# Xiang Li

1E University Pkwy, Unit 1011, Baltimore, MD, 21218
Phone:(443) 563-4838 | email: xli150@jhu.edu

## EDUCATION & HONORS

**Johns Hopkins University**, Baltimore, US                                            **2016-2020**
Whiting School of Engineering
Combined **B.S.** & **M.S.** in **Computer Science**; **B.S.** in **Applied Mathematics and Statistics**
Junior-year election to Tau Beta Pi and Upsilon Pi Epsilon
Dean's List
**GPA: 3.99/4.00**

## PULICATIONS

- [1] A Generative Model for Punctuation in Dependency Trees
  **Xiang Lisa Li**, Dingquan Wang, and Jason Eisner. 2019.
  *Transactions of the Association for Computational Linguistics (TACL).*
  *[Oral Presentation, Journal paper]*

- [2] Specializing Word Embeddings (for Parsing) by Information Bottleneck
  **Xiang Lisa Li** and Jason Eisner. 2019.
  *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*
  ***Best Paper Award** [Oral Presentation, long paper]*

- [3] Posterior Control of Blackbox Generation
  **Xiang Lisa Li** and Alexander Rush. 2020.
  *Prepared for Submission.*

  Order Publications in Medical Imaging:
- M. Shao, S. Han, A. Carass, **X. Li**, B.E. Dewey, A.M. Blitz, J.L. Prince, and L.M. Ellingsen, *"Shortcomings of Ventricle Segmentation Using Deep Convolutional Networks"*, Deep Learning Fails Workshop in conjunction with the 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018), Granada, Spain, September 16 - 20, 2018. (doi)
- M. Shao, A. Carass, **X. Li**, B.E. Dewey, A.M. Blitz, J.L. Prince, and L.M. Ellingsen, "*Multi-atlas segmentation of the hydrocephalus brain using an adaptive ventricles atlas*", Proceedings of SPIE Medical Imaging (SPIE-MI 2018), Houston, TX, February 10 - 15, 2018. (doi)
- J. Glaister, M. Shao, **X. Li**, A. Carass, S. Roy, A.M. Blitz, J.L. Prince, and L. M. Ellingsen, "*Deformable model reconstruction of the subarachnoid space*", Proceedings of SPIE Medical Imaging (SPIE-MI 2018), Houston, TX, February 10 - 15, 2018. (doi)
- A. Carass, M. Shao, **X. Li**, B. Tran, B.E. Dewey, A.M. Blitz, S. Roy, J.L. Prince, and L.M. Ellingsen, "*Whole Brain Parcellation with Pathology: Validation on Ventriculomegaly Patients*", 3rd International Workshop on Patch-based Techniques in Medical Imaging in conjunction with the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2017), Quebec, Canada, September 10 - 14, 2017. (doi)

## FELLOWSHIP & AWARDS

| | |
|---|---:|
| Best Paper Award in EMNLP-IJCNLP 2019 | **Nov. 2019** |
| Gerald M. Masson Research Fellowship | **Aug. 2019** |
| Michael J. Muuss Research Award | **May 2019** |
| Best Insight, Best Visualization, and Best Use of Outside Data Award | |
| *(American Statistical Association Data Fest)* | **Apr. 2019** |
| Provost's Undergraduate Research Award | **Nov. 2018** |
| Research Experience for Undergraduate Fellowship | **May 2018** |
| William Huggins Summer Fellowship | **May 2018** |
| Summer Training and Research Fellowship | **May 2017** |

## RESEARCH EXPERIENCE

***Research Assistant***                                                      **Mar. 2018 - Present**
**Area of Focus:** Natural Language Processing, grammar induction, Bayesian non-parametric model, structural induction, structured prediction, information bottleneck, and syntax.
Advised by **Prof. Jason Eisner,**
*Center for Language and Speech Processing, Johns Hopkins University*
- **A Generative Model for Punctuation in Dependency Trees** [1]:

We propose a latent variable model for punctuation. Our probabilistic generative process contains two steps: the first step reconstructs the underlying punctuation to be compatible with the syntactic tree structure; the second step rewrites the underlying punctuation to surface punctuation via a probabilistic finite state transducer. Our model admits efficient dynamic programming to maximize the incomplete data likelihood.

Our model (on 5 languages) outperforms neural network baselines in punctuation restoration, lets us appropriately render the surface punctuation when we syntactically transform the sentence. The induced rules are consistent with linguists' analysis of English.

- **Specializing Word Embeddings (for Parsing) by Information Bottleneck** [2]**:**
  Given that pre-trained word embeddings contain rich syntactic and semantic information, we propose a variational information bottleneck (VIB) method to nonlinearly compress and specialize these embeddings. Our specialized discrete tags capture most of the information in traditional part-of-speech tag annotations, but our tags can be parsed more accurately at the same level of tag granularity. Our specialized continuous tags yield more accurate parsers in 8 out of 9 languages, outperforming other compression methods like PCA.

- **Grammar induction Project**: [Ongoing project]
  Due to the cost of human-labeled syntactic trees, developing an automated method of forming syntactic trees (also called grammar induction) is an important NLP task.   We propose to address the grammar induction problem from a novel perspective: by defining a training objective based on Contrastive Estimation. We propose to use the latent tree to explain grammatical agreement, subcategorization, and selectional preference, instead of the traditional approach of just part-of-speech tagging.

- **Bayesian Non-parametric Web Scraping**: [Ongoing project]
  There is a vast amount of relational data on the web, available in HTML format. Reliably extracting all relational data is instrumental as a source of world knowledge for natural language understanding. The challenge is that the relational data is embedded in the HTML format via different scripts, and the content covers an unbounded number of topics and domains, leading us to use a Bayesian non-parametric framework. The goal is to infer the latent program and latent database for a given website by inverting the generative process.

*Research Summer Intern*                                                                 **Jun. 2019 – Sep. 2019**

Area of Focus: Natural Language Generation, structured autoencoders, and discrete latent variable models.
Advised by **Prof. Alexander Rush**
*NLP group, Harvard University*

- **Posterior Control of Blackbox Generation** [3]:
  Many tasks in natural language processing require high precision output that obeys dataset-specific constraints. This level of fine-grained control can be difficult to obtain in large-scale neural network models. We propose a structured latent-variable approach that adds discrete control states within a standard autoregressive neural paradigm (e.g. recurrent neural network). This approach allows us to provide arbitrary grounding of internal model decisions, without sacrificing any representational power of neural models.  In **conditional NLG** tasks: our method improves over standard benchmarks, while also providing fine-grained control.

*Research Assistant*                                                                          **Jan. 2017 - Aug. 2018**

Area of Focus: Medical Imaging, Visualization and Image Segmentation.
Advised by **Prof. Jerry Prince**
*Image Analysis and Communication Lab, Johns Hopkins University*

- **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.

## SERVICE

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

- **Introduction to Probability**: led session by facilitating discussions and lecturing, held office hours to answer questions and provided hints on assignments. (Fall 2017, Spring 2018, Fall 2018, Spring 2019)
- **Natural Language Processing**: led recitation to help with conceptual understanding, held office hours to help debugging code, and graded assignments. (Fall 2018, Fall 2019)

*Secondary Reviewer* --- Association for Computational Linguistics (ACL)                           **Jan. 2019**