Shutian Ma

mashutian0608@hotmail.com ◇ (408) 242-4697 ◇ shutian.me ◇ San Jose, CA, USA

Education

Nanjing University of Science and Technology — Nanjing, China
Ph.D. Information Science — 2014 - 2021

Indiana university Bloomington — Bloomington, USA
Visiting scholar, Department of Information and Library Science — 2017 - 2018

Nanjing University of Science and Technology — Nanjing, China
B.M. Information Management System — 2010 - 2014

Professional Experience

Axon — remote, USA
Software Engineer, Contractor — 10/2025 - Present

- Audio to Text Summarization: Conduct Audio LLM research and experiments, including audio data processing and model evaluation.

Meta — remote, USA
GenAI Language Annotation Specialist, Contractor — 05/2025 - Present

- Data Annotation: Rated and compared generative AI model responses based on linguistic standards. Contributed Simplified Chinese LLM data annotation to support research that led to a Meta AI publication.

Freelance — remote, USA
AI Pipeline Deployment Engineer and Research Mentor — 07/2024 - 05/2025

- AI Model Deployment: Deployed and customized ComfyUI on Linux servers and Runpod serverless platform. Built tailored Docker images with automation scripts to streamline client-specific deployment pipelines.
- Research Mentor: Academic mentorship in LLMs, multimodal AI, and data mining.

Tencent — Shenzhen, China
Data Scientist, Full Time — 07/2020 - 08/2023

- Text classification: Implemented the user answer classification using fastText. This project provided labeled data for the company's annual internal IT user research for 3 consecutive years.
- Named Entity Recognition: Led a team of 4 outsourced members in data annotation, implemented brand and snack name recognition in UGC and supported the Smart Retail department in product selection.
- LLM Fine-tuning: Constructed the training dataset for fine-tuning and developed a conversational model by applying LoRA to fine-tune ChatGLM-6B on domain-specific texts.

ByteDance — Beijing, China
Machine Learning Engineer, Intern — 11/2018 - 01/2019

- Ads recommendation: Developed a user query classification system integrating user click behavior to recommend relevant ads. The algorithm was deployed in the product, improving ad targeting efficiency.

Selected Publications

1. Zhang, K., R. Zhu, **S. Ma**, J. Xiong, Y. Kim, F. Murai et al. *Kedrec-lm: A knowledge-distilled explainable drug recommendation large language model.* arXiv preprint arXiv:2502.20350. (2025). Accepted by SIGIR 2025 NIP@IR.
2. **Ma, S.**, C. Zhang, H. Zhang, and Z. Gao. *Citation recommendation based on argumentative zoning of user queries.* Journal of Informetrics 19, no. 1 (2025): 101607.
3. **Ma, S.**, H. Zhang, C. Zhang, and X. Liu. *Chronological citation recommendation with time preference.* Scientometrics 126 (2021): 2991–3010.
4. **Ma, S.**, C. Zhang, and X. Liu. *A review of citation recommendation: from textual content to enriched context.* Scientometrics 122, no. 3 (2020): 1445–1472.
5. **Ma, S.**, H. Zhang, T. Xu, J. Xu, S. Hu, and C. Zhang. *IR&TM-NJUST@ CLSciSumm-19.* BIRNDL@ SIGIR 2414 (2019): 181–195.

6. Börner, K., O. Scrivner, M. Gallant, **S. Ma**, X. Liu, K. Chewning, L. Wu, and J. A. Evans. *Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy.* Proceedings of the National Academy of Sciences 115, no. 50 (2018): 12630–12637.
7. **Ma, S.**, J. Xu, and C. Zhang. *Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset.* Scientometrics 116 (2018): 1303–1330.
8. **Ma, S.**, Y. Zhang, and C. Zhang. *Using multiple Web resources and inference rules to classify Chinese word semantic relation.* Information Discovery and Delivery 46, no. 2 (2018): 120–126.
9. **Ma, S.**, and C. Zhang. *Using Full-text Academic Articles and Wikipedia to Find Alternative Free Bioinformatics Software.*
10. **Ma, S.**, H. Zhang, J. Xu, and C. Zhang. *NJUST@ CLSciSumm-18.* BIRNDL@ SIGIR 2018 (2018): 114–129.
11. **Ma, S.**, J. Xu, J. Wang, and C. Zhang. *NJUST @ CLSciSumm-17.* In Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017), Tokyo, Japan, 2017: 16–25.
12. **Ma, S.**, and C. Zhang. *Using Full-text to Evaluate Impact of Different Software Groups.* In ISSI, pp. 1666–1667. 2017.
13. **Ma, S.**, and C. Zhang. *Document representation and clustering models for bilingual documents clustering.* Proceedings of the Association for Information Science and Technology 54, no. 1 (2017): 499–502.
14. **Ma, S.**, C. Zhang, and D. He. *Document representation methods for clustering bilingual documents.* Proceedings of the Association for Information Science and Technology 53, no. 1 (2016): 1–10.
15. **Ma, S.**, X. Zhang, and C. Zhang. *NLPCC 2016 Shared Task Chinese Words Similarity Measure via Ensemble Learning Based on Multiple Resources.* In Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24, pp. 862–869. Springer International Publishing, 2016.
16. **Ma, S.**, and C. Zhang. *Automatic Collection of the Parallel Corpus with Little Prior Knowledge.* In International Symposium on Natural Language Processing Based on Naturally Annotated Big Data, pp. 95–106. Cham: Springer International Publishing, 2014.

Selected Services

PC Member

- JCDL 2020-2024, EEKE-AII 2023-2024, EEKE 2021-2022, JDIS

Peer Reviewer

- WSDM 2023-2025, AAAI 2023, SIGIR 2022, ICME 2022, PACIS 2021, DLP 2023, DLP-KDD 2020-2021, IEEE BigData 2020, SDP workshop at EMNLP 2020, JIST 2019, PACLIC 33, BIRNDL-2017, Scientometrics, Information Processing and Management, PLOS One, The Electronic Library