Xin Su

CONTACT INFORMATION

University of Arizona
School of Information
1103 E. 2nd St.,
Tucson, AZ 85721

Phone: (503) 383-4379
Email: xinsu@arizona.edu
Website: xinsu.name

RESEARCH INTERESTS

Natural language processing (question answering, retrieval-augmented generation, information extraction), machine learning, artificial intelligence

EDUCATION

University of Arizona, Tucson, AZ — 2020-present
Ph.D. in Information Science, Advisor: Steven Bethard

Loyola University Chicago, Chicago, IL — 2020
M.S. in Computer Science, Advisor: Dmitriy Dligach
M.S. in Supply Chain Management
Business Data Analytics Certificate

Western Oregon University, Monmouth, OR — 2016
B.S. in Business

RESEARCH & INDUSTRY EXPERIENCE

Intel Labs, USA — May - Aug. 2022, May 2023 - present
Artificial Intelligence Research Intern

- Large language model reasoning project: developed a method that can effectively integrate external knowledge sources to significantly improve the reasoning abilities of both open-source (e.g., LLaMA2) and proprietary (e.g., GPT-4) large language models in knowledge-intensive tasks.
- Temporal reasoning project: extracted temporal graphs from the text and combined them with language model representations. Developed a question-answering system and achieved state-of-the-art results on a benchmark dataset (4.6 absolute F1 improvements over the previous state-of-the-art system).

Computational Language Understanding Lab, University of Arizona 2020-present
Graduate Research Associate

- Temporal information extraction project: built an end-to-end neural network model to extract and normalize the temporal information in the text.
- Semantic parsing project: built parsers to automatically translate natural language into structured query language (SQL) and Cypher (graph query language).
- Source-free domain adaptation project: adapted trained models to new domains using self-training, active learning, and data augmentation methods without access to the original training data.

Loyola Natural Language Processing Lab, Loyola University Chicago 2018-2020
Graduate Research Assistant

- Long document representation learning project: designed a framework to encode long documents using large-scale pre-trained language models (BERT and DistilBERT).
- Clinical semantic textual similarity: built an XLNet and BERT ensemble model to compute the similarity of snippets from electronic health records.
- Computable phenotyping project: worked with doctors and statisticians to develop neural network models (CNN and LSTM) and machine learning models (Logistic Regression, Support Vector Machines, Decision Trees, and XGBoost) to identify cohorts of patients that match a predefined set of criteria from electronic health records, such as patients with symptoms of ARDS.
- Language models domain-adaptive pre-training project: continued to pre-train the BERT model on 10 years' worth of electronic health records from the Loyola Medical Center.
- Clinical data preprocessing project: used clinical Text Analysis and Knowledge Extraction System (cTAKES) to build natural language processing pipelines (sentence boundary detection, tokenization, part-of-speech tagging, and entity recognition) to extract medical concepts from years of electronic health records from the Loyola Medical Center.

PUBLICATIONS

[1] **Xin Su**, Tiep Le, Steven Bethard, Phillip Howard, Semi-Structured Chain-of-Thought: Integrating Multiple Sources of Knowledge for Improved Language Model Reasoning. *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies 2024*

[2] **Xin Su**, Phillip Howard, Nagib Hakim, Steven Bethard. Fusing Temporal Graphs into Transformers for Time-Sensitive Question Answering. *Findings of the Association for Computational Linguistics: EMNLP 2023*

[3] **Xin Su**, Yiyun Zhao, Steven Bethard. A Comparison of Strategies for Source-Free Domain Adaptation. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). 2022.*

[4] **Xin Su**, Yiyun Zhao, Steven Bethard. The university of arizona at semeval-2021 task 10: Applying self-training, active learning and data augmentation to source-free domain adaptation. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval). 2021.*

[5] Egoitz Laparra, **Xin Su**, Yiyun Zhao, Ozlem Uzuner, Timothy Miller, Steven Bethard. Semeval-2021 task 10: Source-free domain adaptation for semantic processing. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval). 2021.*

[6] Sujay Kulshrestha, Dmitriy Dligach, **Xin Su**, Richard Gonzalez, Cara Joyce, Matthew M Churpek, Majid Afshar. Classification of Chest Injury Severity Using Clinical Documents. *American Medical Informatics Association (AMIA) Informatics Summit. 2021.* (peer-reviewed abstract)

[7] **Xin Su**, Timothy Miller, Xiyu Ding, Majid Afshar, Dmitriy Dligach. Classifying Long Clinical Documents with Pre-trained Transformers. *arXiv:2105.06752. 2021.*

[8] **Xin Su**, Timothy Miller, Majid Afshar, Dmitriy Dligach. Learning Hierarchical Transformer-based Representations of Clinical Notes. *American Medical Informatics Association (AMIA) Symposium. Chicago, IL. November 2020.* (peer-reviewed abstract)

[9] Anoop Mayampurath, Matthew Churpek, **Xin Su**, Sameep Shah, Elizabeth Munroe, Bhakti Patel, Dmitriy Dligach, Majid Afshar. External Validation of an Acute Respiratory Distress Syndrome Prediction Model Using Radiology Reports. *Critical Care Medicine. 2020.*

[10] **Xin Su**, Anoop Mayampurath, Matthew Churpek, Sameep Shah, Bhakti Patel, Dmitriy Dligach, Majid Afshar. External Validation of an Acute Respiratory Distress Syndrome Prediction Model Using Clinical Text. *American Thoracic Society (ATS) International Conference 2020. Philadelphia, Pennsylvania, May 2020.*

[11] **Xin Su**, Timothy Miller, Farig Sadeque, Majid Afshar, Dmitriy Dligach. Using Transformer-based Approaches for Measuring Semantic Similarity. *National NLP Clinical Challenges (N2C2) Workshop at AMIA 2019 Annual Symposium. Washington, D.C., November 2019.* (abstract)

TEACHING

ISTA 322: Data Engineering. University of Arizona, Teaching Associate, Spring 2023.

ISTA 457 / INFO 557 Neural Networks. University of Arizona, Teaching Associate, Fall 2022.

SERVICE

Association for Computational Linguistics Rolling Review (ARR). Reviewer. 2023-present.

Forty-first International Conference on Machine Learning (ICML). Reviewer. 2024.

The Twelfth International Conference on Learning Representations (ICLR). Reviewer. 2024.

Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS). Reviewer. 2023.

The Conference on Empirical Methods in Natural Language Processing (EMNLP). Program Committee. 2023.

The 61st Annual Meeting of the Association for Computational Linguistics (ACL). Program Committee. 2023.

The Conference on Empirical Methods in Natural Language Processing (EMNLP). Program Committee. 2022.

American Medical Informatics Association (AMIA) Symposium. Reviewer. 2022.

American Medical Informatics Association (AMIA) Symposium. Reviewer. 2020.

TECHNICAL SKILLS

Languages: Python, Java, C, C++, R, SQL, Bash, JavaScript, LaTeX

Tools: Pytorch, Tensorflow, Keras, Scikit-learn, cTAKES, Git, React

HONORS & AWARDS

Edsger W. Dijkstra High Achievement Award (CS), Loyola University Chicago	2020
Travel Award to attend N2C2 Workshop at AMIA, Loyola University Chicago	2019
Beta Gamma Sigma, Loyola University Chicago	2018
Merit Scholarship, Loyola University Chicago	2016
Dean's Honor Roll, Western Oregon University	2016