Freda Shi (a.k.a., Haoyue Shi)

Assistant Professor, University of Waterloo
200 University Ave W., Waterloo, ON, Canada, N2L 3G1

Faculty Member, Vector Institute
661 University Ave., Toronto, ON, Canada, M5G 1M1

E-mail: fhs@uwaterloo.ca

Last updated: Friday 4th October, 2024

Research Interests

Computational linguistics, natural language processing, and machine learning: compositional semantics, grounded language acquisition, unsupervised and semi-supervised representation learning, structured prediction, narrative understanding, and information theory for natural language processing.

Appointments

Assistant Professor, University of Waterloo, Ontario, ON, Canada 2024-

Faculty Member and **Canada CIFAR AI Chair**, Vector Institute, Toronto, ON, Canada 2024-

Education

Toyota Technological Institute at Chicago, Chicago, IL, USA 2018-2024
Ph.D. in Computer Science (Ph.D. thesis of distinction; Master's degree awarded Sept. 2020)
Thesis: Learning Language Structures through Grounding
Advisors: Kevin Gimpel and Karen Livescu
Thesis Committee: Kevin Gimpel, Karen Livescu, Roger Levy and Luke Zettlemoyer

Peking University, Beijing, China 2013-2018
B.S. in Intelligence Science and Technology (Computer Science Track), *summa cum laude*
Minor in Sociology
Thesis: On Multi-Sense Word Embeddings via Matrix Factorization and Matrix Transformation
Advisor: Junfeng Hu

Non-Degree Academic Experience:

Visiting Student with Roger P. Levy 2024
Massachusetts Institute of Technology, Cambridge, MA, USA

Visiting Student with Samuel R. Bowman 2017
New York University, New York City, NY, USA

Visiting Student with Alexander G. Hauptmann 2016
Carnegie Mellon University, Pittsburgh, PA, USA

Selected Honors and Awards

Canada CIFAR AI Chair 2024
Thesis of Distinction, Toyota Technological Institute at Chicago 2024
Nomination for the Best Paper Award, ACL (*with D. Chen, A. Argawal, J. Myerston and T. Berg-Kirkpatrick*) 2024
Highlighted Reviewer, ICLR 2022
Google Ph.D. Fellowship ($\approx$ USD $220,000 for tuition and stipend in 3 years) 2021
Finalist, Facebook Ph.D. Fellowship 2021
Nomination for the Best Paper Award, ACL-IJCNLP (*with L. Zettlemoyer and S. I. Wang*) 2021

Nomination for the Best Paper Award, ACL (*with J. Mao, K. Gimpel and K. Livescu*) 2019
Best Undergraduate Dissertation Award, School of EECS, Peking University 2018
Robin Lee Scholarship (top 2 out of 400, CNY ¥20,000), Peking University 2016
WeTech Qualcomm Global Scholarship (USD $5,000) 2016
Arawana Scholarship (top 4 out of 400, CNY ¥10,000), Peking University 2015
Gold Medalist (*with Tianshi Li and Chengxian Mo*), ACM-ICPC Chengdu Site 2013

Research Internships

Google Brain (Hybrid Internship), Waterloo, ON, Canada → Chicago, IL, USA Jun. 2022-Dec. 2022
Host: Denny Zhou

Meta (Facebook) AI Research (Remote Internship), Seattle, WA, USA Aug. 2021-Dec. 2021
Mentors and collaborators: Luke Zettlemoyer, Sida Wang, Daniel Fried, and Marjan Ghazvininejad

Facebook AI Research (Remote Internship), Seattle, WA, USA Jun. 2020-Dec. 2020
Mentors: Sida Wang and Luke Zettlemoyer

ByteDance AI Lab, Beijing, China Mar. 2018-Aug. 2018
Mentors: Hao Zhou and Lei Li

Megvii (Face++) Research, Beijing, China Oct. 2017-Mar. 2018
Mentors: Yuning Jiang and Jian Sun

Microsoft Research Asia, Beijing, China Sep. 2016-Feb. 2017
Mentors: Zhongyuan Wang and Jun Yan

