Li “Harry” Zhang

zharry.com
zharry@seas.upenn.edu

Last updated: Feburary 2024

RESEARCH INTERESTS

Artificial Intelligence, Machine Learning, Natural Language Processing
Large Language Models, Planning and Reasoning, Human-Centered AI, etc.

EDUCATION

University of Pennsylvania, Philadelphia, PA — Aug 2019 – May 2024
Ph.D. Computer and Information Science — GPA: 3.96/4.00
Adviser: Prof. Chris Callison-Burch
Thesis: *Structured Event Reasoning with Large Language Models*
Committee: Prof. Dan Roth (chair), Prof. Rada Mihalcea, Prof. Graham Neubig, Prof. Mark Yatskar, Dr. Marianna Apidianaki

University of Michigan, Ann Arbor, MI — Aug 2015 – Dec 2018
B.S.E. Computer Science, summa cum laude — GPA: 3.82/4.00
Mentors: Prof. Rada Mihalcea and Prof. Dragomir Radev

PUBLICATIONS

27 total papers ⊆ 20 accepted to top NLP/AI conferences and workshops
⊆ 11 first-authored by self & 4 first-authored by mentored students.
Total citations: 1134; h-index: 10 (source: Google Scholar)
(*Equal contribution; ^Mentored students)

Preprints:

[27] Q. Lyu, K. Shridhar, C. Malaviya, **L. Zhang**, Y. Elazar, N. Tandon, M. Apidianaki, M. Sachan and C. Callison-Burch. Calibrating Large Language Models with Sample Consistency. In preprint.

[26] Y. Lal, B. Dalvi, **L. Zhang**, F. Brahman, B. Majumder, Peter Clark and N. Tandon. One Size Does Not Fit All: Customizing Open-Domain Procedures. In preprint.

[25] B. Majumder, B. Dalvi, P. Jansen, O. Tafjord, N. Tandon, **L. Zhang** and C. Callison-Burch, Peter Clark. CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization. In preprint.

Peer-reviewed papers:

[24] Z. Hou^, **L. Zhang** and C. Callison-Burch. *Choice-75: A Dataset on Decision Branching in Script Learning*. In LREC-COLING 2024.

[23] **L. Zhang**, H. Xu^, A. Kommula, N. Tandon and C. Callison-Burch. *OpenPI2.0: An Improved Dataset for Entity Tracking in Texts*. In EACL 2024.

[22] **L. Zhang***, L. Dugan*, H. Xu*^ and C. Callison-Burch. *Exploring the Curious Case of Code Prompts*. In preprint. In the 1st Natural Language Reasoning and Structured Explanations Workshop at ACL 2023.

[21] T. Zhang^, I. Tham, Z. Hou^, Jia. Ren, L. Zhou, H. Xu^, **L. Zhang**, L. Martin, R. Dror, S. Li, H. Ji, M. Palmer, S. Brown, R. Suchocki, C. Callison-Burch. *Human-in-the-Loop Schema Induction*. In preprint; in ACL 2023 Demos.

[20] Q. Lyu*, S. Havaldar*, A. Stein*, **L. Zhang**, D. Rao, E. Wong, M. Apidianaki and C. Callison-Burch. *Faithful Chain of Thought Reasoning*. In IJCNLP-AACL 2023.

[19] **L. Zhang***, H. Xu*^, Y. Yang, S. Zhou, W. You, M. Arora and C. Callison-Burch. *Causal Reasoning of Entities and Events in Procedural Texts*. In Findings of EACL 2023.

[18] **L. Zhang** and C. Callison-Burch. *Language Models are Drummers: Drum Composition with Natural Language Pre-Training*. In 1st Workshop on Creative AI across Modalities at AAAI 2023.

[17] Y. M. Cho^, **L. Zhang** and C. Callison-Burch. *Unsupervised Entity Linking with Guided Summarization and Multiple Choice Selection*. In EMNLP 2022.

[16] S. Gehrmann, …, **L. Zhang**, …, H. Zhu, S. Brahma, Y. Li, ... *GEMv2: Multilingual NLG Benchmarking in a Single Line of Code*. In EMNLP 2022.

[15] A. Srivastava, …, **L. Zhang**, Q. Lyu and C. Callison-Burch, … *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.* In TMLR.

[14] **L. Zhang**. *Reasoning about Procedures with Natural Language Processing: A Tutorial*. In preprint.

[13] A. Panagopoulou, M. Arora, **L. Zhang**, …, C. Callison-Burch, M. Yatskar. *QuakerBot: A Household Dialog System Powered by Large Language Models*. In Alexa Prize TaskBot Challenge Proceedings.

[12] Q. Lyu, H. Zheng, D. Li, **L. Zhang**, M. Apidianaki, and C. Callison-Burch. *Is "my favorite new movie" my favorite movie? Probing the Understanding of Recursive Noun Phrases*. In NAACL 2022.

