Yu Yin

Email: yinyu201906@gmail.com

EDUCATION BACKGROUND

Imperial College London(IC), UK 09/2023 – 09/2024
MSc in Applied Computational Science & Engineering

University of Liverpool (UoL), UK 09/2021 – 07/2023
BSc in Computer Science | First Class Degree

Xi'an Jiaotong-Liverpool University (XJTLU), China 09/2019 – 07/2021
BSc in Information Computer Science

RESEARCH INTERESTS

Natural Language Processing, Deep Learning, Machine Learning

PUBLICATIONS

[1] Y. Yin, Q. Zhang, and M. Fang, Transfer Learning for Biomedical Named Entity Recognition without Biomedical Resources. (Under Review)
[2] Y. Yin, J. Zhao, Q. Zhang, M. Pechenizkiy, and M. Fang, CovidNews: Answering Questions on Multilingual Covid News. (Under Edition)

RESEARCH EXPERIENCES

Analyzing and Enhancing the Chain of Thought (CoT) on GPT3.5 | Research, UoL 04/2023 – 07/2023

- Conducted an empirical investigation into the impact of zero shot Chain of Thought (CoT) and few shot CoT approaches on the performance of the GPT-3.5 model. Controlled variables to account for various factors, systematically explored the model's response to different conditions, and sought strategies to optimize the GPT-3.5 model's performance:
  - Conducted an extensive review of the literature about zero-shot and few-shot approaches to sort out and integrate the various factors that impact the performance of large language models
  - Randomly partitioned 25% of the dataset for validation purposes to enable the assessment of various factors' effects on the model, while the entire dataset was employed as a test set to evaluate the model performance
  - Multiple CoT prompts were manually created for each of the collated influences, subsequently input into the GPT3.5 model, and assessed for performance against the validation dataset
  - Collected the experimental results with different influences, comparing the results with the standard prompt (without CoT), and selected the influences with large differences in performance for further experiments
  - Entered the CoT prompts corresponding to the selected influences into the GPT-3.5 model to assess its performance using the test set. Analyzed the experimental results and summarized the extent to which each factor affected the performance of the GPT3.5 model
  - Manually generate novel CoT prompts based on the experimental findings, with the objective of enhancing the performance of the large predictive model

Applying Transfer Learning Method for Biomedical Named Entity Recognition | Research, UoL 01/2023 – 04/2023

- Proposed a transfer learning method to train the biomedical named entity recognition (NER) model using non-biomedical resources. The proposed method can significantly enhance the performance of the models on biomedical NER with limited data and reduce the reliance on extensive biomedical NER datasets:
  - Conducted a background review of the field of biomedical named entity recognition (NER), including extensive literature reading, which suggests that current state-of-the-art biomedical NER models perform suboptimal on limited-size biomedical datasets
  - Implemented a transfer learning approach, devised the model training framework, and employed the pre-trained Bio-LM model as the shared backbone model
  - Selected the widely used generic dataset CoNLL2003 as the non-biomedical NER dataset and standardize the format of various datasets to follow the BIO format. Additionally, several state-of-the-art NER models from previous research were chosen as baselines
  - Separated the annotations of LOC, ORG, and PER within the CoNLL2003 dataset into three distinct CoNLL2003 sub-datasets. Subsequently, utilized these three CoNLL2003 sub-datasets along with the same biomedical NER dataset for training. Furthermore, evaluated the model's performance, and identified the best-performing sub-dataset from CoNLL2003 for use in subsequent experiments
  - The chosen CoNLL2003 sub-dataset was employed for transfer learning training along with various biomedical NER datasets. Collected experimental results and conducted a comparative analysis of the performance of the baseline model and the proposed transfer learning method on multiple biomedical NER datasets

- The proposed transfer learning method has demonstrated its effectiveness in significantly enhancing the model's performance across multiple biomedical NER datasets, particularly in cases involving small-scale biomedical NER datasets

Creating an Answering Questions System on Multilingual Covid News | Research, UoL 10/2022 – 12/2022

- Implemented a multilingual COVID-19 news question-answering system to improve the accessibility of reliable and credible COVID-19 information for users across different languages.
  - Conducted an extensive review of the literature about COVID-19 question-answering systems and the COVID-19 pandemic. The analysis revealed a widespread disinformation regarding COVID-19, with a notable scarcity of COVID-19 question-answering systems focusing on factual news content
  - Proposed a COVID-19 news question-answering system structure and leveraged the MM-COVID (Multilingual and Multidimensional COVID-19 news) dataset as the comprehensive multilingual news corpus
  - Employed mDPR (multilingual dense passage retriever) to retrieve the top-k most relevant news passages in the same language as the input query from all news articles in different languages containing multidimensional information such as news content, agency and URL
  - Designed an web interface to present multilingual and multidimensional information about the COVID-19 news returned by mDPR

WORK EXPERIENCES

Algorithm Engineer Intern | Google, Shanghai, China 06/2022 – 08/2022

- Built, trained, and validated a model to predict customer loss and analysed the factors that might affect the decision-making of YouTube users:
  - Imported Excel data with the R language and divided the data into training, verification and testing sets with the adjustment of data types
  - Employed Logistics Regression to build the targeted model, covering 1) constructing regression equations after determining independent and dependent variables, 2) estimating regression coefficients, 3) carrying out significance testing through the analysis of variance tables and 4) assessing the model's performance; concluded that the purchasing power level, user-level, and promotional sensitivity were important factors of customer loss
  - Built an XGBoost algorithm model and obtained the different levels of importance of each variable in the model
  - Introduced the confusion matrix and ROC curve to assess the performance of the two models and found that the difference was not obvious though the one based on XGBoost with relatively better results

- Applied the two models to locate the users with the tendency of loss more accurately, comprehensively analyse the specific causes of customer loss and manage customers with lower costs and enhanced user experience

Software Engineer Intern | Chengdu Yingnuo Industrial Co., Ltd., Sichuan, China 07/2021 - 08/2021

- Designed and implemented a set of user interfaces based on customer requirements
- Finished the writing of user research reports and feasibility analysis reports by collecting and analysing the relevant information on user characteristics, emotions, habits, psychology and needs
- Implemented the functional points of user interfaces and improved interactive experience with a reasonable layout and user-friendly interactive modes
- Conducted a number of test cases to debug each function and ensure its normal work; passed the user acceptance testing successfully for the final product delivery
- Connected the user interfaces with industrial cameras for the identification of shadow points

SKILLS

Programming Languages: C/C++ | Java | Python | R | MySQL
Libraries: NumPy | SciPy | Pandas | Scikit-Learn | PyTorch | Keras | TensorFlow