

Junjie Hu Curriculum Vitæ

Gates and Hillman Centers 5503
5000 Forbes Ave
Pittsburgh, PA 15213

Website: <https://www.cs.cmu.edu/~junjieh/>
Email: junjieh@cs.cmu.edu
Phone: +1 (412) 916-2796

Research Interest

Natural Language Processing, Machine Learning, Artificial Intelligence, Biomedical Data Science

Education

Carnegie Mellon University

Pittsburgh, United States

Ph.D./M.S. in Language Techniques, School of Computer Science

08/2015–05/2021 (expected)

Thesis: *Indirect Supervision for Language Processing over Diverse Domains, Languages, and Modalities*

Advisors: Jaime Carbonell (1953–2020) and Graham Neubig

Committee: Yulia Tsvetkov, Zachary Lipton, and Kyunghyun Cho

Chinese University of Hong Kong

Shatin, Hong Kong

M.Phil. in Computer Science & Engineering

08/2013–07/2015

Thesis: *Kernelized Online Imbalanced Learning*

Advisors: Michael R. Lyu and Irwin King

Committee: Lai-Wan Chan, Kevin Yip, and Qin Lu

South China University of Technology

Guangzhou, China

B.Eng. in Computer Science and Technology, GPA: 3.73/4.00 (Ranked 1/30 in the Gifted Class) 09/2010–07/2013

Research and Industry Experience

Language Technologies Institute

Pittsburgh, United States

Graduate Research Assistant

08/2019–present

- Research on domain adaptation, multilingual NLP, and multimodal machine learning
- Advisors: Graham Neubig and Jaime Carbonell

Google AI Research

Mountainview, United States

Research Intern & Student Researcher

05/2019–12/2019

- Worked on the evaluation and alignment of multilingual models.
- Host: Melvin John and Orhan Firat

Microsoft Research & AI

Redmond, United States

Research Intern

05/2018–08/2018

- Worked on long text generation and visual storytelling.
- Mentors: Yu Chen, Zhe Gan, Jingjing Liu, and Jianfeng Gao

Web Intelligence & Social Computing Lab, CUHK

Shatin, Hong Kong

Research Intern

08/2013–07/2015

- Worked on online learning algorithms, kernel methods, and imbalanced learning problems.
- Advisors: Michael R. Lyu and Irwin King

Noah's Ark Lab, Huawei Technology

Science Park, Hong Kong

Research Intern

06/2014–09/2014

- Worked on the semantic search of users' queries on Huawei app store using deep learning models.
- Advisors: Hang Li and Xin Jiang

Machine Learning Group, SCUT

Undergraduate Research Assistant

Guangzhou, China

09/2011-07/2013

- Worked on active learning and imbalanced learning for classification problems.
- Advisors: Wing W. Y. Ng and Daniel S. Yeung

