孙涛

199-6723-0102 | ASC_8384@foxmail.com | ASC8384.github.io | ASC8384

教育经历

北京航空航天大学（985，**双一流**）| 计算机技术，计算机学院 | *硕士研究生* 2023.09—至今

获夏令营优秀营员，师承智能信息处理研究所副所长的**李舟军**教授，研究方向为自然语言处理（NLP）与大语言模型（LLM），主要关注代码大模型。

湘潭大学（**双一流**）| 计算机科学与技术，计算机学院 | *工学学士* 2019.09—2023.06

GPA: 3.695/4.0(排名 2/88)，获推荐免试研究生资格，连续三年综合测评第一名，曾获**中国大学生程序设计竞赛银奖**、湘潭大学三好学生标兵（校级 Top 2‰）等奖项。

科研经历

UniCoder: 使用通用代码扩展 Code 大模型能力 | *一作*，*CCF-A* ACL 2024 主会

- **Tao Sun***, Linzheng Chai*, Jian Yang*, Yuwei Yin, Hongcheng Guo, Jiaheng Liu, Bing Wang, Liqun Yang, Zhoujun Li. UniCoder: Scaling Code Large Language Model via Universal Code.
- 提出了适用于大模型的通用代码框架，利用思维链明确任务与代码之间的转换过程，并构造相应数据集，增强了代码大模型的生成和理解能力，在多个 Code 任务上具有良好表现。

ND: 基于深度学习的轻量级分组密码差分区分器 | *一作*，*CCF-C* ICONIP 2022 Oral

- **Tao Sun**, Dongsu Shen, Saiqin Long, Qingyong Deng, Shiguo Wang. Neural Distinguishers on TinyJAMBU-128 and GIFT-64. Neural Information Processing: 29th International Conference, ICONIP 2022.
- 针对差分分析，利用 MLP 和 LSTM，对 TinyJAMBU−128 和 GIFT−64 两种密码构造了神经单差分区分器和神经多面体差分区分器，性能达到最优，为差分分析提供了新的攻击思路。

RepoFix: 针对模型提问、定位和修复能力的代码仓库级别能力评估 | *一作* 论文在投

- RepoFix: Repository-Level Code Evaluation: From Issue Detection to Bug Localization and Fixes.
- 在代码仓库级别的现实场景中，首次提出了以 Bug 代码为核心的综合修复流程评估。该评估流程涉及模型对完整含有 Bug 代码仓库的处理，考察其发现有效问题、精准定位相关代码位置并最终修复 Bug 的能力。修复效果通过单元测试进行验证，确保修复成功。

McEval: 超大规模的多语言代码评估 | *负责数据集的构造与模型评测* 论文在投

- Linzheng Chai*, Shukai Liu*, Jian Yang*, Yuwei Yin, Ke Jin, Jiaheng Liu, **Tao Sun**, Ge Zhang, Changyu Ren, Hongcheng Guo, Zekun Wang, Boyang Wang, Xianjie Wu, Bing Wang, Tongliang Li, Liqun Yang, Sufeng Duan, Zhoujun Li. McEval: Massively Multilingual Code Evaluation. arXiv 2406.07436.
- McEval 覆盖了 40 种编程语言，包含 16K 个测试样本，用于全面评估代码模型在代码生成、解释和补全等任务上的多语言能力。在 McEval 上对 20 多个现有大模型进行了系统评估，揭示了开源模型与闭源模型之间的性能差距。

xCOT: 跨语言思维链推理的指令调优 | *负责模型训练* 论文在投

- Linzheng Chai, Jian Yang, **Tao Sun**, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, Zhoujun Li. xCOT: Cross-lingual Instruction Tuning for Cross-lingual Chain-of-Thought Reasoning. arXiv 2401.07037.
- 提出了跨语言指导微调框架，利用思维链，通过多语言指令训练数据和跨语言上下文少样本学习，成功将高资源语言的知识转移到低资源语言，实现了不同语言之间的语义对齐，为大型语言模型的多语言推理能力提供了有效的增强方法。

RoleAgent: 从剧本中构建、交互和测试角色扮演代理 | *负责测试集构建与模型评测* 论文在投

- Jiaheng Liu*, Zehao Ni*, Haoran Que*, **Tao Sun**, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, Zhaoxiang Zhang. RoleAgent: Building, Interacting, and Benchmarking High-quality Role-Playing Agents from Script.
- 提出了角色智能体框架，通过从原始戏剧或剧本生成高质量的智能体，利用分级记忆机制，实现记忆检索、缓存和遗忘功能，达成可信的人类行为模拟，同时引入了系统性评估基准，验证了该框架的有效性。

REALM: 基于大模型检索增强生成驱动的多模态电子健康档案分析增强 | *负责实体识别* 论文在投

- Yinghao Zhu*, Changyu Ren*, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, **Tao Sun**, Long He, Zhoujun Li, Xi Zhu, Chengwei Pan. REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. arXiv:2402.07016.

