YIYUAN (Bill) LI

(510)-717-1939 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 and Understanding, Applied Stochastic Process

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 in Topic Analysis of Restaurant Reviews

RESEARCH EXPERIENCE

Researcher Assistant – Low-resource Language Study 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 interactive spelling suggestion in low-resource languages with nearly zero-knowledge.
- Developed a prototype with HTML, JavaScript and Flask, and recurrent neural network model with PyTorch.
- Achieved more than 50% accuracy with only hundreds of correct words in morphological-rich languages and OCR output of endangered languages, submitted a paper as first author to *DeepLo-EMNLP, 2019*.

Low-resource Mandarin-Shanghainese Code-Switching Study

- Conducted low resource code-switching study in Chinese mandarin and dialect (Shanghainese).
- Identified topics by k-means clustering and word ranking identification; mined relations of code-switch rate, temporal and semantic factors.
- Built morpheme alignment transferred from character alignment by generating by training translation model of pinyin sequences and accumulating attention information.

Researcher Assistant – Text Summarisation October 2016 – June 2018
Natural Language Processing Group, Nanjing University, supervisor: Professor Xinyu Dai

Unsupervised Long Academic Document Summarization (undergraduate thesis)

- Proposed an unsupervised hierarchical model for abstractive summarization of long documents.
- Collected and cleaned 10 thousand papers from ScienceDirect; built unsupervised labelling for sentences in abstracts; conducted context selection by section ranking.
- Built a hierarchical summarization model with Tensorflow to encode sentence-level and section-level information separately; employed context vector for semantic consistence in generation.
- Achieved ROUGE-L of 0.36 in abstract generation.

Research Assistant – Information Extraction September – December 2017
Lab of New Generation Network Technology & Application, Tsinghua University, supervisor: Professor Yongfeng Huang

Unsupervised Hierarchical Aspect Extraction

- Proposed an unsupervised model for fine-grained hierarchical aspect extraction in open corpus with little domain knowledge.
- Extracted opinion and target words using double propagation; constructed hierarchical structure by hierarchical clustering with mutual information-based adaptation.
- Outperformed Glove embedding pretrained on 6 billion corpora by using co-occurrence embedding from unlabeled in-domain data in golden structure matching; submitted a paper as first author to *WWW 2019.*

SELECTED PROJECTS

Tagging-Reinforced Code Generation February – May 2019

- Proposed Code-Tagging Policy Gradient (CTPG) model to incorporate documentation programming knowledge into code generation.
- Developed the reinforcement model in Theano; embedded reward from tagging in natural language description into generation of decoding tree in Seq2Seq model.
- Employed failure recovery to leverage Abstract Syntax Tree (AST) failures and increased the success of code snippet generation 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 supported content request for text file, HTML file and large video file by partial content delivery.
- 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; achieved 0.38 Root Mean Square Error (MSE) in prediction of stars in reviews.
- Provided geographical impact analysis in restaurant reputation by distribution mode of locations and average stars in Las Vegas.

SKILLS

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