Peer-reviewed Conference Publications

- [C.1] **Junjie Hu**, Melvin Johnson, Orhan Firat, Aditya Siddhant and Graham Neubig. *Explicit Alignment Objectives for Multilingual Bidirectional Encoders*. Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2021).
- [C.2] Po-Yao Huang*, Mandela Patrick*, **Junjie Hu***, Graham Neubig, Florian Metze and Alexander Hauptmann. *Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models* Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2021), * indicates equal contributions.
- [C.3] **Junjie Hu***, Sebastian Ruder*, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation*. In International Conference on Machine Learning (ICML 2020, * indicates equal contributions)).
- [C.4] Han Zhao, **Junjie Hu**, and Andrej Risteski. *On Learning Language-Invariant Representations for Universal Machine Translation*. In International Conference on Machine Learning (ICML 2020).
- [C.5] Po-Yao Huang, **Junjie Hu**, Xiaojun Chang, and Alexander Hauptmann. *Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting*. The 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020).
- [C.6] **Junjie Hu**, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. *What Makes A Good Story? Designing Composite Rewards for Visual Storytelling*. Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020, **spotlight**)
- [C.7] Zi-Yi Dou, **Junjie Hu**, Antonios Anastasopoulos, Graham Neubig. *Unsupervised Domain Adaptation for Neural Machine Translation with Domain-Aware Feature Embeddings*. Conference on Empirical Methods in Natural Language Processing (EMNLP 2019).
- [C.8] Chunting Zhou, Xuezhe Ma, **Junjie Hu**, Graham Neubig. *Handling Syntactic Divergence in Low-resource Machine Translation*. Conference on Empirical Methods in Natural Language Processing (EMNLP 2019).
- [C.9] Ming Jiang, **Junjie Hu**, Qiuyuan Huang, Lei Zhang, Jana Diesner, Jianfeng Gao. *REO-Relevance, Extraness, Omission: A Fine-grained Evaluation for Image Captioning*. Conference on Empirical Methods in Natural Language Processing (EMNLP 2019).
- [C.10] Liu Yang, **Junjie Hu**, Minghui Qiu, Chen Qu, Jianfeng Gao, Bruce Croft, Xiaodong Liu, Yelong Shen and Jingjing Liu. *A Hybrid Retrieval-Generation Neural Conversation Model*. The 28th ACM International Conference on Information and Knowledge Management (CIKM 2019).
- [C.11] **Junjie Hu**, Mengzhou Xia, Graham Neubig, Jaime Carbonell. *Domain Adaptation of Neural Machine Translation by Lexicon Induction*. The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).
- [C.12] Han Zhao, **Junjie Hu**, Zhenyao Zhu, Adam Coates, Geoff Gordon. *Deep Generative and Discriminative Domain Adaptation*. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019).
- [C.13] Graham Neubig, Zi-Yi Dou, **Junjie Hu**, Paul Michel, Danish Pruthi, Xinyi Wang. *compare-mt: A Tool for Holistic Comparison of Language Generation Systems*. Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2018) Demo Track.
- [C.14] **Junjie Hu**, Wei-Cheng Chang, Yuexin Wu, Graham Neubig. *Contextual Encoding for Translation Quality Estimation*. 3rd Conference on Machine Translation (WMT 2018).

- [C.15] Graham Neubig, **Junjie Hu**. *Rapid Adaptation of Neural Machine Translation to New Languages*. Conference on Empirical Methods in Natural Language Processing (EMNLP 2018).
- [C.16] Craig Stewart, Nikolai Vogler, **Junjie Hu**, Jordan Boyd-Graber, Graham Neubig. *Automatic Estimation of Simultaneous Interpreter Performance*. Annual Meeting of the Association for Computational Linguistics (ACL 2018)
- [C.17] **Junjie Hu**, Desai Fan, Shuxin Yao, Jean Oh. *Answer-Aware Attention on Grounded Question Answering in Images*. AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration.
- [C.18] **Junjie Hu***, Rui Liu*, Wei Wei*, Zi Yang*, Eric Nyberg. *Structural Embedding of Syntactic Trees for Machine Comprehension*. 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017, * indicates equal contributions)
- [C.19] Zhilin Yang, **Junjie Hu**, Ruslan Salakhutdinov, William W. Cohen. *Semi-Supervised QA with Generative Domain-Adaptive Nets*. Annual Meeting of the Association for Computational Linguistics (ACL 2017).
- [C.20] Zhilin Yang, Bhuwan Dhingra, Ye Yuan, **Junjie Hu**, William W. Cohen, Ruslan Salakhutdinov. *Words or Characters? Fine-grained Gating for Reading Comprehension*. International Conference on Learning Representations (ICLR 2016).
- [C.21] **Junjie Hu**, Jean Oh, Anatole Gersham. *Learning Lexical Entries for Robotic Commands via Paraphrasing* 2016 AAAI conference on Human Computation (HCOMP 2016).
- [C.22] **Junjie Hu**, Haiqin Yang, Irwin King, Michael R. Lyu, Anthony Man-Cho So. *Kernelized Online Imbalanced Learning with Fixed Budgets*. Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015 **Oral**).
- [C.23] Cuiyun Gao, Hui Xu, **Junjie Hu**, Yangfan Zhou. *AR-Tracker: Track the Dynamics of Mobile Apps via User Review Mining*. IEEE Symposium on Service-Oriented System Engineering (SOSE 2015).
- [C.24] Tong Zhao, **Junjie Hu**, Pinjia He, Hang Fan, Irwin King, Michael R. Lyu. *Exploiting Homophily-based Implicit Social Network to Improve Recommendation Performance*, 2014 International Joint Conference on Neural Networks (IJCNN 2014).
- [C.25] **Junjie Hu**, Daniel S. Yeung. *Incremental Learning using Error and Sensitivity Analysis of MCS for Image Classification*. International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR 2013).
- [C.26] **Junjie Hu**. *Active Learning for Imbalance Problem Using L-GEM of RBFNN*. International Conference on Machine Learning and Cybernetics (ICMLC 2012).

