CONTEXTUAL AI ASSISTANT USING TRANSFORMERS AND NLP

Anant Kumar Gupta[1] Aaroosh Agarwal[2] Abhishek Sharma[3] Anshul Tyagi[4] Arpita Grover[5]

[1] *Student, Computer Science JSSATEN, Noida, Uttar Pradesh, India. 201301.*
[2] *Student, Computer Science JSSATEN, Noida, Uttar Pradesh, India. 201301.*
[3] *Student, Computer Science JSSATEN, Noida, Uttar Pradesh, India. 201301.*
[4] *Student, Computer Science JSSATEN, Noida, Uttar Pradesh, India. 201301.*
[5] *Assistant Professor, Computer Science JSSATEN, Noida, Uttar Pradesh, India. 201301.*

ABSTRACT

Building proper answering of the questions that can communicate naturally with humans is a difficult and interesting problem in real world. Rapid advancement in this field is typically constrained by the persistent issue of lack of data. As a result of the expectation that these systems will learn grammar, reasoning, decision-making, and syntax from insufficiently large task-specific datasets. Pre-trained language models, recently presented have the potential to fill the data gap and generate contextualized word embeddings with great benefits. Considered the NLP version of ImageNet, these models have been shown to capture different aspects of language, including hierarchical relationships, long-term dependencies, and emotions. Recent developments in the field of pretrained language models are discussed. We also consider how to use the strengths of these language models to design more engaging and eloquent conversational agents. As a result, the purpose of this work is to explore whether these pre-trained models can address the issues that responding systems face as well as how their architecture can do so.

Keywords— *Natural language processing · Intelligent agents · Transformers*

I. INTRODUCTION

Natural Language Processing (NLP) is a branch of Computer Science and Artificial Intelligence that focuses on the computational treatment of human language with the core intent of making machines understand and generate human languages. NLP's favored applications, such as translation systems, search engines, natural language assistants, sentiment, and opinion analysis, are resolving societal issues at an unprecedented rate. Although data sources are available in diverse forms, i.e., images, voice/speech, body language, texts, the use of texts has recently gained traction due to the enormity of text data available because of the emergence of Web 2.0 and the social media. Textual data is sequential; hence the order of words and their relation in a sentence, i.e., context, play a vital role in deriving complete meaning from the sentence. Not only does the traditional unsupervised machine learning models disregard the occurring order or relation- ship of words in written texts but are also limited by the relatively small fixed input size. Thus, the motivation for applying computationally deep approaches to textual data. The Recurrent Neural Network (RNN) is a sequential model capable of handling sequential data but limited by the slow training time and the long-range sequence dependencies. However, a variant of RNN, the Long Short-Term Memory (LSTM), is paramount in mitigating some of these limitations. The LSTM provided a fair solution to the long-range sequence dependency problem but disregarded parallel computations and was even slower than the RNN.

A different number of strategies have been introduced over a period to address this issue. One of the standard methods of designing an NLP based project is to utilize word embeddings, pre-trained on a huge amount of unlabeled data using distributed word representations such as Glove and Word2Vec, to initialize the first layer of the neural network. The rest of the layers are then trained on a task-specific dataset. However, these techniques failed to capture the correct context of the word used in a sentence. For example, "an apple a day, keeps the doctor away" and "I own an Apple Mac- book, two "apple" words refer to very different things but they would still share the same word embedding vector. Recently, the concept of pre-trained language modeling is introduced. The word embeddings generated by these language models are pre-trained on large corpora and are then utilized as either distributed word embeddings or fine-tuned according to the specific task needs. This comes under the category of transfer learning and has achieved state-of-the-art results in various NLP tasks. The introduction of pre-trained language models has given the field of conversational AI a new direction especially to question answering. Thus, we focus more on the question answering systems than the other two dialogue systems as this field have been totally transformed by these pre-trained language models.

This describes various pre-trained language modeling techniques that have been introduced so far. We then extend the discussion to the implementation of these models in dialogue systems with a particular focus on question-answering systems. Finally, we present open challenges related to these language models that need to be addressed.

.

II. PRE-TRAINED LANGUAGE MODELING

The Dataset available for training are generally small and to avoid less data problem which makes training difficult. Pre-training a model requires large dataset which helps it to train then data of a specific task is provided which is then tested on a training dataset.

