基于继续训练的古汉语语言模型

阎覃 迟泽闻 北京理工大学

前言

- 预训练古文RoBERTa
- 继续训练在语言模型迁移中的应用
- RoBERTa + CRF微调

预训练数据

- 训练数据：殆知阁古代文献藏书
  - 17 亿字古文，共15,694古文书籍，总大小6.63GB
  - 包括古文诗词、小说、四书五经等诸多种类古文文本。

- 数据预处理
  - 字形转换：将表示同一个字的繁体字、异体字转化成统一的形式
  - 分词：古汉语单字表达的语义往往比现代汉语更为丰富，因此直接使用单个字作为分词结果。
  - 词表构建：使用出现频率最高的23,287个字作为词表

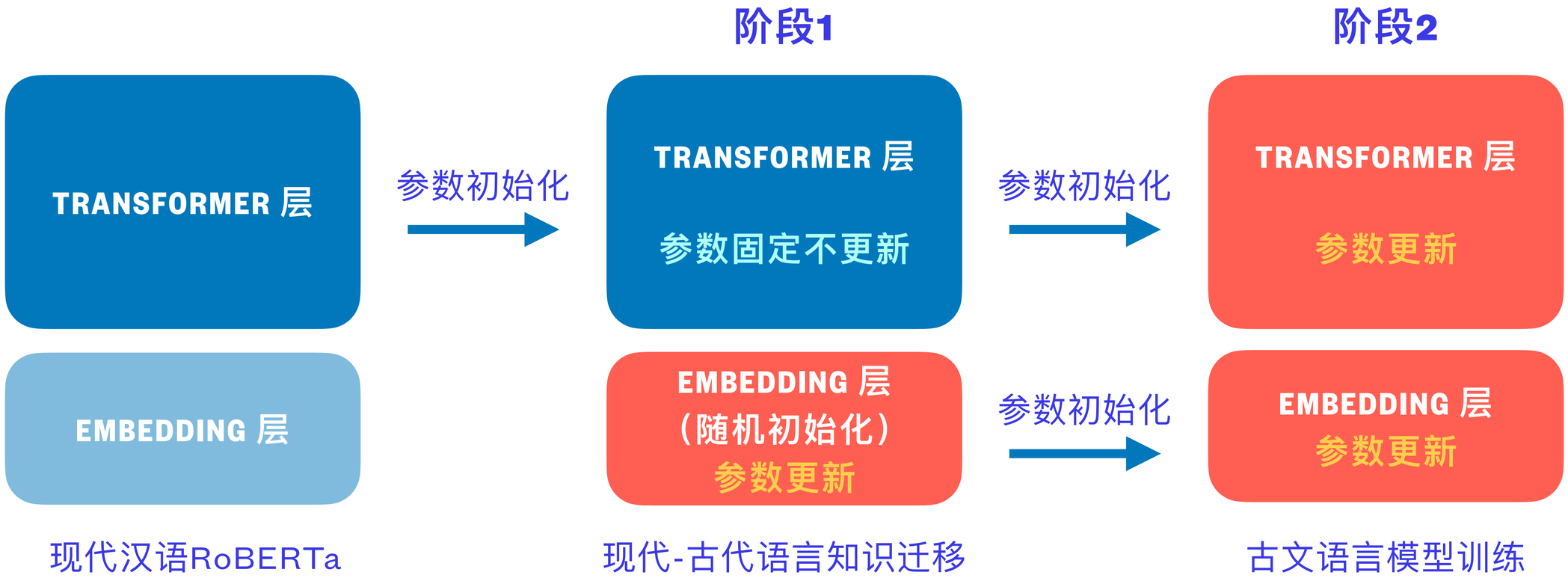
模型详情

- Data: daizhige (1.7B characters)
- Batch size: 1,048,576 words (2048 sequences * 512 words)
- Training steps: 120K steps (20K for step#1, 100K for step#2)
- Optimizer: Adam, 2e-4 learning rate, linear decay, 5K warmups
- guwenbert-base：12-layer，768-hidden
- guwenbert-large：24-layer，1024-hidden
- Trained on 4 V100 GPUs