Peer-reviewed Journal Publications

- [J.1] **Junjie Hu**, Haiqin Yang, Irwin King, Michael R. Lyu, Anthony Man-Cho So. *Online Nonlinear AUC Maximization with Fixed Budgets for Big Data*. IEEE Transactions on Neural Networks and Learning Systems 2017.
- [J.2] Wing N.Y. Ng⁺, **Junjie Hu***, Daniel S. Yeung, Shaohua Yin, Fabio Roli. *Diversified Sensitivity based Undersampling for Imbalance Classification Problems*, IEEE Transactions on Cybernetics 2015. (+ undergraduate advisor, * indicates my undergraduate thesis work)
- [J.3] Jiana Hu, **Junjie Hu**, Haibin Lin, Xiaoping Li, Changlin Jiang, Xianhuan Qiu, Weishan Li. *State-of-Charge Estimation for Battery Management System Using Optimized Support Vector Machine for Regression*. Journal of Power Sources 2014.

Peer-Reviewed Workshop Publications

- [W.1] Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, **Junjie Hu**, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, Sylwia Tur. *TICO-19: the Translation Initiative for COvid-19*. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020
- [W.2] **Junjie Hu**, Haiqin Yang, Yuxin Su, Michael R. Lyu, Irwin King. *Accelerated Information-Theoretic Metric Learning*. Scalable Data Analytics: Theory and Applications at WSDM 2015.
- [W.3] Haiqin Yang, **Junjie Hu**, Michael R. Lyu, Irwin King. *Online Imbalanced Learning with Kernels*. NIPS workshop on Big Learning (BigLearning 2013).

Pre-print

- [P.1] Han Zhao, Zhenyao Zhu, **Junjie Hu**, Adam Coates, Geoff Gordon. *Principled Hybrids of Generative and Discriminative Domain Adaptation* (arXiv:1705.09011).
- [P.2] Aditi Chaudhary, Siddharth Dalmia, **Junjie Hu**, Xinjian Li, Austin Matthews, Aldrian Obaja Muis, Naoki Otani, Shruti Rijhwani, Zaid Sheikh, Nidhi Vyas, Xinyi Wang, Jiateng Xie, Ruochen Xu, Chunting Zhou, Peter J. Jansen, Yiming Yang, Lori Levin, Florian Metze, Teruko Mitamura, David R. Mortensen, Graham Neubig, Eduard Hovy, Alan W Black, Jaime Carbonell. *The ARIEL-CMU systems for LoReHLT18*. (arXiv:1902.08899)

