# Li "Harry" Zhang

zharry.com
zharry@seas.upenn.edu
(734)834-7882

| | |
|---|---|
| **RESEARCH INTERESTS** | Artificial Intelligence, Machine Learning, Natural Language Processing<br>Semantics, Natural Language Understanding, Reasoning, Events |

**EDUCATION**

| | | |
|---|---|---|
| **University of Pennsylvania**, Philadelphia, PA | | Aug 2019 – May 2024 |
| Ph.D. Computer and Information Science | | GPA: 3.92/4.00 |
| **University of Michigan**, Ann Arbor, MI | | Sept 2015 – Dec 2018 |
| B.S.E. Computer Science, *summa cum laude* | | GPA: 3.82/4.00 |

**PUBLICATIONS**

[17] **L. Zhang**\*, H. Xu\*, Y. Yang, S. Zhou, M. Arora, W. You and C. Callison-Burch. *Entity Tracking with Multi-hop Reasoning*. In submission.

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

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

[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] S. Zhou\*, **L. Zhang**\*, 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] Q. Lyu\*, **L. Zhang**\* 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] Q. Lyu\*, **L. Zhang**\* 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. (*Equal contribution)

**RESEARCH EXPERIENCE**

### Procedures: Reasoning about Goals and Steps [6,9,15]        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.

### Procedures: 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.

### Procedures: Application in Dialogs; Alexa Prize [7]        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 effort in *the Alexa Prize TaskBot Challenge 2021*, building a production-ready dialog system for household tasks or cooking.

• Our dialog system is powered by a mixture of large language models and rule-based models, ending up in the semi-finals.

### Procedures: Script Construction [8]                Sept 2020 – Jan 2021

Input: "How do I host a party?" Output: "Send out invitations. Clean your house…"

*Advised by Prof. Chris Callison-Burch*

• We develop two pipelines to construct all the steps given a goal, one based on language model generation (T5), one based on previous goal-step inference models.

• Our pipeline is trained on wikiHow and available in 18 languages.

• The constructed steps are mostly reasonable but also show notable weaknesses.

**Procedures: Event Hierarchy** [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.

**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.

**ACADEMIC SERVICES**

**Session Chair**

• Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)    2020

**Paper Reviewing**

| | |
|---|---|
| • 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 |

**TEACHING EXPERIENCE**

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

• Held weekly office hours and answered questions online for students.

• Helped design course contents such as homework and quizzes.

• Gave supplementary lectures.

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

• Held weekly office hours and answered questions online for 135 students.

• Helped design course contents.

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

• Led weekly lab sessions with more than 30 students to review course materials and guide them through hands-on coding challenges.

• Held weekly office hours and answered questions online for over 1,000 students.

• Helped design course contents, projects and exams.

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

• Hosted weekly walk-in tutoring for an introductory chemistry course.

• Compiled review materials to help students prepare for exams.

**INDUSTRY EXPERIENCE**

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

• Conducted NLP research; see more in Research Experience.

**Intern Analyst**      May 2017 – Aug 2017
*Goldman Sachs Group, Inc.*      Jersey City, NJ

Worked in GS App Store, the firm's software management and delivery platform

• Developed a highly scalable recommender system using collaborative filtering to suggest personalized app recommendations for each user.

• Designed features such as "frequently installed together", "users who installed this also installed", "trending apps" and "top apps by business unit".

• Implemented end-to-end interfaces using C#, JavaScript, AngularJS and Elasticsearch.

**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 |

**SKILLS**

**Programming Skills**

Python, PyTorch

**Language Skills**

Chinese (native), English (fully proficient), French (conversational)

**TEST SCORES**

**GRE** (May 2017): Verbal 162 91%, Quantitative 170 97%, Analytical Writing 4.5 82%

**SAT** (Dec 2015): Reading 750, Writing 790, Math 800

**TOEFL** (Dec 2013): Reading 28, Listening 29, Speaking 29, Writing 28

**HOBBIES**

**MUSIC**

I play, record, and produce music regularly. My tracks are distributed under the name Haz Studio on YouTube, Bilibili, Spotify, Apple Music and all streaming services.

**Billiards and Pool**

| | |
|---|---|
| 1st Place out of 12, 9-Ball Scotch Doubles, 1st Midwest Invitational | Feb 2018 |
| 7th Place out of 32, 9-Ball Scotch Doubles, 14th UM Team Pool Championship | Nov 2017 |
| Captain, Michigan Billiards Team | Sept 2017 – Present |
| Membership Chair, Michigan Billiards Club | Sept 2016 – Present |

**E-Sport**

| | |
|---|---|
| Top 32 out of 828 teams, TESPA Hearthstone Collegiate Championship | Apr 2018 |
| #29 out of approx. 5 million in Ranked Play, Hearthstone, North America | Mar 2018 |
| Member, Michigan Hearthstone Team | Sept 2016 – Present |