Yilun **Zhu**

Linguistics · Research Intern

202-247-1701 | allen.yl.zhu@gmail.com | https://yilunzhu.github.io/ | y-l-zhu | yl-zhu

Education

Georgetown University — *Washington D.C.*

PhD in Linguistics — *Aug. 2017 - May 2022*

- *Selected coursework:* ***LING/COSC-572*** *Empirical methods in NLP* ***ANLY-550*** *Data structs & algorithms* ***LING-765*** *Comp discourse modelling* ***ANLY-590*** *Neural nets & deep learning* ***LING-461*** *Speech processing*

Georgetown University — *Washington D.C.*

M.S. in Linguistics — *Aug. 2017 - May 2019*

Nanjing University — *Nanjing, China*

B.A. in English, Linguistics concentration — *Sep. 2013 - Jun. 2017*

- *Selected coursework: Calculus, Discrete Mathematics, Basics of Programming (C++)*

Skills

NLP & ML	TensorFlow, Pytorch, Keras, Scikit-learn, Numpy, Pandas, StanfordNLP, NLTK
Programming	Proficient in Python, familiar with Java & C/C++, Bash
Miscellaneous	Linux, Google Cloud, Git, SQL, LaTeX
Languages	Mandarin Chinese, English, intermediate French

Internship

Microsoft — *Beijing, China*

Software Engineer Intern — *Jun. 2019 - Aug. 2019*

- Worked on a Question Answering dataset QuaRel for qualitative reasoning in a research team.
- Built a joint-learning model by using Pytorch. The first part of the model extracted the correct entities, and then the model generated the logical form customized for QuaRel given those entities.

China Telecom Co., Ltd. — *Beijing, China*

NLP Intern — *May 2018 - Jun. 2018*

- Built a sentiment analysis system, which assists clients to know the feedback of each attributes of car models from customers.
- Prompted an automatic web crawler in Python to collect 110,000+ pieces of positive and negative evaluations of 513 types of cars from online comments; Parsed comments by using StanfordNLP and built a corpus about terminologies in motor vehicles.
- Trained RNN by using TensorFlow, with pre-trained word embeddings and the corpus as input, to predict the sentimental polarity of each comment. The decent F-score 86.74 assists to filter out those car models with high qualities.

Publications & Presentations

GumDrop at the DISRPT2019 Shared Task: A Model Stacking Approach to Discourse Unit Segmentation and Connective Detection

Y. Yu, Y. Zhu, Y. Liu, Y. Liu, S. Peng, M. Gong, A. Zeldes

Proceedings of the Workshop on Discourse Relation Parsing and Treebanking (DISRPT) at NAACL-HLT, 2019, Minneapolis, MN

Adpositional Supersenses for Mandarin Chinese

Y. Zhu, Y. Liu, S. Peng, A. Blodgett, Y. Zhao, N. Schneider

Proceedings of the Society for Computation in Linguistics (SCiL) at LSA 2019 Annual Meeting, 2019, New York, NY

Extreme Predicative Adjectives in Mandarin Chinese

Y. Zhu

Presented at 8th International Conference on Formal Linguistics (ICFL-8), 2018, Hangzhou, China

Research

Georgetown University Multilingual Discourse Region Partitioner (GUMDROP)

Washington D.C.

Researcher

Jan. 2019 - Mar. 2019

- Developed a DNN model with wide (linguistic features) and deep (pre-trained word embeddings) programming for sentence splitting and discourse unit segmentation, assisting the system to generalize and have better predictions in large corpora.
- Prompted an ensemble by using a Gradient Boosting classifier to blend lookup Wikipedia frequency and a bi-LSTM/CNN-CRF sequence labelling framework for connective detection, increasing the predicting accuracy by .04 in 15 datasets in average.

Evaluation on UCCA and USim for Paraphrase Detection

Washington D.C.

Researcher

Nov. 2018 - Mar. 2019

- Used TUPA parser parsing each pair of paraphrases in MSRP corpus (Microsoft Research Paraphrase Corpus) to generate UCCA structures for analyzing whether UCCA and USim accurately reflect semantic similarities.
- Built UCCA structure via Tree RNN and add pertained word embeddings as input, evaluating whether semantic structure contributes to paraphrase detection.

Externally configurable reference and non-named entity recognizer (xrenner)

Washington D.C.

Research Assistant | Supervisor: Dr. Amir Zeldes

Feb. 2018 - Nov. 2018

- Established benchmark entities (names, gazetteer, etc.) for a rule-based model in the Chinese subsystem.
- Developed a Logistic Regression classifier with model stacking to predict named entities that are unseen the corpus by blending rule-based and CRF models, increasing the accuracy by .07 for coreference prediction.

Semantic network of adposition and case supersenses (SNACS) for Chinese

Washington D.C.

Research Assistant | Supervisor: Dr. Nathan Schneider

Jun. 2018 - Aug. 2018

- Annotated 20 Chapter of The Little Prince and demonstrated the adaptability for SNACS annotation to Chinese, which can further support automatic disambiguation of adpositions in Chinese.

Honors & Awards

2019	**Linguistics Department Conference Travel Grant: $246**, Georgetown University	*Washington D.C.*
2018	**Linguistics Department Conference Travel Grant: $700.17**, Georgetown University	*Washington D.C.*
2017	**College Graduate Excellence Award**, Nanjing University	*Nanjing, China*
2015	**The Pacesetter Youth Volunteer**, Nanjing University	*Nanjing, China*
2015	**Renmin Scholarship**, Nanjing University	*Nanjing, China*
2014	**Renmin Scholarship**, Nanjing University	*Nanjing, China*

Other computer science coursework

MOOC	**Introduction to Computer Systems**, [certificate – Nanjing University], Grade:A
Coursera	**Graph Search, Shortest Paths, and Data Structures**, [certificate – Stanford University], Grade: 100%
Coursera	**Divide and Conquer, Sorting and Searching, and Randomized Algorithms**, [certificate – Stanford University], Grade: 100%