匡盟盟

邮箱：mmkuang@connect.hku.hk | 电话：(+86)132-7015-3586

GitHub：https://github.com/kuangmeng | 主页：https://mkuang.tk

工作经历

2023.11 至今 | **推荐算法工程师** *(2-2)*，DATA-抖音，字节跳动，上海

- **抖音系 App 的推送时机流控**：挖掘特征、构建 Uplift 模型决定抖音系 App 推送的频次和时机。

2021.03 - 2023.11 | **应用研究** *(T8)*，搜索技术中心，微信事业群，腾讯科技，广州

- **大模型在搜索系统中的应用**：(1) 构建大模型 RLHF 的数据集，基于自研 self-instruction 构建出整套单/多轮对话数据集生成策略。为了为大模型提供事实依据，我们研究了微信龙珠大卡以 Plugins 的形式接入大模型预测的可行性，同时提出 Memory 机制有效缓解了大模型输入上下文长度限制的问题。(2) 利用大模型对评论、描述等文本进行关键词/短语提取，丰富文档特征。
- **账号类搜索项目的 Doc 理解**：(1) 从零开始搭建微信搜索的数据增强系统。通过对文档标题等数据进行增强，扩大召回范围并提高相关性模块计算的准确性。(2) 对账号文档的描述、认证实体和子页面进行分析，以理解或挖掘出有利于召回和相关性计算的文本。同时，从账号菜单和外部 URL 中挖掘出可用的子服务以提升账号的召回率。
- **账号类搜索项目的向量检索召回**：为了使用户查询能匹配更合适的文档，我们在原有的文本检索基础上引入了一种关键词加权的双塔模型，训练出高质量的向量检索队列，以此扩充召回范围，同时从效率考虑我们采用聚类量化方式线上部署。

研究课题

用于在线检索系统的通用数据增强系统 | *2022.08 - 2023.11*

为了提高多数据源检索系统的通用性，以及对文档实体进行更全面、准确的描述，同时为了解决搜索改写难以解决的行业定制化表达的难题，我们提出一种通用数据增强系统。该系统能对多来源、异构数据进行统一增强和理解，并且输出结构统一的干净数据用于检索、排序。该系统采用分层多队列设计，从底往上分别由数据源、数据湖、低层增强策略、高层增强策略、策略融合、后验与知识储备等组成。该系统能综合使用向量检索、抽取模型、生成模型、回译、同义词、知识图谱等策略，从人工词典、搜索 Session、外部数据、文档自身数据等来源中获取高质量的增强数据。该增强数据最终会以一定的格式作用于检索系统，提升整合系统的召回率。目前应用了该系统的微信账号类搜索的召回率达到 95% 以上。

文本分类数据样本优化（样本去噪/噪声适应性学习） | *2021.03 - 2022.06 - 2023.11*

数据质量一直是深度学习模型性能提升的瓶颈。为了提高现有人工标注数据的准确性，我们提出了一种两阶段噪声标签检测的样本优化策略。首先，我们使用 BERT 在要去噪的数据上训练一个初步分类器，并生成噪声候选集。然后，我们使用剩余的干净候选集样本在该分类器上进行深度训练。接下来，我们通过该分类器对噪声候选集中样本各类别的预测概率，得出针对各类别的置信矩阵，并根据这些矩阵过滤掉每个类别中的不可信样本（即噪声样本）。实验结果表明，该方法能将标签净化率提升至 96% 以上。目前，这种样本优化方法已应用于微信搜索的各类样本清洗任务。我们正在不断迭代优化去噪算法，并尝试将类似的方法应用于大型模型的有监督微调（SFT）数据集去噪。

风格不变的文本生成器 | *2021.11 - 2022.11*

【上述数据增强系统的一个算子】自然语言生成方法（例如 Seq-to-Seq 模型）在文本生成任务中通常面临泛化能力不足和语义偏移的问题，这成为了在线任务准确率要求较高场景的难题。在本项目中，我们提出了一种基于预训练模型 BERT 的文本生成方法，通过学习待改写或待增强文本的固有语言模式（即语言风格）来克服这些问题。我们设计了一种基于层叠式自适应规范化 (SAdaIN) 模块的文本生成方法，该方法在中文基准数据集（LCQMC 数据集和私有数据集）上的 BLEU-4、Rouge-L 和 SARI 指标表现出色。此文本生成算法已成功应用于微信搜索某些类目的数据增强任务。

关键词加权的向量检索 | *2021.03 - 2021.07*

为了更准确地检索匹配度较高的文档，我们需要精确识别查询和文档中的关键字。为此，我们设计了一种领域自适应的多任务学习模型，通过联合训练一个双塔语义匹配模型和一个关键字识别模型，以精确获取查询与文档间的关联性。双塔语义模型将查询与文档分别编码为语义向量，并在相似度计算阶段将它们结合起来。此外，我们引入了一个关键字识别模型，用于自动检测查询与文档中的关键字权重。我们在公开的 MS MACRO 数据集和微信搜索数据集上进行了实证研究，结果表明，我们的方法优于其他语义检索模型。此框架已成功应用于微信搜索的服务搜索任务。

以数据为中心的多序列比对算法研究 | *2018.09 - 2020.06*

为了提升蛋白质家族多序列比对（MSA）构建的质量，特别是针对“低相似性”家族，我们提出了一种基于两阶段序列建模的 MSA 方法。该方法利用 Transformer 模型为不同类型的蛋白质家族寻找最合适的算法为核心的构建策略。在测试了 711 个“低相似度”蛋白质家族的构建任务后，我们发现相较于其他算法，该方法可以在平均构建准确率上实现 2.8% 的提升。

领域快速移植的中文分词系统 | *2017.11 - 2018.06*

为了提高中文分词的准确性和领域适应性，我们采用了基于条件随机场（CRF）和维特比算法的方法，在人民日报 1998 年的人工分词数据上训练和开发了一套具有领域适应能力的中文分词系统。该系统利用启发式规则和特定领域规则进行优化，从而增强了其在特定领域（如医学、法律和金融）的性能。经测试，该分词系统在这些特定领域的文本分词任务中的平均准确率（F_1 值）高达 97% 以上。

研究论文

[C1] Efficient two-stage label noise reduction for retrieval-based tasks. (WSDM 2022)

[C2] Multi-task learning based Keywords weighted Siamese Model for semantic retrieval. (PAKDD 2023)

[J1] MLProbs: A Data-centric Pipeline for better Multiple Sequence Alignment. (IEEE TCBB)

教育经历

2018.09 - 2020.12 **研究型硕士**，计算机科学，**香港大学**
导师：Hing-fung Ting 教授
毕设：以数据为中心的多序列比对算法研究
研究方向：多序列比对、机器学习、深度学习

2014.09 - 2018.06 **工学学士**，计算机科学与技术，**哈尔滨工业大学**
指导教师：赵铁军教授
毕设：面向领域快速移植的高精度汉语分词系统研究
成绩：89.6/100

奖励和证书

工作期间

2023.01 腾讯业务突破奖：搜一搜用户增长与商业化规模双突破项目

研究生期间

2021.03 李嘉诚奖学金（提名）
2021.03 港大优秀研究毕业生（提名）
2020.11 华为认证 ICT 工程师 – 人工智能
2018.11 Certificate of Teaching and Learning in Higher Education
2018.09 全额研究型硕士奖学金

本科生期间

2018.06 企业奖学金
2015.12 三好学生
2015.11, 2016.11 国家励志奖学金