MSc Project Report

An Approach to Improve Machine Reading Comprehension based Question Answering

Mohamed Azim

2022

A report submitted as part of the requirements for the degree of

MSc Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

ABSTRACT

Question answering is one of the main concepts that come under natural language processing, and it can be also referred to as a branch of artificial intelligence. As the name describes, question answering systems focus on identifying or constructing an answer for the given question and context through an unstructured data collection in a natural language. With the advancements in the NLP domain, question answering has improved by leaps and bounds over the last few years with high performing models. BERT, XLNet, RoBERTa and ALBERT are a few of the models that were built in hope of improving the performance to exceed the human level performance. Although these models have performed well in general, when a context with complex scenarios was input into these models the performance has reduced.

This research is focused on building a novel approach to conducting question answering by applying tokenize and machine reading comprehension (MRC) which requires a machine to answer questions based on a given context, has attracted increasing attention with the incorporation of various deep-learning techniques over the past few years. BERT handles this task by encoding the question and giving paragraphs into a single sequence of words as the input. Then, it performs the classification task only on the output fragment corresponding to the context. Although BERT shows excellent performance, we argue that there are problems with this approach. Hence, we implement a new accurate model with high performances, changing the architecture and fine tuning a BERT model. In that new model achieved f1 value with 84.71% and exact match with 76.01%.

Keywords: Question Answering, Text Processing, Natural Language Processing (NLP), Machine Reading Comprehension (MRC)

Mohamed Azim | 20200770