Mengmeng KUANG

Email: mmkuang@connect.hku.hk | Phone: (+86) 132-7015-3586

Address: B5, T.I.T. Creative Park, No.397 Xingang Mid. Rd., Haizhu Dist., Guangzhou, China.

GitHub: https://github.com/kuangmeng | Home Page: http://about.meng.uno

EDUCATION

DEC. 2020 **M.Phil** Degree in COMPUTER SCIENCE, **The University of Hong Kong**
Supervisor: Prof. Hing-fung TING
Thesis: Data-centric Approaches for better Multiple Sequence Alignment
Research Interests: Sequential Modeling for Multiple Sequence Alignment
GPA: Not Applicable for Research-based Postgraduate

JUN. 2018 **B.Eng** Degree in COMPUTER SCIENCE, **Harbin Institute of Technology**
Thesis Advisor: Prof. Tiejun ZHAO
Thesis: Cross-domain High Precision Chinese Word Segmentation
Research Interests: Language Analysis Technology and Application
GPA: 89.6/100

EXPERIENCE

Work

Mar. 2021 - Cur. | *Applied Researcher*, TENCENT TECH. (WECHAT GROUP), Guangzhou
Participated in the design and implementation of query rewriting, document understanding, semantic retrieval algorithms and data label noise reduction methods for WeChat Search.

Oct. 2020 - Feb. 2021 | *Research Intern*, SMARTMORE TECH., Hong Kong
Engaged in the data mining, data augmentation, model design and framework implementation of SMore OCR.

Research *[Selected first author papers]*

[C1] Efficient two-stage label noise reduction for retrieval-based tasks. (WSDM 2022)

[J1] MLProbs: A Data-centric Pipeline for better Multiple Sequence Alignment. (IEEE TCBB)

Project

A two-stage label noise detection method for classification datasets | Mar. 2021 - Cur.
Data quality has always been the bottleneck restricting the breakthrough of deep learning models. In order to improve the accuracy of existing artificially labeled data, we proposed an effective two-stage noise label detection method. Firstly, we used BERT to train a rough classifier on all the data to be denoised and generated a noise candidate set. Then we predicted the classification probability on that set and calculate the confidence matrix recognizing which sample is unreliable. Experimental results showed that this method can improve the clean label rate to more than 96%.

A property-continuing text generator for query rewriting | Nov. 2021 - Cur.
Natural Language Generation (such as sequence-to-sequence models) based query rewriting methods usually suffer from severe problems of insufficient generalization and serious semantic shift, which is not suitable for online tasks that require high accuracy. In this project, we propose to complete the query rewriting task based on the pre-trained model BERT by learning the inherent language pattern (i.e., language style) of the query to be rewritten. This method shows outstanding results on BLEU-4, Rouge-L, and SARI metrics on three benchmarks, the QRECC dataset, the LCQMC dataset, and a Private dataset collected by ourselves.

Keywords weighted Siamese model for semantic retrieval | Mar. 2021 - Jul. 2021

To retrieve better-matched documents, it is necessary to identify the keywords in the queries and documents accurately. We proposed a novel domain adaptive multi-task model by joint training a Siamese matching model with a keywords identification model to precisely acquire the query-document relevance. The Siamese model produced query and document semantic vectors independently and coupled them only in the similarity calculation stage. We introduced a keyword identification model to detect keywords from queries and documents automatically. Empirical results demonstrated that our method outperforms other competitive baselines on two semantic retrieval datasets (i.e. MS MACRO and WeChat Search datasets).

Data-centric Approaches for better Multiple Sequence Alignment | Sept. 2018 - Jun. 2020

To improve the quality of multiple sequence alignment (MSA) construction on protein families, especially the "low similarity" ones, we proposed a two-stage sequential modeling-based MSA method by training a decision-making model with CNNs, BiLSTM and Attention mechanism to arrange suitable algorithm-centric pipelines for different categories of the protein families. The average accuracy could be improved by 2.8% on 711 "low similarity" protein families through this method.

Cross-domain High Precision Chinese Word Segmentation | Nov. 2017 - Jun. 2018

To improve the accuracy of Chinese Word Segmentation, a system based on conditional random field and Viterbi algorithm was developed with Java training from the artificial word segmentation results of The People's Daily (1998). In order to further improve the adaptability in specific fields (medicine, law and finance), heuristic rules and specific guidelines were added. Finally, this word segmentation system could get an average accuracy of 97% in these specific fields.

Teaching

Jan. 2020 - Jun. 2020	COMP1117 Computer programming
Feb. 2019 - Jun. 2019	COMP7606 Deep learning
Sept.2017 - Feb. 2018	13SC03100600 Software engineering
Sept. 2017 - Feb. 2018	IR03000900 Semantic mining of Internet text

Scholarships and Certificates

Postgraduate

Mar. 2021	The Li Ka Shing Prize (Nominated)
Mar. 2021	HKU Outstanding Research Postgraduate Student (Nominated)
Nov. 2020	Huawei Certified ICT Associate – Artificial Intelligence
Nov. 2018	Certificate of Teaching and Learning in Higher Education
Sept. 2018	Postgraduate Scholarship

Undergraduate

Jun. 2018	Enterprise Scholarship
Dec. 2015, Dec. 2016	Merit Student
Nov. 2015, Nov. 2016	National Encouragement scholarship

Skills

Languages	Chinese Mandarin (Mother tongue), English (Fluent)
Programming	C/C++, Java, Python etc.
Deep learning frameworks	TensorFlow, PyTorch, Keras, Scikit-learn