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.

Scholar: scholar.google.com/citations?3RHx3dsAAAAJ | Homepage: https://mkuang.hk

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
Design and implementation of deep learning models for query rewriting, data augmentation, semantic retrieval, and data label noise reduction tasks.

Oct. 2020 - Feb. 2021 | *Research Assistant*, IMPERIAL COLLEGE LONDON (NHLI)
Designed a medical imaging analysis project and proposed a 3D embedded attentive RNN segmentor for left ventricle delineation from myocardial velocity mapping, supervised by Dr. Guang Yang.

Jun. 2018 - Aug. 2018 | *Research Assistant*, HARBIN INSTITUTE OF TECHNOLOGY (MI&T LAB)
Developed a medical information retrieval and diagnosis platform based on domain adaptive statistical translation, supervised by Prof. Tiejun Zhao.

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 style-continuing text generator for query rewriting and data augmentation | *Nov. 2021 - Cur.*
Natural Language Generation (such as sequence-to-sequence models) based text generation methods usually suffer from insufficient adaptation problems and serious semantic shift problems, which is not suitable for online tasks that require high accuracy. In this project, we propose to complete the text generation task based on the pre-trained model BERT by learning the inherent language pattern (i.e., language style) of the text to be rewritten or augmented. 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, which will be used on real-life query rewriting and data augmentation tasks.

Label noise detection method for classification datasets in NLP field | Mar. 2021 - Cur.

Data quality has always been the bottleneck restricting the breakthrough of deep learning models. To improve the accuracy of existing artificially labeled text 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 calculated the confidence matrix recognizing which sample was unreliable. Experimental results showed that this method could improve the clean label rate to more than 96%. We are still working on brand-new detection methods.

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 jointly training a Siamese matching model with a keywords identification model to acquire the query-document relevance precisely. 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 sequential models (i.e. Transformers) 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.

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 were developed with Java training from the artificial word segmentation results of The People's Daily (1998). Heuristic rules and specific guidelines were added to improve adaptability in specific fields (medicine, law, and finance). Finally, this word segmentation system could get an average accuracy of 97% in these specific fields.

Teaching

Jan. 2020 - Jun. 2020	(COMP1117[HKU]) Computer programming
Feb. 2019 - Jun. 2019	(COMP7606[HKU]) Deep learning
Sept. 2017 - Feb. 2018	(13SC03100600[HIT]) Software engineering
Sept. 2017 - Feb. 2018	(IR03000900[HIT]) 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
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