IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/41ISM/2023/4115023/
Vol. 2023 (5), Article ID 4115023

FAAQA-QAD: A Frequently Asked Arabic Question Answering Dataset for Diseases Detection System*

Yassine SAOUDI
University of Tunis El Manar, Faculty of Sciences of Tunis, RIADI, Tunis, Tunisia

Mohamed Mohsen GAMMOUDI
University of Manouba, Higher Institute of Arts and Multimedia Manouba, RIADI, Manouba, Tunisia

Taoufik YEFERNY
University of Manouba, Higher Institute of Arts and Multimedia Manouba, LIPAH, Manouba, Tunisia

Correspondence should be addressed to: Yassine SAOUDI; yassine.saoudi@fst.utm.tn

* Presented at the 41st IBIMA International Conference, 26-27 June 2023, Granada, Spain

Abstract

Domain-specific question-answering (QA) systems are a specialized field in Natural Language Processing. Their objective is to generate answers to questions within a specific domain, such as healthcare. The existing Arabic datasets are often characterized by their low quality, resulting from inadequate preprocessing or inaccurate annotation because that translated from another language. To tackle this issue, our work introduces a new Modern Standard Arabic QA dataset, entitled FAAQA-QAD. The dataset was collected in Modern Standard Arabic without the use of translation. For the preprocessing and annotation of the FAAQA-QAD dataset, we employed the Pandas library and an end-to-end comprehension CdQA annotator. We further evaluated our FAAQA-QAD dataset using various pre-trained language models, demonstrating promising results in Arabic NLP tasks. This evaluation allowed us to compare the performance of these models using our dataset and conduct an analysis to comprehend the reasons behind the underperformance of certain models. Through our experimentation, we observed that the ARBERT model exhibited superior performance when compared to other RC span-extraction datasets using the FAAQA-QAD dataset.

Keywords: Arabic QA Dataset; Pre-trained Language Models; fine-tuning; Reading Comprehension; Natural Language Processing.

Cite this Article as: Yassine SAOUDI, Mohamed Mohsen GAMMOUDI and Taoufik YEFERNY, Vol. 2023 (5), "FAAQA-QAD: A Frequently Asked Arabic Question Answering Dataset for Diseases Detection System," Communications of International Proceedings, Article ID 4115023, https://doi.org/10.5171/2023.4115023