Feature based Approaches: It has been a difficult task for model to learn appropriate words representation. It uses embeddings which extracts the context sensitive features and then sending it to task specific model for result.

Fine Tuning: It uses the idea for pre training model for unsupervised data then fine tuning it to supervised task. Some of examples are Open AI's Generative pre-training Transformer (GPT) and Bi-directional Encoder Representation from Transformer (BERT).

III. LANGUAGE MODELING APPROACHES

Although the other two categories are still in infancy in terms of utilizing the pretrained language models, question answering, also known as machine comprehension (MC), has achieved state-of-the-art results when leveraged the strength of these as their base architecture.

1.1 Question Answering System

Question answering systems come under the field of information retrieval (IR) and natural language processing (NLP), which involves building a system intelligent enough to answer the questions asked by the humans in a natural language. The question answering systems can be categorized as:

1.1.1 Single-turn MC. There has been rapid a progress after the introduction of pretrained language models and most of the ques- tion answering systems have achieved human-level accuracy on Stanford Question Answering Dataset (SQuAD).

The SQuAD dataset is a standard benchmark for the machine comprehension problem consisting of Wikipedia articles and questions posed on those articles by a group of co-workers. The system must select the right answer span for the question from all the possible answers in the given passage. Furthermore, contributions towards resolving polarity ambiguation in emotion-conveying words remain limited and are greatly encouraged.

3.1.2 Multi-turn MC.

Multi-turn machine comprehension, also known as conversational machine comprehension (CMC), combines the elements of chit-chat and question answering. The difference between MC and CMC is that questions in CMC form a series of conversations and require the proper modeling of history to comprehend the context of the current question correctly. Highquality conversational datasets such as QuAC and CoQA have provided the researchers a great source to work deeply in the field of CMC. The first BERT based model for QuAC was based on history answer embeddings to provide extra information to input tokens. Later, improved accuracy by introducing the last two contexts when answering the current question. Introduced the reasoning process in BERT based architecture that improved the accuracy on the leader board drastically as compared to the previous models. Currently, the top scores QuAC leaderboard 2 are of BERT-based question answering models. The BERT and XLNet based models that tested the accuracy on SQuAD dataset also evaluated their models on CoQA. The top positions on CoQA leaderboard 3 are occupied by pre-trained language models. Figure 2 shows how to adapt a BERT-based model for MC or CMC tasks. The input to the model is a question and a paragraph, and the output is the answer span in the given paragraph. A special classification token [CLS] is added before the given question. Then, the question is concatenated with the paragraph into one sequence using [SEP] token. The sequence is provided as an input to the BERT- based model with segment and positional embeddings. Finally, the hidden state of BERT is then converted into the probabilities of start and end answer span by a linear layer and SoftMax function.

3.2 Other Dialogue Systems

Recent work in large-scale pre-training (such as GPT-2 and BERT) on large text corpus using transformers has attracted significant interests and achieved great empirical success.

However, leveraging the strength of massive publicly available colloquial text datasets to build a full-fledge conversational agent (i.e., task-oriented and chat- oriented) is still progressing. The use of these pre-trained language models is in its inception phase and not much work has been done. In this section, we briefly discuss the research carried out in the two areas:

3.2.1 Task-Oriented Dialogue System. A traditional task-oriented model consists of four modules namely: i) natural language under- standing, ii) dialogue state tracking (DST), iii) policy learning, and iv) natural language generation. The goal of such systems is to assist the user by generating a valuable response. This response generation requires a considerable amount of labeled data from the training purpose. A question that comes naturally to mind is: Can we take advantage of transfer learning through pre-trained language models to enable the modeling of taskoriented systems. The question has been addressed by which introduces a GPT based framework to evaluate the ability to transfer the generation capability of GPT to task-specific multi-domains. They used multi-domain dataset MultiWoz to learn and understand the domain-specific tokens which make it easier to adapt to unseen domains. Chao & Lane recently utilize the strengths of BERT in improving the scalability of DST module. The DST module is use to

maintain the state of user's intentions throughout the dialogue. The key component of the model is BERT dialogue context encoding module which generates contextualized representations of the words which are very effective for mining slot values from the contextual patterns.