Relevant Research Grants

- **DARPA: Low Resource Languages for Emergent Incidents (LORELEI)**
 - PI: Jaime Carbonell. 2015-2019
 - I was the key member contributing to the machine translation sub-task, and we achieved the **highest BLEU score in three out of four language-checkpoint pairs** in the MT evaluation in 2018 based on my work [C.15]
- **Structured Models for Massively Multilingual Language Processing**
 - PI: Graham Neubig. \$181,190 2020-2021
 - I led the writing of the proposal, prepared the progress reports, and got \$181,190 for supercomputing resources.
- **Google Cloud Credit Grant: Domain Adaptation of Neural Machine Translation for COVID-19**
 - PI: Junjie Hu. \$5100 2020-2021
 - This grant funded the computational resources for future research based on my work [C.11, W.1].
- **Google Faculty Award Grant: Towards Human+Machine Translation of Wikipedia in Every Language**
 - PI: Graham Neubig. \$100,526 2020-2021
 - This grant supported my research on translating English Wikipedia articles to many other languages.

Conference and Invited Talks

- **Cross-lingual Generalization, Alignment, and Applications**
 - European Commission Workshop: Crisis Response-Language tools to the rescue 12/2020
 - NLP Seminar at Georgia Institute of Technology 11/2020
- **Pre-training of Multilingual Encoder for Crosslingual Transfer**
 - Research Talk at Google Translate Team 08/2019
- **XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization**
 - LTI Summer Seminar Series at Carnegie Mellon University 07/2020
 - International Conference on Machine Learning 07/2020
- **Pre-training of Multilingual Encoder for Crosslingual Transfer**
 - Research Talk at Google Translate Team 08/2019
- **Cross-Lingual and Cross-Domain Transfer for Neural Machine Translation**
 - European Commission Workshop: Multilingual Models and Unsupervised NMT 6/2019
 - AI Seminar at Carnegie Mellon University 4/2019
 - Research talk at Facebook AI Research Lab, Paris 6/2019
- **Rethinking Visual Storytelling: What Makes A Good Story?**
 - Research talk at Microsoft 365 AI Research, Redmond 8/2018
 - AAAI Conference on Artificial Intelligence 2/2020
- **Machine Reading Comprehension via Structural Tree Embeddings**

- Seminar at the Chinese University of Hong Kong 3/2018
- **Lorelei: Understanding Low Resource Languages**
 - Jointly with Pat Littell, Shruti Rijhwani, and Ruochen Xu. LTI Colloquium at Carnegie Mellon University 9/2017
- **Natural Communication for Human-Robot Collaboration**
 - Symposium on Natural Communication for Human-Robot Collaboration 11/2017

Teaching Experience

- **Guest Lecturer** at Carnegie Mellon University Spring 2020, Spring 2019, Fall 2018
Machine Translation in 11-411/611 Natural Language Processing
- **Guest Lecturer** at Carnegie Mellon University Spring 2018
Multi-task Multi-lingual Learning Models in 11-747 Neural Networks for NLP
- **Teaching Assistant** at Carnegie Mellon University Spring 2018
11-747 Neural Networks for NLP
- **Teaching Assistant** at Carnegie Mellon University Fall 2018
11-731 Machine Translation and Sequence-to-Sequence Models
- **Teaching Assistant** at the Chinese University of Hong Kong Spring 2014, Spring 2015
CSCI3100 Software Engineering
- **Teaching Assistant** at the Chinese University of Hong Kong Fall 2013
CSCI3170 Introduction to Database System
- **Teaching Assistant** at the Chinese University of Hong Kong Fall 2014
CSCI3170 Introduction Retrieval and Web Search

Scholarships and Awards

- CMU Graduate Student Assembly Dissertation Writing Group Grants 2020
- CMU Graduate Student Assembly Conference Travel Grant 2020
- NAACL 2019 Best Demonstration Paper Nomination 2019
- Graduate Research Scholarship, Carnegie Mellon University 2015–2021
- Graduate Research Scholarship, Chinese University of Hong Kong 2013–2015
- Certificate of Merit for Teaching Assistantship, Department of CSE, Chinese University of Hong Kong 2013–2014
- IBM Outstanding Student Scholarship (1 of 77 winners in China) 2012–2013
- Outstanding Undergraduate Awards by China Computer Federation (1 of 99 winners in China) 2012–2013
- National Scholarship, the Ministry of Education in China 2010–2011, 2011–2012
- Second Prize in “China Software Cup” Final (6/1600+ teams nationally) 08/2012

