A Lightweight Linguistic-Enhanced Word Embedding Framework for Medical Q&A Chatbot

Zhou, Jiayi (School: Shanghai High School International Division)

The cost of computing resources to train the Large Language Model (LLM) is very high. In addition, LLMs cannot meet the requirements under medical Q&A scenario, which is requested to provide precise answers verified by authorities and be able to trace the data source. To resolve the challenges, we use Semantic Text Similarity (STS) to select the answer from reliable Q&A dataset by searching for the most similar pair of question and answer. We propose a new Word Embedding & Siamese LSTM & Retrieval-Augmented Generation framework. More specially, we derive three practical algorithms: we create a new algorithm called “Chiyori Word Embedding (CWE)”. We leverage the existing relationships between Chinese characters to represent the words and the relationships between different words. With this algorithm, we have innovated a new way to convert texts into word embeddings. In addition, to reflect the information of word sequencing, we use Siamese LSTM algorithm which uses the word embeddings generated by the CWE algorithm as input. Finally, to provide more user-friendly response to the users, we use Retrieval-Augmented Generation to generate the response by adding the pair of question and answer as input. Our extensive experiments on the LCQMC dataset demonstrate that the new framework achieves better performance than Sentence-BERT (LLM fine-tuned for STS task) on accuracy and time spending for prediction on STS task, with much less consumption of computing resources. In our future research, we will evaluate the performance with other ideographic characters and combine with Transformer which is the common model for LLMs for achieving better performance.

Awards Won: