Yong Hu 胡勇

nghuyong@163.com · (+86) 188-1020-9302 · nghuyong · nghuyong.top

Education

M.S. in Computer Science, Beijing Institute of Technology, Beijing, China 2018.9 – Present

- Research focusing on natural language processing, supervised by Heyan Huang and Xianling Mao.
- Currently working on pre-trained language models, question answering system, and social computing.

B.S. in Internet of Things, Jiangnan University, Wuxi, China 2014.9 – 2018.7

- A student in Honour School, which is the cradle of top-notch innovative talents training in Jiangnan University and only admits the **top 3%** of whole first-year students.
- **National Scholarship** and **Honorary Degree**.

Abilities & Skills

Love trying and challenging new things, keep doing what I care about to the extreme.

- **Strong self-learning ability**: A strong ability to explore and quickly get started in new areas independently; Good at using Google, stackoverflow, GitHub to solve problems.
- **Strong engineering ability**: Proficient in Python, understand Go and js; Proficient in building deep learning models by PyTorch and Keras.
- **Keen on knowledge sharing**: Writing technical articles and contributing to open source projects.
- **Good product sense**: Have innovative thinking and love product design, research, and implementation.
- **Proficient in English**: Reading and writing English documents fluently (IELTS: 6.5, Reading: 7.0).

Work Experience

Researcher, Effyic, Beijing, China 2019.7 – Present

- Effyic is an AI startup, dedicated to building intelligent dialogue platform.
- **Zero to One**: Lead the team to build a FAQ-based question answering system and applied it into the intelligent customer service product, which has been successfully landed in many companies such as Daojia.
- Design, develop and deploy a series of model services:
  - ES-based and embedding-based recall service
  - BERT-based listwise ranking service
  - UniLM-based similar text generation service
  - BERT-based sentiment classification service
  - Deploy the model services with flask, tensorflow-serving and ELK on k8s.
- Based on this QA system, **50%+** of the queries are solved by robots independently and the satisfaction of robot's answers is **80%+**, which significantly improves the efficiency.

Research Intern, WeChat AI, Tencent, Beijing, China 2019.3 – 2019.7

- In charge of the research and engineering of the first-level category classification on WeChat public accounts platform including nicknames (short texts) and articles (long texts).
- Applied BERT (short texts) and TextCNN (long texts) on this task, and the final recall rate was **80%+** and the precision rate was **97%+** on **more than 40 categories**, **much higer (12%)** than the initial demand.
- Deployed the model on Hadoop and classified all nicknames and articles (**20M+**) every day.

Software Engineer Intern, Jiuheng Technology, Beijing, China 2018.7 – 2018.9

- Jiuheng Technology is a digital currency quantitative fund.
- **Zero to One**: Built and maintained a **stable** and **real-time** digital currency transaction data collection system from scratch, which is the **basis** of quantitative trading research.
- The system is based on distributed architecture (MQ, docker-compose) and can collect and process **10M+** transaction data (Bitcoin, Huobi, Litecoin) per hour.

Open Source Projects

WeiboSpider (2.1K+ Stars & 580+ Forks) 2018.9 – Present

- **Continuously** maintaining an open source Weibo data collection system.
- **Rich collection content**: User information, tweet content and comments and social relataionship.
- **Distributed architecture**: **100M+** data collection per day, based on docker, account pool, and ip pool.

ERNIE-Pytorch (450+ Stars & 60+ Forks) 2020.7 – Present

- Converted Baidu's ERNIE (paddlepaddle version) to huggingface's format (PyTorch version), making this powerful Chinese pre-trained language model easy-to-use.
- Contributed this series of models (ERNIE-1.0, 2.0, tiny) to **huggingface/transformers** repository.

Research Experience

COVID-19 Data Collection on Social Media 2020.5 – 2021.1

- Paper: Weibo-COV: A Large-Scale COVID-19 Social Media Dataset from Weibo accepted by the 1st Workshop on NLP for COVID-19 (Part 2)@**EMNLP2020** (**first author**).
- Proposed a novel strategy to construct Weibo public opinion datasets, which can build **large-scale** datasets with **high efficiency** for the **first time**.
- Released Weibo-COV, a first large-scale Chinese social media dataset (**40M+**), focused on COVID-19 and containing rich field information. Released Weibo-COV V2 later with longer time span, bigger data size (**65M+**) and more refined keyword filtering method.
- Received **200+** dataset applications from worldwide research institutions and has supported **30+ publications** and **100+ programs**.

Multi-Modal Classification on Social Media 2020.3 – 2020.5

- Paper: A Cross-Modal Classification Dataset on Social Network accepted by **NLPCC2020** with the **oral presentation** (**first author**).
- Applied adversarial filter to build a **high-quality and high-difficulty** cross-modal weibo posts classification dataset with 85K+ posts, three modalities of text, image and video, and 18 general categories.
- Implement classical cross-modal baselines for tweets classification and empirical results show that the classification over this dataset is challenging enough.

Adaptive Learning on Second Language Learning 2019.1 – 2019.3

- Paper: Multi-task Learning for Low-resource Second Language Acquisition Modeling accepted by **APWeb-WAIM 2020** (**first author**).
- Proposed a novel second language acquisition modeling method, which learns the latent common latent patterns among different language-learning datasets by multi-task learning.
- Our method performed **much better** than SOTA baselines in low-resource scenarios and also slightly improved in non-low-resource scenarios (**refresh 2018 Duolingo shared task**).