Engineering Internships

4th Paradigm Inc., Beijing, China Mar. 2017-Jun. 2017
Mentors: Weiwei Tu and Yuqiang Chen

Google Inc., Beijing, China Jul. 2015-Dec. 2015
Mentors: Xiaoyi Ren and Jie Mao

Teaching

Instructor at the University of Waterloo
CS 784 Computational Linguistics Winter 2025
CS 486/686 Introduction to Artificial Intelligence Autumn 2024

Instructor at Toyota Technological Institute at Chicago and the University of Chicago Autumn 2023
TTIC 31190 Natural Language Processing

Guest Lecturer at the University of Chicago
MPCS 53113 Natural Language Processing Instructor: Amitabh Chaudhary Summer 2021

Teaching Assistant at Toyota Technological Institute at Chicago
TTIC 31220 Unsupervised Learning and Data Analysis Instructor: Karen Livescu Winter 2021

Teaching Assistant at School of EECS, Peking University
Practice of Programming in C&C++ Instructor: Wei Guo Spring 2018
Programming & Algorithms (MOOC on Coursera) Instructor: Wei Guo Fall 2016
Practice of Programming in C&C++ Instructor: Jiaying Liu Spring 2015

Volunteer Lecturer in Mathematics Summer 2014
Rongxian High School Summer Camp, Guangxi, China

Referred Conference Publications

[1] **Freda Shi**, Kevin Gimpel, and Karen Livescu. *Structured Tree Alignment for Evaluation of (Speech) Constituency Parsing*. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2024.

[2] Danlu Chen, **Freda Shi**, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. *LogogramNLP: Comparing Visual and Textual Representations of Ancient Logographic Writing Systems for NLP*. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. ***Best Paper Nominee***. 2024.

[3] **Freda Shi**, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. *Large language models can be easily distracted by irrelevant context*. In: *Proceedings of the Fortieth International Conference on Machine Learning (ICML)*. 2023.

[4] **Freda Shi**, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. *Language models are multilingual chain-of-thought reasoners*. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. 2023.

[5] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, **Freda Shi**, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. *InCoder: A Generative Model for Code Infilling and Synthesis*. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. 2023.

[6] **Freda Shi**, Kevin Gimpel, and Karen Livescu. *Substructure Distribution Projection for Zero-Shot Cross-Lingual Dependency Parsing*. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2022.

[7] **Freda Shi**, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. *Natural Language to Code Translation with Execution*. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2022.

[8] **Haoyue Shi**, Karen Livescu, and Kevin Gimpel. *Substructure Substitution: Structured Data Augmentation for NLP*. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

[9] **Haoyue Shi**, Luke Zettlemoyer, and Sida I. Wang. *Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment*. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. ***Best Paper Nominee***. 2021.

[10] Jiayuan Mao, **Freda Shi**, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. *Grammar-Based Grounded Lexicon Learning*. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[11] **Haoyue Shi**, Karen Livescu, and Kevin Gimpel. *On the Role of Supervision in Unsupervised Constituency Parsing*. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[12] **Haoyue Shi**, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. *Visually Grounded Neural Syntax Acquisition*. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. ***Best Paper Nominee***. 2019.

[13] **Haoyue Shi**, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. *Learning Visually-Grounded Semantics from Contrastive Adversarial Samples*. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018.

[14] **Haoyue Shi**, Hao Zhou, Jiaze Chen, and Lei Li. *On Tree-Based Neural Sentence Modeling*. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2018.

[15] **Haoyue Shi**, Xihao Wang, Yuqi Sun, and Junfeng Hu. *Constructing High Quality Sense-specific Corpus and Word Embedding via Unsupervised Elimination of Pseudo Multi-sense*. In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*. 2018.

[16] **Haoyue Shi**, Jia Chen, and Alexander G. Hauptmann. *Joint Saliency Estimation and Matching using Image Regions for Geo-Localization of Online Video*. In: *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval (ICMR)*. 2017.