[11] **L. Zhang**, I. Jindal and Y. Li. *Label Definitions Improve Semantic Role Labeling*. In NAACL 2022.

[10] **L. Zhang***, S. Zhou*, Q. Lyu, Y. Yang, G. Neubig and C. Callison-Burch. *Show Me More Details: Discovering Event Hierarchies from WikiHow*. In ACL 2022.

[9] Y. Yang, A. Panagopoulou, Q. Lyu, **L. Zhang**, M. Yatskar and C. Callison-Burch. *Visual Goal-Step Inference using wikiHow*. In EMNLP 2021; presented at the 2nd Workshop on Advances in Language and Vision Research at NAACL 2021.

[8] **L. Zhang***, Q. Lyu* and C. Callison-Burch. *Goal-Oriented Script Construction*. In INLG 2021.

[7] **L. Zhang**, Q. Lyu and C. Callison-Burch. *Intent Detection with WikiHow*. In AACL-IJCNLP 2020.

[6] **L. Zhang***, Q. Lyu* and C. Callison-Burch. *Reasoning about Goals, Steps, and Temporal Ordering with WikiHow*. In EMNLP 2020; presented at Workshop on Enormous Language Models at ICLR 2021.

[5] **L. Zhang**, H. Zhu, S. Brahma and Y. Li. *Small but Mighty: New Benchmarks for Split and Rephrase*. In EMNLP 2020.

[4] **L. Zhang**, S. R. Wilson and R. Mihalcea. *Multi-Label Transfer Learning for Semantic Similarity*. In *SEM 2019 and presented at NAACL 2019.

[3] **L. Zhang**, S. R. Wilson and R. Mihalcea. *Direct Network Transfer: Transfer Learning of Sentence Embeddings for Semantic Similarity*. In preprint and presented at IC2S2 2018.

[2] L. Burdick, S. R. Wilson, O. Ignat, C. F. Welch, **L. Zhang**, M. Wang, J. Deng and R. Mihalcea. *Entity and Event Extraction from Scratch Using Minimal Training Data*. In TAC 2018.

[1] C. Finegan-Dollak, J. K. Kummerfeld, **L. Zhang**, K. R. D. Ramanathan, S. Sadasivam, R. Zhang and D. Radev. *Improving Text-to-SQL Evaluation Methodology.* In ACL 2018.

External Funding

Alexa Prize TaskBot Challenge ($250,000) 2021 - 2022

Amazon Seattle, WA

• Primarily authored, applied, and received a stipend award of $250,000 to lead University of Pennsylvania's effort in the Alexa Prize TaskBot Challenge 2021.

INDUSTRY EXPERIENCE

Research Intern Apr 2023 – Current
Allen Institute for Artificial Intelligence (AI2) Seattle, WA

• See *An Improved Dataset for Entity Tracking* and *Language Models Meet Classical Planning* in Research Projects.

Research Intern Apr 2019 – Jun 2019; May 2021 – Aug 2021
IBM Research San Jose, CA

• See *Label-Aware Semantic Role Labeling with Definitions* and *Split and Rephrase: Evaluation Benchmarks and Metrics* in Research Projects.

Software Engineer Intern May 2017 – Aug 2017
Goldman Sachs Group, Inc. Jersey City, NJ

RESEARCH PROJECTS

Language Models Meet Classical Planning Jan 2023 – Present
"Converting wikiHow procedures to planning language representations"
Advised by Prof. Chris Callison-Burch, in collaboration with AI2

• We are concerned with translating natural language descriptions of an environment and a task to a structured representation in the PDDL language, using language models.

• A successful translation is not only semantically reasonable, but also allows search-based planners to find a plan accordingly.

• The symbolic nature of the task challenges language models to plan in open domains in an interpretable and grounded manner.

An Improved Dataset for Entity Tracking Feb 2023 – Jun 2023
"Set oven to 350F for 1 hour" causes the *temperature* of the *tray* change to *hot*
Work done as an intern at AI2, mentored by Dr. Niket Tandon

• We propose OpenPI2.0, built upon OpenPI, a dataset for entity tracking, we add two critical annotations: canonicalization and salience.

• We show that canonicalization benefits evaluation and that salience benefits downstream tasks such as question answering and classical planning.

Entity Tracking with Multi-hop Reasoning [17] Jan 2022 – Oct 2022
"Food sticks" before "adding oil to the hot pan" because "it is not greased"
Advised by Prof. Chris Callison-Burch

• Entities rapidly change in procedures; reasoning about state changes is important.

• We propose a dataset for challenging entity tracking in procedures.

• Our dataset also includes multi-hop reasoning questions based on the entity states to test models' ability to apply entity tracking in realistic question answering.

• Pre-trained large language models fail on our task except GPT3 and Codex.

