Junyu Luo

luoshuiti@outlook.com • soap117.github.io/junyu.github.io/

Education & Awards

Academic Qualifications

- **Pennsylvania State University** — **Ph.D.**
  Information Sciences and Technology, GPA: 3.91/4.00 — *2020.08–2024*
- **Sichuan University** — **Bachelor**
  Computer Science, GPA: 3.86/4.00, Major GPA: 3.924/4.00 — *2020.06*

Honors & Awards

- **The Award of Excellence, MSRA Internship Program** — **2020**
- **National Scholarship** — **(THREE TIMES) 2015-2018**
- **The First Prize Scholarship of Sichuan University** — **(THREE TIMES) 2015-2018**
- **The First Prize in Sichuan Province Lanqiao Programming Contest** — **2017**
- **The First Prize in Sichuan University Mathematics Competition** — **2016**

Skills

- Experience in processing different kinds of data (Image, Text, Web Data, Audio).
- Experience in all kinds of deep learning frameworks, including Transformers, LLMs, diffusion models, GAN, graph neural networks, information retrieval frameworks, and object detection frameworks.
- Experience in Natural Language Processing and Computer Vision related topics.
- Experience in building web pages and mobile applications for machine learning models.
- Master in Python (PyTorch, TensorFlow, Keras) and familiar with C#, C++, Java, and JavaScript.

Research and Work Experience

- **Research Assistant on Machine Learning for Healthcare** — **Dr. Fenglong Ma**
  Pennstate University IST, Pennsylvanian, USA — *Feb 2020–Now*
  - **Multi-modality Pre-training of EHR Data**
    Paper: Hierarchical Pretraining on Multimodal Electronic Health Records.
    Summary: *Developing a novel, multi-modal, and unified pretraining framework called MEDHMP for multi-modality health data pre-training.*
    Used Skills: Multi-modality, Pre-training, Pre-trained Language Model, Self-supervised Learning, Representation Learning, EHR, ICD Codes
  - **Automatic ICD Coding based on Diagnosis Text**
    Paper: Fusion: Towards Automated ICD Coding via Feature Compression.
    Summary: *Using information compression to reduce the clinical note noise and improve the speed of automatic ICD coding.*
    Used Skills: Transformers, NLP, ICD Coding
    Paper: CoRelation: Boosting Automatic ICD Coding Through Contextualized Code Relation Learning.
    Summary: *Improving ICD coding performance through modeling contextualized code relations through graph network.*
    Used Skills: Bi-LSTM, Graph Attention Network, Synonym Fusion, ICD Coding
  - **Medical Text Simplification**
    Paper: Benchmarking Automated Clinical Language Simplification: Dataset, Algorithm, and Evaluation.
    Summary: *Designing a controllable medical term simplification pipeline for using external medical dictionary knowledge.*
    Used Skills: Neural Network Pipeline, NLP, Question Answering, Constrained Generation, External Knowledge Injection
  - **Electric Health Record Mining**
    Paper: HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records.
    Summary: *Using two-level transformers to model the complex EHR code sequential data to predict future diseases.*
    Used Skills: Transformers, Time-aware Attention, EHR, ICD Codes, Disease Prediction
- **Research Intern on Natural Language Processing** — **Dr. Danica Xiao**
  Relativity, USA — *June 2023–Aug 2023*
  - **Designing Algorithm for Preventing Hallucination for Large Language Models (LLMs).**
    Paper: Zero-Resource Hallucination Prevention for Large Language Models
    Summary: *Using prompt engineering to perform self-evaluation under the zero-resource setting to test the understanding of LLMs to the instructions.*
    Used Skills: Neural Network Pipeline, NLP, Large Language Models, Constrained Beam Search, Prompt Engineering