预训练任务

- 使用最经典的掩码语言模型(MLM)作为预训练任务，使模型学会利用上下文信息预测被掩盖的词。
  - 例：会当凌绝顶[MASK]一[MASK]众山小。

预训练流程



阎覃 迟泽闻 北京理工大学

评测数据划分及模型融合

- 超参数搜索 / 对照实验
  - 80% 训练集 10%开发集 10%测试集
- 最终提交
  - 8-fold cross-validation
  - 87.5% 训练集 12.5% 开发集(early stop)
  - 训练8个模型并Ensemble
    - 1. 平均8个CRF的参数得到融合的CRF模型
    - 2. 使用融合的CRF模型对8个RoBERTa的输出平均值解码得到最佳序列

阎覃 迟泽闻 北京理工大学

微调详情

- 模型：RoBERTa + CRF（CRF接RoBERTa最后一层）
- 超参数搜索：贝叶斯优化(Bayesian Optimization)
- Batch Size
  - Base: 48 sequences * 512 tokens
  - Large: 8 sequences * 8 gradient accumulation steps * 512 tokens
- Learning Rate
  - Base: 3e-5(RoBERTa) 5e-3(CRF)
  - Large: 5e-5(RoBERTa) 5e-3(CRF)
- Optimizer: Adam, 100 warmup steps, linear decay

实验 总体表现

- 数据：初赛训练集+初赛测试集*2+复赛训练集
- 所有结果经过4次实验，取平均值

模型 (base)	F1 Score	模型 (large)	F1 Score
chinese-roberta-wwm-ext	85.05	chinese-roberta-wwm-large-ext	86.40
guwenbert-base	90.43	guwenbert-large	90.60
guwenbert-base-crf	91.30	guwenbert-large-crf	91.53