- 提出了 REALM 框架，使用 LLM 对长文本临床笔记进行编码，并使用 GRU 模型来编码电子健康档案（EHR）数据。通过 LLM 提示，有效得抽取了与任务相关的医学实体，并将其与专业标注的外部知识图谱中的实体进行匹配，设计了一个自适应多模态融合网络，整合抽取的知识与多模态 EHR 数据。

SVIPTR: 基于视觉置换提取器的快速高效场景文本识别 | 复现基线模型效果 论文在投

- Xianfu Cheng, Weixiao Zhou, Xiang Li, Jian Yang, Hang Zhang, **Tao Sun**, Wei Zhang, Yuying Mai, Tongliang Li, Xiaoming Chen and Zhoujun Li. SVIPTR: Fast and Efficient Scene Text Recognition with Vision Permutable Extractor.
- 提出了一种基于自注意力机制的单视觉模型，设计了专门针对文本解析优化的特征提取模块，使其能够适应并准确识别任意长度的文本图像输入，具有跨语言识别的通用性。

XFormParser: 具有多模态和多语言知识的半结构化表单解析器 | 负责模型训练 论文在投

- Xianfu Cheng*, Hang Zhang*, Jian Yang, Xiang Li, Weixiao Zhou, Kui Wu, Fei Liu, Wei Zhang, **Tao Sun**, Tongliang Li and Zhoujun Li. XFormParser: Semi-structured Form Parser with multimodal and multilingual knowledge. arXiv: 2405.17336.
- 基于一个全面的预训练语言模型，并创新性地将语义实体识别（SER）和关系提取（RE）整合到一个统一框架中。还开发了一个名为 InDFormBench 的开创性基准数据集，专门针对多语言表单在各种工业场景中的解析需求。

实习经历

北京三快科技有限公司（美团） | 大模型算法实习生 2024.03—至今

- 参与**美团自研大模型**的研发与测试工作，主要负责提升和评估代码专家模型在代码仓库级别的补全和修复能力，探索大模型在软件工程下的应用场景，涉及模型的长文本能力和代码规划与测试能力。

深圳智能思创科技有限公司 |*NLP* 算法实习生 2022.10—2023.05

- 独立负责命名实体识别业务的设计、开发、测试和部署。
- 通过爬虫等技术，获取互联网上的公开金融领域的调研报告等内容，对原始文本数据进行清洗和分割，并利用进行 Label Studio 命名实体标注。研究了词汇信息增强技术在命名实体识别中的应用，结合了 LEBERT 模型，设计了 BiLSTM、CRF 等下游网络，引入了 FGM 对抗训练，与基线模型相比，识别精度提升了 10.09%。将 PyTorch 训练模型转换为 ONNX 格式，在公司相关平台部署模型推理。

项目经历

Windrecorder | **捕风记录仪** | 维护者，贡献者，*Github **2581 Stars*** 2023.11—至今

- 一款可以持续记录屏幕画面、通过关键词搜索等方式随时回溯过去与记忆的工具。
- 增加了自定义录屏选项、自定义视频压缩编码格式、多屏幕记录等功能。

个人项目 | 个人开发 2018.02-至今

- **myRime**，利用 rime 引擎自定义小鹤双拼输入法，目前 Github 获 **102 Stars**。
- **打卡提醒机器人**，通过 QQ、短信等方式对健康打卡、青年大学习等的打卡情况进行监控提醒，在 EL-ADMIN 的基础上设计了后台可视化系统，使本科学院的健康打卡率从 74% 提升到了 97%。
- **RepoTime**，设计了一款 Visual Studio Code 插件，用于自动记录使用 VSCode 时的编码时长、编程语言等信息。

荣誉奖项

奖学金

- 北京航空航天大学新生学业奖学金二等奖 2023
- 京东奖学金 2022
- 湘潭大学伟人之托奖学金（校级 Top 2‰） 2021
- 湘潭大学甲等奖学金（校级 Top 7%） 2020 & 2021 & 2022

竞赛

- 中国大学生程序设计竞赛（CCPC）全国邀请赛**银奖** 2021.06
- 国际大学生程序设计竞赛（ICPC）上海站铜奖 2021.11
- 中国大学生程序设计竞赛（CCPC）桂林站铜奖 2021.11
- 全国大学生数学建模竞赛省一等奖 2021.10
- 蓝桥杯国家二等奖 2022.06
- 中国大学生计算机设计大赛省一等奖 2022.06