Junjie **Chen**

junjiechen.chris@outlook.com | https://chrisjjc.com/ | junjiechen-chris | junjie-chen-b6a075240

Summary

CS PhD with 5 years of experience in NLP. Specializing in LLM-based unsupervised parsing to extract natural language information from large scale data. Broad knowledge of AI/ML methodologies. Worked on LLM-based book translation and LLM Sparse AutoEncoder-based personal information identification.

Education

The University of Tokyo, Advised by Yusuke Miyao — *Tokyo, Japan*

Ph.D. in Computer Science — *Oct. 2021 - Mar. 2025*

- Supported by Japan Society for the Promotion of Science (JSPS) Fellowship
- PhD research published in 3 top-tier venues

The University of Tokyo, Advised by Yusuke Miyao — *Tokyo, Japan*

M.S. in Computer Science — *Oct. 2019 - Sep. 2021*

- Master research published in ACL 2022 (Oral)

University of Liverpool, Advised by Danushka Bollegala — *Liverpool, UK*

Visiting Scholar — *Oct. 2023 - Mar. 2024*

- Research published in ACL 2024 Findings

Northeastern University — *Shenyang, China*

B.S. in Software Engineering — *Sep. 2015 - Jun. 2019*

- GPA: 3.46/4.0 (3.78/4.0 during 6 month exchange at Hokkaido University)

Selected Publications

Improving Unsupervised Constituency Parsing via Maximizing Semantic Information — *ICLR 2025 (Spotlight, top 5.1%)*

Junjie Chen, Xiangheng He, Yusuke Miyao, and Danushka Bollegala — *2025*

Constituents are Frequent Word Sequences among Sentences with Equivalent Predicate-Argument Structures: Unsupervised Constituency Parsing by Span Matching — *ACL 2024 Findings*

Junjie Chen, Xiangheng He, Danushka Bollegala, and Yusuke Miyao — *2024*

Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints — *NAACL 2024*

Junjie Chen, Xiangheng He, and Yusuke Miyao — *2024*

Syntactic-Semantic Dependency Correlation in Semantic Role Labeling: A Shift in Semantic Label Distributions — *Journal of Natural Language Processing*

Junjie Chen — *2022*

Modeling Syntactic-Semantic Dependency Correlations in Semantic Role Labeling Using Mixture Models — *ACL 2022 (Oral)*

Junjie Chen, Xiangheng He, and Yusuke Miyao — *2022*

ProsodyFM: Unsupervised Phrasing and Intonation Control for Intelligible Speech Synthesis — *AAAI 2025 (Oral)*

Xiangheng He, **Junjie Chen**, Zixing Zhang, Björn W Schuller — *2025*

Task Selection and Assignment for Multi-modal Multi-task Dialogue Act Classification with Non-stationary Multi-armed Bandits — *ICASSP 2024*

Xiangheng He, **Junjie Chen** and Bjorn W. Schuller — *2024*

An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation — *Interspeech 2021*

Xiangheng He, **Junjie Chen**, Georgios Rizos, and Björn W. Schuller — *2021*

A system for worldwide COVID-19 information aggregation — *Workshop on NLP for COVID-19*

Akiko Aizawa, Frederic Bergeron, **Junjie Chen**, Fei Cheng, Katsuhiko Hayashi, Kentaro Inui, Hiroyoshi Ito, Daisuke Kawahara, Masaru Kitsuregawa, Hirokazu Kiyomaru, Masaki Kobayashi, Takashi Kodama, Sadao Kurohashi, Qianying Liu, Masaki Matsubara, Yusuke Miyao, Atsuyuki Morishima, Yugo Murawaki, Kazumasa Omura, Haiyue Song, Eiichiro Sumita, Shinji Suzuki, Ribeka Tanaka, Yu Tanaka, Masashi Toyoda, Nobuhiro Ueda, Honai Ueoka, Masao Utiyama, Ying Zhong — *2020*

A pattern-based method for medical entity recognition from Chinese diagnostic imaging text — *Frontiers in Artificial Intelligence*

Zihong Liang, **Junjie Chen**, Zhaopeng Xu, Yuyang Chen, Tianyong Hao — *2019*

A Bibliometric Analysis of the Research Status of the Technology Enhanced Language Learning — *Emerging Technologies for Education*

Xieling Chen, Juntao Hao, **Junjie Chen**, Songshou Hua, Tianyong Hao — *2018*

Work Experience

Amazon — *Tokyo, Japan*

Amazon Science Fellow — *Aug. 2025 - Feb. 2026 (Est.)*

- Early member and core contributor of a multimodal book translation project
- Designed and implemented a human evaluation process.
- Designed and implemented a LLM-centric AI translation pipeline. Work including text translation, image segmentation, and auto-eval.

