Dongxu Zhang

Research Scientist at ASAPP, Inc.
413-210-5459 zhangdongxuu@gmail.com

I am a researcher building LLM-based customer service agents at ASAPP, Inc. Before that, I obtained my Ph.D. in computer science at the University of Massachusetts Amherst, working with Professor Andrew McCallum. My current research and production focuses on safety, evaluation, and prompt optimization of LLMs. I also share broad interests and expertise in general machine learning and knowledge discovery and representation in the era of LLMs.

Education

University of Massachusetts Amherst **2017 – 2022**
Ph.D. in Computer Science *Amherst, MA*

Beijing University of Posts and Telecommunications **2009 – 2016**
B.E. and M.S. in Information and Communication Engineering *Beijing, China*

Experience

ASAPP **Spring 2023 – Present**
Research Scientist

- **Safety Guardrails**: I am 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. The input safety layer blocks unsafe messages from end users such as injection, manipulation, and out of scope requests. And the output safety layer handles unsafe LLM outputs including hallucinations, information leaking, etc. Multiple approaches have been explored, including prompt-based methods, fine-tuning, evidence grounding, etc. For data safety, we built an internal PII redaction system to avoid customer data leakage.
- **Generative Agent**: I am a member of the modeling team for the development of the Generative Agent, one of ASAPP's core products. Beyond safety guardrails, I contributed mostly to the core capability such as evaluation, tool use, reflection and automatic prompt tuning, etc.

UMass Amherst **Fall 2018 – Winter 2022**
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 **Summer 2016**
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 **Winter 2014 - Spring 2016**
Research Assistant *Advisor: Dong Wang*

- **Information Extraction**: Proposed to use RNNs on the sentence-level RE task, and evaluated on a new dataset KBP37.
- **Semantic Representation** Worked on distributional entity representation using multiple resources such as ontology knowledge base, raw corpus, and Wikipedia pages.

Beijing University of Posts and Telecommunications **Fall 2013 - Fall 2014**
Research Assistant *Advisor: Weiran Xu*

- **Information Extraction**: Developed a large-scale slot filling system for the Knowledge Base Acceleration track in NIST's Text Retrieval Conference (this system performed the best among all participants).

Working Experience

Google AI **Summer 2019**

Research Intern *Mentor: Sara McCarthy and Chris Welty*

- **Taxonomy Alignment**: 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 **Summer 2018**

Applied Scientist Intern *Mentor: Subhabrata Mukherjee and Luna Dong*

- **Relation Extraction**: Proposed 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 **Summer 2013**

Algorithm Intern *Mentor: Xiaojie Yu*

- **Speech Recognition**: Trained a language model of Samsung's **speech recognition** system with colleagues. Developed a *fast parallel* k-means clustering module over acoustic features for their internal usage.

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), *Bharath Narasimhan*, *Yuchen Zeng*.

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.