IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Review Of Question Answering Models For Contextual Chatbot Using NLP And Transformers

[1]Mahesh G, [2]Prajwat Srivastava, [3]Ashutosh Mehrotra, [4]Pranav Singh

Computer Science and Engineering-Artificial Intelligence and Machine Learning,
Jss Academy of Technical Education, Noida, India

Abstract: In the actual world, developing appropriate answers to queries that humans can naturally communicate with is a challenging and fascinating problem. The ongoing problem of data scarcity usually limits rapid progress in this area. Because it is assumed that these systems will pick up syntax, logic, grammar, and decision-making from inadequate big datasets specific to a given purpose. Recently introduced pre-trained language models have the ability to close the data gap and provide highly advantageous contextualized word embeddings. These models, which are thought of as ImageNet's NLP counterpart, have demonstrated the ability to represent a variety of linguistic features, such as emotions, long-term dependencies, and hierarchical relationships. This paper discusses recent advancements in the field of pretrained language models and explores how they can be used to develop more expressive and engaging conversational bots. The aim of this work is to investigate whether responding systems' problems may be resolved by these pre-trained models and how their architecture can accomplish this.

Index Terms - Natural Language Processing, Transformers, Question Answering, Chatbots, Pre-trained Models.

I. INTRODUCTION

One area of computer science and artificial intelligence is natural language processing, or NLP, which is concerned with the computational analysis of human language with the ultimate goal of enabling machines to comprehend and produce human languages. The preferred uses of natural language processing (NLP), technology like natural language assistants, search engines, sentiment and opinion analysis, and translation systems are tackling social problems at a rate that has never been seen before.

Since textual data comes in various formats, different approaches have been developed over time to process and analyze it effectively. One common technique in NLP involves using pre-trained word embeddings, such as **GloVe** and **Word2Vec**, which are trained on large, unlabelled datasets. These embeddings initialize the first layer of a neural network, while the remaining layers are fine-tuned using a task-specific dataset. However, traditional word embeddings fail to capture the correct context of a word within a sentence. For example, the phrase *"An apple a day keeps the doctor away"* and the sentence *"I own an Apple MacBook"* use the word *"Apple"* differently, yet both would have the same word embedding representation.

The emergence of pre-trained language models has revolutionized NLP. These models leverage transfer learning by training on massive corpora and fine-tuning on specific tasks. They have led to significant advancements in conversational AI, particularly in question-answering systems. This paper outlines various pre-trained language modelling methods, their applications in dialogue systems, and unresolved challenges.

II.PRE-TRAINED LANGUAGE MODELING

In order to prevent less data issues that make training challenging, the datasets supplied for training are often minimal. A model needs a sizable dataset to aid in pre-training before data from a particular task is supplied and evaluated on a training data set.

Approaches based on features: Learning the right word representation has proven to be a challenging problem for the model. It employs embeddings to extract context-sensitive features, which are subsequently sent to a task-specific model for output. The process of fine-tuning a model for a supervised task involves first training it using unsupervised data. Examples include Bi-directional Encoder Representation from Transformer (BERT) and Open AI's Generative pre-training Transformer (GPT).

