Mengtian Guo

mtguo@email.unc.edu ◇ Homepage ◇ GitHub

Education

University of North Carolina at Chapel Hill — Chapel Hill, NC
Ph.D. in Information and Library Science, Focus: HCI, Information Retrieval — Aug. 2020 - Present

University of Michigan - Ann Arbor — Ann Arbor, MI
Bachelor of Science in Data Science: 3.88/4.0 — Sept. 2018 - May. 2020

Shanghai Jiao Tong University — Shanghai, China
Bachelor of Science in Electrical and Computer Engineering: 3.8/4.0 — Sept. 2016 - Aug. 2020

Internship Experience

Amazon Search — Seattle, WA
Applied Scientist Intern, Enhancing Semantic Search with Generative Query Rewriting — May. 2023 – Aug. 2023

- Designed LLM (Flan-XL) prompts to extract query rewrite pairs from customer search behavior data.
- Built a large-scale data generation workflow using Spark, a token classification model based on BERT and LLM (Flan-XL), producing a query rewriting dataset with 2.3 million instances.
- Developed and fine-tuned a transformer-based seq2seq model (FELIX) in TensorFlow for query rewrite generation.
- Achieved a 3.17% improvement in search result precision through offline testing, validating the effectiveness of the query rewriting model.

Amazon Search — Seattle, WA
Applied Scientist Intern, Data Augmentation for Semantic Search Model — May. 2022 – Aug. 2022

- Implemented three data augmentation methods on a customer search behavior dataset, domain adaptation, back translation, and text paraphrasing using T5, leveraging Spark and AWS.
- Fine-tuned a DSSM-style bi-encoder model within a production-level codebase to evaluate the impact of different data augmentation methods. Derived insights in semantic matching model characteristics compared to PECOS, including lower sensitivity to data size and higher generalizability to unknown search queries.

Publications

How Does Imperfect Automatic Indexing Affect Semantic Search Performance? — HealthNLP at ICHI 2023
Mengtian Guo, David Gotz, Yue Wang

Human-Computer Collaboration for Visual Analytics: an Agent-based Framework — EuroVis 2023
Shayan Monadjemi, Mengtian Guo, David Gotz, Roman Garnett, Alvitta Ottley

GRAFS: Graphical Faceted Search System to Support Conceptual Understanding in Exploratory Search — ACM TiiS 2023
Mengtian Guo, Zhilan Zhou, David Gotz, Yue Wang

A Design Space for Surfacing Content Recommendations in Visual Analytic Platforms — IEEE VIS 2023
Zhilan Zhou, Wenyuan Wang, Mengtian Guo, Yue Wang, David Gotz

Explainable prediction of text complexity: The missing preliminaries for text simplification — ACL-IJCNLP 2021
Cristina Garbacea, Mengtian Guo, Samuel Carton, Qiaozhu Mei

Research Projects

NLP and Information Retrieval

VACLab, UNC — Chapel Hill, NC
Research Assistant, Advisor: Prof. David Gotz, Prof. Yue Wang — Jan. 2021 - May. 2021

- Trained multi-class classification models based on Logistic Regression and BERT to assign 183 MeSH (Medical Subject Heading) terms to paper abstracts based on semantics. Applied the models to index 1.6 million PubMed articles and evaluated the performance of different semantic indexing models on 27 real-world Boolean queries.

- Demonstrated the negative impact of imperfect automatic indexing and proposed a human-machine collaborative indexing strategy that achieved 95% precision and recall with 21.33% human indexing effort. This work was published at the 6th International Workshop on Health Natural Language Processing at ICHI 2023.

Foreseer Group, University of Michigan Ann Arbor, MI

Research Assistant, Advisor: Prof. Qiaozhu Mei Mar. 2019 – Feb. 2020

- Trained and evaluated various machine learning models, including Naive Bayes, SVM, and Random Forest for text complexity classification using bag-of-words features, lexical features, and syntactic features.
- Generated complexity explanation of model predictions using LIME. This work was published at ACL-IJCNLP 2021

HCI

VACLab, UNC Chapel Hill, NC

Research Assistant, Advisor: Prof. David Gotz, Prof. Yue Wang Jan. 2022 - Present

- Designed and developed an interactive data modeling and analysis tool with a Flask API and a dynamic web-based interface to facilitate Machine Learning problem formulation.
- Implemented a Python backend for concurrent training of hundreds of regression models, supporting automatic data preprocessing, model training, and evaluation with Scikit-learn.
- Built a web-based interactive interface using JavaScript, Bootstrap, and D3.js for interactive data visualizations and model exploration.

VACLab, UNC in collaboration with LAS, NCSU Chapel Hill, NC

Research Assistant, Advisor: Prof. David Gotz, Prof. Yue Wang Jan. 2021 - Jan. 2022

- Designed and developed GRAFS (Graphical Faceted Search System) that leverages data mining methods to facilitate sensemaking and learning during exploratory search.
- Implemented a data mining algorithm to extract and summarize informative concepts from search results leveraging vocabulary lookup, clustering, and maximal marginal relevance.
- Developed a Solr search engine by indexing 33 million PubMed abstracts and a faceted search interface using Javascript and D3 with visualization of search results and concepts.
- Designed and conducted a human-subject study to demonstrate the effectiveness of the system on users' sensemaking and learning of the search topic. This work was published at ACM TiiS.

Applied Data Science

VACLab, UNC in collaboration with EPA Chapel Hill, NC

Research Assistant, Advisor: Prof. Yue Wang Jan. 2024 - Present

- Performed geospatial analysis, e.g. spatial auto-correlation and clustering, to uncover PFAS detection patterns.
- Extracted geospatial features, e.g. proximity to PFAS emission points, land cover, and hydrology, from multiple data sources. Trained Random Forest models to predict PFAS detection across US, achieving 80% recall and 73% precision.

College of Pharmacy, University of Michigan Ann Arbor, MI

Research Assistant, Advisor: Prof. Mike Dorsch Jun. 2019 - May. 2020

- Developed a Google function for automatic ASCVD (Atherosclerotic Cardiovascular Disease) risk evaluation by encoding medical guidelines into programs.

Technical Skills

- Programming Languages: Python, C/C++, Javascript, R
- LLM: PEFT (Parameter-Efficient Fine-Tuning), prompt engineering
- ML Frameworks: TensorFlow, Keras, PyTorch, Scikit-learn, Pandas, NumPy