钱泓锦

13051011770 ian@ruc.edu.cn chien.io

教育经历

中国人民大学 2020.09 – 2024.06
人工智能 高瓴人工智能学院 博士研究生 导师：窦志成 & 文继荣 北京

悉尼大学 2017.07 – 2019.01
数据管理与分析 硕士研究生 澳大利亚，悉尼

南开大学 2013.09 – 2017.06
电子信息科学与技术 理学学士 天津

工作经历

泊松实验室，华为 2022.04 – 至 今
实习生 北京
探索搜索引擎的未来形态——对于事实性Query，实现从Web中自动寻找相关知识，并生成一个描述性的事实型文章，作为搜索结果返回。相关落地产品WebBrain已于HDC2022正式发布，相关工作成果已投稿至SIGIR2023。

- 开放领域事实型文本生成
- 检索增强的文本生成

大数据管理与分析方法研究北京市重点实验室，中国人民大学 2020.09 – 至 今
博士研究生 北 京
主要研究方向为自然语言处理和信息检索，入学以来致力于探索未来人机信息交互的多种形态，已在相关领域顶级学术会议上发表多篇文章，并提交多项专利申请。

- 对话式搜索 (SIGIR2022, EMNLP2022, WWW2023)
- 基于深度学习的信息检索 (SIGIR2022)
- 基于用户建模的对话机器人 (EMNLP2020, SIGIR2021, CIKM2021, ECIR2023)

北京智源人工智能研究院 2020.06 – 2021.03
算法工程师 导师：刘占亮 北京
团队致力于开创政务领域多任务问答助手，聚焦北京多部门政务场景。在项目申报阶段，主要参与技术计划书的编写；在项目启动阶段，负责算法团队研发工作。项目期间，提交专利申请15个，已授权14个（第一专利发明人5个），曾带领团队获得2020年度CCF政务疫情问答比赛Top3。

- 政务领域问答助手系统
- 搭建政务垂域混合搜索引擎
- 开放领域细粒度命名实体识别

一览群智 2019.01 – 2020.09
算法工程师 导师：刘占亮 北京
负责从零搭建基于深度学习的多语种NLP工具包LensNLP，任务包括文本分类、命名实体识别、关系抽取、语义解析等，功能上实现数据融合、一键训练、一键部署、分布式加速等，形态上实现pip包，Docker file构建，Docker包等。负责支持NLP技术在金融、公安、媒体等领域的落地实现，工作内容包括售前技术展示、需求对接、定制化开发，以及项目后期的整体交付和技术总结。负责探索前沿人工智能技术实现，快速搭建demo，并评估落地价值、数据可用性和领域适配性。

- 中文、英文、维吾尔语自然语言处理多任务工具包
- 自然语言处理技术在多领域的应用和落地
- 探索特定领域预训练语言模型的训练和应用

已发表论文

* indicates equal contribution.

(1) **Learning Denoised and Interpretable Session Representation for Conversational Search. WWW 2023** (CCF-A)
Kelong Mao, **Hongjin Qian**, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, Zhao Cao

(2) Topic-Enhanced Personalized Retrieval-based Chatbot. **ECIR 2023** (CCF-C)
Hongjin Qian and Zhicheng Dou

(3) Explicit Query Rewriting for Conversational Dense Retrieval. **EMNLP 2022** (CCF-B)
Hongjin Qian and Zhicheng Dou

(4) ConvTrans: Transforming Web Search Sessions for Conversational Dense Retrieval. **EMNLP 2022** (CCF-B)
Kelong Mao, Zhicheng Dou, **Hongjin Qian**, Fengran Mo, Xiaohua Cheng and Zhao Cao

(5) Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation. **NAACL 2022** (CCF-B)
Hanxun Zhong, Zhicheng Dou, Yutao Zhu, **Hongjin Qian** and Ji-Rong Wen

(6) Webformer: Pre-training with Web Pages for Information Retrieval. **SIGIR 2022** (CCF-A)
Yu Guo, Zhengyi Ma, Jiaxin Mao, **Hongjin Qian**, Xinyu Zhang, Hao Jiang, Zhao Cao and Zhicheng Dou

(7) Curriculum Contrastive Context Denoising for Few-shot Conversational Dense Retrieval. **SIGIR 2022** (CCF-A)
Kelong Mao, Zhicheng Dou and **Hongjin Qian**.

(8) Learning Implicit User Profile for Personalized Retrieval-Based Chatbot. **CIKM 2021** (CCF-B)
Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen

(9) Pchatbot: A Large-Scale Dataset for Personalized Chatbot. **SIGIR 2021** (CCF-A)
Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou and Ji-Rong Wen

(10) Speaker or Listener? The Role of a Dialog Agent. **Findings of EMNLP 2020** (CCF-B)
Yafei Liu*, **Hongjin Qian***, Hengpeng Xu, Jinmao Wei