Professional Service

- **Conference Reviewer/Program Committee Member**
 - Reviewer, Annual Meeting of the Association for Computational Linguistics 2020
 - Reviewer, Empirical Methods in Natural Language Processing (EMNLP) 2019, 2020
 - Reviewer, North American Chapter of the Association for Computational Linguistics (NAACL) 2019, 2021
 - Reviewer, International Conference on Computational Linguistics (COLING) 2020
 - Reviewer, Conference on Machine Translation (WMT) 2019
 - Reviewer, The 6th Workshop on Noisy User-generated Text (W-NUT) 2020

- Program Committee, AAAI Conference on Artificial Intelligence (AAAI) 2017, 2018, 2020, 2021
- External Reviewer, International AAAI Conference on Weblogs and Social Media (ICWSM) 2014
- External Reviewer, IEEE International Conference on Big Data 2014
- External Reviewer, International Conference on Web Search and Data Mining (WSDM) 2014
- **Journal Reviewer**
 - Reviewer, Neural Networks 2018
 - Reviewer, Neurocomputing 2017–2019
- **Organizing Committee**
 - Co-organizer, Machine Translation and Multilingual Reading Group at Carnegie Mellon University 2017–2019
 - Server Manager, Neulab at Carnegie Mellon University 2018–2020
- **Volunteer**
 - ACM Recommender Systems Conference (RecSys) 2013
 - IEEE International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR) 2013
 - IEEE International Conference on Machine Learning and Cybernetics (ICMLC) 2012

Selected Press Coverage

- **Google AI Blog:** “XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization” 2020
- **Synched Review:** “CMU, DeepMind & Google’s XTREME Benchmarks Multilingual Model Generalization Across 40 Languages” 2020
- **VentureBeat:** “Google releases benchmark to spur development of multilingual AI models” 2020
- **AIM Review:** “Google Releases Xtreme To Induce Development of Multilingual AI Models” 2020
- **NewVoice (Portuguese):** “IA: Xtreme estimula criação de modelos multilíngues” 2020
- **Slator:** “Simultaneous Interpreters May Soon Get Real-Time Help Just When They Need It” 2018
- **Wired article:** “Covid-19 Is History’s Biggest Translation Challenge” 2020
- **Facebook AI Blog:** “Using AI to help health experts address the COVID-19 pandemic” 2020
- **TWB:** “TWB partners with tech leaders to develop COVID-19 language technology for 37 languages” 2020
- **Slator:** “TWB, Big Tech, Academia Take on Covid-19 Language Challenge” 2020
- **DevEx:** “Opinion: COVID-19 calls for disrupting the way we communicate in a crisis” 2020
- **Language Magazine article:** “COVID Translation Initiative Takes Off” 2020

Languages

- **Mandarin, Teochew:** Native
- **English, Cantonese:** Bilingual proficiency

References

- **Graham Neubig**
Associate Professor
Language Technologies Institute, Carnegie Mellon University
gneubig@cs.cmu.edu
- **Micheal R. Lyu**
Choh-Ming Li Professor, ACM/IEEE/AAAS Fellow, Croucher Senior Research Fellow

Computer Science and Engineering Department, Chinese University of Hong Kong
lyu@cse.cuhk.edu.hk

- **Jingjing Liu**

Senior Principal Researcher and Research Manager
Microsoft Research
jingjl@microsoft.com

- **Orhan Firat**

Research Scientist
Google AI Research
orhanf@google.com