README

Assignment 7-中文语法纠错

使用 ELMo 预训练语言模型获取词向量

使用 LSTM 或者 GRU 构建 Encoder-Decoder 模型，处理语法错误纠正问题

本次实验使用数据集是 NLPCC 2018 Task 2，语料来自语言学习网站 Lang-8，其中母语人士可以选择学习者的文章进行更正，某些输入句子可能有不同的更正，其他细节可见于官方提供的 Guidelines 文件 taskgline02.pdf

测试集来源：https://github.com/YingyWang/NLPCC_2018_TASK2_GEC/blob/master/CS2S+BPE+Emb/data/gold.01，`./data/processed/seg.tx` 是处理后的无标签测试数据，用于dataloader加载

本次实验使用的预训练模型来自 HIT-SCIR 发布的兼容AllenNLP 的 ELMoForManyLangs 的中文版本，使用方法与上一次实验相同，由于预训练模型较大，同学们可以自行前往 Github 查看安装要求及预训练模型下载：https://github.com/HIT-SCIR/ELMoForManyLangs

使用 Pytorch 中的提供的 RNN 模块构建相应的深度神经网络，以上文获取到的词向量作为输入进行训练，以 MaxMatch 作为评价指标，具体实现可参考 https://github.com/shibing624/pycorrector, https://github.com/nusnlp/m2scorer

Description

- 下载模型及数据
- 运行 `./data/process.py` 得到 `./data/processed/seg.train`
- 请补充完成 `metrics/maxmatch.py`，`model` 文件夹下encoder-decoder模型以及BeamSearch解码部分
- 运行 `python -u main.py`

Tips

预训练语言模型

同学们也可以自由选择预训练语言模型完成本次实验，不过必须在报告中仔细说明预训练语言模型的来源和训练方法。

建议使用的中文预训练语言模型：中文 BART

语法纠错评估

MaxMatch 采用词级别评估，这可能与你的模型词表不一致，所以本次作业允许同学使用中文 ERRANT 进行评估，获取链接为：https://github.com/HillZhang1999/MuCGEC/tree/main/scorers/ChERRANT

参考结果

使用中文 BART-large 进行训练，在中文 ERRANT 下的评估结果为：P: 48.24, R: 33.64, F: 44.39

通常来说，MaxMatch 的评估结果应与中文 ERRANT 相近。