沙磊

北京市海淀区学院路37号北京航空航天大学新主楼B1030
(+86) 13521038522 ⋄ shalei@buaa.edu.cn

教育经历

北京大学 *2013.09 - 2018.07*
直博，自然语言处理，导师：穗志方
研究方向：基于语言模型的事件抽取及生成

工作经历

北京航空航天大学 2023.02 至今
教授 人工智能研究院

北京航空航天大学 2022.04 - 2023.02
副教授 人工智能研究院

- 研究可解释自然语言模型，可控制自然语言生成，面向科学人工智能，大规模信息抽取。获2022海外优青资助。

牛津大学（英国） 2020.03-2022.03
副研究员 计算机系智能系统组

- 用于可控文本生成，解释模型预测结果和机器学习模型推理的可解释，健壮且可靠的深度学习模型的研究
- 研究用于个性化对话生成的可解释模型。

苹果公司（美国） 2018.08 -2020.4
资深 *NLP* 研究科学家 *Siri Understanding* 组

- 提升对话领域分类性能
- 基于分类的聊天对话：我们提出了一教师-学生网络将正则表达式形式的规则知识（教师模型）提炼到神经分类器（学生模型）中。该方法可以为神经模型提供许多关于规则的提示，从而可以提高性能。
- 关于聊天对话生成的研究。我正在指导我的实习生完成关于兴趣追踪聊天对话生成的研究。我们准备在端到端对话模型中应用队列架构来跟踪从每个用户的话语中提取的兴趣词。

微软亚洲研究院 2017.07 - 2018.02
Intern 系统组

任务式对话系统的研究。我们提出了一种主动意图切换模型，可以在对话转换期间检测和切换意图。我们的模型使用混合切换方法来决定是否应该更改当前意图以及应该切换到哪个目标意图。

微软亚洲研究院 2014.11 - 2015.06
Intern 知识挖掘组

做关于事件抽取方面的研究。阅读事件抽取相关的文献，实现自己的想法，并且完成论文。

主持项目

国家自然科学优秀青年基金（海外）项目 2023.03 - 2026.03
项目题目：可解释自然语言模型（*100*万） 主持人：沙磊

研究内容：申请人的研究工作着眼于如何使得深度模型“白盒化”，即令人类可以完全了解深度模型内部的运行机制，从而可以使得模型的运行更加的健壮，结果更加可控。以便于放心地应用在一些生命攸关的场景中，如智能医疗，智能法律等等。

国家自然科学青年基金项目 2024.01 - 2027.01

项目题目：面向约束的可控文本生成（*30*万） 主持人：沙磊

研究内容：申请人的工作着眼于面向约束的可控文本生成,从而使得生成的文本可以符合指定的长度、风格、内容要求等多种约束条件,可以应用在对输出形式有强约束的场景中,如自动文案生成、智能法律助理、文本内容审核等方面。

北航概念验证项目 2023.06 - 2025.06

项目题目：基于多组学生物信息*+AI*处理的泛癌早筛技术（*50*万） 主持人：沙磊

研究内容：本项目开发“第二代”泛癌早期筛查、微小残留病灶的一体化解决方案。项目合作者均来自国内外知名高校院所和科研机构，是中国首批通过cfRNA组学（非DNA）+ AI人工智能进行癌症早筛的团队。核心技术包含三个高级自研cfRNA及cfDNA生物信息学和人工智能分析算法。在六批独立计算实验、五种癌症，共计约400例样品中，达成了约90%灵敏性、约100%特异性的跨代际优异性能。技术性能约10倍优于传统cfDNA方法，单样品估算成本仅为cfDNA方法的1/10。在癌症筛查领域突破了以Grail为首的欧美癌症早筛企业对该行业的垄断。