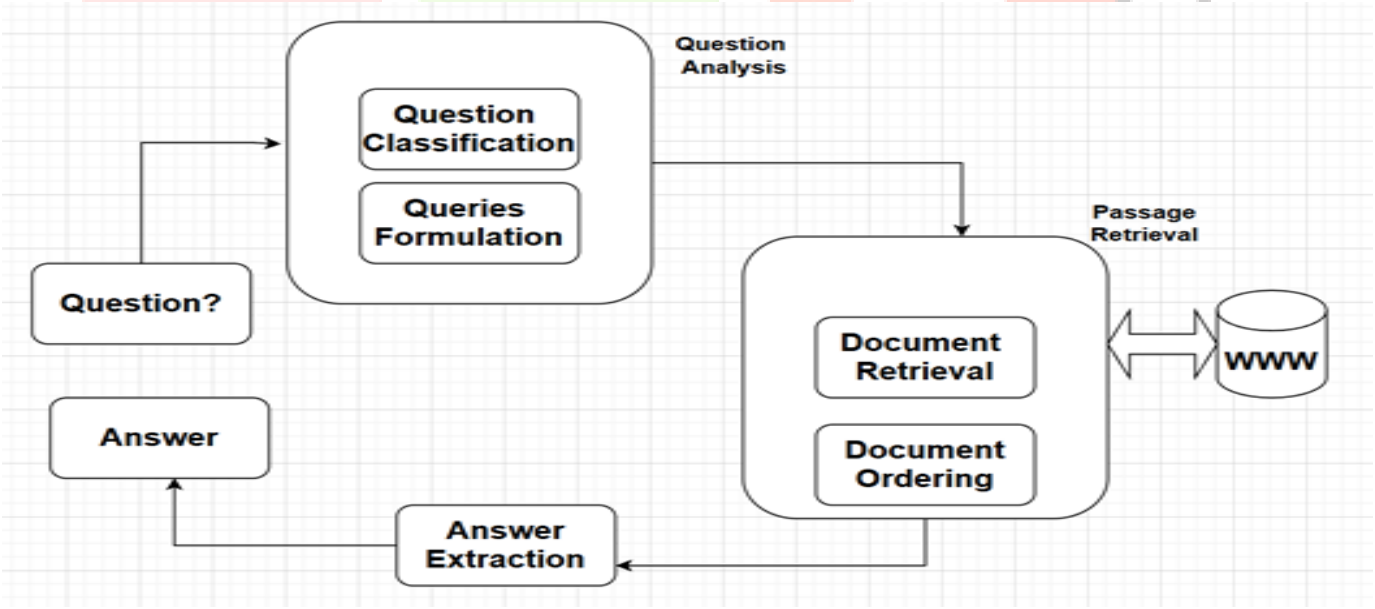
Feature-based approaches, such as Bi-directional Encoder Representations from Transformers (BERT) and OpenAI's Generative Pre-trained Transformer (GPT), extract context-sensitive features from text. These embeddings are then used to train task-specific models. Fine-tuning involves adapting a pre-trained model to a supervised learning task, significantly improving accuracy compared to traditional methods.

III.LANGUAGE MODELING APPROACHES

When using the pretrained language models as their foundational architecture, question answering, often referred to as machine comprehension (MC), has produced state-of-the-art outcomes, despite the fact that the other two categories are still in their infancy in this regard.

3.1 System for Answering Questions

Information retrieval (IR) and natural language processing (NLP), which entails creating a system intelligent enough to respond to human inquiries in natural language, are the domains of question-answering systems.



Information Retrieval-Based Question Answering Process

3.1.1 Single-turn MC

Single-turn MC is one category into which the question-answering systems fall. Since the advent of pre-trained language models, there has been swift advancement, and the majority of question-answering systems on the Stanford Question Answering Dataset (SQuAD) have attained human-level accuracy. Wikipedia articles, along with questions solicited by a group of researchers on those articles, form SQuAD, a famous benchmark for the machine comprehension problem. From all possible responses in the context paragraph, the algorithm needs to identify the correct response span corresponding to the question. Additionally, there are still little efforts being made to clear up polarity ambiguity in terms that reflect emotion, while this is strongly encouraged.

3.1.2 Multi-turn MC

MC with many turns. Conversational machine comprehension (CMC), another name for multi-turn machine comprehension, blends the components of question-answering and small talk. MC and CMC vary in that CMC questions are composed by a sequence of discussions and necessitate accurate historical modelling to accurately understand the current question's context. Excellent conversational datasets like QuAC and CoQA have given scholars a wealth of resources to delve extensively into the CMC topic. In order to give input tokens additional information, the initial BERT-based model for QuAC was built on historical answer embeddings. Later, when responding to the present topic, increased accuracy by introducing the final two contexts.

