Xiang Li

9 E 33 ST 408B, Baltimore, MD, 21218
(443) 563-4838 | xli150@jhu.edu

EDUCATION:

Johns Hopkins University, Baltimore, US
Whiting School of Engineering, **Expected Graduation: 2020**
B.S. in **Computer Science** and **Applied Mathematics.**
MSE. in Computer Science
GPA: 3.99/4.0

RESEARCH EXPERIENCE

Research Assistant **Mar. 2018 - Present**
Center for Language and Speech Processing, Johns Hopkins University
Advised by Dr. Jason Eisner

- **A Generative Model for Punctuation in Dependency Trees**: (Published to TACL 2019)

  Abstract: Treebanks traditionally treat punctuation marks as ordinary words, but linguists have suggested that a tree's "true" punctuation marks are not observed. These latent "underlying" marks serve to delimit or separate constituents in the syntax tree. When the tree's yield is rendered as a written sentence, a string rewriting mechanism transduces the underlying marks into "surface" marks, which are part of the observed (surface) string but should not be regarded as part of the tree. We formalize this idea in a generative model of punctuation that admits efficient dynamic programming. We train it without observing the underlying marks, by locally maximizing the incomplete data likelihood (similarly to the EM algorithm). When we use the trained model to reconstruct the tree's underlying punctuation, the results appear plausible across 5 languages, and in particular are consistent with Nunberg's analysis of English. We show that our generative model can be used to beat baselines on punctuation restoration. Also, our reconstruction of a sentence's underlying punctuation lets us appropriately render the surface punctuation (via our trained underlying-to-surface mechanism) when we syntactically transform the sentence.

- **Specializing Word Embeddings (for Parsing) by Information Bottleneck:** (Accepted by EMNLP 2019)

  Abstract: Pre-trained word embeddings like ELMo and BERT contain rich syntactic and semantic information, resulting in state-of-the-art performance on various tasks. We propose a variational information bottleneck (VIB) method to nonlinearly compress these embeddings, keeping only the information that helps a discriminative parser. We compress each word embedding to either a discrete tag or a continuous vector. In the discrete version, our automatically compressed tags form an alternative tag set: we show experimentally that our tags capture most of the information in traditional POS tag annotations, but our tag sequences can be parsed more accurately at the same level of tag granularity. In the continuous version, we show experimentally that moderately compressing the word embeddings by our method yields a more accurate parser in 8 of 9 languages, outperforming other compression methods like PCA that perform dimensionality reduction.

- **Bimachine Project**: Worked on formal language proof: a hypothesis that any stochastic Finite State Transducer are compositions of two Probabilistic Finite State Transducers (PFST).

- **Grammar induction Project**: Working on doing grammar induction conditioned on compressed POS tags, word embeddings.

Research Summer Intern **Jun. 2019 - Present**
NLP group, Harvard University
Advised by Dr. Sasha Rush

- Designed the Variational Autoencoder Model (VAE) with a structured encoder as a Semi-Markov-CRF, and a Recurrent Neural Network (RNN) autoregressive decoder. Our method improves performance on conditional generation tasks such as data2text, and also induces a meaningful latent template structure.

Research Assistant **Jan. 2017 - Aug. 2018**
Image Analysis and Communication Lab, Johns Hopkins University
Advised by Dr. Jerry Prince

- **Automatic Paranasal Sinus Segmentation Project**: Implemented and tuned a convolutional neural network with LSTM features to segment a CT scan into multiple labels and deal with sinusitis (blockage in sinus) in order to aid medical diagnosis.
- **Visualization Project**: Rendered 3D visualization of ventricular system for brain patients, in order to better visualize the volume change and shape transform over time.

PUBLICATION:

- A Generative Model for Punctuation in Dependency Trees
  Xiang Li, Dingquan Wang, and Jason Eisner (2019 in TACL)
- Specializing Word Embeddings (for Parsing) by Information Bottleneck
  Xiang Li and Jason Eisner (2019 in EMNLP)

TEACHING EXPERIENCE

Teaching Assistant **Sep. 2017 - Present**
Johns Hopkins University

- **Introduction to Probability**: lead session by facilitating discussions, hold office hour to answer questions and provide wise hints. (Each semester)
- **Natural Language Processing**: lead recitation, hold office hour to help debugging codes, and grade assignments. (Fall 2018, Fall 2019)

HONORS & AWARD

Masson Fellowship	**Aug. 2019**
Michael J.Muuss Research Award	**May. 2019**
Best Insight, Best Visualization, and Best Use of Outside Data Award (from ASA Data Fest)	**April. 2019**
Provost's Undergraduate Research Award (PURA)	**Nov. 2018**
Research Experience for Undergraduate (REU) Fellowship	**May. 2018**
Fellowship William Huggins Summer Fellowship	**May. 2018**
Summer Training and Research (STAR) Fellowship	**May. 2017**
Member of Tau Beta Pi	**Nov. 2018 - Present**
Member of Upsilon Pi Epsilon	**April 2019 - Present**
Dean's List	**Sep. 2016 - Present**

SKILLS

Language: Proficient in Python, Java, C++, Matlab, and R.
Tools: Proficient in PyTorch, Keras, Latex, and OpenGL.
Selected Courses: Machine Learning, Optimization, Sequence Modeling, Deep Learning, Stochastic Processes, Graph Theory, Causal Inference, Approximation Algorithm, Machine Translation, and Natural Language Processing.