YIYUAN (Bill) LI

5107171939 yiyuanli@andrew.cmu.edu nativeatom.github.io

EDUCATION

Carnegie Mellon University (Pittsburgh, PA) Expected December 2019
M.S. Electrical and Computer Engineering GPA 3.64/4.0
Coursework: Foundation of Computer Systems, Computer Networks, Network, Neural Network for Natural Language Processing, Speech Recognition Understanding

Nanjing University (Nanjing, China) July 2018
B.S. Electronic Information Science and Technology GPA 4.41/5.0
Coursework: Artificial Intelligence, Algorithm, Operating Systems, Database, Parallel Processing, Data Science and Innovation (TA)

University of California, Berkeley (Berkeley, CA) January – June 2017
Exchange Student Department of Electrical Engineering and Computer Science GPA 4.0/4.0
Coursework: Statistical Machine Learning, Optimization, Independent Study

Peking University (Beijing, China) July 2016
Exchange Student College of Engineering GPA 3.94/4.0
Coursework: Compliant Robotics

RESEARCH EXPERIENCE

Student Researcher September 2018 - Present

Language Technology Institute, Carnegie Mellon University, supervisor: Professor Alan W Black

Low-resource Online Spelling Correction System

- Developed an online spelling correction system for spelling suggestion starting from zero-information of any language using HTML, JavaScript, Flask and PyTorch; achieved more than 50% accuracy with only hundreds of correct words, submitted a paper as first author to *DeepLo-EMNLP, 2019*.

Low-resource Mandarin-Shanghainese Code-Switching Study

- Analyzed topics of Mandarin-Shanghainese dataset via kmeans clustering; mined relation of code-switching rate with time and setting.
- Transferring character alignment to morpheme alignment by generating pinyin sequences pairs and trained a Seq2Seq model to collect attention information.

Student Researcher October 2016 – June 2018

Natural Language Processing Group, Nanjing University, supervisor: Professor Xinyu Dai

Unsupervised Long Academic Document Summarisation (undergrad thesis)

- Collected and cleaned 10 thousand papers from ScienceDirect; built unsupervised alignment of sections and sentences in abstract for context ranking.
- Proposed an unsupervised hierarchical Seq2Seq model and context vector in abstract generation of scientific papers; achieved ROUGE-L of 0.36.

Student Researcher September – December 2017

Lab of New Generation Network Technology & Application, Tsinghua University, supervisor: Professor Yongfeng Huang

Unsupervised Hierarchical Aspect Extraction

- Proposed an unsupervised method in hierarchical fine-grained aspects structured extraction by affinity propagation clustering and adaptation strategy via mutual information; outperformed Glove embedding pre-trained on 6 billion corpus; submitted a paper as first author to *WWW 2019*.

Student Researcher January – August 2017

Berkeley Artificial Intelligence Research Laboratory(BAIR), University of California, Berkeley, supervisor: Professor Laurent El Ghaoui

Topic Extraction and Recommendation in Chinese Restaurant Reviews

- Built web crawler in BeautifulSoup to get reviews from Dianping.com; employed hierarchical latent topic extraction and customized topic recommendation from restaurant reviews; employed word2Vec visualization of topic distribution of different restaurants.

PROJECTS

Tagging-Reinforced Code Generation February – May 2019

- Proposed Code-Tagging Policy Gradient (CTPG) model in Theano to incorporate documentation programming knowledge into Seq2Seq code generation of python; proposed failure recovery mechanism to leverage Abstract Syntax Tree (AST) failures by 30%.
- Achieved 0.73 in accuracy and 0.85 in BLEU at Django dataset, outperformed pervious retrieval-based result.

HTTP Server Development March – April 2019

- Developed a HTTP server in Java that support content request for text file, HTML file and large video file.
- Stood pressure test of 5000 concurrent requests with 95% served within 500ms in Apache benchmark.

Yelp Data Challenge March – June 2017

- Analyzed restaurant reviews in Yelp using text processing; proposed a punctuation boosted bag-of-words model; improved feature importance of opinion words of less frequency in Random Forest.
- Designed prediction pipeline, PCA and Time Series analysis for the team; achieved 0.38 Root Mean Square Error (MSE) in prediction of stars in the reviews.
- Provided geographical impact analysis in restaurant reputation by distribution mode of locations and average stars in Las Vegas.

SKILLS

Programming Languages: C, Python, Java, Matlab, R, SQL, HTML, JavaScript, CSS, Terraform
Systems and Softwares: Tensorflow, PyTorch, Theano, Keras, AWS, SAS, Git, Linux, Win, OS X
Additional: Natural Language Processing, Information Extraction, Text Summarisation