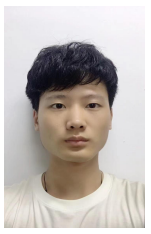
尤元岳

个人信息

- **出生日期：**1996 年 10 月 19 日
- **电话：**+8613572426306
- **电子邮箱：**1292008164@qq.com
- **地址：**辽宁省，朝阳市，凌河街，凌河湾御景苑

✧ **个人照片：**



教育背景

本科：西安电子科技大学（211&双一流） 09/2016-6/2020

- **院系：**通信工程学院
- **专业：**通信工程专业
- **平均学分绩点：**3.9/4.0
- **相关课程：**数字信号处理，电磁学，高频电路，时间频率分析，射频移动通信系统，移动通信原理
- **学位证书:**西安电子科技大学学士学位

研究生：爱丁堡大学(The University of Edinburgh)（QS 世界排名 22） 09/2020-11/2021

- **院系：**信息学院(Informatic college)
- **专业：**人工智能专业(Artificial Intelligence)
- **相关课程：**机器学习，深度学习，自然语言处理，语音识别

职业技能

- 熟悉使用 Linux 系统，了解 Linux 下的基本指令和操作
- 熟悉 python 编程语言，了解 C++，Java，C 等编程语言，熟悉数据结构和算法（已经刷 leetcode 200+题目）
- 熟悉自然语音处理，语音识别，了解计算机视觉，了解使用 pytorch 对深度神经网络的构建和训练，deepspeed 使用以及 llama factory 框架的使用，提示工程，检索增强的 RAG 模式
- 了解使用 docker 对服务进行部署
- 了解使用 postgresql，hive 数据库

工作经验

北京知呱呱科技服务有限公司 07/2022-至今

- 担任 NLP 算法工程师进行专利领域的 NLP 算法开发以及服务搭建

算法方面

项目一：专利行业的困难点，法条 A2（专利的方案不符合自然规律），A25 条款判断（专利不属于保护客体）的判断

- 项目背景：开发算法模型，协助专利审查员进行 A2（专利所描述的技术方案不符合自然规律），A5（专利所保护的实体，方法不属于保护的客体）相关专利条款的判断，减少审查员的工作量
- 项目难点：专利文本较长，平均一篇专利要 1-2w 字左右，所要判断的专利条款较为模糊且抽象。
- 方法及成果：
- 针对于长文本分类，尝试了将文本截断，文本分块并且均值化的方法输入至 bert 模型中，用 cls 进行分类，由于信息的损失，效果均不理想。
- 使用了两阶段的长文本分类训练方法 cogltx，第一阶段通过 ernie 对文本段的选取，第二阶段基于选取的文本段再输入至第二个 ernie 中进行预测，使用去除一句话查看 loss 增加程度来实现一阶段无监督的训练，整体的预测准确率得到了提升。
- 为了进一步提升整体模型的效果，针对于任务特点，基于 cogltx 的长文本分类模型，修改了原本 cogltx 的框架，将原本半监督的双阶段训练方式修改为监督方式的双阶段训练，先通过监督的方式训练第一阶段的文本选取模型，选取好的文本段送至第二阶段的文本判断模型，将专利审查中的两项任务 A2 条款判断（专利的方案不符合自然规律），A25 条款判断（专利不属于保护客体）的分类分别做到了 89%和 90%的准确率，达到了行业领先的水平。