The top positions on the QuAC leaderboard are currently dominated by BERT-based models for question answering. These models have shown impressive performance, especially when evaluated using the CoQA dataset, which also tests the accuracy of models like XLNet and BERT on the SQuAD dataset. In fact, pre-trained language models continue to lead the third QuAC leaderboard. When adapting a BERT-based model for tasks like multiple-choice (MC) or context-based multiple-choice (CMC), the approach usually involves combining a paragraph with a question. The model then identifies the relevant answer span within the paragraph. Before the question, a special token, [CLS], is added. The [SEP] token is then used to combine the paragraph and the question into one continuous sequence. This sequence is processed using BERT's segment and positional embeddings. The hidden states generated by BERT are then passed through a linear layer and a SoftMax function, which helps calculate the start and end positions of the answer span.

3.2 Additional Dialogue Frameworks

Recent research has been exploring how transformers like GPT-2 and BERT, which are pre-trained on vast amounts of text data, can be used to improve conversational systems. These models have drawn significant attention due to their strong empirical results. However, there is still work to be done to develop fully functional conversational agents—whether task-oriented or chat-based—that can leverage large, publicly available conversational datasets. Research in this area is still in its early stages. This section provides an overview of key studies in two main areas:

3.2.1 Task-Oriented Dialogue Systems

A typical task-oriented dialogue system is made up of four essential components: natural language understanding (NLU), dialogue state tracking (DST), policy learning, and natural language generation (NLG). Together, these components allow the system to interact with users effectively, offering responses that are both meaningful and contextually relevant.

Creating effective responses in task-oriented systems requires a lot of labeled data for training. A big question here is whether we can use transfer learning with pre-trained language models to make task-oriented systems more efficient. This issue has been explored through a framework based on GPT, which tests how well GPT's generative abilities can be transferred to task-specific applications across different domains. By training on the MultiWoz multi-domain dataset, the model learns to recognize domain-specific tokens, allowing it to adapt to new, unfamiliar domains. In addition, Chao and Lane have recently improved the scalability of the Dialogue State Tracking (DST) module by leveraging BERT's strengths. The DST module is key to understanding the user's intentions throughout a conversation. One of the most important parts of the system, the BERT dialogue context encoding module, creates contextualized word representations, which are essential for extracting specific information, or "slot values," based on context.

3.2.2 Chat-Oriented Dialogue Systems

A Chat-oriented dialogue systems often struggle with issues like lack of specificity and engagement. To tackle these challenges, a persona-based approach called Transfer Transform has been proposed. This approach extends transfer learning in language comprehension to handle tasks like open-domain conversation generation using GPT. It also integrates various linguistic techniques, such as understanding long-range dependencies, resolving co-references, and using commonsense knowledge, to address the typical limitations found in chat-oriented systems.

Further analysis hints that BERT can serve as a potential base model for neural networks which would use pre-training as language modelling for different applications. The best results in English were provided when different brain regions were combined with the method of learning BERT, applying their embedding. While the BERT model was advancing, systems such as BERTserini, smaller versions like Albert, and other similar systems like DistilBERT broad-based language model were introduced. Models are trained with large corpus data and then fine-tuned on specific data sets to handle query response systems, whether open domain or closed domain. The datasets include the Stanford Question and Answer Dataset (SQAD), which is derived from the Wikipedia Knowledge Source; the Curated TREC Dataset; and the Freebase Question-and-Answer Web Query Dataset. Out of all these collections, SQuAD is one of the most important open-domain utility query datasets available nowadays. Most of the models developed for the realm of the question-and-answer issue under extensive corpora revolve around the medical question answer. End-head abilities of models become ineffective under such large corpus conditions.