3.2.2 Chat-Oriented Dialogue System. Chat-oriented dialogue systems are known to have several issues such as they are often not very engaging and lack specificity. To address these problems, Transfer Transform, a persona-based model, is introduced. They have extended the transfer learning from language understanding to generative tasks such as open-domain dialogue generation using GPT and addressed the above-mentioned issues by combining many linguistics aspects such as common-sense knowledge, co-reference resolution, and long-range dependency.

IV. DATASET USED

The Stanford Question Answering Dataset (SQUAD) dataset which provides various articles and data all over from internet is used for the experiment of the question answering. This analysis is needed so model is trained of various questions.

Immune system Sample Questions:

The immune system protects organisms against what?

Which part of the immune system protects the brain?

What acquired condition results in immunodeficiency in humans?

V. LITERATURE REVIEW

Several tasks have been proposed in recent years to challenge global knowledge by integrating search variables based on Bigram hashing, TF-IDF pairing, and machine reading comprehension. This is the start of the system response to this inquiry. QAS Transformers' most current representation is a two-way encoder (BERT). It pre-trains massive corporate data using neural models such as transformers. Many NLP categories, such as question-answer, text summary, and problem solving, benefit from such refining. In addition to BERT, research has recently proved the potential of neural models employing pre-training linguistic modelling for a wide range of applications by using BERT as the basis model. By merging multiple brain structures with the BERT learning algorithm and exploiting its embedding, state-of-the-art results in English were produced. End-to-end interactive chatbot systems such as BERTserini, the milder form of BERT known as Albert, and certain systems such as the all-purpose language model DistilBERT were introduced as BERT model research progressed. After being pre-trained with massive corpus data, models are trained on specialized datasets to answer queries in open-domain or closed-domain query response systems. The Curated TREC Dataset, the Question-and-Answer Web Query Dataset from Freebase, and the Stanford Question and Answer Dataset (SQAD) based on the Wikipedia Knowledge Source are some datasets for question-and-answer system. SQuAD is one of the most important open-domain utility query datasets that is now accessible in all these collections. Several models have been presented to address the question-and-answer problem from the huge corpus, the majority of which focus on the medical-question-answer problem, while the model's end-heading capabilities are ineffectual in the vast corpus. Our model, on the other hand, employs a retriever-reader dual algorithmic system, which improves the step-by-step response efficiency of our suggested model.

One of the researchers generated the COVIDQA dataset, which includes 124 pairs of query papers connected to COVID-19. The architecture begins with keyword-based retrieval, which defines the collection of relevant candidate documents and then gives the most relevant documents for the machine-learning models top rank. The model then queries and records the data to discover the most relevant path. However, due to the small number of data items in the Covid QA dataset, supervised training of the QA model is not achievable. In the dataset, there is not even a "no response" example, which is an impractical approximation in actual situations. Training and test data are utilised for the same task and have the same distribution in conventional machine learning models. Transfer practise, on the other hand, tries to transfer previously gained information from one source task (or domain) to another, where the source and target functions and domains are similar or distinct but connected. It is beneficial in the field of logic mining from a variety of angles. First and foremost, linguistic prudence was obviously appreciated. Second, transfer exercises can assist address, or at least alleviate, the scarcity of labelled datasets, which is one of the most significant issues in the field of logic mining. Third, the accessible datasets are frequently limited in size and highly domain and job related. They can use various citations, logical systems, and feature locations. This means that in every feasible application of logic mining, we require logic specialists to label adequate data for the work, which is a time consuming and labour - intensive activity. So, we created it for the Logic Detection phase by transferring the pre-trained Knowledge Learning to a larger dataset to present another challenge. Only two research published in 2020 imply a transfer learning model for logical problems, according to the literature. The first point is that methods that train authors for taxonomy on sentence embedding retrieved from multiple pre-trained transformers are discriminating evidence. The second is a logic detection technique that employs BERT. Our supervised learning model is based on a distilled model of BERT, which is 40% smaller in size and has 97 percent linguistic capacity compared to the 60 percent quicker BERT model.