北航敢为面上项目 2024.09 - 2026.06

项目题目：领域大模型的高效构建与动态更新方法研究（*50*万） 主持人：沙磊

研究内容：本项目聚焦于推动大模型技术的普适化与高性能应用，致力于构建面向行业实际需求的轻量高效可控大模型解决方案。针对当前大模型存在的算力瓶颈、知识错误和部署困难等关键问题，项目提出了三项核心技术路径：轻量化构建领域大模型、高效长序列量化调优框架、以及知识错误的系统性编辑与修正机制。相关技术已在多个真实任务场景中开展验证。该技术体系填补了国内大模型可控编辑与精调部署能力的空白，有望打破当前国际通用大模型主导的技术壁垒，为我国自主构建高可靠AI基础设施提供坚实支撑。

小米揭榜挂帅项目 2024.07 - 2025.07

项目题目：基于大模型与知识库融合的幻觉缓解技术研究（*20*万） 主持人：沙磊

研究内容：本研究旨在通过融合知识图谱来缓解大模型幻觉问题。在预训练阶段混合知识图谱信息，微调阶段引入知识对齐监督，推理阶段修正模型内部知识表征。研究将探索优化目标、算法和评估指标，建立幻觉缓解技术体系。

郑州市揭榜挂帅重点研发专项项目 2025.02 - 2027.02

项目题目：*AI*驱动的医疗健康应用场景分析模型的研究（*300*万） 主持人：沙磊

研究内容：本项目面向医疗服务系统中智能化语音识别、语义解析和健康管理的需求与挑战，针对医疗场景语音识别优化、领域语义解析与知识对齐以及健康管理智能化提升这三个关键问题展开研究。首先，通过语音识别阶段专业词库融合与多方言适配技术，研究如何在复杂医疗环境中提高语音识别的精准性与鲁棒性；然后，基于医疗领域大模型的语义解析与知识对齐优化框架，探索如何通过上下文理解与领域知识引入实现任务意图的准确解析与知识融合；最后，针对健康管理服务中个性化不足的问题，我们将研究多源数据整合与健康风险预测模型，建立智能化的健康管理与干预体系，从而应对医疗系统在智能化应用中面临的挑战，全面提升医疗服务的效率与质量。

研究兴趣

自然语言理解 当前的NLP模型倾向于使用越来越复杂的深度神经网络来记忆任务的特定模式。但是，如果没有真正理解自然语言，基于深度理解和基于推理的任务就无法真正解决。因此，自然语言理解方法（例如语义解析，神经-符号结合）是非常有价值的研究课题。

可解释大语言模型 大型语言模型(LLM)如GPT-3模型缺乏解释性和可控性- 难以理解它们的内部推理并指导其文本生成。本项目旨在使大型语言模型更具解释性和可操纵性。我们将专注于设计使解释模型预测和指导文本生成向期望属性的模型架构和训练技术。潜在的方法包括从训练数据中检索相关段落,突出显示重要的输入标记,生成自然语言解释,并公开像情感和风格。但这些方法无法实际帮助大模

型做出更好的判断。我们会研究更加重要的问题，比如为什么模型在某种输入案例之下会失败，失败的原因是什么，如何debug模型让其能做对。这样建立了透明和可控制的LLM将可以更安全地在现实世界应用中部署。进而会加速LLM在生物医疗和法律等人命关天的领域的采用。

可控文本生成 文本生成通常与神经机器翻译，神经摘要生成，表格到文本生成等结合在一起。在某些特定应用程序中，我们通常希望根据一些外部信息更改生成的文本。例如，将情绪从正面变为负面，或将关键信息更改为其他内容。通常，很难获得这种任务的平行语料训练数据。因此，用小样本学习的方法生成可控文本是非常值得研究的问题。

获得奖项

2020 EMNLP 2020 杰出审稿人

2018 北京大学优秀毕业生

2017 李惠荣奖学金

2016 北京大学三好生

2016 五四奖学金

2015 微软亚洲研究院明日之星

2015 北京大学博士奖学金

已发表论文(*代表通讯作者)

[1] DiffusionAttacker: Diffusion-Driven Prompt Manipulation for LLM Jailbreak
Hao Wang, Hao Li, Junda Zhu, Xinyuan Wang, Chengwei Pan, Minlie Huang, **Lei Sha***. In Proceedings of EMNLP 2025 Conference, 2025. (EMNLP 2025).

