Qisheng Liao

Phone: +971 585933506, +1 908 793 9748, +86 135 0935 0177
E-mail: Qisheng.Liao@mbzuai.ac.ae
Address: Masdar City, Abu Dhabi, UAE

EDUCATION

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), United Arab Emirates 2022-2024
Master of Science in Natural Language Processing
Supervised by Tim Baldwin, and Muhammad Abdul-Mageed,

New York University (NYU), United States 2020-2022
Master of Science in Computer Science
Member of Music X Lab at NYU Shanghai supervised by Gus Xia

University of California Santa Cruz (UCSC), United States 2016-2020
Bachelor of Science in Computer Science (Highest Honor)

PUBLICATION

Zhang, C., Doan, K. *, **Liao, Q. ***, & Abdul-Mageed, M. The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages. *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
* Equal contribution

Liao, Q., Xia, G., & Wang, Z. Calliffusion: Chinese Calligraphy Generation and Style Transfer with Diffusion Modeling. *Proceedings of the 14th International Conference on Computational Creativity (ICCC)*, 2023.
The video proposal of this work is accepted to Neural IPS 2023 Creative AI Track.

Liao, Q., Lai, M., & Nakov, P. MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval)*, 2023

TEACHING EXPERIENCE

Teaching Assistant & Lab Session Instructor
- Advanced Natural Language Processing. Spring 2023, MBZUAI
- Data Science for Soc. and Info. Networks. Spring 2022, NYU Shanghai
- Network Analytics. Fall 2021, NYU Shanghai

RESEARCH EXPERIENCE

Language Identification
- Unknown Language Identification (Ongoing)
  We try to design a model that can not only classify more than 100 main languages but classify unknown languages when a language that is not included in training is passed to the model. It is a hard problem because there are lots of similar languages (language used same scripts or language from the same language family). How to deal with them is the main part for this research project.

- An Investigation into ChatGPT's Language Identification Ability (Ongoing)
  We try to evaluate more than 500 languages with ChatGPT and investigate the abilities of ChatGPT for low resources languages. We also compare the performance of ChatGPT with other pretrained language identification tools.

- Contrastive Learning for Unseen Language Identification & Arabic Dialect Identification**.**
  We finetuned RoBERTa model in flore-101 & NADI 2021 datasets with contrastive loss. In contrastive learning, we consider different languages as negative samples. While for positive samples, we used the same setting as SimCSE paper that the same sentence with different random dropout are two positive sentences. We also consider languages from the same language family to be hard negative examples and add more loss to them.

Semantic Matching
Syntax Relations in Semantic Matching Tasks
- Modified BERT structure and added information from syntax trees to the model.
  - We designed a new Graph Convolution Network that can represent syntax tree information to get syntax matrix. We

used cross-attention and gates module to calibrate the weight of semantic matrix. The modification is in the bottom three transformer layers of BERT.
- Experimental results on 10 standard benchmarks demonstrate that our model performs better in semantic matching tasks.

This project is submitted to ICASSP 2024.

Sociopragmatic Understanding

- A Study of Sociopragmatic Understanding in LLMs
  - We present an extensive multilingual benchmark specifically designed for sociopragmatic meaning understanding, which includes 169 datasets in 64 languages from 12 language families, 16 types of scripts.
  - We did extensive zero-shot evaluation on 13 models and finetuning evaluation on 4 models. The results are compared in different ways such as confusion matrix, etc.

  The result is published: *The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages.*

Deep Learning for Chinese Calligraphy and Handwriting Texts

- Diffusion Models for Chinese Calligraphy and Handwriting Texts (Ongoing)
  - We designed diffusion models for Chinese Calligraphy generation conditioned with characters, styles, and authors.
  - We have several settings for this model and one of them is a Denoising Diffusion Probabilistic Model with natural language text as the condition. The text is used to describe characters, scripts, and styles of the calligraphy we want to generate.
  - We tried Low-Rank Adaptation with our pretrained model. The results show that any symbols from other languages or untrained characters can be adapted to Chinese Calligraphy with by one-shot or few-shots finetuning with our model.
  - We did subjective and objective evaluation. In objective evaluation, we used our own model for Chinese calligraphy recognition, the results showed, the performance of our generated model is similar to the performance of real samples. The result of the subjective survey showed that people cannot distinguish our generated characters with real samples.

  The initial result is published: *Calliffusion: Chinese Calligraphy Generation and Style Transfer with Diffusion Modeling*

- Chinese Calligraphy Scripts and Characters Recognition.
  - We designed a multitask model to predict scripts and characters jointly. We firstly predicted the scripts then combined the representation of scripts to the representation of characters to make the final decision. The top-1 script accuracy is 0.90 and the top-1 character accuracy is 0.83.
  - We designed algorithms to recognize the whole artwork instead of a single character. By this algorithm, in the best case, both the script and character accuracy can be improved to 0.99.
  - We designed a supervised contrastive model for artwork generation based on existing data. An artwork can be generated based on the contrastive loss between each calligraphy, because similar calligraphy would have small contrastive loss.

Propaganda Framing

- Propaganda Framing for Data from European Union (Ongoing)

  We try to do propaganda framing to 2M data from European Union. We used the contrastive model that is used to get good results for SemEval Shared Task.

- SemEval 2023 Shared Task, Task3, Sub-Task 2: Framing Detection
  - We designed a multilabel multitask contrastive learning model. Based on SimCLR and SimCSE, we modified the loss function to make it fit for multilabel task setting.
  - Our system was ranked first on the official test set and on the leaderboard for five of the six languages.

  The result is published: *MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning*

Text to Speech

- Text to Speech System with Diffusion Models (Ongoing)

  We try to use diffusion models to do text to speech. Previously, the main method to do text to speech is an architecture with phoneme, Mel-spectram and HiFi-GAN. Our proposed approach is a new architecture that we use alternative representation instead of Mel-spectram to do speech generation.