CHENGXUAN (SHELTON) XIA

669-666-4614 | cxia17@ucsc.edu | https://www.linkedin.com/in/chengxuanxia/

SUMMARY

I am a graduate student with over 2 years of industry experience in an AI company, and over 2 years of academic research experience in Computer Science. I am passionate about AI/ML algorithms and general software engineering.

TECHNICAL SKILLS

- Languages: Python, Java, R, Matlab, SQL
- Development and Deployment: Git, Docker, Github, Vim, Linux/Unix, Sublime Text, SVN
- Framework & tools: PyTorch, TensorFlow, Spark, Pandas, Numpy, Matplotlib, Flask, LaTeX, MySQL, Oracle, SQLite

EDUCATION

M.S. in **Natural Language Processing** Expected Graduation: Mar. 2024

University of California, Santa Cruz San Jose, CA

- **Current GPA: 3.9/4.0**
- **Courses:** Natural Language Processing (NLP), Data Science and Machine Learning Fundamentals, Deep Learning for NLP, Advanced Machine Learning for NLP, Linguistic Models of Syntax & Semantics for Computer Scientists

M.S. in **Computer Science (Research-based)** Sept. 2019 - Jun. 2022

University of Chinese Academy of Sciences Beijing, China

- **Core Courses:** Data Structure, Algorithm design and analysis, Data Mining, Optimization Methods in Algorithms, Knowledge Graph and Semantic Computing, Text Data Mining Seminar, Big Data Analysis
- **Thesis:** *ADTMAN: Accurate Descriptive Term Matching (WWW 2023 submitted)*

B.S. in **Physics** Sept. 2014 - Jun. 2019

University of Chinese Academy of Sciences Beijing, China

- **Core Courses:** Linear Algebra, Calculus, General Physics, Probability and Mathematical Statistics

PROFESSIONAL EXPERIENCE

Software Engineer (Part-time) Sept. 2020 - Jul. 2022

Beijing Paoding Technology Co. Beijing, China

- Implemented an application with Flask API to extract and match the value-description pairs from financial statements, with SQLite Database, which was productized and sold to 16 audit firms, on average reducing 30 hours manual work per month
- Developed a keyword-based label generator with the weak-supervised method on Gitlab cloud service, generated 6 million high-quality labeled data, equivalent to 6 months' work and cost of $20,000 from human annotation
- Built a Python library for data processing, extracting information from documents that size over 60k sentences and producing Excel and JSON strings, these results will be used in downstream APIs
- Constructed a bi-directional LSTM RNN model to accurately identify and extract elements such as tables, paragraphs, and pictures from unstructured documents like contracts and financial statement

ACADEMIC EXPERIENCE

Research Assistant Jul. 2020 - Jun. 2022

Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences

- Built a model with 12 embedding layers, a bi-LSTM as transformer and combined with a financial KG and context message as external knowledge; achieved 0.963 F1 score
- Extracted 4 logical relations in financial texts and expand application scope to traditional Semantic Textual Similarity tasks
- Reformed the pipeline approach of a Named-Entity Recognition Algorithm to a span-based method to extract structured information, improved accuracy to 0.96
- Built a GRU RNN based language model to generate a continuation in Chinese (in PyTorch); achieved 22.4 perplexity on a test set of 30,000 sentences

Team Leader Nov. 2019 - Jan. 2020

CCF Big Data & Computing Intelligence Contest

- Built a machine learning model based on BERT to analyze emotional preference of Internet news, achieved 0.817 accuracy
- Utilized a python framework Scrapy to collect information from news website, built a keywords-to-emotions dictionary to increase the model performance by 8.6% and finally outperformed 98% of competitors