Rakuten — *Tokyo, Japan*

Research Intern — *Nov. 2024 - Feb. 2025*

- Investigated the business application of LLM Sparse AutoEncoder (SAE)
- Improved SAE-based Personal Information Identification (PII) system from 72% to 87%.

National Institute of Informatics — *Tokyo, Japan*

Research Assistant — *Jun. 2024 - Feb. 2025*

- Developed evaluation pipeline for Japanese LLM of various sizes (180M to 175B)
- Maintained 800-GPUs clusters for LLM training

The University of Tokyo — *Tokyo, Japan*

Teaching Assistant — *Oct. 2022 - Feb. 2023*

- Mentored 3 student teams on their final NLP projects (one published at JNLP-2023 conference)

Research Projects

Amazon

Whold-book Translation Project — *2024-2025*

- Developed an MQM-style guideline tailored to evaluating book translation.
- Developed an LLM-centric translation system that jointly performs visual grounding, speech bubble detection, story comprehension, and translation.
- Improved annotation consistency from 0 (no agreement) to 0.3 (fair agreement) and translation quality from non-publishable (MQM score 80) to near-publishable (MQM score 95)
- Improved the fluency and story accuracy of the translation.
- Improved the performance of speech bubble detection.

University of Tokyo, PhD Research Project - 2

Unsupervised Constituency Parsing via Discovering Semantic Information — *2024-2025*

- Developed a paraphrasing-based bag-of-substring model to estimate substring-level semantic information
- Developed a calibrated training objective for PCFG-based unsupervised constituency parsing
- The training objective significantly improved parsing accuracy by 8% absolute on average across four languages
- The semantic information estimate is effective in identifying semantic arguments

University of Tokyo, PhD Research Project - 1

Enhancing Language Model-based Syntactic Dependency Parsing with Grammatical Prior *2023-2024*

- Developed a Metropolis-Hasting based algorithm for estimating word-word mutual information
- Developed a sub-sampling method to incorporate grammatical prior into mutual information calculation
- The grammatical prior achieved aggregated accuracy improvement by over 5%
- The prior enabled over 10% accuracy uplift in predicting semantic-related dependencies

University of Tokyo, Master Research Project

Enhancing Syntax-aided Semantic Dependency Parsing with Mixture-of-Expert Models *2021-2022*

- Developed a mixture-of-experts models that model semantic dependencies by their syntactic patterns
- Applied variational inference techniques to automatically cluster similar syntactic patterns
- Significantly improved parsing accuracy compared to state-of-the-art baselines

University of Tokyo, Extracurricular Project

Extracurricular Projects *2024 - 2025*

- Improved Japanese glyph generation with paired positive-negative prompt engineering
- Improved character-level language modeling with a novel distance-to-whitespace loss function

Grants & Fellowships

2023 **DC2 Fellowship**, Japan Society for the Promotion of Science
2024 **Special Allowance for Outstanding Student**, Japan Society for the Promotion of Science
2025 **Travel Grant ($2000)**, Association for the Advancement of Artificial Intelligence
2022 **IST-RA Fellowship**, The University of Tokyo

Skills

Research Experience	Unsupervised Natural Language Parsing with LLM, MultiModal book Translation
Generative Modeling	Language Models, Diffusion, Normalizing Flow, Variational AutoEncoder
Machine Learning Frameworks	Pytorch, Tensorflow, Lightning, Hydra, vLLM
Distributed Computing	Slurm, Huggingface Accelerate, Megatron, Lightning
DevOps	AWS, Docker, Cloudflare, Brazil, Docker
Programming Skills	Python, Typescript, Bash
Language	Chinese (Native), Japanese (N1, Business), English (C-1 equivalent, Business Level)