个人优势

- **技术能力：精通 Python，熟悉前端、C/C++（可看懂并修改代码）**
- **工作经验：研究生方向为 NLP 领域（机器翻译），目前已 4-5 年工作经验**
- **过往绩效：近三年绩效分别为 3、4、3（5 分制）**
- **软性素养：在校期间担任过班长职务，有一定的组织、团结协作能力**

个人简历

基本信息

姓　名：黄佳跃

性　别：　男

出生年月：1994 年 12 月

手机号码：18896727208

邮箱地址：121123953@qq.com

教育背景

2017年9月—2020年7月	苏州大学	软件工程（NLP 方向）	硕士
2012年9月—2016年7月	苏州科技大学	环境科学	本科

备注：中间 Gap 一年参见以下 2016.07 — 2017.07 的工作经验，本硕均为全日制

工作经验

2022.03 — 至今	智慧芽信息技术有限公司	高级自然语言处理工程师
2020.07 — 2022. 02	中移软件技术有限公司	Python 自然语言处理工程师
2016.07 — 2017.0 7	苏州科大环境发展股份有限公司	OA 系统研发工程师

项目经验

- **技术问题中的主题词、属性、词关系抽取**
  - 如针对技术问题描述（输入即为一句问题描述）：
    - “传统的传动轴密度大”，抽取“传动轴”（主题词）、“密度”（属性）、“大”（对属性的描述），词关系则可表述为“传动轴的密度大”
    - “导致同批次高温退火后的磁阻的一致性较差”，抽取“磁阻”、“一致性”、“差”，词关系表述为“磁阻的一致性差”
  - 基于 Bert + CRF 的架构结合基于 BIO 标注法标注的 5k 条标注数据进行 Finetune
  - 项目亮点：
    - 实验对比了多种 NER 技术方案（如 Bert + BiLSTM + CRF），最终经过学习率调参后选用 Bert + CRF 架构（其中 Bert 经过专利领域数据进行 continue pretrain）

- 数据增强：得到初版模型后进一步做数据增强，使用模型对未标注数据进行标注，并抽样观测拟定规则、筛选高质量标注数据并进一步核验后进行语料扩充，优化原有模型的 F1 值至约 90%

➢ 专利文本中的技术功效段抽取

- 任务：抽取一篇专利文本中的功效段描述；功效段通常指文本片段（一句或相邻的多句），供下游应用（可进一步抽取功效段描述中的主题词、属性、词关系）
- 专利文本属于比较高度结构化的文本，一篇专利通常包含技术领域、背景技术、发明内容、附图说明、实施例等部分的描述，通常专利的功效段描述包含在发明内容中，且部分功效段表述有着比较固定的文本描述范式（此部分可基于文本表述规则通过正则匹配获取），但也有一些功效描述比较灵活多变。
- 训练数据的获取：基于内部已有的功效段描述数据 + 通过规则匹配获取的功效段描述，作为训练数据，并通过下采样的方式构造非功效段数据，使得配比平衡（1：1）
- 模型架构：基于 Bert + TextCNN 架构进行文本分类判断、抽取功效段
- 项目亮点：
  - 采用 Bert 的多种输出（最后一层 [CLS] 的 embedding、多层取平均）作为 TextCNN 的输入进行实验，得出最优方案
  - 数据增强：训练初版模型结果后抽取功效段，通过分析挑选 badcase 并将其作为训练数据、进一步进行数据扩充训练

➢ 生物医药领域论文检索

- 针对 query 进行关键词抽取（分词、词性标注、停用词过滤、NER），基于关键词构建查询语句请求搜索接口得到初召回结果（含 BM25 得分，该得分将于后续的语义得分进行加权）
- 基于 BERT 架构 + 对比学习进行无监督训练，得到语义模型后对基于关键词搜索初召回结果进行重排序（精排），提升搜自身、Top-100 的相关度（因为该任务没有标准的测试数据，Top-100 相关度主要通过测试团队人工 check）
- 无监督对比学习过程的数据构造：同一文本片段经过 2 次 BERT 编码的结果（由于每次均有 dropout，因此两次编码结果并不相同）为一对正样例对，负样例则采用 in-batch negative 的方式构建，损失函数使用 InfoNCE 交叉熵损失
- 项目亮点：
  - 训练初版语义模型时，通过梯度重计算（Gradient Checkpointing）优化训练时中间激活的显存占用，尽可能增大 batch size（为 128）让 in-batch negative 中负样例的训练更充分
  - 训练得到初版语义模型后，人为构造 hard negative 样例进一步优化优化模型效果；hard negative 样例来自于初版模型的效果评估（分析 badcase）以及从 top-100 中随机抽取得来（抽取结果会排除目标正样例、并经过分析人员再次 check）

- **专利检索（语义检索 Rerank）**
  - 搜索初召回结果的获取方式同以上“生物医药领域论文检索”
  - 语义模型预训练（pretrain）：利用大批量（（2000 万+）专利文本数据，使用 Auto-encoder 架构（参考 RetroMAE 思路，即 BGE 预训练思路）进行预训练
  - 模型微调（finetune）：使用预训练架构中的 Encoder 进行文本向量化，随后基于对比学习进行微调（一篇专利与其对应的审查员推荐的相似专利作为有效的标注数据，将该数据作为一对正样例，负样例同样采样 in-batch negative 方式构建，损失函数为 InfoNCE
  - pretrain + finetune 后，对粗召回结果进行精排能使得 Top-100 的结果提升 10%（如果使用开源模型进行 finetune，而没有垂域数据的 pretrain，效果会低 3% 左右）
  - 项目亮点：
    - 优化 RetroMAE 预训练的过程中加载数据的方式，使得数据量巨大时也不会出现 OOM 问题
    - 结合专利特点，额外构造负样例进行对比学习训练（筛选出审查员推荐的相关但不相似的专利数据作为负样例）

- **大模型预训练、RLHF**
  - 基于 Chinese LLaMA、QWen 进行专利领域大模型的二次预训练（continue pretraining）、SFT，参考 DeepSpeed-Chat 的一套流程进行 RLHF（RM 训练 + PPO），训练模型进行专利领域的相关问答（如关键信息抽取、文本摘要等）

- **大模型 SFT、DPO**
  - 项目背景：公司内部大模型团队基于多类 base 模型（LLaMA、Qwen、Mixtral 8*7B）尝试训练了多个垂直领域大模型（通用专利领域、生物医药领域、新能源汽车领域、通信领域等），而针对
  - 基于基础专利大模型进行 SFT、DPO，训练模型参加临床执业医师考试（单选题考试）
  - 基于基础专利大模型进行 SFT、DPO，训练模型进行意图识别（判断用户输入的 query 适合使用具体哪一类垂直领域大模型进行请求解答）

- **大模型工程应用**
  - 结合专利大模型 + 向量知识库进行 AI 智能问答（RAG）：针对用户所提问题，先通过语义检索得到结果，并拟定 Prompt 调用专利大模型，生成 summary 回答并返回
  - AI 搜索增强（大模型辅助短文本搜索）：构造 prompt 模板结合大模型接口调用，对用户输入的 query 进行意图识别、query 改写、AI 扩词（对关键词进行同义词、简写、缩写扩充），并拼接成搜索接口指定的请求格式进行搜索