实验4

机器翻译

前言

机器翻译是在无人工参与的情况下，使用机器自动将文本从一种语言翻译成另一种语言的流程。

- # 实验数据

创新工场、搜狗、今日头条联合举办的“AI challenger全球AI挑战赛”，数据集为英中机器文本翻译。

该数据集包含英中机器翻译提供的高达1千万的中英双语句对语料，这个量级，在开放的中英语料里仅次于联合国平行语料库。

数据集介绍

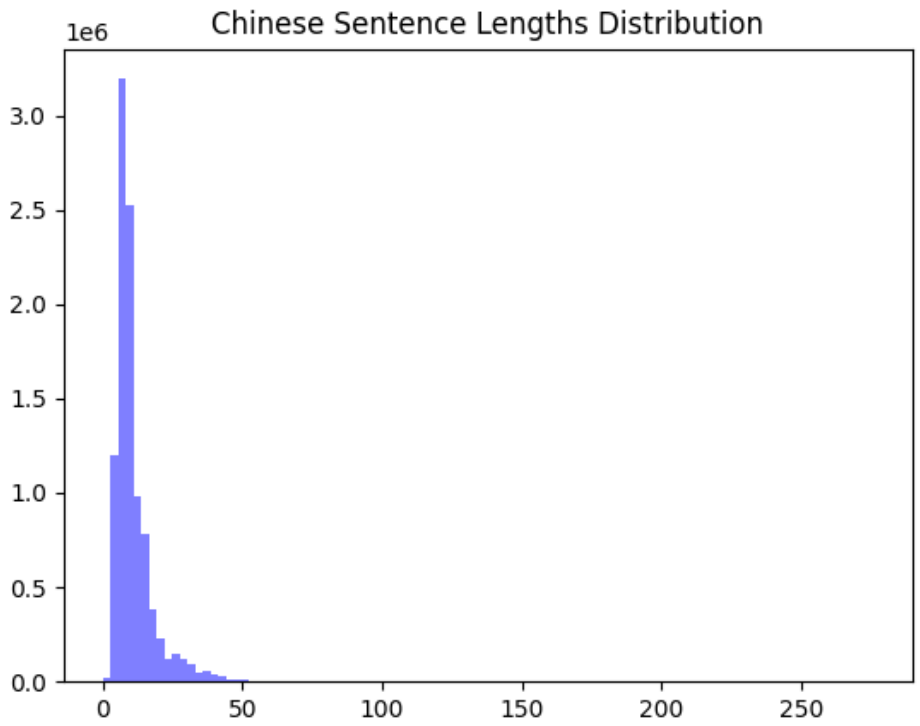
规模最大的口语领域英中双语对照数据集。提供了超过1000万的英中对照的句子对作为数据集合。所有双语句对经过人工检查，数据集从规模、相关度、质量上都有保障。

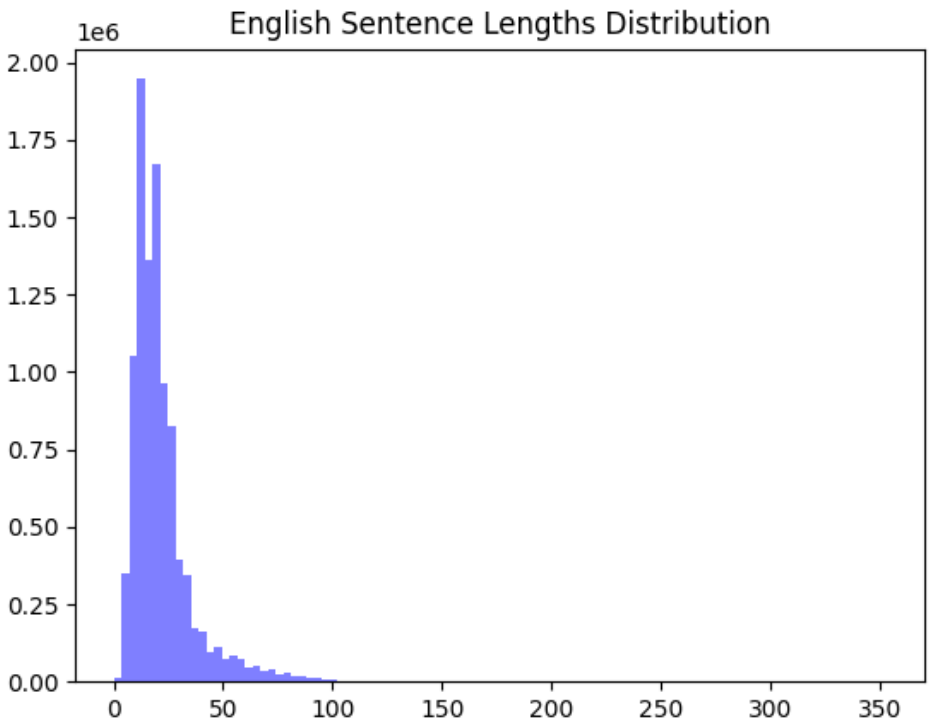
下载地址：https://tianchi.aliyun.com/dataset/174937

```
With fruit growing all year round, this is indeed a paradise for birds.
一年都有水果生长，这确实是鸟的天堂。

I dropped Henry at your office an hour ago.
一小时前我开车送海瑞去了你办公室。

Father and son, two bricklayers, are sitting in a cafe arguing about a
car.
一对父子，都是泥水匠，他们坐在一家咖啡馆里为一辆汽车争吵不休。
```

