Dongxu Zhang

PhD candidate at UMass Amherst
413-210-5459 dongxuzhang@cs.umass.edu

I am a fifth-year Ph.D. candidate at the University of Massachusetts Amherst, working with Professor Andrew McCallum. My current research focuses on using region-based embeddings for graph representational learning. I also share broad interests and expertise in information extraction, knowledge graph construction, recommender systems, and natural language processing using machine learning and deep learning methods.

Education

University of Massachusetts Amherst **2017 – Feb, 2023 (expected)**
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*

Research Experience

UMass Amherst **September 2018 – Present**
Research Assistant *Advisor: Andrew McCallum*

- **Representing Graphs with Box Embeddings (Dissertation Work in Progress)**:

  1. Proposed to model graphs by representing each node as a box region (a Cartesian product of intervals) where directed edges are captured by the relative containment of one box in another [5].

  2. Theoretically proved that box embedding can model any directed acyclic graphs [5].

  3. Showed promising empirical results over the massive amount of graphs distributions, where box embeddings allow for relationships involving longer paths (e.g., transitivity) to be modeled easily [5].

  4. Generalized box embeddings to capture cycles in the graph, making the model more robust and flexible to real-world graphs [1].
- **Box Embeddings**: Proposed box embeddings to represent hierarchically structured data. In this line of research, we focused on improving the training procedure by smoothing/randomizing the landscape of boxes, to avoid bad local minimal. I mostly contributed to methodology discussions and applied our method to a market basket dataset [6, 10].
- **Relation Extraction**: Created two distantly supervised relation extraction (RE) datasets.

  1. *ChemDisGene* is a collection of biomedical research abstracts annotated with mentions of Chemical, Disease, and Gene/Gene-product entities, and pairwise relationships between those entities. In comparison to other biomedical RE datasets, *ChemDisGene* is both larger and cleaner; it also includes annotations linking mentions to their entities [2].

  2. *StaRE* is a state-change relation extraction dataset. The training dataset is created via time stamp alignment between Wikidata relationships and Gigawords corpus, along with manually-annotated development and test sets. It is shown that the addition of state-change information can not only be used to train a system that detects a change of state in relations but also reduce noises when used for static relation extraction [4].
- **Structured Prediction**: Proposed a structured prediction framework that leveraged predefined constraints as light supervision. In this project, I contributed baseline implementations such as reinforcement learning-based and greedy-based methods [7].

Peking University **May 2017 – July 2017**
Research Assistant *Advisor: Sujian Li*

- **Scientific Summarization**: Developed a model for scientific article summarization using citation sentences which usually contain brief descriptions or comments of the cited paper [11].

Rensselaer Polytechnic Institute **April 2016 – June 2016**
Visiting Scholar *Advisor: Heng Ji*

- **Low-resource NLP:** Focused on projecting named entities from English to other languages using bilingual parallel corpora. Our framework consists of two parts, a dictionary-based string match module to generate high-precision seed training data, and a deep neural network module to propagate labels to increase coverage [13].

Tsinghua University **October 2014 - March 2016**
Research Assistant *Advisor: Dong Wang*

- **Relation Extraction**: Proposed to use RNNs on the sentence-level RE task. Evaluated on a new dataset KBP37 [14].

- **Sementic Representation Learning**: Worked on distributional entity representation using multiple resources such as ontology knowledge base, raw corpus, and Wikipedia pages [15].
- **Knowledge Distillation**: Proposed a distillation method from well-trained topic models to a fully connected neural network for efficient inference [12].

Beijing University of Posts and Telecommunications **September 2014 - September 2015**
Research Assistant *Advisor: Weiran Xu*

- **Large-scale Relation Extraction**: Developed a slot filling system for the Knowledge Base Acceleration track in NIST's Text Retrieval Conference. Our system performed the best among all participants [16].
  1. Combined manually annotated process and boot-strapping steps to improve both precision and recall.
  2. Employed Elastic Search to handle Tigabyte-level corpus.

