Mike Zhang

Aalborg University, Department of Computer Science (Data, Knowledge, and Web Engineering Group) &
Pioneer Cenre for Artificial Intelligence (P1)
Copenhagen, Denmark
mikejj.zhang@gmail.com | jjzha.github.io | github.com/jjzha

ACADEMIC BACKGROUND

Postdoctoral Researcher 2024–2026
Aalborg University Copenhagen, DK

- Researching at the intersection of NLP and Education.
- Mentored by Johannes Bjerva and Euan Lindsay.

Postdoctoral Researcher 2024–2024
IT University of Copenhagen Copenhagen, DK

- One-month postdoc, researched Danish Foundation Models.
- Mentored by Rob van der Goot.

Ph.D. Natural Language Processing 2020–2024
IT University of Copenhagen & Copenhagen, DK
Ludwig Maximilian University of Munich Munich, DE

- Researching Information Extraction for Job Market Understanding, published papers at top-tier NLP venues (ACL, EMNLP, NAACL, EACL);
- Managed a team consisting of a research assistant and annotator;
- Released several open-source language models on HuggingFace and datasets related to Job Market Understanding;
- Advised by Barbara Plank and Rob van der Goot, member of NLPnorth and MaiNLP.

M.A. Information Science 2018–2020
University of Groningen Groningen, NL

- Focus areas: Computational Linguistics, Semantics, Evaluation.
- Thesis: The Effect of Translationese in Machine Translation Test Sets. Advised by Antonio Toral Ruiz.

B.Sc. Information Science 2015–2018
University of Groningen Groningen, NL

- Focus areas: Computational Linguistics, Linguistics, Web Technology.
- Thesis: From Relations to Falsehood: Labeled Bilexical Dependencies for Deception Detection. Advised by Barbara Plank
- Exchange Semester at The Chinese University of Hong Kong (CUHK), HK.

WORK EXPERIENCE

Ph.D. Research Visitor Oct. 2023 – Nov. 2023
Swiss Federal Institute of Technology Lausanne (EPFL) Lausanne, CH

- Kickstarted research on NLP applications for Job Market Understanding;
- Visited EPFL's NLP Lab, hosted by Syrielle Montariol and Antoine Bosselut.

Ph.D. Research Visitor Feb. 2023 – Jul. 2023
National University of Singapore (NUS) Singapore, SG

- Conducted research with respect to NLP and IR related to job descriptions and related text sources;
- Applied Nearest Neighbor Language Model methods for Skill Extraction;
- Visited the WING Group, advised by Min-Yen Kan.

Ph.D. Research Intern Sep. 2022 – Dec. 2022
NEC Laboratories Europe Heidelberg, DE

- Improved precision of neural OpenIE systems by 17% using data-centric approaches, while keeping recall;
- Investigated Curriculum Learning and Data Augmentation for OpenIE;
- Developed a data pre-processing and cleaning pipeline for large-scale OpenIE data in Python.
- Conducted research with the Human-Centric AI team.

Data Engineer / Research Engineer Intern Aug. 2019 – Aug. 2020
Dataprovider.com Groningen, NL

- Improved predicting company identification codes by 40 F1, using Python and Scikit-Learn;
- Spearheaded a data cleaning pipeline for structuring 72M addresses with Python and Geocoding;
- Added 10+ features into a proprietary search engine resulting in more information about companies, using Python and Elasticsearch.

PEER-REVIEWED PUBLICATIONS

Equal contribution

18. Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, **Mike Zhang**, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, Sara Hooker. "Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning". *Under Review*. 2024
17. **Mike Zhang**, Rob van der Goot, Min-Yen Kan, and Barbara Plank. "NNOSE: Nearest Neighbor Occupational Skill Extraction". *In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. 2024.
16. **Mike Zhang**, Rob van der Goot, and Barbara Plank. "Entity Linking in the Job Market Domain". *In Findings of the Association for Computational Linguistics: EACL 2024*. 2024.
15. Khanh Cao Nguyen, **Mike Zhang**, Syrielle Montariol, and Antoine Bosselut. "Rethinking Skill Extraction in the Job Market Domain using Large Language Models". *In Proceedings of the 1st Workshop on Natural Language Processing for Human Resources*. 2024.
14. Antoine Magron, Anna Dai, **Mike Zhang**, Syrielle Montariol, and Antoine Bosselut. "JobSkape: A Framework for Generating Synthetic Job Postings to Enhance Skill Matching". *In Proceedings of the 1st Workshop on Natural Language Processing for Human Resources*. 2024.