- 实验数据分析

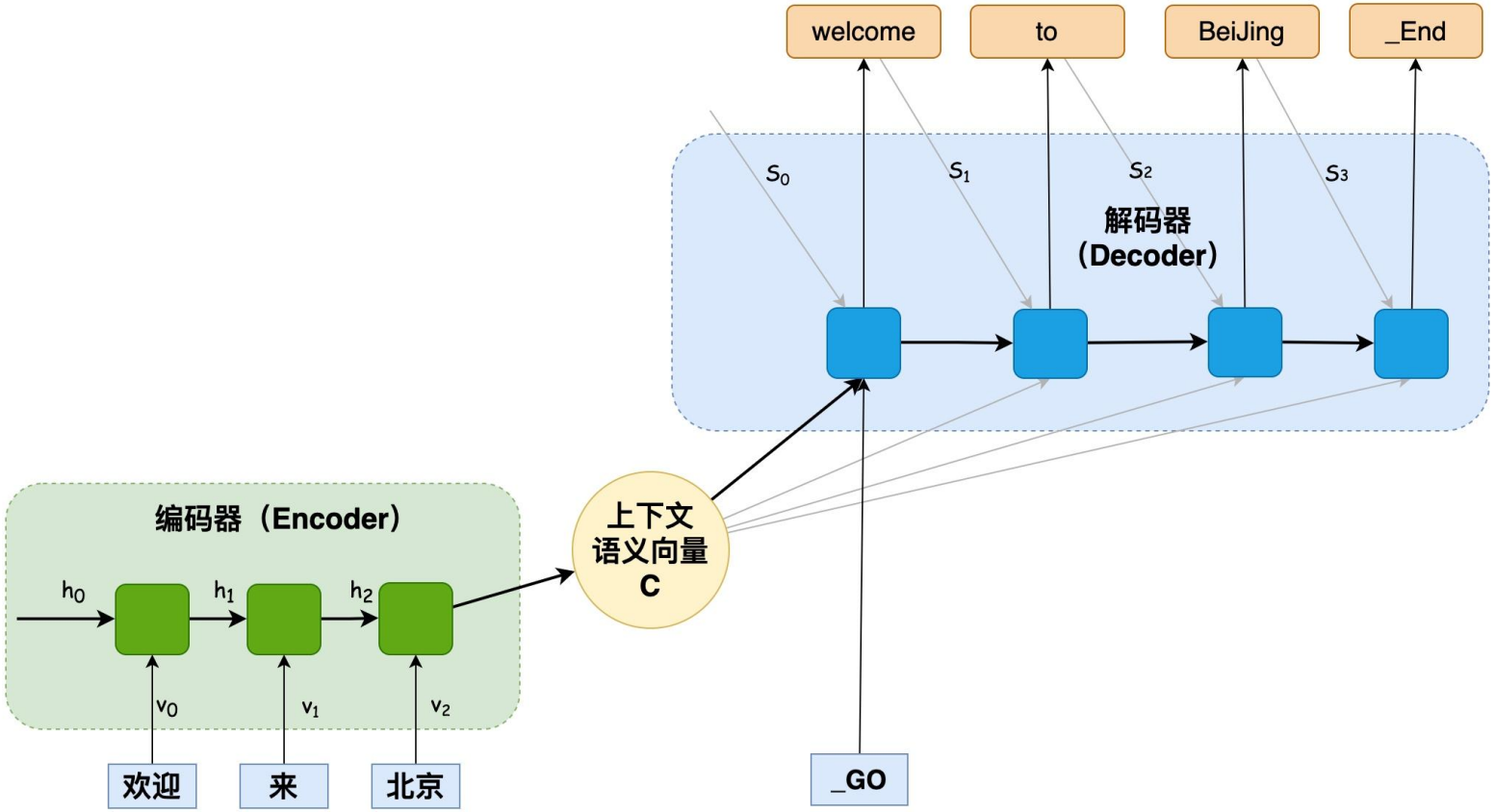
analyze_data.py





- # 实验理论解读

标准的机器翻译算法是 **编码-解码**（Encoder-Decoder） 网络，也叫 **序列到序列**（sequence to sequence） 网络，其可以由RNN或者Transformer结构来实现。



- 实验实现

  - 数据预处理：提取训练和验证样本pre_process.py
  - 训练：train.py

    要想可视化训练过程，在终端中运行：

    $ tensorboard --logdir path_to_current_dir/logs
  - 运行demo.py查看翻译效果。

```
> 你一直有工作。
= you ve always worked .
< you ve been working working .
> 我肯定这个洞是越来越小了。
= i m sure that hole s getting smaller .
< i m sure this is a little little .
> 当他回来时...
= and when he returned . . .
< when he came back . . .
```

- **资源下载**

下载代码，成功复现，可以解释代码（6-7分）

- **完善代码**

  补充翻译结果评估（每个评估1分）可参考bleu_score.py

  - BLEU (Bilingual Evaluation Understudy)：基于N元组（n-gram）匹配的指标，评估翻译结果和参考翻译之间的相似度。
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering)：考虑了词义、词形变化和词序的指标
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)：常用于评估摘要质量，也适用于翻译评估，侧重于召回率。
  - TER (Translation Edit Rate)：评估翻译结果需要多少编辑才能达到参考翻译的质量。

- **完善代码**

  模型优化（每个优化1分）

  - 使用更深的编码器和解码器层数，提升模型的表示能力。
  - 引入混合注意力机制，结合多头自注意力和其他类型的注意力机制。
  - 利用预训练的语言模型（如BERT、GPT、mBART等）作为编码器或解码器的初始化，提高模型的理解和生成能力。
  - 其他优化。

谢谢

FOR YOUR LISTENING

陈雅妮.