

Semantic Bot: A Semantic Question Answering System for Raw Documents

Anchelo Adrian

A dissertation submitted in partial fulfilment of the requirement for Bachelor of Science
(Honours) degree in Computer Science

School of Computing
Informatics Institute of Technology, Sri Lanka in collaboration
with University of Westminster, UK

2023

ABSTRACT

Semantic searches allow users to retrieve relevant information from raw data. Advancement in embeddings such as transformer models have made the semantic retrieval more efficient. However, the use of transformers requires more resources than a lexical search. And retrieval systems with lexical search tend to be inaccurate. Therefore, creating a retrieval pipeline for a raw document with a decent speed and accuracy is difficult.

Sentence transformers are the advancement in transformers that are more suited for semantic search. Combination of several transformer architectures will further improve the accuracy of retrieval. Integrating sentence encoders along with word-based architectures will improve the speed of retrieval architectures.

This research presents ensembled semantic retrieval technique to search a raw PDF with low latency. A novel combination of word embeddings with sentence encoders to retrieve information is introduced by this research.

Keywords: Question Answering Systems, Semantic Textual Similarity, Sentence Encoders, Cross Encoders

Subject Descriptors

Information systems s → Information retrieval → Document Representation → Data Encoding

.