Working Experience

Google AI **June 2019 - August 2019**
Research Intern *Mentor: Sara McCarthy and Chris Welty*

- **Taxonomy Alignment**: We leveraged box embeddings to tackle taxonomy alignment between anatomy and disease taxonomies to predict which body parts each disease has effects on.
  1. Constructed the ground truth conditional probability from disease taxonomy to anatomy taxonomy based on the frequencies of terminology co-occurrence among PubMed articles.
  2. Each vertex in both taxonomies is then represented as a high dimensional box embedding, which is learned by approximating the distribution of ground truth conditional probabilities using box intersection volumes.

Amazon **June 2018 - August 2018**
Applied Scientist Intern *Mentor: Subhabrata Mukherjee and Luna Dong*

- **Relation Extraction**: Given a pair of entities and a list of associated textual expressions, this project aimed at predicting relationships between the entity pair.
  1. Represented each entity as an aggregation over its associated textual context across the corpus.
  2. This entity representation was used directly for relation prediction, as well as the query for attention mechanism to select most relevant text expressions of the given entity pair [8].

Naturali **February 2017 - April 2017**
Algorithm Intern *Mentor: Dekang Lin*

- **Query Expansion** Developed a cold-start query expansion system using word co-occurrence.

Samsung Telecommunication R&D Center **May 2013 - Sep 2013**
Algorithm Intern *Mentor: Xiaojie Yu*

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

Teaching and Mentorship

CompSci 105: Computer Literacy **Fall 2017, Spring 2018**
Teaching Assistant *Amherst, MA*

- This is an introductory course in computer literacy. As a teaching assistant, I was holding office hours to answer students' questions, grading exams, and instructing laboratory assignments.

Independent Study **2019 - 2021**
Mentorship *Amherst, MA*

- *Brian Dang*, PubMed document labeling from MeSH hierarchies. This project led to a UMass Honored Thesis.
- *Jui Shah*: State-change relation extraction. This work got accepted by LREC 2022.
- *Bharath Narasimhan*: Interpretable relation extraction.
- *Yuchen Zeng*: Procedural recipe generation.
- Co-advised many excellent master students with researchers from Amazon, Bloomberg, and IBM for their course projects.

Professional Services

Workshop Co-organizer: SciNLP 2021
Conference Committee: ICLR 2021, EMNLP 2021, EMNLP 2022
Conference Reviewer (Nbr. of papers reviewed): TKDE 2018 (1), VLDB 2019 (1), TKDD 2019 (1), NLPCC 2020 (4), ICLR 2021 (3), NLPCC 2021 (4), ACL 2021 (4), EMNLP 2021 (4), ACL Rolling Review since October 2021 (7), EMNLP 2022 (2).

Publications

* indicates equal contributions.

[1] **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).

[2] **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.

[3] EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, **Dongxu Zhang**, and Andrew McCallum. "Event-Event Relation Extraction using Probabilistic Box Embedding". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (**ACL**)*. 2022, pp. 235–244.

[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] Amirmohammad Rooshenas, **Dongxu Zhang**, Gopal Sharma, and Andrew McCallum. "Search-guided, lightly-supervised training of structured prediction energy networks". In: *Advances in Neural Information Processing Systems (**NeurIPS**)* (2019).

[8] **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.

[9] **Dongxu Zhang** and Zhichao Yang. "Word embedding perturbation for sentence classification". In: *arXiv preprint arXiv:1804.08166* (2018).

[10] 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.

[11] **Dongxu Zhang** and Sujian Li. "Pku@ clscisumm-17: Citation contextualization". In: *BIRNDL@SIGIR*. 2017.

[12] **Dongxu Zhang**, Tianyi Luo, and Dong Wang. "Learning from LDA using deep neural networks". In: *Natural language understanding and intelligent applications*. Springer, 2016, pp. 657–664.

[13] **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.

[14] **Dongxu Zhang** and Dong Wang. "Relation classification via recurrent neural network". In: *arXiv preprint arXiv:1508.01006* (2015).

[15] **Dongxu Zhang**, Bin Yuan, Dong Wang, and Rong Liu. "Joint semantic relevance learning with text data and graph knowledge". In: *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*. 2015, pp. 32–40.

[16] Yuanyuan Qi, Ye Xu, **Dongxu Zhang**, and Weiran Xu. "BUPT_PRIS at TREC 2014 Knowledge Base Acceleration Track". In: *The Twenty-Third Text REtrieval Conference (**TREC 2014**) Proceedings* (2014).