已授权专利

(1) **一种基于拉丁字母的维吾尔语处理方法和系统** (CN111428509A)
钱泓锦, 黄真, 窦志成, 刘占亮

(2) **一种开放领域问答任务中长文本检索的方法和电子设备** (CN111881264A)
钱泓锦, 刘占亮, 刘家俊, 窦志成

(3) **一种融合语言模型的光学字符识别方法、装置和电子设备** (CN111738251A)
钱泓锦, 刘占亮, 窦志成, 刘家俊

(4) **一种维吾尔语实体识别的方法、装置和电子设备** (CN111814433A)
钱泓锦, 刘占亮, 窦志成, 刘家俊

(5) **一种多层级长文本向量检索方法、装置和电子设备** (CN112988952A)
钱泓锦, 刘占亮, 窦志成, 文继荣, 曹岗

(6) **一种基于规则与学习的语义解析方法、装置和电子设备** (CN112347793A)
钱泓锦, 李晓桐, 刘占亮, 杨玉树, 窦志成, 曹岗, 文继荣

(7) **一种基于索引数据的自然语言处理方法和系统** (CN111488423A)
刘占亮, **钱泓锦**, 窦志成, 刘家俊

(8) **基于语义规则和多维模型的多数据源NL2SQL系统** (CN112559550A)
李智, **钱泓锦**, 刘占亮

(9) **一种政务FAQ知识库自动构建方法、装置和电子设备** (CN112784022A)
郭司绪, **钱泓锦**, 杨玉树, 刘占亮, 窦志成, 曹岗, 文继荣

(10) **基于复杂数据类型的faq知识库自动生成方法和装置** (CN112800177A)
郭司绪, **钱泓锦**, 杨玉树, 刘占亮, 窦志成, 曹岗, 文继荣

(11) **一种文本检索结果评分方法、检索方法和装置** (CN111930928A)
张宇, **钱泓锦**, 刘占亮, 窦志成

(12) **一种图像文档的文本抽取方法、装置及电子设备** (CN112036406A)
黄园园, **钱泓锦**, 刘占亮, 窦志成

(13) **一种语义检索方法、装置及电子设备** (CN112035730A)
周阳, **钱泓锦**, 刘占亮, 窦志成

(14) **基于向量化语义规则快速实现NL2SQL的方法和装置** (CN112001188B)
肖超峰, 李智, **钱泓锦**, 刘占亮

(15) **一种接口性能测试方法、装置及电子设备** (CN111881060A)
张欢, 李智, **钱泓锦**, 刘占亮, 窦志成
(16) **基于表格数据的FAQ知识库自动构建方法和装置** (CN112800032A)
郭司绪, 杨玉树, **钱泓锦**, 刘占亮, 窦志成, 曹岗, 文继荣

申请中专利

(1) **一种基于汉语拼音的词表征方法及装置** (CN110162789A)
窦志成, **钱泓锦**, 黄真
(2) **一种检索式个性化对话方法与系统** (CN113901188A)
窦志成, **钱泓锦**

项目经历

WebBrain: 探索未来搜索引擎形态 2022.04 – 2022.12
与华为泊松实验室合作，探索对于事实型查询，自动从互联网搜索相关知识，并生成带引用的描述文档，作为搜索结果返回。在项目中，我们从零开始构建了大规模数据集、定义了任务形态、提出了一个领先的技术框架。

基于文澜多模态预训练模型的视觉障碍辅助工具 2021.11 – 2022.01
支持某头部手机厂商OS内置视障辅助功能：利用多模态预训练模型为图片生成自然语言描述。主要负责算法代码的工程化、并发性能调优、容错性能优化。

智能政务信息助手 2020.05 – 2021.03
依托智源人工智能研究院，研究新一代人工智能信息助手，并重点立足于政务领域用户的实际需求。在项目中负责开发基础文档检索服务、语义理解服务和上层的基于文档的问答服务。

政务领域项目POC 2019.04 – 2020.06
支持业务方进行多个项目的POC：太极融媒体NLP服务、公安部部标文本服务接口、基于语义解析的公安领域NL2SQL、地址数据标准化等项目。

金融领域项目POC 2019.04 – 2020.06
支持业务方进行多个项目的POC：建信金科小微企业贷款反欺诈、建信金科国际结算智能审单、建信金科运营流程知识图谱、招银科技外汇智能审单等项目。

多语言NLP工具包LensNLP 2019.02 – 2019.06
整合团队内部分散的NLP模型，制定训练数据、模型代码、推理API、模型部署方式的标准，对效果差的工具重新开发。实现NLP工具包多种方式 (pip, Docker, etc.) 一键式训练、快速部署等功能。

维吾尔语 Core NLP Toolkit 2019.01 – 2019.05
从零开始爬取维吾尔语文本，招募人工标注团队进行数据标注 (得到业界最大的维吾尔语NER数据集)，训练维吾尔语语言模型，开发NLP core task模型，最终进行Demo搭建。

专业技能

- 编程语言: Python, C, Go, SQL, Shell
- 框架&工具: Pytorch, Tensorflow, ElasticSearch, Kibana, Docker
- BI工具：Mixpanel, Tableu, Google Data Studio, Segment.io
- 语言：英语 (熟练)

获奖情况

[Top10%] 2022年中国人民大学拔尖人才创新资助计划 2023.01
[Top10%] 2021年中国人民大学拔尖人才创新资助计划 2022.01
[Top30%] 一等学业奖学金 2021.11

其他

- PC Reviewer: EMNLP 2021, EMNLP 2022, EACL 2023, ACL 2023
- 助教: 《数据结构》, 《学术写作》