Mike Zhang

IT University of Copenhagen, Department of Computer Science (NLPnorth)
Rued Langgaards Vej 7, DK 2300, Copenhagen S, Denmark, Building 3E13
+45-7218-5204
mikz@itu.dk — jjzha.github.io — github.com/jjzha

PRINCIPAL INTERESTS

Automated high-quality Information Extraction from unstructured text with use cases that have societal impact.
Natural Language Processing, Machine Learning, Active Learning, Distant Supervision, Transfer Learning, annotation, technical writing, data analysis and exploration.

ACADEMIC BACKGROUND

Ph.D. Natural Language Processing 2023
IT University of Copenhagen (ITU), Copenhagen, DK

- Ph.D. research in Information Extraction advised by prof. Barbara Plank.
- Overseeing a small research team with a research assistant and annotator.

M.A. Information Science 2020
University of Groningen (RUG), Groningen, NL

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

B.Sc. Information Science 2018
University of Groningen (RUG), Groningen, NL

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

WORK EXPERIENCE

Data Engineer Jan. 2020 – Aug. 2020
Dataprovider.com B.V., Groningen, NL

- Implemented requested features into the proprietary search engine by maintaining code quality with PEP8 standards, basic unit- and integration tests, Git CI/CD, and code reviews;
- Built a classifier predicting SIC codes for company names via character-level features, using Python and scikit-learn, improved predictions by 70% accuracy;
- Spearheaded improvement of the preprocessing steps for structuring 72M unstructered company addresses with Python and Google's Geocoding API.

Data Scientist Intern Aug. 2019 – Dec. 2019
Dataprovider.com B.V., Groningen, NL

- Developed five text generation algorithms in Python for the generation of undiscovered country code top-level domains for five countries: Found ∼15K undiscovered domain names/hr;
- Obtained in total 1M+ existing, but undiscovered country code top-level domains for indexing in the proprietary search engine;
- Integrated machine learning models into Web APIs using NGINX and saved results into Elasticsearch.

PEER-REVIEWED PUBLICATIONS
***equal contribution**

4. **Mike Zhang** and Barbara Plank. "Cartography Active Learning." *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021.
3. Kristian Nørgaard Jensen,* **Mike Zhang**,* and Barbara Plank. "De-identification of Privacy-related Entities in Job Postings." *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. 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.

SERVICES

- **Reviewer**: ARR (2021, 2022), W-NUT (2021), CoNLL (2021), ACL (2019)
- **Volunteer**: EMNLP (2021), CLIN (2019)
- **Talks**:
    1. Talk at the CL group of the University of Groningen, GroNLP (2021)

TEACHING

- *Research Project, M.Sc. Computer Science (KIREPRO1PE)*, Supervision 2021
- *Communicating SOTA NLP Research to a Broader Audience (Ph.D. Course)*, Co-Organizer 2021
- *Introduction to Natural Language Processing and Deep Learning (BSSEYEP1KU)*, Senior Teaching Assistant 2021
- *Learning from Data (LIX016M05)*, Head Teaching Assistant 2019
- *Machine Learning (SOMINDW07)*, Head Teaching Assistant 2019, 2020
- *Social Media (LIX017B05)*, Teaching Assistant 2019

STUDENT SUPERVISION

1. Pedro das Neves Rodrigues Mateus Cristóvão, M.Sc. Research Project, *Converting Job Requirements into Skills*, Fall 2021

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)