13. Elena Senger, **Mike Zhang**, Rob van der Goot, and Barbara Plank. "Deep Learning-based Computational Job Market Analysis: A Survey on Skill Extraction and Classification from Job Postings", *In Proceedings of the 1st Workshop on Natural Language Processing for Human Resources*. 2024.
12. Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, **Mike Zhang**, and Rob van der Goot. "Can Humans Identify Domains?". *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 2024.
11. **Mike Zhang**, Rob van der Goot, and Barbara Plank. ESCOXLM-R: Multilingual Taxonomy-driven Pre-training for the Job Market Domain. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2023.
10. Elisa Bassignana, Max Müller-Eberstein, **Mike Zhang**, and Barbara Plank. Evidence > Intuition: Transferability Estimation for Encoder Selection. *In Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing*. 2022
9. Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, **Mike Zhang**, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental Standards for Deep Learning in Natural Language Processing Research. *In Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022.
8. **Mike Zhang**, Kristian Nørgaard Jensen, Rob van der Goot, and Barbara Plank. Skill Extraction from Job Postings using Weak Supervision. *RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems*. 2022
7. **Mike Zhang**, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. SkillSpan: Hard and Soft Skill Extraction from Job Postings. *In Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2022.
6. **Mike Zhang**, Kristian Nørgaard Jensen, and Barbara Plank. Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning. *In Proceedings of the 13th Edition of its Language Resources and Evaluation Conference*. 2022.
5. Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, **Mike Zhang**, Christian Hardmeier, and Barbara Plank. "A Common Set of Experimental Standards for Deep Learning Research: A Natural Language Processing Perspective." *Machine Learning Evaluation Standards workshop at ICLR 2022*. (Non-Archival) 2022. **Outstanding Paper Award**
4. **Mike Zhang** and Barbara Plank. Cartography Active Learning. *In Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021.
3. **Mike Zhang**, Kristian Nørgaard Jensen, and Barbara Plank. De-identification of Privacy-related Entities in Job Postings. *Proceedings of the 23rd Nordic Conference on Computational Linguistics*. 2021.
2. **Mike Zhang** and Antonio Toral. The Effect of Translationese in Machine Translation Test Sets. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. 2019.
1. **Mike Zhang**, Roy David, Leon Graumans, and Gerben Timmerman. Grunn2019 at SemEval-2019 Task 5: Shared Task on Multilingual Detection of Hate. *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019.

OPEN-SOURCE CONTRIBUTIONS

Language Ambassador (The Aya Project) 2023
I was the language ambassador for Dutch in the Aya project by Cohere for AI. By creating high-quality instruction data for large multilingual language models we make sure no language is left behind. I communicated with the region ambassadors on the progress of high-quality Dutch instruction examples.

GRANTS & AWARDS

Research Stay Abroad (IT University of Copenhagen) 2022
Received **3,500 EUR** for a research visit abroad in an external institution. The destination is the National University of Singapore (NUS).

Outstanding Paper Award (ML Evaluation Standards Workshop) 2022
Awarded **2,000 USD** for an outstanding paper award at the ML Evaluation Standards Workshop at ICLR 2022, see paper [5].

Student Scholarship Award (EMNLP 2021, 2022) 2021-2022
Received free attendance to EMNLP 2021 (eq. **300 USD**) in Punta Cana, DR.
Received free attendance to EMNLP 2022 (eq. **325 USD**) in Abu Dhabi, UAE.

Marco Polo Fund (University of Groningen) 2018
Awarded **1,000 EUR** for studying abroad outside of Europe. For this grant there is one awardee per semester. Destination was The Chinese University of Hong Kong (CUHK). Courses in Java Programming, Politics, Computer Networks.

SERVICES

- **Chair**:
  EACL SRW (2023–2024)
  LREC (Session; 2022)
- **Program Committee**:
  ACL (2019–)
  EMNLP (2021–)
  NAACL (2022–)
  EACL SRW (2023)
  NAACL SRW (2022)
  ARR (2021–)
  W-NUT (2021)
  CoNLL (2021–)
  LREC (2022)
  RecSysHR (2022–)
  Wise-Supervision (2022)
- **Invited Talks**:
  Talk at NLP Workshop on Linguistic Variation (2023).
  Talk at WING (National University of Singapore; 2023).
  Talk at GroNLP (University of Groningen; 2021).
- **Volunteer**:
  EMNLP (2021–2022)
  CLIN (2019)

TEACHING

Supervision:

- *BSc Thesis, Computational Linguistics* 2023
  "Skill Extraction for Vietnamese Job Market Understanding",
  T. Nguyen.

- *BSc Thesis, Computational Linguistics* 2023
  "Unsupervised Skill Extraction for Job Market Understanding",
  A.S. Barwig.
- *MSc Thesis, Computer Science/Software Design* 2022
  "Matching University Curricula to Occupations",
  S. Pasham & P. das Neves Rodrigues Mateus Cristóvão.
- *Research Project, Computer Science (KIREPRO1PE)* 2021
  "Converting Job Requirements into Skills",
  P. das Neves Rodrigues Mateus Cristóvão.

Courses:

- *Seminars in Data Science (KSSEDAS1KU)*, Lecturer 2022, 2023
- *Introduction to Natural Language Processing and Deep Learning (BSSEYEP1KU)*, Senior Teaching Assistant, Lecturer 2021, 2022
- *Communicating SOTA NLP Research to a Broader Audience (Ph.D. Course)*, Co-Organizer 2021
- *Machine Learning (SOMINDW07)*, Head Teaching Assistant 2019, 2020
- *Learning from Data (LIX016M05)*, Head Teaching Assistant 2019
- *Social Media (LIX017B05)*, Teaching Assistant 2019

CORE COMPETENCES

- **Programming**: Python (PyTorch, scikit-learn, Pandas, NumPy), R, Java, HTML
- **Tools**: Git, Elasticsearch, Agile, Scrum, LaTeX
- **Skills**: Scientific writing, research management, public speaking, data analysis
- **Languages**: Dutch (Native), English (Fluent), Spanish (Basic), Mandarin (Basic), Danish (Basic)