YIYUAN (Bill) LI

(510)-717-1939 bill.lyy.nisioptimum@gmail.com https://nativeatom.github.io/

EDUCATION

Carnegie Mellon University (Pittsburgh, PA) December 2019
M.S. Electrical and Computer Engineering GPA 3.69/4.0
Coursework: Foundation of Computer Systems, Computer Networks, Network, Neural Network for Natural Language Processing, Introduction to Machine Learning, Speech Recognition and 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 – May 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

Multilingual Grammatical Error Correction with Minimal Supervision

- Proposed a two-stage identification-prediction model to utilize grammatical information captured by pre-trained model like BERT to extend grammatical error correction to languages lacking enough annotations.
- Achieved more than 0.7 precision in FCE dataset without fine-tuning based on pre-annotation of errors and 0.54 in error span detection; employed partial masking for information leak; worked on the error fertility module and mitigating supervision in the multilingual setting.

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 0.84 accuracy with hundreds of correct words in morphological-rich languages and endangered languages in OCR.

Researcher Assistant – Text Summarization 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 based on average ROUGE score.
- 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 to select the representative aspects and aspect elimination for pruning.
- Achieved up to 0.83 aspect-match accuracy in clothing domain, outperformed Glove embedding pretrained on 6 billion corpora by using co-occurrence embedding from unlabeled in-domain data in golden structure matching; preparing the manuscript for submission.

PUBLICATIONS

Yiyuan Li, Antonios Anastasopoulos, Alan W Black. Towards Minimal Supervision BERT-based Grammar Error Correction. *AAAI 2020* (to appear)
Yiyuan Li, Antonios Anastasopoulos, Alan W Black. Comparison of Interactive Knowledge Base Spelling Correction Models for Low-Resource Languages. *Natural Language, Dialog and Speech (NDS) Symposium, 2019*

SELECTED PROJECTS

Code Switching Instance Identification Boosted Speech Recognition September – December 2019

- Proposed a multi-task framework for code switch speech recognition by joint learning with code switch instance identification based on LAS model.
- Achieved 0.268 CER, outperformed LAS on SEAME dataset; employed edit module to for further correction; analyzed result by sentimental switching, part-of-speech tagging for better understanding of code-switching instances.

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; parsed code snippets into Abstract Syntax Tree (AST) and extracted programming elements, embedded action-wise rewards from natural language description tagging into generation of decoding tree in Seq2Seq model.
- Employed failure recovery to leverage 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 GBs video file by partial content delivery.
- Stood pressure test of 5000 concurrent requests with 95% served within 500ms in Apache benchmark.

SKILLS

Programming Languages: Python, C, Java, R, HTML, SQL, JavaScript, CSS, Matlab, Terraform
Softwares: PyTorch, Tensorflow, Keras, Theano, BERT, SRILM, AWS, Apache, Spark, SAS, Git, Scikit-learn, Pandas, Matplotlib, Spacy