Kai-Chou, Yang

As a Kaggle Competition Master and a **winner of international data science challenges**, I am experienced in machine learning, deep learning and related frameworks such as PyTorch.

My research focuses on **natural language processing (NLP)**, where I have released 11 open-source projects such as MianBot **(700+★ on Github)** and presented certain academic papers on top conferences like ACL, AAAI, CIKM, **and** WSDM.

International Awards

For the following achievements, I am the **first author** as well as the **team leader**.

2nd Place, CIKM Cup: Cross-lingual Short-text Matching Challenge

- Proposed two densely-connected architectures, CPRNN and DACNN, for sentence pair modeling.
- Fused semantic features from different levels to create diversity intra-models.
- The solution has been oral presented on **CIKM 2018 in Turin, Italy.**

3rd Place, WSDM Cup: Fake News Classification Challenge

- Implemented various NLI networks like **ESIM** and injected world knowledge using **BERT**.
- Proposed a disagreement-aware model based on the single-word attention.
- The paper has been oral presented on **WSDM 2019 in Melbourne, Australia.**

4th place, Google AI: Gendered Pronoun Resolution Competition

- Leveraged the information redundancy from **BERT** and extracted features from the optimal layer.
- Proposed a **multi-heads Siamese semantic scorer** for answer selection.
- The paper has been presented on **ACL 2019 in Florence, Italy.**

Kaggle Competition Master, Ranked top 0.2% (233/114,366)

- Top 1% (4/838), Gendered Pronoun Resolution Competition.
- Top 1% (27/4,550), Toxic Comment Classification Challenge.
- Top 3% (30/1,449), CareerCon 2019 - Help Navigate Robots.
- Top 4% (103/3,165), Jigsaw Unintended Bias in Toxicity Classification.
- Top 6% (223/3,633), CommonLit Readability Prize.
- Top 10% (384/3,946), TalkingData AdTracking Fraud Detection Challenge.

Work Experience

Taiwan AI Labs, Machine Learning Engineer, Sep 2019 ~ Now

Question Answering System

- Propose a conditional question generator with mT5 for controllable QA data augmentation and as the base of dense retrieval, which improves recall@50 from baseline model by 16%.
- Build a generative pseudo labeling pipeline using a open-domain passage retriever and machine reader, which improve the nDCG@10 by 4.2 - 9.7, on various domains.
- Build an efficient passage re-ranker based on tiny-bert with a time-series based clustering framework for effective negative passage sampling.
- Leverage FinBERT on QA analysis and slot filling for fintech dialogue system.

Natural Language Understanding

- Implement a document encoder with self-contrastive learning and a document clustering algorithm, which is scalable for million scale of streaming data.
- Implement a GROVER-like generator as the backbone for topic detection, article rewriting, and tag generation.
- Propose a semi-automatic framework for fake-news identification, which gathers evidence from event properties, user behavior and textual features.
- Propose a SOTA Chinese typo correction system based on a boosting loop of automatic speech recognition and text to speech for weak supervision.
- Build a general-purpose NLP training pipeline for team use involving data augmentation, data regularization, and unsupervised domain adaptation.

Education

Master in Department of Computer Science, NCKU **GPA: 4.30**

- Honorary member of the Phi Tau Phi Scholastic Honor Society. (Ranked 1st among all graduates.)
- As a teaching assistant for Introduction to Data Science, Data Mining and Discrete Mathematics.
- As a speaker / teaching assistant for introduction lectures of machine learning.

Bachelor in Department of Computer Science, NCKU **GPA: 3.92**

- Academic excellence awards 2016.
- Academic excellence awards 2015.
- Honorable mention on the graduation exhibition.
- Research assistant on a question answering system project for the Ministry of Science and Technology.

Side Projects

I list some of my project experiences. You can refer to my Github for the other interesting ideas.

Mianbot

- Got 700+ stars and 200+ forks on Github.
- Implemented the hierarchical keywords matching using **word2vec**.
- Implemented the IR-based searching module to support chit-chat.
- Allow user to define customized scenarios with JSON.
- The extracted QA pairs were released in PTT-Gossiping-Dataset, a widely-used Chinese chit-chat corpus.

NCKU Smart-Life LineBot

- A Linebot that helps solve trivial matters such as restaurant recommendation.
- The dialogue system is based on LUIS for intent classification.
- The backend was built with Django / Flask (new version) and host on Heroku.
- The backend is connected with Line server using the web API.

Knowledge & Skills

- General Machine Learning
  - Classification, Regression, Clustering, Boosting, Feature Engineering.
- Natural Language Processing
  - Sentence Pair Modeling: Natural language Inference, Machine Reading Comprehension, Sentence Similarity
  - Text Classification / Regression / Clustering
  - Deep contextual representation (ELMO / BERT / XLNet / ELECTRA / RoBERTa / ERINE2.0 / BigBird / T5)
- Recommendation System
  - Factorization: Matrix Factorization, Factorization Machine, DeepFM
  - Graph Embedding: DeepWalk, Node2Vec, item2Vec

Publication

1. Fake News Detection as Natural Language Inference. **Kai-Chou Yang;** Timothy Niven; Hung-Yu Kao. WSDM Cup 2019
2. Fill the GAP: Exploiting BERT for Pronoun Resolution. **Kai-Chou Yang;** Timothy Niven; Tzu Hsuan Chou; Hung-Yu Kao. ACLWS'19
3. Generalize Sentence Representation with Self-Inference. **Kai-Chou Yang;** Hung-Yu Kao. AAAI 2020
4. The Prevalence and Impact of Fake News on COVID-19 Vaccination in Taiwan: A Retrospective Study of Digital Media. Yen-Pin Chen; Yi-Ying Chen; **Kai-Chou Yang;** Feipei Lai; Chien-Hua Huang; Yun-Nung Chen; Yi-Chin Tu. JMIR