Automated Evaluation Approach for Time-Dependent Question Answering Pairs on Web Crawler Based Question Answering System

Authors : Shraddha Chaudhary, Raksha Agarwal, Niladri Chatterjee

Abstract : This work demonstrates a web crawler-based generalized end-to-end open domain Question Answering (QA) system. An efficient QA system requires a significant amount of domain knowledge to answer any question with the aim to find an exact and correct answer in the form of a number, a noun, a short phrase, or a brief piece of text for the user's questions. Analysis of the question, searching the relevant document, and choosing an answer are three important steps in a QA system. This work uses a web scraper (Beautiful Soup) to extract K-documents from the web. The value of K can be calibrated on the basis of a trade-off between time and accuracy. This is followed by a passage ranking process using the MS-Marco dataset trained on 500K queries to extract the most relevant text passage, to shorten the lengthy documents. Further, a QA system is used to extract the answers from the shortened documents based on the query and return the top 3 answers. For evaluation of such systems, accuracy is judged by the exact match between predicted answers and gold answers. But automatic evaluation methods fail due to the linguistic ambiguities inherent in the questions. Moreover, reference answers are often not exhaustive or are out of date. Hence correct answers predicted by the system are often judged incorrect according to the automated metrics. One such scenario arises from the original Google Natural Question (GNQ) dataset which was collected and made available in the year 2016. Use of any such dataset proves to be inefficient with respect to any questions that have time-varying answers. For illustration, if the query is where will be the next Olympics? Gold Answer for the above query as given in the GNQ dataset is “Tokyo”. Since the dataset was collected in the year 2016, and the next Olympics after 2016 were in 2020 that was in Tokyo which is absolutely correct. But if the same question is asked in 2022 then the answer is “Paris, 2024”. Consequently, any evaluation based on the GNQ dataset will be incorrect. Such erroneous predictions are usually given to human evaluators for further validation which is quite expensive and time-consuming. To address this erroneous evaluation, the present work proposes an automated approach for evaluating time-dependent question-answer pairs. In particular, it proposes a metric using the current timestamp along with top-n predicted answers from a given QA system. To test the proposed approach GNQ dataset has been used and the system achieved an accuracy of 78% for a test dataset comprising 100 QA pairs. This test data was automatically extracted using an analysis-based approach from 10K QA pairs of the GNQ dataset. The results obtained are encouraging. The proposed technique appears to have the possibility of developing into a useful scheme for gathering precise, reliable, and specific information in a real-time and efficient manner. Our subsequent experiments will be guided towards establishing the efficacy of the above system for a larger set of time-dependent QA pairs.

Keywords : web-based information retrieval, open domain question answering system, time-varying QA, QA evaluation
Conference Title : ICKRQA 2022 : International Conference on Knowledge Representation and Question Answering
Conference Location : Barcelona, Spain
Conference Dates : October 20-21, 2022