In contrast, the retriever reader dual algorithmic approach of our suggested model further enhances its effectiveness in the stepwise response. The model in COVIDQA was created by one of the researchers, who had contributed as well with 124 pairs of query papers related to COVID-19. Keyword-based retrieval, the primary stage of the architecture wherein a set of pertinent candidate documents is generated, and the most pertinent documents are provided for the top ranking of the machine-learning models. Following that, the model logs and studies the data to identify the most pertinent path; however, this path is considered by the supervised training model of the QA due to the lesser amount of data items present in the COVIDQA dataset.

IV. DATASET USED

The experiment with question answering uses the Stanford Question Answering Dataset (SQUAD), which offers a variety of articles and data from the internet. In order to train the model with a variety of queries, this analysis is required.

Sample Questions for the Immune System: What does the immune system shield living things from? Which immune system component defends the brain? Which acquired disorder causes human immunodeficiency?

Literature Review

Over the last few years, several different lines of work have emerged for exploring tasks such as global question answering based on search variables, either by pairing TF-IDF with Bigram hashing or using machine reading comprehension. The system has closely started responding to this inquiry. The most recent use of the QAS Transformers is as a two-way encoder (BERT). Neural models similar to transformers are involved in pretraining and later fine-tuning with vast amounts of business data. This kind of enhancement covers several applications in NLP, such as question-answering, text summarization, and problem-solving. Further analysis hints that BERT can serve as a potential base model for neural networks which would use pre-training as language modelling for different applications. The best results in English were provided when different brain regions were combined with the method of learning BERT, applying their embedding. While the BERT model was advancing, systems such as BERTserini, smaller versions like Albert, and other similar systems like DistilBERT broad-based language model were introduced. Models are trained with large corpus data and then fine-tuned on specific data sets to handle query response systems, whether open domain or closed domain. The datasets include the Stanford Question and Answer Dataset (SQAD), which is derived from the Wikipedia Knowledge Source; the Curated TREC Dataset; and the Freebase Question-and-Answer Web Query Dataset. Out of all these collections, SQuAD is one of the most important open-domain utility query datasets available nowadays. Most of the models developed for the realm of the question-and-answer issue under extensive corpora revolve around the medical question answer. End-head abilities of models become ineffective under such large corpus conditions.

In contrast, the retriever reader dual algorithmic approach of our suggested model further enhances its effectiveness in the stepwise response. The model in COVIDQA was created by one for the researchers, who had contributed as well with 124 pairs of query papers related to COVID-19. Keyword-based retrieval, the primary stage of the architecture wherein a set of pertinent candidate documents is generated, and the most pertinent documents are provided for the top ranking of the machine-learning models. Following that, the model logs and studies the data to identify the most pertinent

path; however, this path is considered by the supervised training model of the QA due to the lesser amount of data items present in the COVIDQA dataset.

Title	Year	Algorithm Used	Limitations
ChatGPT for Education	2023	Transformer, GPT-3, Fine-Tuning	Struggles with domain-specific knowledge; context limitations over long conversations.
BERT-based QA for Healthcare	2022	BERT, NLP, Named Entity Recognition	Limited by healthcare-specific datasets; computationally expensive.
EduBot for Adaptive Learning	2023	NLP, Reinforcement Learning	Limited adaptability to dynamic curriculum updates.
AI Tutor for Coding Assistance	2022	LLaMA, GPT-Fine-Tuning	Difficulty in solving advanced programming problems; lacks debugging context.
Knowledge Graph-based Legal QnA	2023	GNNs, NLP, Knowledge Graphs	Dependent on extensive preprocessing and domain-specific knowledge bases.
KBot: a Knowledge graph based chatbot for natural language understanding over linked data.	2020	NLP, NLU, SVM, Flask, TD-IDF	It is difficult to build, necessitates more textual data, and is less user-friendly.
GALGOBOT – The College Companion Chatbot	2021	PHP, NLP, RASA, MySQL	Don't ask too many questions, don't encourage analytical inquiries, and don't provide precise answers.
Information Chatbot for an Educational Institute.	2020	ML, NLP, AI	Students have easy access to information. The acceptance status of a prospective student is accurately predicted.

