WDPS Assignment

https://github.com/Looong01/WDPS-GPT

Bo-Chian Chen
Vrije Universiteit Amsterdam
b.chen4@student.vu.nl

Yingda Song
Vrije Universiteit Amsterdam
y.song4@student.vu.nl

Loong Li
Vrije Universiteit Amsterdam
z.li14@student.vu.nl

Yida Tao
Vrije Universiteit Amsterdam
y.tao2@student.vu.nl

I. Problem Definition

The goal of the assignment is to implement a program that post-processes the output of large language models to improve the quality of its answers. To end this, first, we choose LLaMa as our large language model. Second we use Bing as our aid of existing knowledge bases from web. Lastly, we select SentenceTransformer for comparing the answers and extract entities.

II. Choosing LLM

I have selected LLaMA-13b-chat-hf,hereinafter referred to as LLaMa, as my preferred large language model based on a careful evaluation of its balanced advantages that align with the specific requirements of my project. In contrast to GPT-3, characterized by a higher parameter count and larger dataset but accompanied by concerns related to computational resources and data quality, LLaMA exhibits a commendable equilibrium, offering a substantial parameter count without compromising computational efficiency. In comparison to PaLM, renowned for its enhanced memory capacity and factual coherence, LLaMA distinguishes itself through its superior number of parameters and access to a more extensive training dataset. The decision to opt for LLaMA2-13b-chat-hf (13b is the 13 billion parameter and chat is the chat-specific tune-tuning model) is underpinned by its capacity to deliver a comprehensive and well-rounded solution, meeting the nuanced demands of the project at hand.

III. Choosing existing knowledge bases

I have selected the Bing as an essential tool for refining answers generated by large language models, recognizing its capacity to deliver the most up-to-date information from the web. The Bing functions as a robust tool, encompassing a wide array of capabilities that enable us to articulate both straightforward and complex inquiries. Diverging from conventional web searches, Bing is strategically designed to curtail extraneous information and streamline tasks, facilitating a more efficient decision-making process for us.

IV. Sentence representation model trails

To assess correctness, two candidate methods, namely SentenceTransformer and BERT, were employed. To gauge the accuracy of each method, distinct confidence thresholds of 0.85 and 0.9 were independently applied to responses obtained from both LLaMa and Bing for identical queries.

A. Case: Entity extraction

There are large difference between them. SentenceTransformer can always perform the extraction normally. On the other hand, BERT will always mistake d, p, u as an entities and thus extract them.

B. Case: 0.85 similarity

In this trial, correctness was determined by considering answers with over 85% similarity for identical queries from LLaMa and Bing. The precision of SentenceTransformer reached 103/200, and BERT achieved a precision of 100/200, indicating their respective abilities in accurately identifying correct responses.

C. Case: 0.90 similarity

In this trial, correctness was determined by requiring answers from LLaMa and Bing to have over 90% similarity for identical queries. Under this criterion, SentenceTransformer achieved a precision of 50/200, whereas BERT maintained a precision of 100/200, showcasing their respective performance in identifying correct responses.

V. Manual verification

After conducting a thorough analysis of confidence levels, a notable discrepancy in correctness across various confidence levels became apparent. To pinpoint the origins of these errors, a meticulous examination of queries and manual inspection of test results were undertaken. First and foremost, scrutiny was applied to the LLaMa extracted answer, revealing a surprising alignment with the ground truth in a majority of instances. Subsequently, attention shifted to the examination of entities extracted from Bing. Remarkably, over 90% of these entities proved to be relevant and corroborated the answers obtained from LLaMa, as validated through a meticulous line-by-line verification process. Consequently, it is discerned that the primary bottleneck within our system lies in the correctness model governing the alignment of answers from both models, specifically the SentenceTransformer. This observation operates under the assumption that the responses from both LLaMa and Bing approximate the ground truth with a high degree of fidelity.

