Xiaoyin Chen

xc177@duke.edu | chenyn66.github.io | Google Scholar

Education

Duke University Aug. 2021 – May 2023
Master of Science, Computer Science, Advisor: Prof. Sam Wiseman *GPA: 4.00/4.00*

University of California, Irvine Sept. 2016 – Mar. 2020
Bachelor of Science, Computer Science, Minor in Statistics *GPA: 3.94/4.00, Major GPA:3.99/4.00*

Research

Language Model Augmented Information Retrieval Aug. 2022 – Present
Duke University | DukeNLP *Mentor: Prof. Sam Wiseman*

- Proposed a query reformulation model by leveraging pretrained language models;
- Derived a differentiable loss function based on BM25 scoring to enable end-to-end training and direct optimization;
- Reduced the average retrieval latency from 2 minutes to 4.5 seconds compared to the previous SOTA augmentation method (GAR) while achieving competitive performance;
- Will submit to ACL 2023.

Commonsense Reasoning via Knowledge Infused Text Generation Oct. 2022 – Dec. 2022
Duke University *Course Project for Neuro-symbolic AI*

- Improved and implemented a general framework for applying arbitrary non-differentiable constraints to text generation inspired by cognitive Dual-System approach;
- Formulated text generation as a tree search problem and applied a modified Monte Carlo Tree Search algorithm;
- Utilized GPT-3 for fact checking to ensure the generated sentences are consistent with commonsense;
- Improved the average constraint satisfaction rate from 90.1% to 98.4% compared to the baseline.

Evaluating Logical Reasoning Capability with Syllogisms Oct. 2022 – Dec. 2022
Duke University *Course Project for NLP*

- Proposed and implemented a pipeline for automatically generating logical questions without any human labeling;
- Developed an algorithm that samples logical questions in symbolic form by composing all 24 valid syllogisms;
- Written 100+ templates for verbalizing symbolic expressions to ensure the variety of utterances;
- Demonstrated that GPT-3 is unable to consistently infer syllogisms and generalize to a greater depth, even when all rules are given in the prompt.

Applied Machine Learning in Political Science Oct. 2021 – Sept. 2022
Duke University | DevLab@Duke (now DevLab@Penn) *Mentors: Dr. Jeremy Springman, Prof. Erik Wibbels*

- Proposed a machine learning framework for predicting civic space events from a large, noisy, machine-generated data;
- Created a pipeline for data pre-processing, model training, predicting and interpreting;
- Improved the interpretability of a tree-based boosting model by visualizing its decision function and decision path;
- Achieved 70%+ precision and 50%+ recall in predicting certain events in several countries;
- Developed a time-series model inspired by the sequential data structure.

Semi-supervised Learning with VAEs Aug. 2019 – Dec. 2020
University of California, Irvine | Learning, Inference, & Vision Group *Mentor: Prof. Erik Sudderth*

- Designed and implemented a baseline dataset to easily visualize and debug semi-supervised learning performance of models;
- Proposed and tested variants of model architectures, loss functions, sampling methods, and prior distributions;
- Implemented and tested semi-supervised methods from related papers;
- Evaluated and improved disentanglement properties of the latent representations.

Networks Alignment Analysis in Bioinformatics Sept. 2018 – Jan. 2020
University of California, Irvine *Mentor: Prof. Wayne Hayes*

- Derived conditions for optimal network alignments by information theory;
- Designed and performed empirical analyses of derived conditions using the Hungarian Algorithm;
- Proposed the first theory that explains the alignability of networks as a function of the input graph size.

Probabilistic Graphical Models for Noise Detection Sept. 2019 – Jan. 2020

University of California, Irvine *Mentor: Prof. Wayne Hayes*

- Collaborated with Prof. David Mobley of the Department of Chemistry;
- Proposed a probabilistic graphical model for correcting energy transition errors in chemical reaction networks constrained by the law of conservation of energy;
- Applied Gibbs sampling to infer the maximum a posteriori estimator of node energy levels;
- Reduced MSE from 1.63 to 1.12 (30% improvement) compared to the previous method.

Publications & Preprints

End-to-End Query Augmentation for Sparse Retrieval

Xiaoyin Chen, Sam Wiseman.

In preparation, Preprint (Available in January 2023), 2022.

An Early Warning System for Democratic Resilience: Predicting Shocks to Civic Space

Xiaoyin Chen, Jeremy Springman, Erik Wibbels.

In preparation, Preliminary report, 2022.

Learning Consistent Deep Generative Models from Sparse Data via Prediction Constraints

Gabriel Hope, Madina Abdrakhmanova, **Xiaoyin Chen**, Michael C. Hughes, Erik B. Sudderth.

4th Symposium on Advances in Approximate Bayesian Inference, 2022. arxiv.org/abs/2012.06718.

On the Current Failure – But Bright Future – of Topology-driven Biological Network Alignment

Siyue Wang, **Xiaoyin Chen**, Brent J. Frederisy, Benedict A. Mbakogu, Amy D. Kanne, Pasha Khosravi, Wayne B. Hayes.

Advances in Protein Chemistry and Structural Biology: Protein interaction networks, Volume 131.
https://arxiv.org/abs/2204.11999.

Cross-species Prediction of Protein Function by Global Network Alignment (Presentation)

Siyue Wang, **Xiaoyin Chen**, Brent J. Frederisy, Benedict A. Mbakogu, Amy D. Kanne, Pasha Khosravi, Giles R.S. Atkinson, Wayne B. Hayes.

28th Conference on Intelligent Systems for Molecular Biology (protein prediction track), 2020.
https://www.iscb.org/cms_addon/conferences/ismb2020/tracks/functioncosi.

Using Cycle Closure Constraints to Estimate and Correct for Errors in Calculated Relative Binding Free Energies

Xiaoyin Chen, David L. Mobley, Wayne B. Hayes.

Unpublished manuscript, 2020

Work Experience

Deep Learning Engineer Intern Jul. 2018 – Sept. 2018

Tencent, Guangzhou, China

- Worked on deep learning for semantic matching and information retrieval;
- Implemented and tested 8+ models from related papers;
- Proposed a novel padding method by dynamical expanding;
- Developed a joint model CNN and LSTM that achieved the best performance on the internal dataset.

Teaching Experience

Undergrad Tutor & Grader Mar. 2017 – Dec. 2017 & Sept. 2018 - Dec. 2018

University of California, Irvine

- ICS-33 (Intermediate Programming in Python) on regular expression, recursion, class inheritance, etc.;
- ICS-46 (Data Structure Implementation and Analysis in C++) on binary tree, hashing and graph algorithm etc.;
- Hold lab sessions of 40+ students for 6 hours per week to help students with homework and lead them to find the answer. Graded assignments and exams.

Project

InstantQuiz [SD Hacks 2018 Top 5] | *Natural Language Processing* https://devpost.com/software/instantquiz

- Ranked top 5 out of 109 submissions;
- Built an application that automatically generates quiz questions for any article;
- Implemented TextRank algorithm to extract important sentences;
- Developed a sequence to sequence model with LSTM to transform sentences into questions.