Haau-Sing (Xiaocheng) Li

Curriculum Vitae

Center for Data Science
New York University
☎ *(+1) 929-299-8072.*
✉ *xl3119@nyu.edu*
My Webpage
Github

Education

2019–2021: **Master of Science, Data Science (NLP Track).**, *Center for Data Science, New York University*, New York, United States.

CGPA : 4/4

Courses : Deep Learning, Inference and Representations, Natural Language Understanding (93.5, A), Big Data (96.5, A), Machine Learning (A), Deep Learning with NLP (98.3, A), Probability and Statistics (98.1, A).

2009–2013 : **Bachelor of Arts, English Language**, *Renmin University of China*, Beijing, China.

CGPA : 3.7/4

Courses : **linguistics:** Introductory Linguistics, Computational Linguistics, Intensive Reading (2-year training of phonetics, phonology, etymology and (occasionally) syntax).

Computer Science: Programming, Discrete Mathematics, Data Structures, Database and Data Mining, Operations Research.

Mathematics: Statistics, Stochastic Process, Probability, Calculus.

Publications and Drafts

2020 Yian Zhang, Alex Warstadt, **Haau-Sing Li**, and Samuel R. Bowman. When do you need billions of words of pretraining data? *ArXiv*, volume abs/2011.04946, 2020.

2020 Alex Warstadt, Yian Zhang, **Haau-Sing Li**, Haokun Liu, and Samuel R Bowman. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 2020. Association for Computational Linguistics.

2020 Haau-Sing Li, Kyunghyun Cho, and Narges Razavian. BEnTo: Transformers language models for icd coding. *Draft [link]*, 2020.

Research Projects

Grossman School of Medicine, New York University

Sept,2020 – present **Capstone Project: Pretraining Clinical BERT and LongFormer.**.
Pretrained clinical BERT and Longformer models with and without self-defined vocabulary and word embeddings. Performing layer-wise probing on ICD coding.

June,2020 – present ***ICD Coding with BERT.***.
Experimented on ICD coding with clinical notes. Implemented BERT with n-gram features and snippet features and ensembled predictions. Implement multi-GPU training.[link to paper draft]

Working with : **Dr. Narges Razavian**, *Assistant Professor, Department of Population Health and Department of Radiology*, NYU School of Medicine (*webpage*)

Dr. Kyunghyun Cho , *Associate Professor, Department of Computer Science and Center for Data Science*, New York University (*webpage*)

Center for Data Science, New York University

Feb,2020 – present ***Probing Inductive Bias of RoBERTa models.*.**

Probed pretrained RoBERTa models on synthetic dataset to test feature preference learning. Paper accepted at EMNLP2020. Probing RoBERTa models OpenSubtitles dataset to compare human and RoBERTas in language acquisition.

Probed pretrained RoBERTa models on unsupervised grammaticality judgement. Preprint.

Testing compressed BERT models with knowledge distillation with edge-probing.

Working with : **Dr. Samuel R. Bowman** , *Assistant Professor, Department of Linguistics, Center for Data Science and Department of Computer Science*, New York University (*webpage*)

Department of English Studies, Renmin University of China

Oct,2018 – Apr,2019 ***Undergraduate Thesis: Exploring Lexical Variation of World Englishes: a Lectometric Perspective.*.**

Implemented Multi-dimensional Scaling to study variations on English text with bag-of-word representation. Implemented hierarchical clustering to rank dissimilarities between English variants.

Worked with : **Dr. Xinyue Yao**, *Lecturer, Department of English Studies*, Renmin University of China (*research webpage*)

Department of Statistics, Stanford University

June,2018 – Sept,2018 ***Adaptive Multi-Armed Bandits.*.**

Implemented adaptive multi-armed bandit algorithm. Re-implemented policy-selection rules, simulations, and compared with UCB and Epsilon-greedy bandits.

Worked with : **Dr. Tze-Leung Lai**, *Ray Lyman Wilbur Professor, Department of Statistics*, Stanford University (*faculty webpage*)

Dr. Ka-Wai Tsang, *Assistant Professor, School of Data Science*, Chinese University of Hong Kong, Shenzhen (*faculty webpage*)

Posters & Presentations

Nov,2020 ***Learning which features matter: Roberta acquires a preference for linguistic generalizations.***, *Natural Language, Dialog and Speech Symposium (NDS)*, New York Academy of Sciences.

Aug,2018 ***Adaptive Multi-Armed Bandits.***, *Data Analytics & Statistical Modeling Workshop*, Financial and Risk Modeling Institute, Stanford University.

Fellowships & Awards

2020 ***Summer Reasearch Initiative*** of Moore Sloan Data Sicence Environments (MSDSE).

2019 ***Distinctive Graduate*** of Renmin University of China (5%).

2018 ***Undergraduate Research Scholarship*** of Renmin University of China (10%).

2018 ***Meritorious Award*** of Mathematical Contest in Modeling (MCM)(10%).

2016 – 2018 ***Scholarship for Academic Excellence*** of Renmin University of China (10%, 3 years).

Computer skills

Programming Languages Python, Linux, R (Primary), C (Primary).

Research Frameworks PyTorch, Transformers, MXNet, Fairseq, jiant.

Big Data SQL, MapReduce, Spark.