[2] Reasoning-to-Defend: Safety-Aware Reasoning Can Defend Large Language Models from Jailbreaking
Junda Zhu, Lingyong Yan, Shuaiqiang Wang, Dawei Yin, **Lei Sha***. In Proceedings of EMNLP 2025 Conference, 2025.

[3] Layer-Aware Representation Filtering: Purifying Finetuning Data to Preserve LLM Safety Alignment Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, Rui Li, Jing Shao, **Lei Sha***. In Proceedings of EMNLP 2025 Conference, 2025. (EMNLP 2025).

[4] Towards Harmonized Uncertainty Estimation for Large Language Models.
Rui Li, Jing Long, Muge Qi, Heming Xia, **Lei Sha**, Peiyi Wang, Zhifang Sui. In Proceedings of ACL 2025 Conference, 2025. (ACL 2025)

[5] How Far are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation.
Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, **Lei Sha**, Wenjie Li, Zhifang Sui. (ACL 2025 finding)

[6] LLMs know their vulnerabilities: Uncover Safety Gaps through Natural Distribution Shifts.
Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, **Lei Sha**, Junchi Yan, Lizhuang Ma, Jing Shao. (ACL 2025)[Outstanding Paper]

[7] Be a Multitude to Itself: A Prompt Evolution Framework for Red Teaming.
Rui Li, Peiyi Wang, Jingyuan Ma, Di Zhang, Zhifang Sui, **Lei Sha**. In Proceedings of EMNLP 2024 Conference, 2024. (EMNLP 2024 finding).

[8] ATM: Adversarial Tuning Multi-agent System Makes a Robust Retrieval-Augmented Generator, Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, **Lei Sha***. (EMNLP 2024)

[9] ShieldLM: Empowering llms as aligned, customizable and explainable safety detectors,
Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, **Lei Sha**, Zhifang Sui, Hongning Wang, Minlie Huang. (EMNLP 2024 finding)

[10] ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings,
Hao Wang, Hao Li, Minlie Huang, **Lei Sha***. (EMNLP 2024)

[11] Correcting Flaws in Common Disentanglement Metrics.
Louis Mahon, **Lei Sha**, Thomas Lukasiewicz. In Transactions of Machine Learning Research, 2024. (TMLR 2024).

[12] Text Attribute Control via Closed-Loop Disentanglement.
Lei Sha, and Thomas Lukasiewicz. Transactions of the Association for Computational Linguistics (TACL, CCF-B, IF:17.59).

[13] A Stable Fast and Fully Automatic Learning Algorithm for Predictive Coding Networks.
Tommaso Salvatori, Yuhang Song, Yordan Yordanov, Beren Millidge, **Lei Sha**, Cornelius Emde, Zhenghua Xu, Rafal Bogacz, Thomas Lukasiewicz. The Twelfth International Conference on Learning Representations (ICLR 2024)

[14] Harnessing the Plug-and-Play Controller by Prompting.
Hao Wang, **Lei Sha***. In Proceedings of EMNLP 2023 Conference, 2023. (EMNLP 2023).

[15] Latent Rationalizing Prediction by Adversarial Information Calibration.
Lei Sha, Oana-Maria Camburu and Thomas Lukasiewicz. Artificial Intelligence Journal (AI, CCF-A, IF:16.4).

[16] Controlling Text Edition by Changing Answers of Specific Questions
Lei Sha, Patrick Hohenecker and Thomas Lukasiewicz. In Findings of ACL 2021 Conference, 2021. (ACL finding 2021).

[17] Learning from the Best: Rationalizing Predictions by Adversarial Information Calibration
Lei Sha, Oana-Maria Camburu and Thomas Lukasiewicz. In Proceedings of AAAI 2021 Conference, 2021. (AAAI 2021).

[18] Multi-type Disentanglement without Adversarial Training
Lei Sha and Thomas Lukasiewicz. In Proceedings of AAAI 2021 Conference, 2021. (AAAI 2021).

[19] Gradient-guided Unsupervised Lexically Constrained Text Generation.
Lei Sha. In Proceedings of EMNLP 2020 Conference, 2020. (EMNLP 2020).

