Xuemei **Tang**

INFORMATION SCIENCE · DIGITAL HUMANITIES
No. 5 Yiheyuan Road, Haidian District, Beijing, PRC, 100871
☎ (+86) 18210374514 | ✉ tangxuemei@stu.pku.edu.cn

Education

Peking University *Beijing, China*
PHD INFORMATION SCIENCE *Sept. 2020 - Jun. 2024*

- Research Interests: Natural Language Processing, Digital Humanities

Beijing Normal University *Beijing, China*
MA LINGUISTICS AND APPLIED LINGUISTICS *Sept. 2017 - Jun. 2020*

- Research Interests: Natural Language Processing, Digital Humanities

Tianjin Normal University *Tianjin, China*
BA INFORMATION ENGINEERING & CHINESE LANGUAGE AND LITERATURE *Sept. 2012 - Jun. 2016*

- Main course: C Programming; Computer Network; Computer Graphics; Signals and Systems
- Minor course: Ancient Chinese; Introduction to Literature; Ancient Chinese Literature; Modern Chinese Literature; History of Foreign Literature.

Publications

- **Xuemei Tang**, Qi Su, Jun Wang. 2024. ***Chinese Word Segmentation with Heterogeneous Graph Convolutional Network***.**Natural Language Processing (formerly known as Natural Language Engineering)**, (SSCI, Accepted).
- **Xuemei Tang**, Qi Su, Jun Wang. 2024. ***Incorporating Deep Syntactic and Semantic Knowledge for Chinese Sequence Labeling with GCN***. **Aslib Journal of Information Management**, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/AJIM-07-2023-0263(SSCI).
- **Xuemei Tang**, Zekun Deng, Qi Su, Jun Wang. 2024. ***CHisIEC: An Information Extraction Corpus for Ancient Chinese History***. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (**LREC-COLING 2024**, CCF-B), pp 3192–3202.
- **Xuemei Tang**, Qi Su. 2022. ***That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory***. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (**ACL**, CCF-A)(Volume 1: Long Papers), pp 7830–7840.*
- **Xuemei Tang**, Qi Su. 2022. ***A Metrological Study on the Spatial Narrative of the Qishu Genre: Take A Dream of Red Mansions and Water Margin as examples***. *Workshop on Chinese Lexical Semantics **CLSW** 2022: Chinese Lexical Semantics, pp 326–336.*
- **Xuemei Tang**, Qi Su, Jun Wang, Yuhang Chen, Hao Yang. 2021. ***Automatic Traditional Ancient Chinese Texts Segmentation and Punctuation Based on Pre-training Language Model***. Proceedings of the 20th Chinese National Conference on Computational Linguistics (**CCL**), pp 678–688.
- **Xuemei Tang**, Qi Su, Jun Wang. ***AnChineseNERE: An Ancient Chinese Corpus with Named Entity and Relation Annotation***. (Submitted to Journal of Digital Scholarship in the Humanities, Now Minor Revision)
- **Xuemei Tang**, Qi Su, Jun Wang. 2024. ***An Effective Incorporating Heterogeneous Knowledge Curriculum Learning for Sequence Labeling***. http://arxiv.org/abs/2402.13534.
- **Xuemei Tang**, Jun Wang, Qi Su. 2024. **Small Language Model Is a Good Guide for Large Language Model in Chinese Entity Relation Extraction**. http://arxiv.org/abs/2402.14373.

Participated Grants

The Construction of the Knowledge Graph for the History of Chinese Confucianism *Natural Science Foundation of China (NSFC), China*
Jan. 2020 - PRESENT

- I am responsible for the meticulous work of collecting and cleaning ancient Chinese data. In addition, I developed complex models for automatic text segmentation, punctuation, word segmentation, and relation extraction within the framework of the Intelligent Annotation Platform for Chinese Ancient Texts.

Automatic Knowledge Graph Generation Platform for Canonical Texts *Ministry of Education, China*
Oct. 2021 - Oct. 2022

- Firstly, I investigated the Knowledge Graph platform and wrote part of the project proposal. Second, I developed the relation extraction model for the knowledge graph platform. Finally, I wrote and revised closure reports.

