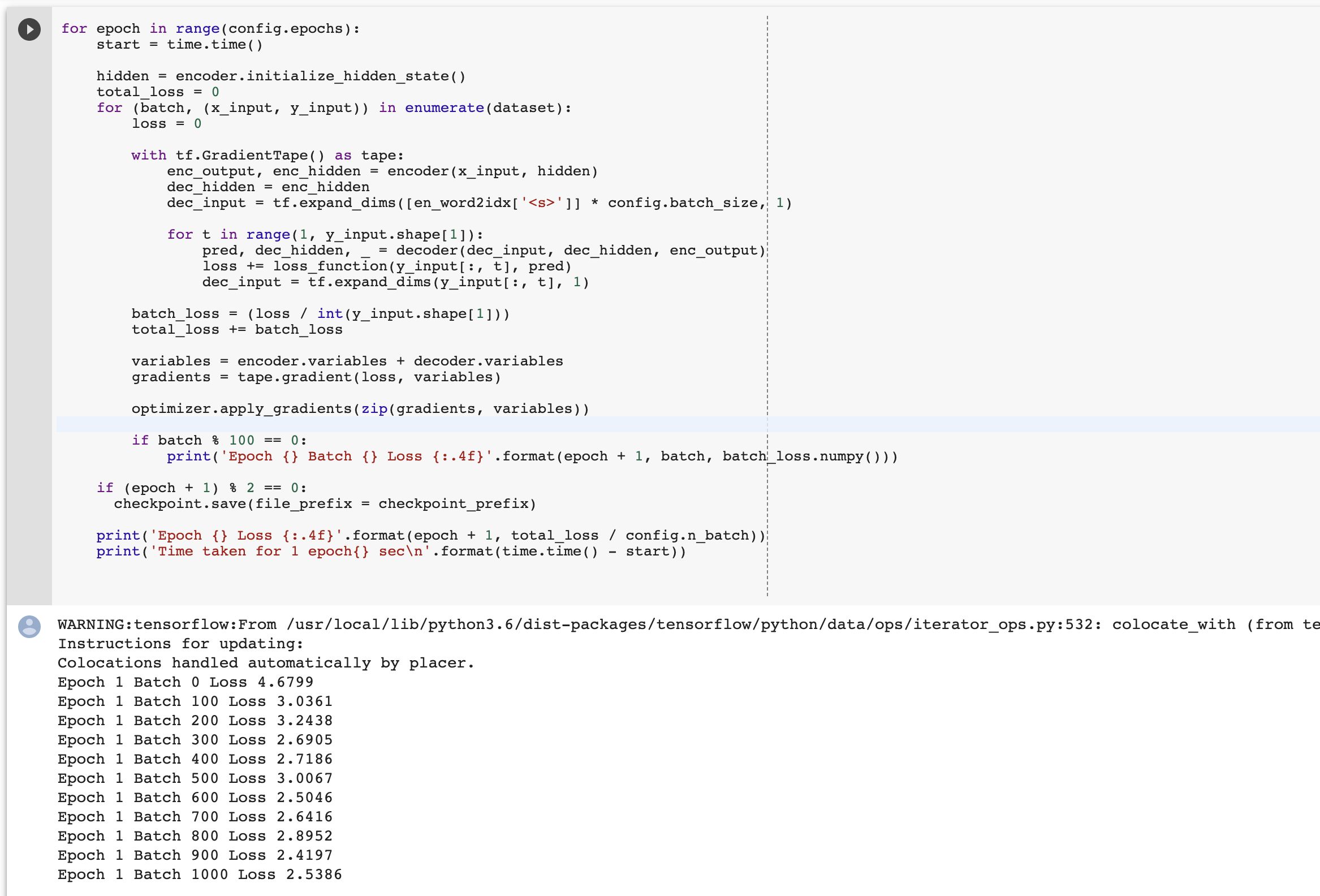
模型保存



数据集提取

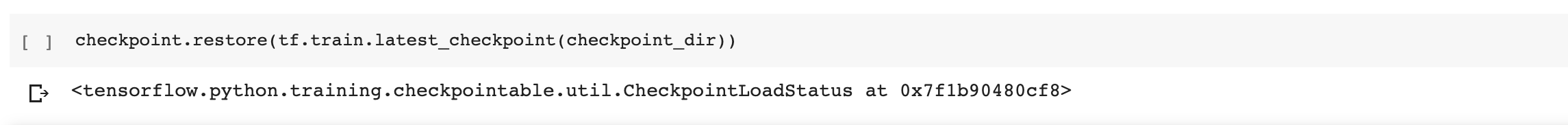


模型训练，由于训练时间超长，该代码在Colab上完成，进行了10轮训练