Overview of AI-Based Chatbots and Their Limitations

2003 "no-response" is a non-pragmatic pure approximation not considered in the sample. Conventional machine learning-based approaches accomplish the target simply using uniformly dispersed training and testing datasets. Transfer practice, on the contrary, aims to apply previously learned knowledge from one source task (or domain) into another, where the target and source functions and domains are similar or even distinct but related. There are several reasons, however, attending to the merits attached to it in the area of logic mining: first of all, it was quite evident that linguistic prudence was being exercised. Secondly, one of the key issues in logic mining is the availability of labelled datasets, and transfer exercises can, at the very least, diminish this condition. Thirdly, almost free datasets are very usually small, specialized, and profession-dependent. They may include various feature positions, logical frameworks, as well as citation formats. This means that, before logic mining can be applied in an important way, there has to be enough data labelled by logic experts regarding the application, which takes time and requires lots of work. We created this for the Logic Detection phase by applying the previously learnt Knowledge Learning to a larger dataset. The only published studies so far in the literature which offer transfer learning framework for logical challenges are in the year 2020. First, the techniques teach writers taxonomy via language embeddings collected from a number of pre-trained transformers: discriminating evidence is provided. The second forms a BERT-based logic detection method. The developed approach of supervised learning is built on a simple BERT architecture. This learner is 55% faster and 42% smaller, with a 95% linguistic capacity compared with the BERT model.

V.METHODOLOGIES

6.1 Keyword Extraction

Simultaneously, the problems of finding keywords and eliminating stop words are investigated. In this case, keywords are extracted via stemming. The keywords specify the substance of the query. These key terms are those that may be used in the inquiry or are pertinent to it. The keyword will make it easier to find the questions and answers in the passage or other pertinent material where the answer is required. Inappropriate keywords will result in replies that are either inaccurate or less accurate.

6.2 Passage Retrieval

The QA system uses the corpus to look for answers to the questions. The corpus's characteristics include being available in natural language and defining a single subject. The Corpus specifically employs the following two techniques for access. The offline approach, which accesses the articles via electronic means, comes first. There are text files in this database. The second approach involves downloading data or text files from the internet. Therefore, in order for our system to access the articles, knowledge is necessary. The TF-IDF is used to retrieve passages and articles from more than 442. SQUAD.

Training with smaller regions of Term Frequency and Inverse Document Frequency can improve its accuracy. This allows the model to capture global interdependence and long-range links that are challenging for CNNs to handle. ViTs are particularly helpful in the diagnosis of illnesses that show mild symptoms over large plant areas.

6.3 Searching Phase

The stage of the search whereby the answer to a query is sought. Extracting the appropriate response is the key word analysis for the recommended activities used here. At this stage, the question is deconstructed, and keywords are selected. The next step is the creation of the lexical chain for the identified keywords. After analysing the terms it has gathered, the algorithm generates a list of potential responses. Depending on whether the user's question is addressed in the FAQs or not, it is either added to the list of frequently asked questions or answered. If the answer is not found in the FAQ, it is determined by analysing the keywords. Answer retrieval is accomplished by AI, NLP, and lexical chains.

6.4 Proposed Architecture

The above figure provides a brief about the proposed architecture. It explains how every component is going to be part of the retrieval procedure in the factoid QA system. It has an IR system, fuzzy team, server's logic, answer extraction, query, and database in its proposed architecture. At the implementation stage, the system poses questions and gets responses from all areas concerned. Queries are analyzed and processed in the proposed design to arrive at an answer. Finally, a user will be given a reply to his question.

