Shaoyang Xu

Tianjin University, No. 92 Weijin Road, Xuefu Street, Nankai District, Tianjin

syxu@tju.edu.cn | August 2000 | Homepage | Github | Google Scholar

Education

Tianjin University (985 Project) *Tianjin, China*

Master (Recommendation for Admission), Computer Science *Sept 2022 - Jan 2025 (expected)*

Supervisor: Prof. Deyi Xiong

Soochow University (211 Project) *Soochow, China*

Bachelor, Artificial Intelligence *Sept 2018 - Jun 2022*

Supervisor: Prof. Peifeng Li & Dr. Feng Jiang
GPA: 3.8/4.0 (rank: 3/70)

Publications

Exploring Multilingual Human Value Concepts in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?
Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, Deyi Xiong
Submitted to The 2024 Conference on Association for Computational Linguistics (ACL 2024)

Mitigating Privacy Seesaw in Large Language Models: Augmented Privacy Neuron Editing via Activation Patching
Xinwei Wu, Weilong Dong, **Shaoyang Xu**, Deyi Xiong
Submitted to The 2024 Conference on Association for Computational Linguistics (ACL 2024)

Language Representation Projection: Can We Transfer Factual Knowledge across Languages in Multilingual Language Models?
Shaoyang Xu, Junzhuo Li, Deyi Xiong
Accepted to The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

Topic Segmentation via Discourse Structure Graph Network
Shaoyang Xu, Feng Jiang, Peifeng Li
Accepted to Journal of Chinese Information Processing, 2021

Research Projects

Multilingual Human Value Concepts in Large Language Models (1st author, submitted to acl2024) *Oct 2023 - Feb 2024*

- Inspired by studies unveiling that LLMs encode representations of human value concepts, proposed a framework to empirically **explore the existence of multilingual human value concepts in LLMs and perform cross-lingual analysis on these concepts**.
- Conducted experiments on **7 human value concepts** (morality, deontology, utilitarianism, fairness, truthfulness, toxicity and harmfulness), across **16 languages** and **3 LLM families** with different patterns of multilinguality (**LLaMA2-chat-7B, 13B, 70B; Qwen-chat-1B8, 7B, 14B; and BLOOMZ-560M, 1B7, 7B1**).
- Empirically substantiated **the existence of multilingual human values in LLMs**, identifying **3 cross-lingual traits of these concepts arising from language resource disparities**: cross-lingual inconsistency, distorted linguistic relationships, and unidirectional cross-lingual transfer between high- and low-resource languages. Also, validated **the feasibility of cross-lingual control over value alignment** capabilities of LLMs.
- Drawing from the above findings, provided prudent **suggestions on the composition of multilingual data for LLMs pre-training**: including a limited number of dominant languages for cross-lingual alignment transfer while avoiding their excessive prevalence, and keeping a balanced distribution of non-dominant languages, which might contribute to enhancing **multilingual AI safety and utility**.

Length Extrapolation for Large Language Models *Jul 2023 - Sep 2023*

- Studied common methods for length extrapolation, including position encodings with extrapolation capabilities like **Rope and Alibi**, transforming extrapolation into interpolation through methods like **Position Interpolation, and nonlinear NTK**.
- Inspired by the human behavior of **summarizing and synthesizing when processing long documents**, initially considered a length extrapolation approach involving **compressing extensive contexts into a few abstract vectors**. Abandoned this idea later on due to overlap with existing published work and other reasons.

Cross-Lingual Factual Knowledge Transfer in Multilingual LMs (1st author, accepted to emnlp2023) *Mar 2023 - Jun 2023*

- Based on prior findings in cross-lingual knowledge transfer, investigated **whether cross-lingual alignment of representation spaces enable factual knowledge transfer across languages**, thus mitigating disparities in factual knowledge between high- and low-resource languages in multilingual pretrained language models.
- Proposed a parameter-free framework, LRP2, comprising two key modules: the **Language-Independent Representation Projection (LIRP)** module that first maps non-English representations into English representation space, and the **Language-Specific Representation Projection (LSRP)** module for projecting them back.
- Conducted experiments on multilingual factual knowledge probing benchmarks, mLAMA and X-FACTR, demonstrating **significant improvements in non-English factual retrieval accuracy**.
- Performed interpretable analysis on LRP2 from the perspective of knowledge neurons, uncovering **the existence of knowledge neurons that are shared across multiple languages** in multilingual pretrained language models and the underlying mechanism of LRP2: increasing the abundance of cross-lingual knowledge neurons.

Zero-Shot Multilingual Machine Translation *Nov 2022 - Feb 2023*

- Conducted an in-depth analysis of multilingual machine translation models (encoder-decoder), unveiling the presence of **inconsistent distribution patterns in representations between English and non-English sentences** at the encoder side. This phenomenon could serve as a source of off-target issues.
- Based on the above observations, proposed **a novel module to disentangle language-specific information from semantic information**. After decoupling, only the language agnostic semantic information from the encoder is preserved and sent to fine-tune the decoder.
- Achieved an **improvement in zero-shot Translation BLEU score from 4.52 to 10.83** on OPUS100 dataset. However, the performance remains below that of English-pivot translation (14.61), indicating room for further improvement.

Achievements

2019	1st Student Scholarship, Academic Excellence Award	*SUDA*
2020	2nd Student Scholarship, Merit Students Award, 3rd Prize of CCSP2020 (East China Division)	*SUDA*
2021	1st Student Scholarship, Merit Students Award, 2nd Prize of National LanQiao Cup	*SUDA*
2022	Excellent Undergraduate Thesis / 1st Student Scholarship	*SUDA / TJU*
2023	2nd Student Scholarship, Advanced Individual Award	*TJU*

Skills

Programming Python, Shell, LaTeX, Pytorch, Tensorflow, Transformers, Fairseq

Languages Mandarin, English (CET-6 Score: 553)