• We use chain-of-thought reasoning to improve model performance.

See more projects at the end of this CV.

ACADEMIC SERVICE

Chair

• Program Chair:
Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) 2023

• Program Chair:
1st Workshop on Data Science with Human in the Loop at EMNLP 2022 2022

• Session Chair:
Asia-Pacific Chapter of the Association of Computational Linguistics (AACL) 2020

Reviewer

I have reviewed 28 papers at top NLP/AI conferences.

- Empirical Methods in Natural Language Processing (EMNLP) 2023
- Association of Computational Linguistics (ACL) 2023
- International Conference on Computational Linguistics (COLING) 2022
- International Conference on Language Resources and Evaluation (LREC) 2022
- Association of Computational Linguistics Rolling Review (ARR) 2021 - Present
- International Conference on Computational Linguistics (COLING) 2020
- Computer Speech and Language (CSL) journal. 2018

MENTORSHIP

During my years as a PhD student, I closely mentored the following master students resulting in many papers in top conferences (EMNLP, ACL, etc.).

Hainiu Xu May 2022 – May 2023

- Published 2 co-first-author papers and 2 non-first paper
- Won the 2023 Penn Engineering Master's Outstanding Research Award
- In the PhD program at King's College London

Tianyi Zhang May 2022 – May 2023

- Published 1 first-author papers and 1 co-first-author papers

Zhaoyi Hou Jan 2022 – May 2023

- Published 1 first-author paper and 1 non-first paper
- In the PhD program at Pittsburg University

Young Min Cho Oct 2021 – May 2022

- Published 1 first-author paper
- In the PhD program at the University of Pennsylvania

TEACHING

Teaching Assistant — Computational Linguistics Jan 2020 – Dec 2020
CIS 530: The graduate level NLP course University of Pennsylvania

Teaching Assistant — Natural Language Processing Sept 2018 – Dec 2018
EECS 595: The graduate level NLP course University of Michigan

Teaching Assistant — Programming and Data Structures Sept 2016 – Apr 2017
EECS 280: An introductory programming course University of Michigan

Tutor — Elementary Chemistry Sept 2016 – Dec 2016
Science Learning Center University of Michigan

COURSES

Graduate

Operating Systems(A+), Independent Research (A+), Machine Learning (A-), Common-sense Reasoning (A), Software Foundations (A-), Big Data (A), Neurolinguistics (A-), Composition of Electronic Music (A)

Undergraduate

Natural Language Processing(A+), Directed Research (A+), Information Retrieval(A), Machine Learning(A), Artificial Intelligence (A), Computer Security(A), Multivariate Calculus(A+), Probability and Statistics (A-), Matrix Algebra (A-)

HONORS

Merit-Based Scholarship of $2,000, UM Engineering Class of 1935 2017 – 2018
James B. Angell Scholar, University of Michigan 2017
University Honors of all semesters, University of Michigan 2015 – 2018

Dean's Honor List of all semesters, UM College of Engineering 2015 – 2018

RESEARCH PROJECTS (cont'd)

Procedure Understanding in Dialogs: Alexa Prize [13] May 2020 – May 2022

User: "Add baking soda and vinegar to my list." Intent: "Cleaning; less likely cooking."

Advised by Prof. Chris Callison-Burch

- An important component of dialog systems is detecting intent from utterances.
- As intents are similar to goals, our models with procedural knowledge achieve state-of-the-art performances on several, multilingual benchmarks.
- I co-led University of Pennsylvania's team as a finalist in *the Alexa Prize TaskBot Challenge 2021*, building a production-ready dialog system for household tasks.
- Our dialog system is powered by a mixture of large language models and rule-based models, ending up in the semi-finals.

Hierarchical Relations of Events [10] Jan 2021 – Sept 2021

To "host a party", one needs to "clean the house"; to do so, one needs to "vacuum"

Advised by Prof. Chris Callison-Burch and Prof. Graham Neubig

- A goal has steps, each of which in turn can itself be a sub-goal with some sub-steps.
- Event hierarchy can decompose steps and provide information upon request.
- We link steps to other wikiHow articles with high precision and recall by exploring various approaches based on semantic similarity, allowing for hierarchical lookups.
- Our hierarchy is shown via crowdsourcing to help users accomplish tasks and improves performance in downstream tasks such as video retrieval.

Goal-Step and Temporal Relations of Events [6,7,9,15] Nov 2019 – May 2020

"Hire an attorney" is a step of a "lawsuit"; so is "appear at court", which happens later

Advised by Prof. Chris Callison-Burch, part of the DARPA KAIROS project

- A procedure consists of a goal and a series of steps, which may be ordered.
- Goal-step and step-step temporal relations are important knowledge for human-centered AI systems, especially for task-oriented dialog systems.
- We collect procedure data to-scale from the how-to website wikiHow for model training and curate a high-quality evaluation benchmark.
- Models pre-trained with our data show strong zero- and few-shot performance on various other tasks, such as story completion, intent detection, and event prediction.
- Similarly, our models can learn to reason about goals and steps represented as image.
- Our multimodal models show strong transfer performance on tasks like video retrieval.

