匡盟盟

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

地址：中国广州市海珠区新港中路 397 号 T.I.T. 创意园 B5 栋

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

教育经历

2020.12 **研究型硕士**，计算机科学，**香港大学**
导师：Hing-fung Ting 博士
毕设：Data-centric Approaches for better Multiple Sequence Alignment
研究方向：多序列比对、机器学习、深度学习

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

工作经历

项目

用于 *Query* 改写和数据增强的风格不变的文本生成器 | *2021.11 至今*

基于自然语言生成（如 Seq-to-Seq 模型）的文本生成方法通常存在泛化性不足和语义偏移的问题，不适合用在对准确率要求较高的在线任务中。在这个项目中，我们提出通过学习待改写或增强的文本的固有语言模式（即语言风格）来完成基于预训练模型 BERT 的文本生成任务。该方法在三个基准数据集（QRECC 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。这种文本生成算法已经逐步用于微信搜索的 Query 改写和数据增强任务。

两阶段分类数据样本优化 | *2021.03 至今*

数据质量一直是制约深度学习模型突破的瓶颈。为了提高现有人工标注数据的准确性，我们提出了一种有效的两阶段样本优化（噪声标签检测）方法。首先，我们依靠 BERT 对所有要去噪的数据训练一个粗略分类器，并生成一个噪声候选集。然后我们通过分类器对这个噪声候选集中样本不同类别的预测分类概率，计算出针对不同类别的置信矩阵，并依次来过滤每个类别中的不可信样本（即噪声）。实验结果表明，该方法可以将标注干净率提高到 96% 以上。该样本优化方法已经运用到微信搜索的各种样本清洗任务中。

用于语义检索的关键词加权双塔模型 | *2021.03 - 2021.07*

为了检索更好匹配的文档，需要准确地识别 Query 和文档中的关键字。我们提出了一种新颖的领域自适应多任务学习模型，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来精确获取 Query 与文档之间的相关性。双塔语义模型独立产生 Query 和文档的语义向量，并仅在相似度计算阶段将它们耦合。我们引入了一个关键字识别模型来自动检测 Query 和文档中的关键字权重。实证结果表明，我们的方法在两个语义检索数据集（MS MACRO 数据集和微信搜索数据集）上优于其他语义检索模型。该框架已经在微信搜索中的服务搜索任务中上线。

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

为了提高蛋白质家族的多序列比对 (MSA) 构建的质量，尤其是“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，基于 Transformer 模型，为不同类别的蛋白质家族安排合适的以算法为中心的构建方式。该方法在 711 个“低相似度”蛋白质家族的构建任务中，相比于其他算法平均构建准确率可以提高 2.8%。

领域快速移植的中文分词系统 | *2017.11 - 2018.06*

为提高中文分词的准确性与领域适应性，在人民日报（1998）的人工分词结果上，基于条件随机场和维特比算法，训练开发了一套领域适应的中文分词系统。启发式规则和特定领域规则，提高了该分词系统在特定领域（医学、法律和金融）的适应性。该分词系统在这些特定领域文本分词任务中的平均准确率（F_1 值）可以达到 97% 以上。

工作

2021.03 至今 | **应用研究**，微信事业群，腾讯科技，广州
参与并负责微信搜索相关的 Query 改写、数据增强、向量检索以及数据去噪等算法的设计与开发。

2020.10 - 2021.02 | **研究助理**，伦敦帝国理工
在杨光博士的指导下，参与一个医疗图像分析项目并提出了一种 3D 嵌入式注意力的 RNN 分割器，用于从心肌速度映射中分割出左心室的结构。

2020.10 - 2021.02 | **研究员实习生**，思谋科技，香港
在余备博士的带领下，参与 SMore OCR 产品的训练数据增强、识别算法的神经网络设计与实现。

2018.06 - 2018.08 | **研究助理**, 哈尔滨工业大学
在赵铁军教授的指导下，基于领域移植的统计翻译开发一款医疗信息检索和诊断平台。

研究 *[挑选的一作论文]*

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

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

助教

2020.01 - 2020.06 (COMP1117[HKU]) 计算机编程（Python）
2019.02 - 2019.07 (COMP7606[HKU]) 深度学习
2017.09 - 2018.02 (13SC03100600[HIT]) 软件工程
2017.09 - 2018.02 (IR03000900[HIT]) 互联网文本语义挖掘

奖学金和证书

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
2016.11 国家励志奖学金
2015.12 三好学生
2015.11 国家励志奖学金

技能

基础 TensorFlow、编程（C/C++、Java、Python 等）、办公软件
中级 RESTful, Keras, Scikit-Learn, PyTorch
掌握 机器学习与深度学习建模、多序列比对