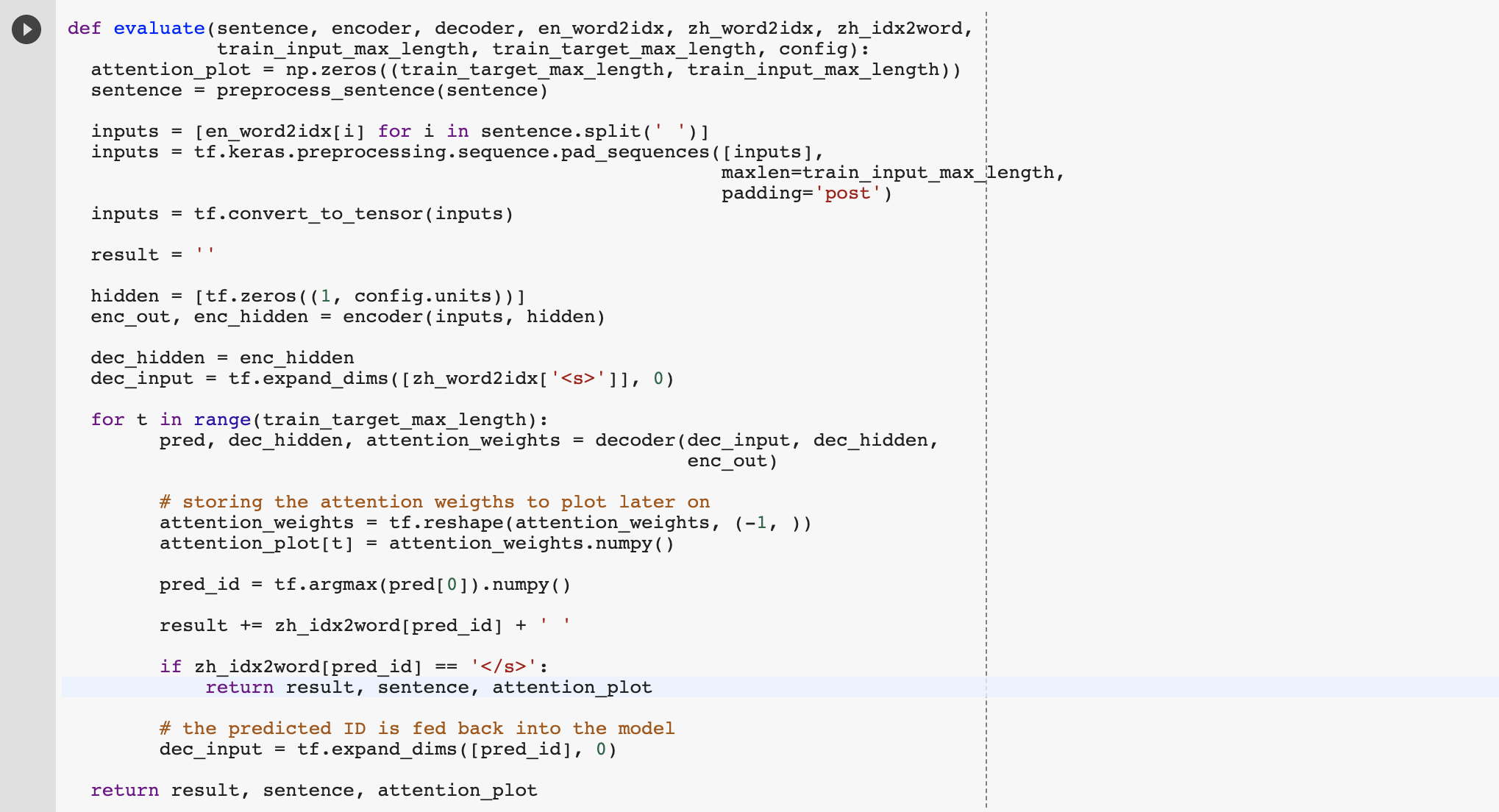


1. **测试**

读取保存的最近版本模型

****

将测试的句子同样做之前的数据处理过程，输入模型测试



输出最终的测试结果，可以看到模型仍有提升的空间中，增加训练轮数有助于改善这种翻译错误。