[20] Estimating Minimum Operation Steps via Memory-based Recurrent Calculation Network.
Lei Sha, Chen Shi, Qi Chen, Lintao Zhang and Houfeng Wang. In Proceedings of IJCNN 2020 Conference, 2020. (IJCNN 2020).

[21] Order-Planning Neural Text Generation From Structured Data.
Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of AAAI 2018 Conference, 2018. (AAAI 2018).

[22] Joint Extracting Event Trigger and Arguments Using dependency bridge Recurrent Tensor Network.
Lei Sha, Feng Qian, Baobao Chang and Zhifang Sui. In Proceedings of AAAI 2018 Conference, 2018. (AAAI 2018).

[23] A Multi-View Fusion Neural Network for Answer Selection.
Lei Sha, Xiaodong Zhang, Feng Qian, Baobao Chang and Zhifang Sui. In Proceedings of AAAI 2018 Conference, 2018. (AAAI 2018).

[24] Will Repeated Reading Benefit Natural Language Understanding?
Lei Sha and Zhifang Sui. In Proceedings of NLPCC 2017 Conference, 2017. (NLPCC 2017).

[25] Reading and Thinking: Re-read LSTM Unit for Textual Entailment Recognition.
Lei Sha, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of Coling 2016 Conference, 2016. (Coling 2016).

[26] Capturing Argument Connection for Chinese Semantic Role Labeling.
Lei Sha, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of EMNLP 2016 Conference, 2016. (EMNLP 2016).

[27] RBPB: Regularization-Based Pattern Balancing Method for Event Extraction.
Lei Sha, Jing Liu, Chin-Yew Lin, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of ACL 2016 Conference, 2016. (ACL 2016).

[28] Joint Learning Templates and Slots for Event Schema Induction.
Lei Sha, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of NAACL 2016 Conference, 2016. (NAACL 2016).

[29] Recognizing Textual Entailment Using Probabilistic Inference.
Lei Sha, Sujian Li, Tingsong Jiang, Baobao Chang, and Zhifang Sui. In Proceedings of the EMNLP 2015 Conference, 2015. (EMNLP 2015).

[30] Recognizing textual entailment via multi-task knowledge assisted LSTM.
Lei Sha, Sujian Li, Baobao Chang, Zhifang Sui. CCF International Conference on Natural Language Processing and Chinese Computing, 2017.

[31] RecInDial: A Unified Framework for Conversational Recommendation with Pretrained Language Models.
Lingzhi Wang, Huang Hu, **Lei Sha**, Can Xu, Daxin Jiang, Kam-Fai Wong. (AACL 2022)

[32] Small Changes Make Big Differences: Improving Multi-turn Response Selection in Dialogue Systems via Fine-Grained Contrastive Learning.
Yuntao Li, Can Xu, Huang Hu, **Lei Sha**, Yan Zhang, Daxin Jiang. (INTERSPEECH 2022)

[33] Associative Memory via Predictive Coding.
Tommaso Salvatori, Yuhang Song, Yujian Hong, Simon Frieder, **Lei Sha**, Zhenghua Xu, Rafal Bogacz, Thomas Lukasiewicz. Advances in Neural Information Processing Systems, 2021. (NeurIPS 2021).

[34] Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning Chen Shi, Qi Chen, **Lei Sha**, Sujian Li, Xu Sun, Houfeng WANG and Lintao Zhang In Proceedings of EMNLP 2018 Conference, 2018. (EMNLP 2018).

[35] Table-to-text Generation by Structure-aware Seq2seq learning.
Tianyu Liu, Kexiang Wang, **Lei Sha**, Zhifang Sui, Baobao Chang. In Proceedings of AAAI 2018 Conference, 2018. (AAAI 2018).

[36] We know what you will ask: A dialogue system for multi-intent switch and prediction.
Chen Shi, Qi Chen, **Lei Sha**, Hui Xue, Sujian Li, Lintao Zhang, Houfeng Wang. CCF International Conference on Natural Language Processing and Chinese Computing, 2019.