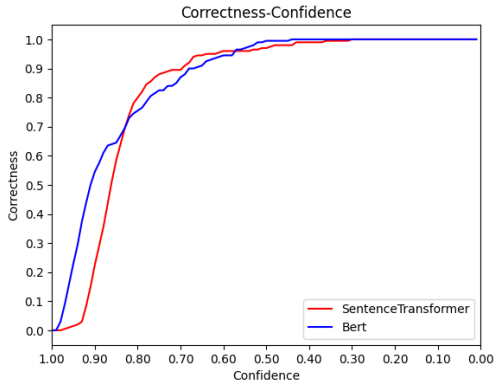

Fig. 1. Correctness-Confidence trend

A. Correctness-Confidence trend

We stabilized the model using SentenceTransformer and constructed a convergence graph. In Figure 1, we depict the convergence of correctness against confidence, ranging from 1 to 0. Notably, the correctness attains a value of 1 when the confidence is 0, signifying a scenario where responses are deemed identical even in cases of stark dissimilarity. However, a crucial observation emerges as we ascertain that the correctness approaches approximately 0.9 at a confidence level around 0.735. This outcome underscores the efficacy of the employed coefficients, indicating our ability to implement strategies that are not only systematically sound but also align with the model's correctness.

B. Decision

In summary, we've chosen SentenceTransformer for its consistent superiority over BERT in precision and correctness verification tasks. Supported by manual verification and the Correctness-Confidence trend graph, SentenceTransformer demonstrates stability and systematic soundness, making it the optimal choice for our language understanding needs.

VI. Implementation

A. Code Structure

1 The question is stored in a txt file.
2 The answers to the questions are obtained through the Bing search engine and then processed using Llama to extract the entities in the web answers.
3 Load the tokenizser and model for the LLAMA-2 pre-trained model. Iterate over the given question, process the question using the tokenizer and convert it to model input.Generate answers using the LLAMA-2 model and write them to the appropriate files.
4 Use the KeyBERT model of Sentence-Transformers and the BERT model to extract entity keywords from the answers and store them in the corresponding files.
5 The similarity between the generated answers and the answers given by the search engine network is compared to determine the correctness of the answers given by the language models.By comparison, we find that Sentence-Transformer has better performance. So we choose the answer generated by it as the results.
6 Write the answers in the specified format to a txt file.

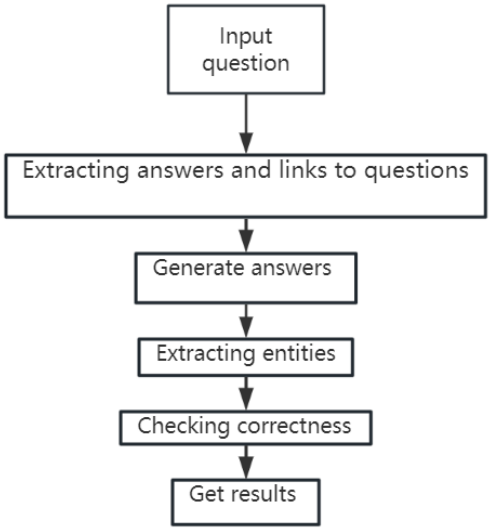

Fig. 2. System Structure

B. Instruction

For instruction, dependencies and more, click here.

VII. Evaluation

A. Test Result

After running on the 200 sample queries, we got 179/200 correctness. That is, around 89.5% accuracy under 0.7 confidence.

B. Improvement

In the responses obtained from LLaMa (Answers_llm.txt), instances of garbled sequences of characters have been identified. While these anomalies do not directly impact the extracted answers, it is prudent to acknowledge their potential influence on the execution of Sentence-Transformers. Therefore, considering a pre-processing step to refine the text of LLaMa answers could contribute to an enhancement in the overall accuracy of the final results. This precautionary measure aims to optimize the performance of Sentence-Transformers by addressing potential disruptions arising from the presence of garbled sequences in the original answers from LLaMa.