[17] Shan Xu, **Haoyue Shi**, Xiaohui Duan, Tiangang Zhu, Peihua Wu, and Dongyue Liu. *Cardiovascular Risk Prediction Method Based on Test Analysis and Data Mining Ensemble System*. In: *Proceedings of the 2016 IEEE International Conference on Big Data Analysis*. 2016.

Referred Journal Publications

[18] Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, **Haoyue Shi**, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. *NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation*. In: *Northern European Journal of Language Technology* 9.1 (2023).

Referred Workshop Publications

[19] Cheng-I Jeff Lai, **Freda Shi**, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David Cox, David Harwath, Yang Zhang, Karen Livescu, and James Glass. *Audio-Visual Neural Syntax Acquisition*. In: *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2023.

[20] Shubham Toshniwal, **Haoyue Shi**, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. *A Cross-Task Analysis of Text Span Representations*. In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. 2020.

[21] Yuqi Sun, Haoyue Shi, and Junfeng Hu. *Implicit Subjective and Sentimental Usages in Multi-sense Word Embeddings*. In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2018.

[22] Haoyue Shi, Caihua Li, and Junfeng Hu. *Real Multi-Sense or Pseudo Multi-Sense: An Approach to Improve Word Representation*. In: *Proceedings of the 1st Workshop on Computational Linguistics for Linguistic Complexity*. 2016.

Theses

[23] **Haoyue Freda Shi**. *Learning Language Structures through Grounding*. PhD thesis. Toyota Technological Institute at Chicago, 2024.

[24] **Haoyue Shi**. *On Multi-Sense Word Embeddings via Matrix Factorization and Matrix Multiplication*. Bachelor's thesis. Peking University. 2018.

Invited Talks

[25] **Freda Shi**. *Towards Computational Multilingualism with Large Language Models*. Invited talk, University of Toronto and Ontario Tech University. Oct. 2024.

[26] **Freda Shi**. *Towards Computational Multilingualism with Large Language Models*. Invited talk, Boston University. May 2024.

[27] **Freda Shi**. *Computational Multilingualism in the Era of Large Language Models*. Invited talk, Vector NLP Workshop. Feb. 2024.

[28] **Freda Shi**. *Learning Syntactic Structures from Visually Grounded Text and Speech*. Invited talk, University of Michigan. Oct. 2023.

[29] **Freda Shi**. *Learning Language Structures through Grounding*. Invited talk, Peking University. Sept. 2023.

[30] **Freda Shi**. *Learning Language Structures through Grounding*. Invited talk, University of Toronto. Aug. 2023.

[31] **Freda Shi**. *Language Models Are Multilingual Chain-of-Thought Reasoners*. Invited talk, Translate Theory Reading Group, Google AI. Oct. 2021.

[32] **Freda Shi**. *Naturally Supervised Parsing: Assumptions, Methods and Evaluation*. Invited talk, Yahoo! NYC Remote Research Seminar. Apr. 2021.

[33] **Freda Shi**. *Visually Grounded Neural Syntax Acquisition*. Invited talk, Remote Seminar, Carnegie Mellon University. Aug. 2020.

[34] **Freda Shi**. *Structures in Natural Language: How to learn it and how to use it?* Invited talk, Remote NLP Seminar, University of Alberta. May 2020.

[35] **Haoyue Shi**. *Visually Grounded Neural Syntax Acquisition*. Invited talk, NLP Seminar, Peking University. Dec. 2019.

Open-Sourced Project Contributions

Implementations that accompany to the publications listed above are open-sourced if permitted, and are not listed below.

1. **NLTK**
   A suite of open source Python modules, data sets, and tutorials.
   `https://nltk.org`
2. **NL-Augmenter**
   A general-purpose data augmentation framework for NLP.
   `https://github.com/GEM-benchmark/NL-Augmenter`
3. **Multimodal concreteness score estimator**
   Implementation of the paper *Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets* (Hessel et al., 2018).
   `https://github.com/victorssilva/concreteness`
4. **Structured self-attentive sentence embeddings**
   Implementation of the paper *A Structured Self-Attentive Sentence Embedding* (Lin et al., 2017).
   `https://github.com/explorerfreda/structured-self-attentive-sentence-embedding`

Advising and Mentoring

Current Students

- Ruoxi Ning, Ph.D. Student in CS at the University of Waterloo (2024–), recipient of Cohere Scholarship.
- Michael Ogezi, Ph.D. Student in CS at the University of Waterloo (2024–).
- Shucheng (Bruce) Huang, Ph.D. Student in MME at the University of Waterloo (2023–), co-advised with Amir Khajepour.

- Hala Sheta, MMath Student in CS at the University of Waterloo (2024–), co-advised with Daniel Brown, recipient of Vector Scholarship.
- Sheng Yao, MMath Student in CS at the University of Waterloo (2024–).

Current Visiting Students

- Yilei Tu, Visiting Master's Student, University of Waterloo (2024–2025). Home institution: ETH Zürich.
- Haojin (Sean) Wang, Visiting Undergraduate Student, University of Waterloo (2024–2025). Home institution: Tongji University.
- Xiaoxi Luo, Visiting Scholar, University of Waterloo (2024–2025). B.S. 2024, Peking University.

Current External Student Collaborators

- Vu Trong Kim, Undergraduate Student at KAIST (2024–).
- Yixuan Wang, Master's Student at the University of Chicago (2024–).
- Jianyu Wang, Master's Student at the University of California, San Diego (2023–).

Current Undergraduate Research Assistants

- Disen Liao, University of Waterloo (2024).
- Andrew Xue, University of Waterloo (2024).

Service

Area Chair for

- COLM, 2024;
- Language Grounding to Vision, Robotics and Beyond, ACL 2023;
- Machine Learning for NLP, EMNLP 2023.

Reviewer for Conferences and Journals in

- Computational Linguistics and Natural Language Processing: TACL (2023–2025), ACL (2019-2022), ACL Rolling Review (2020–2024), COLING (2020, 2022), EACL (2022), EMNLP (2020–2022), LREC (2020), NAACL (2021), NLPCC (2020, 2021);
- Machine Learning: JMLR (2023), TPAMI (2022), ICLR (2020–2025), ICML (2020–2024), NeurIPS (2020–2024);
- Artificial Intelligence: IJCAI (2021), AAAI (2019, *secondary to Hao Zhou*), UAI (2023, *secondary to Lili Mou*);
- Computer Vision: ViGiL Workshop (2021), CVPR (2020, *secondary to Jiayuan Mao*);
- Robotics: ICRA (2024).

Workshop Co-Organizer of

- The 10th Workshop on Representation Learning for NLP, 2025;
- TTIC Student Workshop, 2020.

Institutional Committee Member of

- Undergraduate Academic Plans Committee, School of Computer Science, University of Waterloo, 2024;
- Postdoctoral Scholar Recruiting Committee, Vector Institute, 2024.

Co-Organizer of UChicago-TTIC NLP Reading Group, 2022–2023.

Student Member of the TTIC Student Admission Committee, 2021–2022.

Student Representative at TTIC, 2020–2021.

Student Co-Chair of the Women at TTIC group and **coordinator** with UChicago Graduate Women in CS, 2019–2022.

Peer Mentor for new students at TTIC, 2019–2020, 2022–2023.

Chief of the PKU Guqin Society, 2016–2017.

Skills

Programming Languages:
- Proficient: C/C++, Python(2/3), MATLAB, Pascal, HTML/CSS
- Capable: C#, SCOPE, JavaScript, Java, Scala, Mathematica, Bash, SQL

Natural Languages:
Mandarin (native), English (fluent), classical Chinese (advanced reading & writing), Cantonese (intermediate listening & speaking), Japanese (intermediate), German (beginner), Hebrew (beginner), Spanish (beginner)

Tools & Frameworks: Vim, Caffe, Torch, PyTorch, GDB, Git, LaTeX, CMake, Visual Studio, ssh