Personal Statement

- Currently, NLP mostly focuses on English, while the Chinese NLP community generally focuses on the processing of modern Chinese, and the application of NLP in ancient Chinese is less effective. Therefore, I focus on this research area to develop efficient and accurate NLP models. My key works are as follows.
- **That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory.** The evolution of language follows gradual change rules. Grammar, vocabulary, and lexical semantics shift over time, creating a diachronic linguistic gap. Consequently, texts in different era languages pose challenges for natural language processing tasks like word segmentation and machine translation. Therefore, we propose a cross-era learning framework for Chinese word segmentation (CWS), CROSSWISE, which uses the Switch-memory (SM) module to incorporate era-specific linguistic knowledge. The framework effectively integrates the knowledge of the eras into the neural network.
- **An Effective Incorporating Heterogeneous Knowledge Curriculum Learning for Sequence Labeling.** Sequence labeling models often benefit from incorporating external knowledge. However, this practice introduces data heterogeneity and complicates the model with additional modules, increasing expenses for training a high-performing model. To address this challenge, we propose a two-stage curriculum learning (TCL) framework designed for sequence labeling tasks. The TCL framework enhances training by gradually introducing data instances from easy to hard, aiming to improve both performance and training speed.
- **Incorporating Deep Syntactic and Semantic Knowledge for Chinese Sequence Labeling with GCN.** Little attention has been paid to the utility of hierarchy and structure information encoded in syntactic and semantic features for Chinese sequence labeling tasks. In this paper, we propose a novel framework to encode syntactic structure features and semantic information for Chinese sequence labeling tasks with graph convolutional networks (GCN).
- **CHisIEC: An Information Extraction Corpus for Ancient Chinese History.** Natural Language Processing (NLP) plays a pivotal role in the realm of Digital Humanities (DH) and serves as the cornerstone for advancing the structural analysis of historical and cultural heritage texts. This is particularly true for the domains of named entity recognition (NER) and relation extraction (RE). In our commitment to expediting ancient history and culture, we present the “Chinese Historical Information Extraction Corpus” (CHisIEC). CHisIEC is a meticulously curated dataset designed for the development and evaluation of NER and RE tasks, offering a resource to facilitate research in the field. We also explored the capabilities of large language models (Alpaca2, ChatGLM2, GPT-4) for RE in ancient Chinese.
- **Small Language Model Is a Good Guide for Large Language Model in Chinese Entity Relation Extraction.** Large language models (LLMs) have been successful in relational extraction (RE) tasks, especially in the few-shot learning. An important problem in the field of RE is long-tailed data, while not much attention is currently paid to this problem using LLM approaches. Therefore, in this paper, we propose SLCoLM, a model collaboration framework, to mitigate the data long-tail problem. In the SLCoLM framework, we use the “*Training-Guide-Predict*” strategy to combine the strengths of pre-trained language models (PLMs) and LLMs, where a task-specific PLM framework acts as a tutor, transfers task knowledge to the LLM and guides the LLM in performing RE tasks.
- Throughout my research journey, I have gained a wealth of experience in model design and tuning by learning a variety of techniques such as language models, curriculum learning, and graph neural networks. Over the past year, I have gained some experience in fine-tuning large language models. I fine-tuned the large language models LLama and ChatGLM with instructions and used GPT-3.5 and GPT-4 to explore their performance on specific tasks through In-contextual learning. In the future, I will continue to explore research in the direction of large language models.

Honors & Awards

Domestic

2022	**Merit Student**, Peking University	*Beijing, China*
2022	**Leo KoGuan Scholarship**, Peking University	*Beijing, China*

Skills

Deep Learning Framework	Pytorch, Tensorflow
NLP Technique	Graph Neural Networks, Pre-trained Language Model, Large Language Model, Curriculum Learning
NLP Task	Sequence Labeling, Text Classification, Text Generation, Information Extraction
Others	Gephi, Ontology, Semantic Network, Neo4j, Protege, Topic Model

Other Experiences

2023	**Reviewer**, ACL 2023
2023	**Reviewer**, EMNLP 2023
2023	**PC Member**, EMNLP 2023 Industry Track
2023	**Reviewer**, Journal of Neural Computing and Applications (NCAA)