Mengmeng KUANG

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

Google Scholar: https://scholar.google.com/citations?3RHx3dsAAAAJ

GitHub: https://github.com/kuangmeng | Homepage: https://mkuang.nooa.tech

EDUCATION

DEC. 2020 **M.Phil** 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
GPA: Not Applicable *RANK:* 1/3

JUN. 2018 **B.Eng** in COMPUTER SCIENCE AND TECHNOLOGY, **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 *RANK:* 23/242

EXPERIENCE

Work

Nov. 2023 - Cur. | *Algorithm Engineer (*-*)*, DATA-DOUYIN, BYTEDANCE INC., Shanghai

- **Message recommendations for Douyin apps:** We construct recommendation algorithms for message content within Douyin Apps, encompassing processes of retrieval and ranking.

Mar. 2021 - Nov. 2023 | *Applied Researcher (T8)*, WECHAT GROUP, TENCENT TECH., Guangzhou

- **LLMs for retrieval:** (1) We built a single/multi-turn dialog dataset generation strategy based on "self-instruction" for the RLHF of LLMs. (2) To provide a factual basis for the LLMs, we studied the feasibility and effectiveness of WeChat dragon box accessing the LLMs rediction in the form of "Plugins". (3) We proposed a "share memory" mechanism to effectively alleviate the limitations of the context length.

- **Document understanding:** (1) Implemented a WeChat search data augmentation system from scratch. By augmenting data such as document titles, the scope of retrieval is expanded, and the accuracy of the correlation module is improved. (2) Analysis of descriptions, entities, and subpages of documents to understand texts that facilitate retrieval and correlation calculations.

- **Semantic retrieval:** (1) To retrieve more appropriate documents, we introduced a keyword-weighted Siamese model and trained a high-quality semantic retrieval queue to expand retrieval. (2) Considering the efficiency, we adopt the clustering quantification method for online deployment and inference.

Project

A Universal Data Augmentation System for Online Retrieval Systems | Aug. 2022 - Nov. 2023

We propose a universal data augmentation system to improve the versatility of the multi-data source retrieval system, describe the document entities more comprehensively and accurately, and solve the problem of industry-specific expressions that are difficult to solve by search rewriting. The system can uniformly augment and understand multi-source and heterogeneous data, and output data with uniform structure for retrieval and sorting. The system can comprehensively use techniques such as semantic retrieval, extraction models, generative models, back-translation, synonyms, knowledge graph, etc., to obtain high-quality augmented data from sources such as artificial dictionaries, user search sessions, external crawler data, and document descriptions. The recall rate of the WeChat Search Engine using this system has reached more than 95%.

Label noise detection method for classification datasets | *Mar. 2021 - Nov. 2023*

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 cleaned and generated a noise candidate set. Then we trained the classifier by the data remained, predicted the category probabilities on the noise candidate set, and calculated the confidence matrix to recognize which sample was unreliable. Experimental results showed that this method could improve the clean label rate to more than 96%.

A style-invariant text generator | *Nov. 2021 - Nov. 2022*

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. To fuse the learned text style information, we proposed a stacked adaptive instance normalization (SAdaIN) module. 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 was used on real-life query rewriting and data augmentation tasks in WeChat Search.

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 only in the similarity calculation stage. We also 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 in the evaluation.

Research *[Selected first author papers]*

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

[C2] Multi-task learning based Keywords weighted Siamese Model for semantic retrieval. (PAKDD 2023)

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

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

Work

Jan. 2023 Tencent Business Breakthrough Gold Award

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