林韵

linyun@stu.pku.edu.cn | +86 13425559546

教育经历

北京大学 北京

硕士，计算语言学 GPA: 3.78/4.0 2022.09 – 2025.06

修读课程：机器学习，深度学习，自然语言处理，Python 数据分析原理与应用，语义学。保研成绩排名第 1。

斯坦福大学 美国

暑期交换，计算机科学 GPA: A 2023.06 – 2023.08

修读课程：CS229，CS221，CS224N。课程报告《How do Computers “Think” about Syntax?》A+

中山大学 广州

本科，应用语言学 GPA: 4.2/5.0 2018.08 – 2022.06

修读课程：系统功能语言学、语料库语言学、语篇分析、语言景观

所获荣誉：校级优秀毕业论文、校级优秀毕业生、优秀学生奖学金、弘毅学生干部奖学金。

实习经历

百度 文心一言 北京

自然语言处理部北京组-NLP 实习生 2023.08-2023.11

- **提示工程：**主要负责文心一言模型文本创作能力方向，重点涉及幽默、华丽、华丽、恐怖等 8+语言风格。涉及 case 分析、标准构建、prompt 工程、query 富集全流程，遵循 RLHF 流程提供 SFT-RM-PPO 格式化数据，文心一言 4.0 的文创能力显著提升。
- **工具开发：**根据模型训练需求开发效率工具。训练 ERNIE 模型开发语义识别工具并部署服务，训练后 F1 达 0.97；开发指令进化工具，根据 instruction 富集 query + response，提升组内效率 60%。开发的工具 100%被采用。
- **论文分享：**进行文献阅读与调研 40 篇+，调研该领域前沿的工作成果，每周定期进行论文分享，已成功将 Evol-Instruct、RRTF 等方法运用至工作之中，探索将前沿研究成果与组内落地业务相结合的可行方法。

字节跳动 AI Lab 北京

AI-Lab 智能语音-算法-音频理解-专家 2023.02-2023.06

- **模型训练：**参与 7+项目音频理解模型的训练、测试和优化，独立负责财经通用项目多意图模型的训练任务。根据模型跑测的精确率、召回率、F1 值指标分析模型优化问题，对训练产生的 bad case 进行修复，优化当前数据集。每日可完成优化标签 8-10 个，模型训练效率组内第 1。运用正则表达式辅助捞取数据中正负例，确保训练好的模型按时迭代。
- **数据增强：**编写 Python 程序调用 bert 对模型训练数据进行增强，依据 query 语义内容对文本数据进行同义词替换、删减、增添、回译等操作，增加模型训练数据的多样性及质量的同时减少过拟合风险。该脚本被组内成员所采用，提升组内工作效率 30%。
- **数据处理：**参与番茄小说项目数据生产、体系建立、标注、清洗等工作。与公司 AI 小说大模型按主题对话生产数据，并依据其生成文本的流畅性、相关性、延展性等指标建立完善的标注体系，为模型训练提供高质量的训练数据。

学术经历

北京大学王选计算机研究所

实习学生

- 参与导师重点项目，进行数据污染方向研究，聚焦于结合语言学中句法特征研究 LLMs 测试集中数据是否被包含在训练集中。

论文&项目

- **NLP 项目：**作为主要成员参与北京大学人工智能研究院与数字人文研究中心组织的 CCL2023 年古籍命名实体识别评测项目。通过微调“bert-ancient-chinese”及数据增强技术在《二十四史》上进行古籍 NER 任务，最终在训练集上获得 82.3 的 F1 值。
- **在投论文：**《基于人工智能的外语教学新范式：ChatGPT 在外语教学中的应用能力评测》、*LDA at 20: Applications and Advances of Topic Modeling in the Humanities*、《法律语言学视域下粤港澳三地反家暴法对受害者形象建构的对比研究》
- **毕业论文：**《基于语料库的文献综述中投射语言分析》中山大学校级优秀毕业论文 + 第十五届北京大学研究生论坛优秀论文。

技能

- **语言：**英语（雅思 7.5，专四优秀+专八良好）、法语（辅修）、中文（普通话、粤语、潮汕话，均为母语）
- **编程：**Python, PyTorch, VS Code, Linux, AntConc, UAM Corpus Tool
- **兴趣：**北京大学女篮校队队长、中山大学女篮校队队长（团队协作精神）

YUN LIN

linyun@stu.pku.edu.cn | +86 13425559546

EDUCATION

Peking University **Beijing, China**
Master, Computational Linguistics *GPA: 3.78/4.0* Sep 2022 – Jun 2025
Related Coursework: Machine Learning, Deep Learning, NLP, Big-data Analysis with Python, Semantics

Stanford University **Stanford, CA**
Summer Session Student, CS *GPA: A* Jun 2023 – Aug 2023
Related Coursework: CS221, CS224n, CS229. Final Paper: How do Computers "Think" about Syntax? A+

Sun Yat-Sen University **Guangzhou, China**
B.A., Applied Linguistics *GPA: 4.2/5.0* Aug 2018 – Jun 2022
Related Coursework: Functional Linguistics, Corpus Linguistics, Discourse Analysis
Honors: Excellent Undergraduate Thesis & Outstanding Undergraduate Honor

INDUSTRY EXPERIENCE

Baidu, Inc. **Beijing, China**
ERNIE Bot, NLP Intern Aug 2023 – Nov 2023

- **Prompt Engineering:** Primarily responsible for text generation capabilities of ERNIE Bot, focusing on 8+ language styles Including humor, horror, elegance, etc. Followed RLHF process to provide formatted SFT-RM-PPO data for training.
- **Tool Development:** Trained an ERNIE-based model to do binary semantic classification, achieving 0.97 F1 score. Developed an instruction evolution tool to enrich query + response based on instructions, which improved team efficiency by 60%.
- **Paper Sharing:** Conducted literature review on 40+ papers and provided paper sharing session on a weekly basis to investigate cutting-edge advancements in this field. Successfully applied methods like Evol-Instruct and RRTF to our work.

ByteDance. AI Lab **Beijing, China**
AI-Lab Speech & Audio Team, NLP Expert Intern Feb 2023 – Jun 2023

- **Model Training:** Trained, optimized, and tested NLP models to classify speech2text queries from the app TikTok following established workflows to ensure timely deployment and model update.
- **Bad Case Fixing:** Monitored and evaluated the performance of responsible models in production, fixing bad case problems. Based on the model outcomes, provided feedback on labeling standards and data quality.
- **Data Creating:** Created data for ByteDance in-house ChatGPT-like LLM, participating in the processes of data production, data cleaning, data annotation and data quality control. Explored techniques to fix low quality data in the training sets.

RESEARCH EXPERIENCE

Wangxuan Institute of Computer Technology, Peking University
Student Researcher

- Participating in the advisor's major project related to *data contamination*, with a special focus on combining linguistic syntactic features to study whether data in LLMs test sets are included in their training sets.

Papers & Projects

- **GuNER project:** Fined tune "bert-ancient-chinese" model to do ancient Chinese literature NER task with the use of data augmentation techniques to reach a F1 score of 82.3 on the test set. (CCL 2023 competition)
- **Undergrad Thesis:** *A Corpus-based Analysis of Projection in Literature Review Chapter.* Accepted by Peking University 15th Graduate Student Academic Forum 2023 + Outstanding Graduation Thesis
- **Preparing Papers:** *A New Paradigm for Foreign Language Teaching Based on AI: Application Proficiency Assessment of ChatGPT in Language Education; LDA at 20: Application and Advances of Topic Modeling in the Humanities*

SKILLS

- **Languages:** English - IELTS 7.5 | French – minor | Mandarin, Cantonese, Chaoshan Dialect – native speaker
- **Programming:** Python, PyTorch, VS Code, Linux, AntConc, UAM Corpus Tool
- **Interest:** PKU Women's Basketball Team - Captain; SYSU Women's Basketball Team – Captain (spirit of teamwork)