Project - Question Answering System/ChatBot

Project Team Members- Avinash Ramesh, Charu Cheema, Cory Randolph
Project Team Name- Insight Finders

CMPE 256 - Fall, 2021
11/6/2021

1.1 Purpose of Document

This is a Requirements Specification document for an Open Domain Question Answering chatbot/system that can respond to user queries based on wikipedia articles. The task of identifying answers to natural questions from a huge corpus of data is known as open-domain question answering (QA). The scope, features, requirements, and implementation details of the proposed system are described in this document.

1.2 Background

In an information retrieval system, an open domain question-answering system tries to provide a response to the user's query. The returned response is in the form of short texts rather than a list of related items/documents. The answer spans are a start and end index into the context paragraph, which indicate the start and end of the answer to the question as found in that paragraph. This task is more constrained than open-ended or multiple-choice questions as the answer is found directly in the input paragraph, but it still captures a variety of complex question types. The system uses a mix of computational linguistics, information retrieval, and knowledge representation techniques to obtain answers.

1.3 Project Scope

- Project scope encompasses a Question Answering (QA) chatbot/system that can respond to user questions based on Wikipedia.

- Modeling our system to respond to user questions based on the top N contexts gathered from Wikipedia.

1.4 Functional requirements

- Primary feature requirement is to build a Q&A chatbot/system that responds to user inquiries based on wikipedia pages.
- Based on the search query, the system deduces the semantics of the user inquiry and extracts the top N related contexts from Wikipedia.
- Our NLP model will use these top retrieved contexts to forecast answers.
- The system should provide a user-interactive (UI) for users to ask questions and receive answers along with a reference Wikipedia Page.

1.5 CRISP-DM process:

- **Business understanding:** Question and Answering (Q&A) systems allow users to formulate questions in natural language and receive the most appropriate and concise replies, making it easier for them to find relevant information. QA systems are an efficient way to resolve user queries. When done right, they can provide a very pleasant customer experience.

- **Data understanding and cleaning**:
  Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. There are about 150k questions available in the dataset.

  Data cleansing will be performed on the Wikipedia context to remove punctuation and special characters, as well as on the user's queries to improve model accuracy and reduce background noise.

- **Data preparation**: We will create a phrase matcher to extract only relevant paragraphs based on the user query, and then we will use sentence embedding techniques and cosine similarity to extract only the top N related contexts from the corpus before feeding it to the ML model for computation.

- **Modeling:** Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. We will be primarily using BERT base cased model bert-base-cased-squad2 trained on SQuAD v2 dataset and roberta-base-squad2 for comparison purposes.

- **Evaluation:** We'll be evaluating performance of several models for comparison purposes.

- **Deployment:** We will be deploying our model complete with a fitting chatbot similar UI using StreamLit for end user's access.

1.6 References:

- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2749028.pdf
- https://chatbotslife.com/chatbot-qa-implementing-chatbot-solution-the-chatc-group-49acef43aaa9
- https://medium.com/@christophberns/using-crisp-dm-to-predict-car-prices-f15eb5b14025
- http://web.stanford.edu/class/cs224n/project/default-final-project-handout.pdf
- https://web.cse.ohio-state.edu/~bair.41/616/Project/Example_Document/Req_Doc_Example.html