[37] Topic Medical Concept Embedding: Multi-Sense Representation Learning For Medical Concept.
Feng Qian, Chengyue Gong, Luchen Liu, **Lei Sha**, Ming Zhang. In Proceedings of BIBM 2017 Conference, 2017. (BIBM 2017).

[38] Syntax Aware LSTM Model for Chinese Semantic Role Labeling.
Feng Qian, **Lei Sha**, Baobao Chang, Lu-chen Liu, Ming Zhang. In Proceedings of EMNLP 2017 Conference, 2017. (EMNLP 2017).

[39] A Progressive Learning Approach to Chinese SRL Using Heterogeneous Data.
Qiaolin Xia, **Lei Sha**, Baobao Chang, Zhifang Sui. In Proceedings of ACL 2017 Conference, 2017. (ACL 2017).

[40] Attentive Interactive Neural Networks for Answer Selection in Community Question Answering.
Xiaodong Zhang, **Lei Sha**, Sujian Li, Houfeng Wang. In Proceedings of AAAI 2017 Conference, 2017. (AAAI 2017).

[41] Encoding Temporal Information for Time-Aware Link Prediction.
Tingsong Jiang, Tianyu Liu, Tao Ge, **Lei Sha**, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of EMNLP 2016 Conference, 2016. (EMNLP 2016).

[42] Towards Time-Aware Knowledge Graph Completion.
Tingsong Jiang, Tianyu Liu, Tao Ge, **Lei Sha**, Sujian Li, Baobao Chang and Zhifang Sui. In Proceedings of Coling 2016 Conference, 2016. (Coling 2016).

[43] Multi-label Text Categorization with Joint Learning Predictions-as-Features Method.
Li Li, Houfeng Wang, **Lei Sha**, Xu Sun, Baobao Chang, Shi Zhao. In Proceedings of EMNLP 2015 Conference, 2015. (EMNLP 2015)

[44] Event Schema Induction Based on Relational Co-occurrence over Multiple Documents.
Tingsong Jiang, **Lei Sha** and Zhifang Sui. In Proceedings of Natural Language Processing and Chinese Computing, 2014. (NLPCC 2014).

已发表专利

[1] 沙磊，穗志方；一种新闻事件生成式问答数据集的生成方法；公开日：2019.08.09；申请号：CN201810057805.8；公开号：CN110110050A

[2] 邓司伟，沙磊，金泳成，周天尧；基于统计学与随机森林的游离RNA肝癌早筛方法及系统；申请号：CN202310775133.5；公开号：CN116825177A；公开日：2023.09.29

[3] 沙磊，邓司伟，金泳成，周天尧；基于大规模深度模型的cfRNA泛癌早筛方法及系统；申请号：CN202310509731.8；公开号：CN116580833A；公开日：2023.08.11

[4] 邓司伟，金泳成，周天尧，沙磊；一种发现胰腺导管腺癌相关游离RNA生物标志物的方法；申请号：CN202310513309.X；公开号：CN116463420A；公开日：2023.07.21

[5] 张海楠，钱成，沙磊，王子玮，郑志明；一种利用隐藏状态过滤的防隐私攻击方法及装置；申请号：CN202411512708.5；公开号：CN119377773A；公开日：2025.01.28

[6] 邓司伟，沙磊，金泳成，周天尧；基于机器学习的游离RNA阿兹海默早筛方法及系统；申请号：CN202310775099.1；公开号：CN116959710A；公开日：2023.10.27

[7] 刘潜璞，邓司伟，沙磊；一种基于预训练语言模型的泛癌早筛方法及系统；申请号：202311529611.0；公开号：CN117912668A；公开日：2024.04.19

[8] 沙磊；一种基于统一阈值的多类校准方法; 申请号：CN202411847678.3；公开号：CN120030447A；公开日：2025.05.23

[9] 刘剑，沙磊；基于多尺度细粒度反馈学习的大模型上下文问答回复质量优化方法及系统，申请号：2025111078722

[10] 陈冠桦，沙磊；一种多视角知识密集型检索增强生成方法及系统，申请号：2025111078281