Yang Bai

baiyang94@ufl.edu | 352 284 2346 | Portfolio | Linkedin

SUMMARY

Researcher designing **Machine Learning** algorithms for structured and unstructured data with applications in **Knowledge Graphs** and **Natural Language Processing**.

EDUCATION

UNIVERSITY OF FLORIDA

Gainesville, FL, USA
2019 - 2023(Expected)
PhD. Computer Science
GPA: 3.96/4.0

2016-2018
MSc. Electrical and Computer Eng.
GPA: 3.5/4.0

SICHUAN UNIVERSITY

2012-2016 | Chengdu, China
BSc. Micro Electronics

SKILLS

LANGUAGES

Python :	*Expert*
Java :	*Expert*
SQL/SPARQL :	*Expert*
C/C++ :	*Intermediate*
JavaScript :	*Intermediate*
F# :	*Intermediate*

TOOLS

PyTorch	TensorFlow
Keras	Scikit-learn
Hugging Face	NLTK
SciPy	Pillow
OpenCV	OpenIE
Matlab	NumPy
Pandas	Oracle DB
REST API	Flask
Docker	Akka.NET
Git	Linux
Google Test	JUnit

COURSEWORKS

Elements of Machine Intelligence
Deep Learning for Computer-Graphics
Applied Machine Learning
Trustworthy Machine Learning
Distributed Operating System
Programming Language Principles
Database Management System
Database System Implementation
Analysis of Algorithms
Advanced Data Structures
Computer Networks

WORK EXPERIENCE

University of Florida — Graduate Student Researcher | Sep. 2019-present

- Active Interpretation of Disparate Alternatives
  - Individual contributor and team lead in the DARPA-sponsored project "Active Interpretation of Disparate Alternatives(AIDA)", an alternative hypotheses search engine over event-enteric knowledge graphs. **Our system achieves top performance at the NIST TAC SM-KBP2020 evaluation.**
  - Developed a two-level graph searching algorithm to explore knowledge graphs at both mention-level and cluster-level improving the final F1 score by **25%**.
  - Developed a graph clustering algorithm to differentiate alternative hypotheses by measuring both structural and semantic distance between candidates, which improves the original cluttering quality(v-measure) by **20%**.
- **Multi-answer open-domain question answering with controversial stance mining for query-based large-scale check-worthy claim detection**
  - Constructed a benchmark dataset using the Twitter API with three annotators.
  - Designed the new evaluation metrics, data schema, and annotation instructions.
  - Developed an annotation tool with a user-friendly UI.
  - Developed an end-to-end pipeline to evaluate how different modules along the pipeline (including information retrieval, machine reading comprehension, and distinct answer selection module) affect the final performance.

Nokia Bell Labs — Machine Learning Intern | Jun. 2022-Aug. 2022

- Proposed and implemented a retrieval-based framework to ease ticket root cause analysis by retrieving the most relevant log lines from the attached log files (10-100M log lines/ticket) given ticket information.
  - Conducted data cleaning, processing, visualization, and analysis on massive time-series semistructured system-level log corpus.
  - Developed a dense log retrieval system that finetunes self-pretrained tickets and log encoders through an adaptive multi-model machine learning framework.
  - The best model outperforms a BM25 baseline model by **16.1%**.

SELECTED PUBLICATIONS **Google Scholar**

More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering
Yang Bai,D.Wang **arXiv 2021**

GAIA at SM-KBP 2020 - A Dockerized Multi-media Multi-lingual Knowledge Extraction, Clustering, Temporal Tracking and Hypothesis Generation System
M.Li,...,**Yang Bai**,..., D.Wang **TAC 2020**

GAIA at SM-KBP 2019-A Multi-media Multi-lingual Knowledge...
M.Li,...,**Yang Bai**,..., D.Wang **TAC 2019**