Hui-Syuan Yeh

PhD Student
BioNLP & Knowledge Base

☎ +33774951916
Paris, France
hui-syuan yeh
✉ katherine5393@gmail.com
Dotkat-dotcome
homepage

Summary

3rd year Ph.D. student specializing in BioNLP and Knowledge Bases, seeking industrial R&D experience. Keen interest in applying language processing skills to support healthcare.

Education

2021 - 2024 **PhD in Computer Science** **Université Paris-Saclay, France**

- **Project:** Propmt-based learning approach for biomedical relation extraction, generating answer words associated to each relation with a dependency parser, achieving +13% f1 gained for vanilla finetuning. *Accepted at LREC 2022.*
- **Project:** Accessing the technical levels of medical terms with models of different biomedical specialty, and prompts of different technical levels. *It is documented as an internal pilot study. The project is in progress to scale in terms of annotated samples and into four languages.*
- **Project:** Curating an adverse drug reaction dataset from patients' perspectives in three languages, including French, German, and Japanese. *Accepted at LREC 2024.*
- **Project:** Exploring imbalanced dataset for inter-sentence relation extraction. *On-going.*
- **Project:** Exploring using knowledge base description to help extracting information from non-English corpus. More specifically, I invistigate the question if translating the corpus to English, which expands the coverage of knowledge base to the corpus, improves the knowledge integration. The experiments are conducted with both Bert-based models and Flan-T5. *On-going.*
- Participated in National NLP Clinical Challenges. N2C2. Nov 2022. *Oral presentation at AMIA 2022.*
- Invited talk at MaiAGE, INRAE, France. May 2022.
- Invited talk at LOPE, NTU, Taipei, Taiwan. Jan 2023.
- Visiting Researcher at Social Computing, NAIST, Nara, Japan. Feb 2023 - Apr 2023.
- Visiting Researcher at Speech and Language Technology, DFKI, Berlin, Germany. Jun 2023 - Aug 2023.
- Co-organizing the shared task NTCIR-17 MedNLP-SC. 2023.
- Co-organizing the shared task SMM4H:Cross-Lingual Few-Shot Relation Extraction for Pharmacovigilance in French, German, and Japanese.. 2024.
- On track to complete my PhD by **October 2024**.

2017 – 2020 **M.Sc. - in Computer Science** **Universität des Saarlandes, Germany**

- **IMPRS Scholarship** for Master's Degree in Computer Science by MPI für Informatik.
- **Relevant Coursework**: Statistical learning, Statistical Natural Learning Processing, Neural Network: Theory and Implementation, Information Retrieval, a selection of project-based seminars.
- **Thesis**: Model-Based Generation of Microscopy Images for Training Deep Neural Network. Implemented the generative model synthesizing samples of breast cells in microscopy images with image processing techniques.

2011 – 2015 **B.Sc. - in Mathematics** **National Taiwan Normal University, Taiwan**

- **Scholarship** for Advanced Summer Course in Probability by National Center for Theoretical Sciences 2013.
- **Relevant Coursework**: statistics, probability, linear algebra and calculus.

Work Experience

2020 – 2021 **NLP Research Assistent** **Coli at Saarland University, Germany**

- Carried out experiments according to protocols laid out by primary researchers.
- Conducted statistical analyses on performance and datasets.
- Formatted and collected experiment data.
- Reviewed research literature and online tools.
- Participated in the editing and proofreading of research pre-prints.

2016 – 2017 **NLP Project Manager** **Crowdinsight.Inc., Taiwan**

- Organized 3 projects based on web-crawling and NLP attracting 10+ potential angel investors.
- Designed 5+ data analysis/IOT prototypes and proposals in different domains, including retail, airline.
- Invited talk on Machine Learning at NCCU, Taiwan

2015 – 2016 **Maths Teacher Intern** **Taipei Municipal Jianguo High School, Taiwan**

Technical Skills

- Regular job submissions with **Slurm** to an HPC cluster for LLM training and inference.
- Experience with building information extraction dataset, including strategic planning, iterative updating guidelines.
- **ML Libraries:** PyTorch, HuggingFace (transformers, peft), OpenPrompt, sklearn, Numpy, spaCy, scipy, NLTK, Weights & Biases
- **Programming:** Python
- **Other resources/frameworks:** SQL, Unix/Linux, Git, Github, Brat

Publications

- _H. Yeh* & L. Raithel* et al._ **A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages** LREC-COLING 2024. Awaiting conference proceedings. (link)
- _S. Wakamiya & L. Raithel & H. Yeh et al._ **Ntcir-17 mednlp-sc social media adverse drug event detection: Subtask overview**. Workshop at NTCIR 2023. (link)
- _H. Yeh* & G. H. B. Andrade* & F. W. Mutinda* & L. Raithel* et al._ **KEEPHA at n2c2 2022: Track 1**. n2c2 Shared Task and Workshop at AMIA Fall Symposium 2022. (link)
- _H. Yeh et al._ **Decorate the Examples: A Simple Method of Prompt Design for Biomedical Relation Extraction**. LREC 2022. (link)
- _E. Chang & A. Kovtunova & S. Borgwardt & V. Demberg & K. Chapman & H. Yeh et al._ **Logic-Guided Message Generation from Raw Real-Time Sensor Data**. LREC 2022. (link)
- _E. Chang* & Y. Shiue* & H. Yeh et al._ **Time-Aware Ancient Chinese Text Translation and Inference**. Workshop on Computational Approaches to Historical Language Change associated with ACL 2021.(link)
- _E. Chang & H. Yeh et al._ **Does the Order of Training Samples Matter? Improving Neural Data-to-Text Generation with Curriculum Learning**. EACL 2021. (link)
- _E. Chang & X. shen & H. Yeh et al._ **On Training Instance Selection for Few-Shot Neural Text Generation**. ACL short 2021. (link)

Additional Information

- **Languages**: _Mandarin_ (native), _English_ (fluent), _German_ (beginner), _French_ (beginner).
- **Extracurricular**: dance, cooking.