项目二：垂直领域大模型训练-大模型在专利领域的辅助撰写

- 项目背景：根据用户输入一段创意，将用户的创意扩写成一篇专利的交底书，在给用户提供思路的同时，也可以减轻用户的写作量
- 项目难点：专利涉及的技术十分专业，需要极强的领域知识，且文本特色鲜明，有专利独有的文本特色，并且要生成的整篇专利交底书文字较长，需要根据用户的创意（几百字）扩展生成成千上万字的专利交底书。
- 方法及成果：
- 根据大模型在不进行微调训练时的表现效果评测多个大模型的基座能力，评测出 Yi 的基座能力在专利领域有较为强大的能力。
- 将整个生成专利交底书的任务拆分成多任务，将多任务的数据进行配比，使用 deepspeed 进行单机多卡对 Yi-6B 进行微调 sft 训练，并且为了防止模型过拟合以及保证数据多样性，同时融入了通用领域的数据，以及引入噪音的训练方式。
- 为了进一步提升模型效果，基于模型的生成结果进行了多采样，并且对采样结果进行排序，基于 sft 阶段训练出现的 bad case，设计了对采样数据的进行自动排序的算法，与人为评测有着 80%的一致性。通过排序的采样数据对 reward 模型训练，为了让 reward 模型更加鲁棒，reward 模型训练过程由判断最后一个 token 的评分转为整句话的评分任务，并且用 reward 模型指导进行 sft 后模型的 ppo 训练，以及 dpo 训练，同时为了让 dpo 训练更加稳定，修改了 dpo 的训练 loss，在 dpo 的原本 loss 基础上，添加了对于 good case 样本的交叉熵 loss，既保证了 dpo 最后的训练效果，也减少了

reward 模型的训练。使用 AWQ 量化技术将训练好的模型部署上线，并且进行了类 openai 的接口封装。

✧ 模型效果经过了专利代理老师评测，且现已上线至公司的智能撰写模块，并且为当前公司的核心功能。同时实时跟进最新开源的大模型 Llama2，chatglm3，Yi，Qwen 等模型及其专利领域的效果。

项目三：基于 RAG 的技术的大模型知识问答应用

✧ 项目背景：基于用户输入的技术问题，进行专利领域的知识问答，并且标明出处

✧ 项目难点：外挂知识库，检索的准确性，模型回答的结果与用户的问题相匹配

✧ 方法及成果：

✧ 通过指令工程的方式另模型先对用户输入的问题进行三个方向的改写，要解决的技术问题，技术手段，技术性标题。对三个改写结果进行向量编码，为 query 向量，并且对全量专利数据基于深加工模型模型对应的生成技术问题，技术手段，技术标题，并且进行向量编码以及入库为 doc 向量。令三个字段的 query 向量分别与对应字段的 doc 向量进行检索召回，对召回的结果进行摘要整合，从而展示给前端界面。

项目四：专利领域的数据深加工，核心技术摘要，关键词提取

✧ 项目背景：将一整篇专利（几万字）提取该篇专利的核心技术点，从而方便用户快速阅读了解当前专利主要的技术手段。

✧ 项目难点：专利文本较长，不同的技术点分布在专利的不同部分，且摘要效果要通过专利代理人的评测。

✧ 方法及成果：

✧ 使用了 cogltx 的长文本双阶段训练框架，并且将原本的分类任务修改为摘要任务。并且在训练过程中，为了让模型效果更加的优秀，添加了重要片段重复的机制，及在第一阶段文本段选取的过程中，将部分重要的片段进行重复，可以明显提升模型效果。针对于关键词提取，采用了 bert+crf，bert+bilstm+crf 的对比实验，经过实验验证，最终采用了 bert+crf 的模型。

✧ 其中核心方案任务模型的 rouge-l 达到了 0.8，其余模型也均通过了专业代理老师的评测，现已应用于公司 SAAS 产品平台。

项目五：专利领域核心难点，专利审查的新创性分析

✧ 项目背景：输入为一篇专利，在 200 篇候选专利中挑选哪些专利公开的技术点覆盖了输入的专利技术点

✧ 项目难点：对比专利是否覆盖了输入专利的技术点要结合领域知识，且对比专利是否覆盖本专利的技术要结合技术细节的上下文，使用的同样的技术手段，达到了相同的技术效果，才算技术点覆盖，而技术效果通常并不会在专利当中体现，属于是隐藏的信息，并且技术手段也是并不是文字上的相同。

✧ 方法及成果：

✧ 对数据进行了精细的清洗，基于 BERT+CRF 的框架，将输入专利的技术点以及对比专利的技术点通过 B，I 的标签，修改了传统 BIO 标签识别任务，此任务中输入为输入专利文本以及技术点的 BI 标签，对比专利的文本，任务目标是预测对比专利当中与输入专利技术点相似的片段内容，并打上 BI 标签。为了实现此任务，修改了 BERT+CRF 训练以及 decode 过程的逻辑，保持输入专利部分文本内容的状态以及概率不动，仅对对比专利文本的 CRF 推理进行 loss 计算，把对比专利文本的起始状态与输入专利文本的结束状态相连，从而实现基于任务目标的 loss 修改。