Label-Aware Semantic Role Labeling with Definitions [11] May 2021 – Aug 2021

Instead of tagging semantic roles of "work" as A0, A1, tag them as "employee" and "job"

Work done as an intern at IBM Research, advised by Yunyao Li

- Semantic Role Labeling is a core NLP task, answering the question "who did what to whom, when and how," by labeling tokens in a sentence as arguments of some predicate.
- Instead of using symbolic labels (e.g. A0, AM-TMP) for arguments, we propose to provide models with label definitions, which linguistics used to annotate data
- Models trained on our definition-injected data achieve state-of-the-art performance on the CoNLL09 benchmark given predicate senses, and strong few-shot performance

Annotation Projection for Cross-lingual Event Extraction Oct 2019 – Apr 2021

Convert English data to another language by translation, alignment, and projection

Advised by Prof. Chris Callison-Burch, part of the DARPA BETTER project

- Event extraction deals with identifying entities and events from texts, along with their arguments and relations, given an otology in the domain of interest.

• Given ample labeled data in English and evaluation data in another language, we first translate the sentences using neural machine translation, then word align the tokens using neural aligners, and finally project the labels based on some heuristics.

• Models trained on our data with projected annotations achieve strong performance compared to cross-lingual zero-shot models.

Split and Rephrase: Evaluation Benchmarks and Metrics [5,16] Apr 2019 – Jun 2019

Revamp the evaluation for the text simplification task of splitting long sentences

Work done as an intern at IBM Research, advised by Yunyao Li

• We developed a rule-based model using no training data which performs on par with the current state-of-the-art neural model, showing the evaluation might be flawed.

• We released two new crowdsourced benchmarks with improved quality.

• We conducted a case study on the flaws of BLEU score, and the cost-efficiency of using crowd workers to evaluate model performance.

Transfer Learning in Semantic Similarity [3] Oct 2017 – May 2018

Explore transfer learning methods using sentence embeddings in semantic similarity

Advised by Prof. Rada Mihalcea

• Proposed a new transfer learning method for semantic similarity tasks, achieving state-of-the-art performance on various datasets using various neural networks architectures.

• Compared and analyzed performances of popular transfer learning methods on a collection of mainstream LSTM-based models and semantic similarity datasets.

• Interpreted qualitatively the source of improvement in the domain of human activities.

Multi-label Learning in Semantic Similarity [4] Mar 2017 – Sept 2018

Explore multi-task learning using sentence embeddings in semantic similarity

Advised by Prof. Rada Mihalcea

• Proposed a modification of LSTM architecture for semantic similarity datasets with multiple relations, achieving state-of-the-art in various dimensions.

• Compared with multi-task learning and single-task learning baselines.

Active Interpretation of Disparate Alternatives [2] Jan 2017 – Feb 2019

Use multi-modal news reports to generate hypotheses about real life events

Advised by Prof. Rada Mihalcea, part of the DARPA AIDA project

• Produced knowledge elements using the text from multiple account of the events regarding the Ukrainian-Russian relations.

• Performed keyword extraction and named entity recognition to extract knowledge elements and assign saliency to them.

Natural Language to SQL in Academic Advising [1] Sept 2015 – Apr 2017

Part of the IBM Sapphire project to build a dialogue system for academic advising

Advised by Prof. Dragomir Radev

• Implemented a named entity recognizer specifically on the academic advising ontology to automatically expand training data by permutating entities.

• Designed over 50 semantically distinct and meaningful advising questions as well as their corresponding SQL queries to be used as training data.

• Contributed in building the Advising dataset parallel to the ATIS and GeoQuery datasets that contains more than 300 entries in the academic advising domain.

• Presented in 2016 Michigan Research Community poster symposium.

Text Clustering Based on Humor in Cartoons Jan 2016 – Apr 2016

Dataset from caption submissions for the cartoon section of New Yorker magazine

Advised by Prof. Dragomir Radev

• Restarted and oversaw the project with under-documented codebase.

• Rewrote Perl scripts in Python using state-of-the-art machine learning APIs.

• Experimented with text embeddings such as word2vec and Skip-Thought to compare performances in multiple clustering algorithms such as the Louvain algorithm.

ACL Anthology Network Sept 2016 – Dec 2016

A power taxonomy of papers from top NLP conferences

Advised by Prof. Dragomir Radev

• Implemented distance metrics between papers classified into the same category.

• Fixed display issues on the front end and did QA on the database.

• Presented in 2016 University of Michigan NLP workshop.