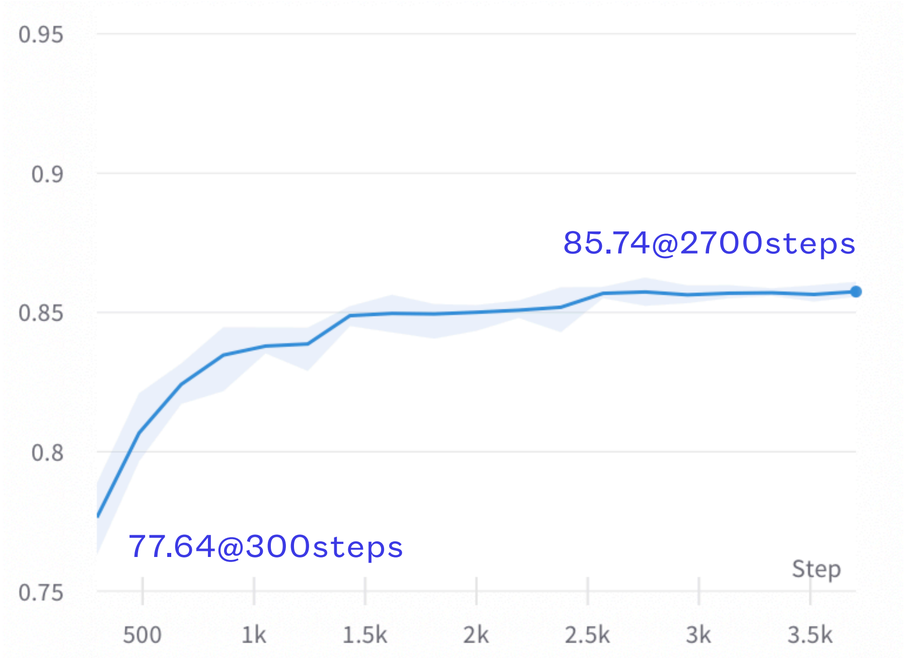
实验 继续训练

- 数据：初赛训练集
- 所有结果经过4次实验，取平均值

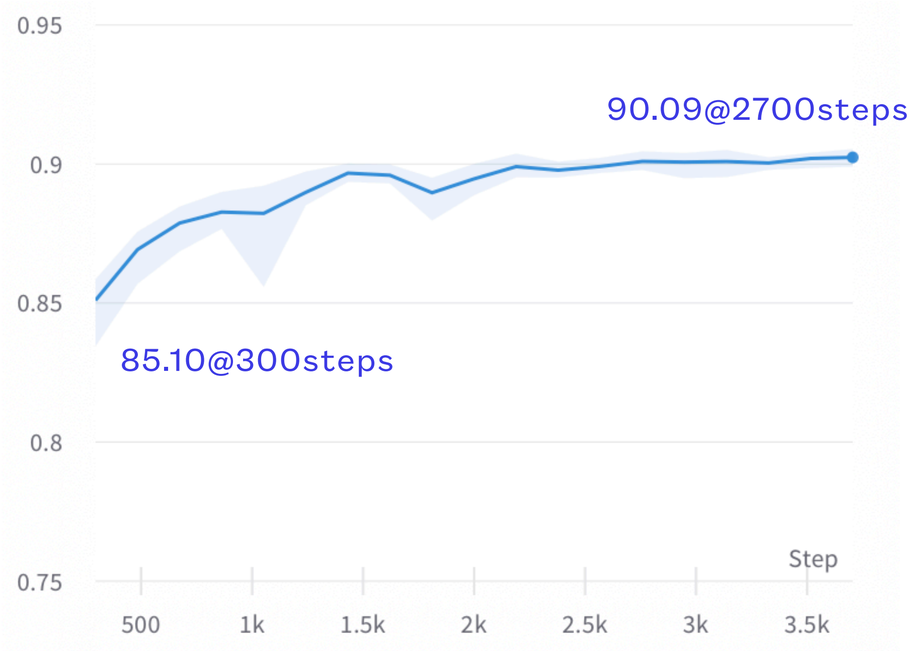
模型 (from scratch)	F1 Score	模型 (cont. train)	F1 Score
fs	88.79	ct	89.69
fs-crf	89.95	ct-crf	90.50

实验 few-shot

chinese-roberta-wwm-large-ext

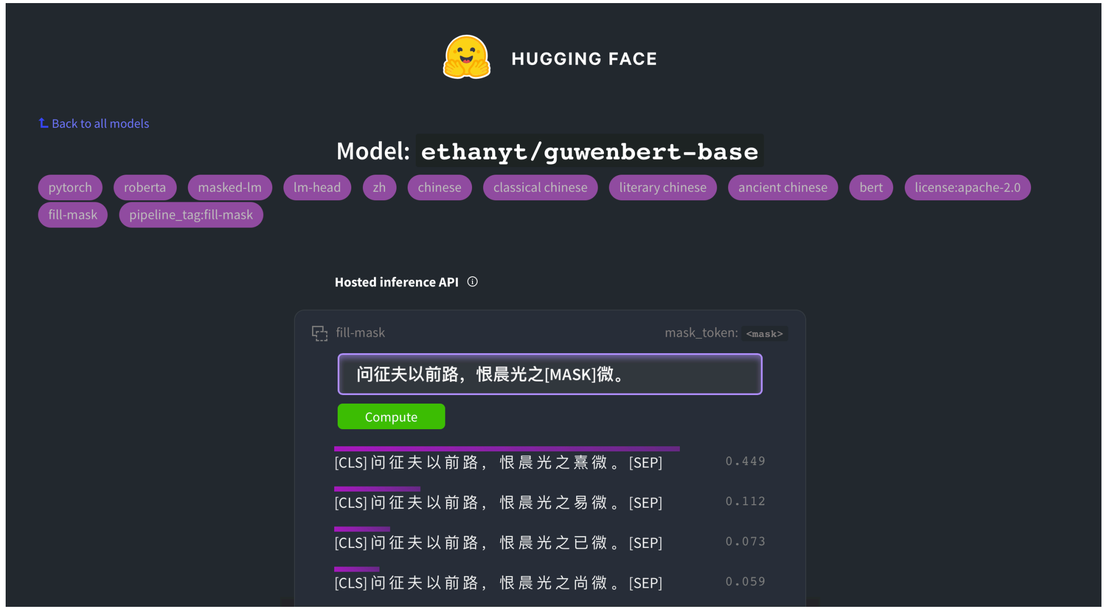


guwenbert-large



开源计划

- 模型已经接入Huggingface Transformers
  - https://huggingface.co/ethanyt/guwenbert-base
  - https://huggingface.co/ethanyt/guwenbert-large
- Github项目主页
  - https://github.com/Ethan-yt/guwenbert



总结

- 我们构建了一个古文预训练语言模型，并且开放下载
- 我们提出的古文模型优于通用现代汉语模型
- 通过结合现代汉语特征，持续训练进一步提升了总体表现

谢谢大家

- 欢迎各位老师提出宝贵意见
- 阎覃 北京理工大学
- ethanyt@qq.com