✧ Top1 准确率达到 20.2%，该任务的 SOTA 的 top1 准确率为 17.96%

工程方面

✧ 对训练好的模型以及开发好的算法进行服务部署，供前后台联调，编写 Dockerfile，封装 docker 容器，部署服务。

✧ Gitlab 的上线流程管理，包括小组成员的上线流程，打标签 tag 的管理，merge request 的管理等。

✧ 对多台服务器批量生产深加工摘要数据，实现进度监控，异常重启，断点继续处理的批量处理数据的系统，完成了全量专利 3000w，4 个字段的批量生产任务。

项目经验

研究生阶段毕业设计--建立多语言语音识别系统-独立开发 05/2021-08/2021

导师： Peter Bell

✧ 项目目标是在 Linux 系统环境下搭建对 Xhosa 语言的语音识别系统

✧ 通过使用 kaldi 工具和脚本的编写对数据进行处理，搭建 TDNN 语言模型

✧ 尝试了传统的 monophone，triphone，lda-mllt，SAT 的 GMM 模型，此外还尝试了用神经网络 TDNN 对单一 Xhosa 语言进行识别，最后使用 Zulu+Xhosa 双语言对 multilingual TDNN 语音识别系统进行训练，相比 SAT 模型有着 10%的识别精确度的提升。

✧ 验证了 multilingual TDNN 模型对小语种语音识别的强大功能。并且将语音识别系统对 Xhosa 识别的错误率降低到了 18.48%。

✧ 项目 GitHub 地址：https://github.com/LYSuperCarrot/Multilingual-tdnn-asr-for-xhosa

NLU 课程项目-机器翻译-参与整个流程 02/2021-03/2021

✧ 该项目是搭建 LSTM 的 seq2seq 网络作为 baseline，从而将法语翻译成英语。

✧ 还在 baseline 基础上尝试了 lexicon 模型使翻译精确度更高，提升了 2.4 的 BLEU 值。

尤元岳

- ✧ 此外又在此项目中应用了 multi-head attention 技术并且搭建了 transformer 结构，比较了多种模型对机器翻译的表现。
- ✧ 项目 GitHub 地址：https://github.com/LYSuperCarrot/NLU_cw2_NMT

实习经验

常州铭赛机器人科技公司 06/2019-08/2019

- ✧ 开发了 Weinview 触摸屏界面，读取和写入触摸屏中的数据，通过搭建 socket 服务其和客户端实现与电脑进行通信
- ✧ 协助产品经理评估产品升级计划

个人账号

- ✧ GitHub：https://github.com/LYSuperCarrot
- ✧ CSDN：https://blog.csdn.net/Mr_Carrot?spm=1001.2101.3001.5343

获得的奖项

- ✧ 连续三年获得校级三等奖学金 校级 2016-2019
- ✧ 获得星火杯创新大赛三等奖 校级 2017
- ✧ 英语四六级
- ✧ 雅思：6.5（听力 6.5 阅读 7.5 写作 6.5 口语 5.5）
- ✧ 《一种基于扩散模型的专利文档查询方法》第一作者的授权专利，《一种在大语言模型生成文本中嵌入及检测数字水印的方法》第三作者的授权专利，另有《一种基于大模型的商标生成方法及系统》和《一种基于扩散模型的科技文献附图生成方法及系统》第一作者的两篇专利正在审查过程当中

个人评价与爱好

- ✧ 申请研究生时，曾获得了帝国理工的通信工程和信号处理专业的 offer，爱丁堡大学人工智能的 offer，因喜欢编程和人工智能算法，研究生方向选择了爱丁堡大学人工智能专业
- ✧ 工作认真负责，工作期间绩效一直保持组内第一第二的排名，并且年终绩效为组内第一。擅长沟通交流，工作期间经常跨部门沟通协作，经常作为部门对外的对接人与其他部门同事沟通并且解决问题。
- ✧ 喜欢有挑战的事情，遇到困难不退缩，做事喜欢条理分明，按照计划做事，不拖延
- ✧ 喜欢跑步，打羽毛球，打乒乓球，踢足球