Question Answering System for Wikipedia dataset and a chatbot for user queries

Krishna Rukmini Puthucode
Georgia State University
Atlanta, Georgia
kputhucode1@student.gsu.edu

Ruthuparna Naikar
Georgia State University
Atlanta, Georgia
rnaikar1@student.gsu.edu

Kausik Amancharla
Georgia State University
Atlanta, Georgia
kamancharla@student.gsu.ed

***Abstract*—Possible opportunities for Question Answering System have been suggested in the previous work, including in the field of Artificial Intelligence.The need of questions and answers is prompted for various purposes,e.g.self-study, academic assessment,and coursework This project proposes to tackle open domain question answering using Wikipedia as the unique knowledge source: the answer to any factoid question. With the application of Natural Language Processing techniques, we propose to build a Question Answering System which can automatically find answers to matching questions directly from documents.**

***Index Terms*—BERT,NLP**

I. Introduction

Search engines have always been effective when dealing with large amounts of data, and they are generally good at retrieving documents that contain the user query.The next logical step will be to try to automate the part of the process where the user searches for a response in the retrieved text or document.Natural Language Processing along with Deep Learning techniques is making possible what was once considered to be a fantastical capacity for computers to answer any and all human questions. Under Natural Language Processing (NLP), Question Answering or QA is a discipline that enables users to retrieve answers from machines for questions posed in natural language.An open domain question answering system in information retrieval aims to provide a response to the user's query. Instead of a list of related documents, the answer is in the form of short texts.

sectionDataset

One of the first things required for any NLP tasks is a corpus. In linguistics and NLP, corpus (literally Latin for body) refers to a collection of texts.The good thing is that the internet is awash in text, much of which has been compiled and organized, even if it still needs to be fine-tuned into a more accessible, precisely structured format. Wikipedia, in particular, is a well-organized source of textual data. It's also a huge repository of information, and an unrestricted mind will conjure up a plethora of uses for such a set of words. What we plan here is build a corpus from a collection of English Wikipedia articles that are freely and easily accessible online.As this is an inferential study we plan to pick 300K articles from 6M articles from Wikipedia public dumps.

II. Related Work

Danqi Chen, Adam Fisch, Jason Weston Antoine Bordes in their paper Reading Wikipedia to Answer Open-Domain Questions propose an approach which combines a search component based on bigram. Using the techniques of hashing and TF-IDF matching with a multi-layer recurrent neural network model trained to detect answers in Wikipedia paragraphs

A second motivation to cast a fresh look at this problem is that of machine comprehension of text, i.e., answering questions after reading a short text or story.

There are a number of highly developed full pipeline QA approaches using either the Web, as does QuASE , or Wikipedia as a resource, as do Microsoft's AskMSR , IBM's DeepQA and YodaQA which serve as an inspiration for this project

III. Methodology

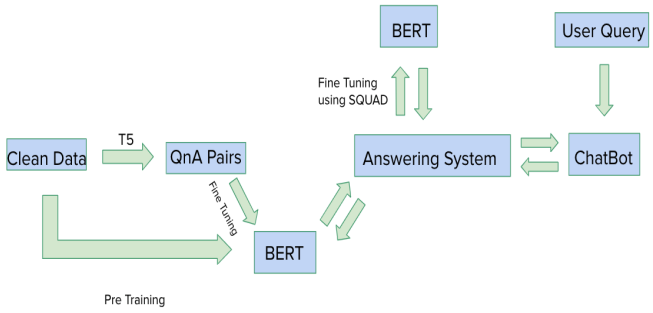

Fig. 1. Ensemble model - extraction performance

A. Data Prepossessing

The process of detecting and correcting (or removing) corrupt or inaccurate information from a record set, table, or database is known as data cleansing, and it involves finding missing, wrong, inaccurate, or irrelevant parts of the data and then replacing, altering, or deleting the dirty or coarse data.Wikipedia data comes with a lot of markup tags which have to be removed.We need format training data that is one sentence per line.

B. Pipeline 1

Question Generation System: The translation of machine-readable,non-linguistic information into human language representation is the task of natural language generation(NLG),which is one of sub study of natural language processing(NLP). Question-answer generation from text is classified as NLG task focused on generating question-answer pairs from unstructured text. Using Transfer Learning Technique, we aim to generate QA pairs. We introduce T5 model which is Text to Text Transfer Transformer Model T5: Text-to-Text-Transfer-Transformer model proposes recasting all NLP tasks as a single text-to-text format, with text strings as the input and output. Because of the formatting, a single T5 model may be used for a variety of purposes

C. Pipeline 2

Answering System: In this project, the Answering system aims to retrieve the answer to the question from the corpus.

1) Question Processing: We process the given query to extract all the important words. These words usually are the Nouns, Proper Nouns and the Adjectives.

2) Document Retrieval: The matching of a user query against a collection of free-text records is known as document retrieval.In order to extract the relevant document we use the Keyword Extraction strategy. We aim to find the list of keywords which represent the whole document. We plan to find most relevant documents by searching the given quey with the list of keywords.

3) Context Retrieval: Context for a given question is a passage or a list of sentences which has the highest possibility of having the answer. Using the top documents retrieved, we aim to extract the context by performing applying a ranking function

D. Pre-Training BERT

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language.BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training.In addition to the masked language model, BERT uses a next sentence prediction task that jointly pre-trains text-pair representations. In this project, we plan to pre-train BERT from scratch for our requirements. There are two steps in BERT: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters.

In this project, we aim to pre-tune the Bert from scratch using the Wikipedia dataset and then fine-tune it with out generated question-answer pairs.

E. Limitations and Challenges

Wikipedia is extremely vast in terms of both number of articles and textual data. Although we are performing an Inferential Study, we believe that dealing with huge amount of textual data can be challenging. Time becomes an factor when we think of prepossessing or model training

The second issue we possibly foresee is generating accurate keywords. Since Wikipedia has huge amount of keywords it is difficult to reply on just a few of them assuming that they represent the entire document.

Pre-training BERT is m

IV. Project Outcomes

In this project, we aim to implement a open-domain Question Answering System using Wikipedia dataset. We also plan to pre-train the BERT from strach. We plan to pre-train it the the wikipedia data which is our copus and we plan to fine-tune it with our generated question-answer pairs.

The authors would like to thank...

References

[1] QA with Wiki: improving information retrieval and machine comprehension by Rohan Sampath,Puyang Ma
[2] Open domain question answering using Wikipedia-based knowledge model Pum-Mo Ryu,Hyunki Kim
[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding byJacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova Google AI Language
[4] Benchmarking question answering systems by Ricardo Usbeck,Michael Röder
[5] End-to-End Open-Domain Question Answering with BERTserini Wei Yang, Yuqing Xie