# Shutian Ma

Email: mashutian0608@hotmail.com | https://shutian.me/ | San Jose, CA | (408) 242-4697

## Brief Bio

**Research Area:** Machine Learning, Deep Learning, Natural Language Processing, Large Language Model
**Programing Language**: Python, Java, SQL, Shell, PyTorch

## Education

**Ph.D., Information Science, Nanjing University of Science & Technology, China** **2014-2021**
- Department of Information Science, Supervisor: Chengzhi Zhang

**Visiting scholar, Indiana university Bloomington, USA** **2017-2018**
- Cyberinfrastructure for Network Science (CNS) Center, Department of Information and Library Science, School of Informatics, Computing, and Engineering, Supervisor: Katy Börner, Xiaozhong Liu

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

## Experience

**GenAI Language Annotation Specialist, Contractor, Meta, USA** **05/2025-now**
- ***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.

**AI Pipeline Deployment Engineer & Research Mentor, Freelance, Remote** **07/2024-05/2025**
- ***Model Deployment***: Deployed and customized ComfyUI on both Linux servers and the Runpod serverless platform. Integrated custom nodes and models (e.g., FLUX via Diffusers). Built tailored Docker images with automation scripts to streamline client-specific deployment pipelines, leveraging the ComfyUI-SD3 image for scalable inference.
- ***Research Mentor***: Academic mentorship in LLMs, multimodal AI, and data mining, guiding research projects, optimizing model performance.

**Machine Learning Engineer, Full-time, Wuxi Lead Intelligent Equipment, China** **10/2023-12/2023**
- ***RAG-based LLM***: Parsed PDF files to construct a full-text dataset and implemented a Retrieval-Augmented Generation approach using Langchain. The system retrieves relevant text blocks from the dataset based on vectorized query results, which are then passed to an LLM for answer generation.
- ***Knowledge based QA***: Constructed a question-answer pair knowledge base/ knowledge graph using unstructured domain-specific texts. The system returns direct answers based on user queries.

**Data Scientist, Full-time, Tencent, China** **07/2020-08/2023**
- ***Classification of answers to open-ended questions***: 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, supporting insights through efficient categorization of responses.
- ***Brand and Snack 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.
- ***ChatGLM with LoRA 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.

**Machine Learning Engineer, Intern, Aegis, China** **03/2020-04/2020**
- ***Legal Entity Recognition***: Deployed an annotation platform using Doccano, implemented NER in legal text using BiLSTM+CRF.

**Machine Learning Engineer, Intern, ByteDance, China** **11/2018-01/2019**
- ***Ads recommendation***: Developed a user query classification system using fastText, integrating user click behavior to recommend relevant ads. The algorithm was deployed in the product, improving ad targeting efficiency.

## Selected Publications (over 500 citations)


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). ***Accpeted by SIGIR 2025 NIP@IR***
2. **Ma, Shutian**, Chengzhi Zhang, Heng Zhang, and Zheng Gao. "Citation recommendation based on argumentative zoning of user queries."Journal of Informetrics 19, no. 1 (2025): 101607.
3. Chaoguang Huo, Shutian Ma, and Xiaozhong Liu. "Hotness prediction of scientific topics based on a bibliographic knowledge graph." Information Processing & Management 59, no. 4 (2022): 102980.
4. Zheng Gao, Chun Guo, **Shutian Ma**, and Xiaozhong Liu. "Improving Community Detection Performance in Heterogeneous Music Network by Learning Edge-Type Usefulness Distribution." In International Conference on Information, pp. 68-78. Cham: Springer International Publishing, 2022.

5. **Shutian Ma**, Heng Zhang, Chengzhi Zhang, and Xiaozhong Liu. "Chronological citation recommendation with time preference." Scientometrics 126 (2021): 2991-3010.
6. Chaoguang Huo, Xiaozhong Liu, and **Shutian Ma**. "How Bibliographic Features Contribute to Scientific Topic Prediction." (2021).
7. Katy Börner, Olga Scrivner, Leonard E. Cross, Michael Gallant, **Shutian Ma**, Adam S. Martin, Lisel Record, Haici Yang, and Jonathan M. Dilger. "Mapping the co-evolution of artificial intelligence, robotics, and the internet of things over 20 years (1998-2017)." PloS one 15, no. 12 (2020): e0242984.
8. Heng Zhang, Lifan Liu, Ruping Wang, Shaohu Hu, **Shutian Ma**, and Chengzhi Zhang. "IR&TM-NJUST@ CLSciSumm 20." In Proceedings of the First Workshop on Scholarly Document Processing, pp. 288-296. 2020.
9. **Shutian Ma,** Chengzhi Zhang, and Xiaozhong Liu. "A review of citation recommendation: from textual content to enriched context." Scientometrics 122, no. 3 (2020): 1445-1472.
10. Chengzhi Zhang, Zijing Yue, Qingqing Zhou, **Shutian Ma**, and Zi-Ke Zhang. "Using social media to explore regional cuisine preferences in China." Online Information Review 43, no. 7 (2019): 1098-1114.
11. Heng Zhang, **Shutian Ma**, and Chengzhi Zhang. "Using Full-text of Academic Articles to Find Software Clusters." In ISSI, pp. 2776-2777. 2019.
12. **Shutian Ma**, Heng Zhang, Tianxiang Xu, Jin Xu, Shaohu Hu, and Chengzhi Zhang. "IR&TM-NJUST@ CLSciSumm-19." BIRNDL@ SIGIR 2414 (2019): 181-195.
13. Katy Börner, Olga Scrivner, Mike Gallant, **Shutian Ma**, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James 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.
14. **Shutian Ma,** Jin Xu, and Chengzhi Zhang. "Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset." Scientometrics 116 (2018): 1303-1330.
15. **Shutian Ma**, Yingyi Zhang, and Chengzhi Zhang. "Using multiple Web resources and inference rules to classify Chinese word semantic relation." Information Discovery and Delivery 46, no. 2 (2018): 120-126.
16. **Shutian Ma**, and Chengzhi Zhang. "Using Full-text Academic Articles and Wikipedia to Find Alternative Free Bioinformatics Software."
17. **Shutian Ma**, Heng Zhang, Jin Xu, and Chengzhi Zhang. "NJUST@ CLSciSumm-18." BIRNDL@ SIGIR 2018 (2018): 114-129.
18. **Shutian Ma**, Jin Xu, Jie Wang and Chengzhi 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.
19. **Shutian Ma**, and Chengzhi Zhang. "Using Full-text to Evaluate Impact of Different Software Groups." In ISSI, pp. 1666-1667. 2017.
20. **Shutian Ma**, and Chengzhi Zhang. "Documents representation for comparable corpora clustering: A preliminary study." iConference 2017 Proceedings (2017).
21. **Shutian Ma**, and Chengzhi 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.
22. **Shutian Ma**, Chengzhi Zhang, and Daqing He. "Document representation methods for clustering bilingual documents." Proceedings of the Association for Information Science and Technology 53, no. 1 (2016): 1-10.
23. **Shutian Ma**, Xiaoyong Zhang, and Chengzhi 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.
24. **Shutian Ma**, and Chengzhi 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