Mengmeng KUANG

Email: kuangmeng@msn.com | Phone: (+86) 132-7015-3586

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: Machine Learning, Deep Learning, Bioinformatics

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
GPA: 89.6/100

EXPERIENCE

Work

Mar. 2021 - Cur. | *Application Researcher*, TENCENT (WECHAT GROUP), Guangzhou
Participated in the design and implementation of semantic retrieval-related algorithms and noisy label detection algorithms for WeChat Search.

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

Project

Keywords weighted Siamese model for semantic retrieval | Mar. 2021 - Cur.
In order 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 language 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, which allowed the document embeddings to be pre-learned offline. We also introduced a keyword identification model to detect the domain 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).

A two-stage label noise detection method for classification datasets | Mar. 2021 - Jun. 2021
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 use BERT to train a rough classifier on all the data to be denoised and generate a noise candidate set through this model. Then we predict the classification probability on that set and calculate the confidence matrix recognising which sample is an unreliable sample. Experimental results showed that this method can improve the clean label rate to more than 96%.

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 deep learning-based MSA method by training a decision-making model with CNNs, BiLSTM, Attention 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. 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.

Research

[C1] A data-centric pipeline using convolutional neural network to select better multiple sequence alignment method. (ACM-BCB 2020)

[C2] DLPAlign: A Deep Learning based Progressive Alignment Method for Multiple Protein Sequences. (CSBio 2020)

[C3] Three-Dimensional Embedded Attentive RNN (3D-EAR) Segmentor for Left Ventricle Delineation from Myocardial Velocity Mapping. (FIMH 2021)

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

[C4] An efficient two-stage label noise detection method for text classification datasets. (Submitted to EMNLP 2021)

[C5] MKSM: Multi-task learning based Keywords weighted Siamese Model for semantic retrieval. (Submitted to CIKM 2021)

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

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 (Mother tongue), English (Fluent)
Programming	C/C++, Java, Python etc.
Frameworks	TensorFlow, PyTorch, Keras, Scikit-Learn