Dongxu Zhang

Machine Learning Scientist at Optum

☎ 413-210-5459 ✉ zhangdongxuu@gmail.com

I am a machine learning scientist at Optum, dedicated to enhancing software and systems in the healthcare domain. Before that, I was a researcher building LLM-based customer service agents at ASAPP. I obtained my Ph.D. in computer science at the University of Massachusetts Amherst, working with Professor Andrew McCallum. My current research focuses on LLMs and AI safety. I share broad interests and expertise in general machine learning, information extraction and knowledge representation.

Education

University of Massachusetts Amherst **2017/09 – 2023/02**
Ph.D. in Computer Science *Amherst, MA*

Beijing University of Posts and Telecommunications **2009/09 – 2017/03**
B.E. and M.S. in Information and Communication Engineering *Beijing, China*

Working Experience

Optum **2025/01 – Present**
Lead Machine Learning Scientist

ASAPP **2023/03 – 2024/10**
Research Scientist

- **GenAgent**: I was a member of the modeling team of the Generative Agent, one of ASAPP's core products. I contributed to the construction of function call description formatting. I improved and simplified the structured reasoning mechanism of the AI agent. I contributed to the end-to-end evaluation of the AI agent via user simulation.
- **Safety Guardrails**: I was a tech lead for AI safety. Our team built multiple safety guardrails for the ASAPP Generative Agent, including input safety, output safety, and data safety. I **finetuned** language models for PII redaction and hallucination detection with enhancement from **synthetic data** generation[1]. I conducted **prompt optimization** for input and output safety layers, including a prompt-based hallucination detector. Our safety layers reached over 95% recalls on detecting unsafe behaviors on internal customers' data, and over 80% F1 for hallucination detection.

Google AI **2019/06 - 2019/08**
Research Intern *Mentor: Sara McCarthy and Chris Welty*

- Leveraged box embeddings to tackle taxonomy alignment between anatomy and disease taxonomies in order to predict which body parts each disease has effects on.

Amazon **2018/06 - 2018/08**
Applied Scientist Intern *Mentor: Subhabrata Mukherjee and Luna Dong*

- Developed a relation *inference* method that aggregated the semi-structured context of each entity across the corpus for entity-entity relation prediction (improving the relation prediction MAP from 69.5% to 81.4%) [7].

Samsung Telecommunication R&D Center **2013/05 - 2013/10**
Algorithm Intern *Mentor: Xiaojie Yu*

- Trained language models for Samsung's **speech recognition** system. Developed a *fast parallel* k-means clustering module over acoustic features for their internal usage.

Academia Experience

UMass Amherst **2017/09 – 2023/02**
Research Assistant *Advisor: Andrew McCallum*

- **Geometric Embedding based Graph Representation**: Proposed a graph representation that embeds each vertex as a box region (a Cartesian product of intervals) and directed edges are captured by the relative containment of one box in another [5] (collaboration with IBM). A following work generalized box embeddings to capture cycles in the graph, making the model more robust and flexible to real-world graphs (increasing link prediction AUC from 93.8% to 97.9%) [2].
- **Information Extraction**: Created a **biomedical domain** relation extraction datasets *ChemDisGene*[3] (one of the largest existing RE dataset in the domain, including 80k abstracts and 18 relation types)(collaboration with CZI). Proposed a distantly supervised relation extraction datasets *StaRE* [4] (the first dataset to detect state-change relations)(collaboration with Bloomberg).

Rensselaer Polytechnic Institute 2016/04 - 2016/06

Visiting Scholar *Advisor: Heng Ji*

- **Low-resource NLP**: Automatic named entity annotation for low-resource languages (Turkish and Uzbek) with bilingual corpora [9](improving F1 of NER on Turkish from 48.3% to 57.6%).

Tsinghua University 2014/11 - 2016/03

Research Assistant *Advisor: Dong Wang*

- Developed an RNN-based relation extraction model.
- Developed an entity representation using multiple resources such as structured KB, semi-structured wiki and raw corpus.

Beijing University of Posts and Telecommunications 2013/02 - 2014/10

Research Assistant *Advisor: Weiran Xu*

- Developed a large-scale (Tegabyte-level corpus) slot-filling system for the Knowledge Base Acceleration track in NIST's Text Retrieval Conference (this system performed 1st among all participants).

Professional Services

Workshop Co-organizer: SciNLP 2021
Conference Reviewer: TKDE'18, VLDB'19, TKDD'19, ICLR'21-22, ACL'21, EMNLP'21-22, NeurIPS'22, ARR'21-22.
Mentorship: *Brian Dang* (UMass Honored Thesis), *Jui Shah* (accepted by LREC), *EunJeong Hwang* (accepted by ACL), *Bharath Narasimhan*, *Yuchen Zeng*.

Skills

Python libraries: pytorch, huggingface transformers, scikit-learn, pandas; **Other**: Linux, pyenv, github, latex

Selected Publications

* indicates equal contributions.

[1] **Dongxu Zhang**, Varun Gangal, Barrett Lattimer, and Yi Yang. "Enhancing Hallucination Detection through Perturbation-Based Synthetic Data Generation in System Responses". In: *Findings of the Association for Computational Linguistics **ACL 2024***. Aug. 2024.

[2] **Dongxu Zhang**, Michael Boratko, Cameron Musco, and Andrew McCallum. "Modeling Transitivity and Cyclicity in Directed Graphs via Binary Code Box Embeddings". In: *Advances in Neural Information Processing Systems (**NeurIPS**)* (2022).

[3] **Dongxu Zhang***, Sunil Mohan*, Michaela Torkar, and Andrew McCallum. "A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes". In: *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (**LREC**)*. 2022.

[4] Jui Shah*, **Dongxu Zhang***, Sam Brody, and Andrew McCallum. "Enhanced Distant Supervision with State-Change Information for Relation Extraction". In: *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (**LREC**)*. 2022.

[5] Michael Boratko*, **Dongxu Zhang***, Nicholas Monath, Luke Vilnis, Kenneth L. Clarkson, and Andrew McCallum. "Capacity and Bias of Learned Geometric Embeddings for Directed Graphs". In: *Advances in Neural Information Processing Systems (**NeurIPS**)* (2021), pp. 16423–16436.

[6] Shib Dasgupta*, Michael Boratko*, **Dongxu Zhang**, Luke Vilnis, Xiang Li, and Andrew McCallum. "Improving local identifiability in probabilistic box embeddings". In: *Advances in Neural Information Processing Systems (**NeurIPS**)* (2020), pp. 182–192.

[7] **Dongxu Zhang**, Subhabrata Mukherjee, Colin Lockard, Xin Luna Dong, and Andrew McCallum. "OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL**)*. 2019, pp. 762–772.

[8] Xiang Li*, Luke Vilnis*, **Dongxu Zhang**, Michael Boratko, and Andrew McCallum. "Smoothing the geometry of probabilistic box embeddings". In: *International Conference on Learning Representations (**ICLR**)*. 2018.

[9] **Dongxu Zhang**, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran Xu. "Bitext name tagging for cross-lingual entity annotation projection". In: *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (**COLING**)*. 2016, pp. 461–470.