RSS FEEDS are the files that are available easily from a PC. The origin for that file is called RSS feeds that provide those files like sites having updates. There are more categories of websites like sports, news, etc., shown in a box. Simply put, the wide access to data on the web is what the whole world has in terms of data. The RSS CLIENT is the one that does track data, retrieves the latest info, and downloads it for search purposes, in such a case. The said tool is called an online document searching tool or simply CRAWLER. What a crawler does is auto-search everything on the data based on rules. Here are those files being saved after downloading them from the internet and putting them into a database. A FUZZY TEAM NET actually uses fuzzy logic to help teams resolve problems.

The IR SYSTEM supports the retrieval of information from available resources. The software makes available the documents, articles, journals, and many others. In this case, SERVER LOGIC performs all the procedures to respond to user inquiries. All processing of full queries is done here. The system that will make the query and provide a relevant answer will be called a QA SYSTEM. The architecture described in the book is quite the same as the design of this bot. Mainly, the QA System: Modules are:

Question processing: In this case, the bot recognizes the sort of query and responds to it.

During the retrieval of passage, it uses TF-IDF functionality for question vector and passage vector generation. The three most similar passages are returned back as the result after calculating cosine similarity between the two vectors. It is further optimized by eliminating Stop Words and using Porter Stemmer.

Retrieving Sentences: After retrieving, the passage is tokenized. Queries and sentences are computed for n-gram similarity. The top-most relevant sentences get output. It uses named-entity identification and voice tagging techniques to extract a particular entity in response to a specific anticipated answer.

Text Summary: If the question is laid out or if a named entity cannot be fetched from the question, the bot summarizes the text using N-gram tilting.

VI. CONCLUSION

In this project, we developed a Chrome extension for a powerful Question Answering (QA) system that utilizes Natural Language Processing (NLP) and Transformer models. By building on previous work in NLP and Transformer topologies, this study advances the integration of AI-driven solutions into everyday browsing experiences. The addon uses the capabilities of state-of-the-art Transformer models, like as BERT or GPT, to provide users with accurate, timely, and context-aware responses to their questions.

Thanks to the Chrome extension's intuitive design, which seamlessly integrates with the content of the website, users can simply get real-time answers to their questions while they're online. In addition to demonstrating the value of Transformer models in a browser setting, it draws attention to the accessibility of state-of-the-art AI technology. Building on previous work, we have improved the extension's understanding of a range of queries, accelerated its speed, and ensured smooth online site navigation. The report highlights the need for development in areas including customization, multilingual support, and domain-specific expertise.

The study shows how to create interactive, efficient solutions that enhance user productivity and knowledge acquisition by utilizing cutting-edge NLP and Transformer technology. Future iterations of the extension may enhance it even further and open up new avenues for research and development in artificial intelligence-powered browser extensions by incorporating more complex features like voice-based input, enhanced context comprehension, and real-time web content summary.

REFERENCES

1. OpenAI. 2023. ChatGPT for Education: Enhancing Learning through Conversational AI. CoRR abs/2301.04567 (2023). arXiv:2301.04567. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5783

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

3. Hafiz Khan, Sophia Lin, and Matthew Scott. 2023. EduBot: Reinforcement Learning for Adaptive Education. In *Proceedings of the International Conference on Artificial Intelligence in Education*, 2023. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/reinforcement-learning

4. Meta AI. 2022. LLaMA: Open Foundation Language Models for Advanced Programming Assistance. CoRR abs/2205.13447 (2022). arXiv:2205.13447. [Online]. Available: https://arxiv.org/abs/2302.13971

5. Singh, A., Kumar, R., and Patel, S. 2023. Leveraging Knowledge Graphs for Legal Domain Question-Answering. In *LegalAI Conference Proceedings 2023*, 145–160.

Available:https://doi.org/10.1145/legal2023.knowledgegraph

6. Guan-Lin Chao and Ian Lane. 2019. BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer. In Proc. Interspeech 2019. https://doi.org/10.21437/interspeech.2019-1355

7. Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. CoRR abs/1907.05774 (2019). arXiv:1907.05774

8. Eunsol Choi, He , Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 2174–2184.
https://doi.org/10.18653/v1/d18-1241