- **Research Intern on Machine Learning for Clinical Data** — **Dr. Cheng Qian**
  IQVIA, USA — *May 2022–Dec 2022*
  - **Designing Clinical Trial Retrieval Algorithm Based on Trial Protocols.**
    Paper: Clinical Trial Retrieval via Multi-grained Group-based Similarity Learning
    Summary: *Designing a hierarchical matching model for trial protocols with novel group-based training loss and 2D word matching.*
    Used Skills: NLP, Transformers, Convolutional Network, Group Loss, Hierarchical Attention, Information Retrieval
  - **Designing Personalized Drug Risk Prediction Model.**
    Paper: pADR: Towards Personalized Adverse Drug Reaction Prediction by Modeling Multi-sourced Data.
    Summary: *Incorporating the patient's EHR modality with the drug molecular level information to predict the potential adverse reaction.*
    Used Skills: Pre-trained Language Models, Transformers, Multi-modality, SMILES Chemical Presentation, EHR, ICD codes, Adverse Event Prediction
- **Research Intern on Knowledge Computing** — **Dr. Jinpeng Wang**
  Microsoft Research Lab - Asia (MSRA), Beijing, China — *Mar 2019–Jan 2020*
  - **Automatic Pattern Recognition from Power Point Design.**
    Summary: *Transforming the pattern matching into a sequential matching problem to discover potential design patterns.*
    Used Skills: Sequential Matching
  - **Object Detection for Special Chart Images.**
    Paper: ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework
    Summary: *Designing a high precision point-based object detection model for chart objects.*
    Used Skills: Computer Vision, Object Detection, Point Detection
- **Research Intern on Natural Language Processing** — **Dr. Min Yang**
  Shenzhen Institutes of Advanced Technology(SIAT), Shenzhen, China — *Sep 2017–Jul 2018*
  - **Developed methods to generate semantic embedding for long sentences and cross-model searching**
    Paper: Cross-modal Image-Text Retrieval with Multitask Learning.
    Summary: *Using back-encoding to ensure the cross-modality relation between learned text and image embeddings.*
    Used Skills: Cross-modality, AutoEncoder, Representation Learning, Information Retrieval

Publications (Selected)

- **Junyu Luo**, Cheng Qian, Xiaochen Wang, Lucas Glass, and Fenglong Ma. 2023. *pADR: Towards Personalized Adverse Drug Reaction Prediction by Modeling Multi-sourced Data.* In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM 23), October 21–25, 2023, Birmingham, United Kingdom.
- **Junyu Luo**, Zhi Qiao, Lucas Glass, Cao Xiao, and Fenglong Ma. 2023. *Clini calRisk: A New Therapy-related Clinical Trial Dataset for Predicting Trial Status and Failure Reasons.* In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM 23), October 21–25, 2023, Birmingham, United Kingdom.
- **Junyu Luo**, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui and Fenglong Ma. *Benchmarking Automated Clinical Language Simplification: Dataset, Algorithm, and Evaluation*. Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022), OCTOBER 12-17, 2022, GYEONGJU, REPUBLIC OF KOREA.
- **Junyu Luo**, Cao Xiao, Lucas Glass, Jimeng Sun and Fenglong Ma. *Fusion: Towards Automated ICD Coding via Feature Compression*. Findings of the 59th Annual Meetingof the Association for Computational Linguistics (Findings of ACL), 2021.
- **Junyu Luo**, Zekun Li, Jinpeng Wang, Chin-Yew Lin: *ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework*. Proceedings of the 2021 Winter Conference on Applications of Computer Vision (WACV), 2021.
- **Junyu Luo**, Muchao Ye, Cao Xiao, Fenglong Ma. *HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records*. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2020.
- **Junyu Luo**, Ying Shen, Xiang Ao, Zhou Zhao, Min Yang. *Cross-modal Image-Text Retrieval with Multitask Learning*. Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM), 2019.
- **Junyu Luo**, Yong Xu, Chenwei Tang, Jiancheng Lv. *Learning Inverse Mapping by AutoEncoder Based Generative Adversarial Nets*. ICONIP (2) 2017: 207-216.

Submissions

- *Zero-Resource Hallucination Prevention for Large Language Models*. AAAI 2024.
- *CoRelation: Boosting Automatic ICD Coding Through Contextualized Code Relation Learning*. EMNLP 2023.
- *Hierarchical Pretraining on Multimodal Electronic Health Records*. EMNLP 2023.
- *Clinical Trial Retrieval via Multi-grained Group-based Similarity Learning*. SDM 2024.