

# Langlin Huang

Email:h.langlin@wustl.edu | GitHub:https://shranso.github.io/

## EDUCATION

---

- **Washington University in St. Louis** St. Louis, U.S.  
Incoming student for Ph.D. in Computer Science Aug. 2024 - present
- **Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)** Beijing, China  
M.E. in Computer Science and Technology -GPA: 3.75/4.0, Outstanding Dissertation Sep. 2021 - Jun. 2024
- **University of International Business and Economics** Beijing, China  
B.E. in Data Science and Big Data Technology -GPA: 3.74/4.0 (Rank:2/147) Sep. 2017 - Jun. 2021

## RESEARCH INTERESTS

---

- Machine Translation & Multilingual Representation Learning
- Language Model Reasoning & Reliable Language Generation

## PUBLICATIONS

---

- **Integrating Multi-scale Contextualized Information for Byte-based Neural Machine Translation**  
Langlin Huang, Yang Feng  
ACL findings, 2024 [Paper][Code]
- **Enhancing Neural Machine Translation with Semantic Units**  
Langlin Huang, Shuhao Gu, Zhuocheng Zhang, Yang Feng  
EMNLP findings, 2023 [Paper][Code]
- **BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models**  
Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangdong Gui, Yunji Chen, Xilin Chen, Yang Feng  
Preprint edition on arXiv. Jun. 2023 [Paper] [Code]
- **Automatic Construction of a Depression-Domain Lexicon Based on Microblogs: Text Mining Study**  
Genghao Li, Bing Li, Langlin Huang, Sibing Hou  
JMIR medical informatics, 2020, Vol 8. Jun. 2020 [Paper]

## PROJECTS

---

- **BayLing: On the Multi-lingual Ability & Multi-turn Interaction of Large Language Models** Apr. 2023 - Jun. 2023  
Exploited the language-aligning potential of translation data for improving the **multi-lingual ability of LLMs**;  
Constructed interactive translation data and leveraged it to enhance LLM's instruction following ability.
  - **Contributions:** Sifted high-quality translation data with statistical and model-based metrics.  
Found the few high-quality translation data magic, efficiently endowing LLaMA with new language capability.
  - **Achievement:** Released BayLing, a multilingual & interactive LLM finetuned with a few data based on LLaMA.
  - **Project link:** <https://github.com/ictnlp/BayLing/tree/main>
- **CVAE-based Label Smoothing for Neural Machine Translation** Feb. 2022 - Aug. 2022  
Proposed a flexible label smoothing for training language models and translation models.
  - **Contributions:** Proposed to replace uniform distributions with predicted real label distributions in label-smoothed cross-entropy loss.  
Proposed to predict real label distribution with a Conditional Variational Auto Encoder(CVAE) module by foreseeing the ground truth word.
  - **Achievement:** Significantly improved translation performance by 1.2 BLEU on En-Ro and Zh-En translation tasks.
  - **Patent Link:** <http://epub.cnipa.gov.cn/patent/CN115455993A>
- **Chinese-Thai Translation System** May. 2022 - Jul. 2022  
Developed strong Chinese-Thai bidirectional machine translation systems.
  - **Contributions:** Proposed a strategy to modify pre-trained language model mBART, without hurting performance.  
Crawled external in-domain texts and augmented training data via back-translation.
  - **Achievement:** Won the **Championship** in the 18th China Conference on Machine Translation(CCMT) Zh-Th track.
  - **Technical Report link:** [http://sc.cipsc.org.cn/mt/conference/2022/papers/test\\_paper/60/60\\_Paper.pdf](http://sc.cipsc.org.cn/mt/conference/2022/papers/test_paper/60/60_Paper.pdf)
- **Automatic Construction of a Depression-Domain Lexicon Based on Microblogs** Jun. 2019 - Jun. 2020  
Constructed a depression-domain lexicon, starting from few seed words, by analyzing Weibo texts.
  - **Contributions:** Crawled a large amount of depression domain texts from microblog (Sina Weibo).  
Leveraged word2vec and label propagation algorithm to enlarge depression lexicon iteratively.
  - **Achievement:** Proposed a depression domain lexicon with more than 500 words, helping significantly improve online depression detection.
  - **Paper link:** <https://medinform.jmir.org/2020/6/e17650>

## TECHNICAL SKILLS

---

**Master:**Python, Pytorch, C, C++, Pandas, Data Analysis & Visualization

**Proficient:**JAVA, R, Shell, LaTeX, Web Scraper