Table 1: Literature Survey

Title	Year	Algorithm Used	Limitation
KBot: a Knowledge graph based chatbot for natural language understanding over linked data.	2020	NLP, NLU, SVM, Flask, TD-IDF	Complex to build, less user interactive, more text-based data is required
GALGOBOT – The College Companion Chatbot	2021	PHP, NLP, RASA, MySQL	Do not support analytical queries, limited to specified no. of questions, do not provide precise answers
Information Chatbot for an Educational Institute.	2020	ML, NLP, AI	Students get easy access to information, Precise prediction of whether a prospective student gets admission or not is provided
An Intelligent Chatbot System Based on Entity Extraction Using RASA NLU and Neural Network	2020	NLP, NLU, RASA, NN, RNN	If long sentences are there then it affects the recognition capability
An Improved Rapid Response Model for University Admission Enquiry System Using Chatbot	2020	AI, NLP, NLU, IBM Watson, Botium	Only limited to admission related queries, do not support analytical queries, it takes time to answer any question

VI. PROPOSED IMPLEMENTATION

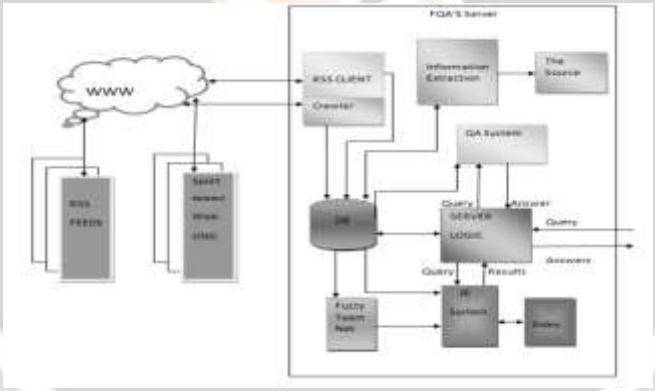

Figure.1 Proposed Architecture

A. Keyword Extraction

The issues of removing stop words and locating keywords are examined simultaneously. Stemming is used to extract keywords in this instance. The question's content is defined by the keywords. These essential phrases are those that are relevant to the inquiry or may appear in it. The keyword will assist in retrieving the questions and responses from the passage or other relevant content where the response is to be given. When inappropriate keywords are found, answers will be less accurate or incorrect. The stop words are also removed here from list of stop words.

B. Passage Retrieval

The Corpus is utilized in the QA system to search for answers to the queries posed. The corpus' features are that it should only define one topic and be available in natural language. For accessing, the Corpus explicitly uses two methods as follows: First, there is the offline method, which uses electronic means to access the articles. The database here contains text files.

The second method is online, in which text files or data are downloaded from the internet. Therefore, expertise is essential for our system in order to access the articles. The TF-IDF is utilised for passage retrieval, as well as articles over 442 that are from SQUAD. Its accuracy can be increased by training using Term Frequency and Inverse Document Frequency

C. Searching Phase

The phase of searching during which a question's answer is looked for. The key word analysis for the suggested actions that have been employed here is the extraction of the right response. The question is broken down at this point, and keywords are chosen. The development of the lexical chain for the discovered keywords is the following stage. The algorithm now evaluates the terms it has collected before producing a list of possible answers. The user's query is either answered or added to the FAQ list depending on whether it corresponds to the FAQs or not. The keywords are analysed to find the answer if it is not included in the FAQ. Answer retrieval is done by lexical chain, NLP, AI.

D. Proposed Architecture

The figure.1 mentioned depicts the suggested architecture. The structure provides information on every single part that is utilized for the factoid QA system's response retrieval. The IR system, the Fuzzy team, the server's logic, the answer extraction, the query, and the database have all been employed in the suggested architecture. The system is questioned during the implementation phase, and all relevant departments respond to the inquiries. In our suggested design, the questions are analyzed, processed, and the response is then retrieved. Finally, the user receives an answer to their query.

The files that can be easily accessed from a computer are RSS FEEDS. This file is obtained by RSS feeds, which offer the files as websites with updates. There is a box that displays additional website categories, such as sports, news, etc. The internet's data can be accessed by using the WWW, or the world wide web. The person who maintains the data, gets the most recent information, and downloads it for the search is the RSS CLIENT. In this instance. CRAWLER is used for document searching on the internet. This crawler performs data searching automatically. The files that are downloaded from the internet and stored here are stored using DATABASE. FUZZY TEAM NET utilizes fuzzy logic to assist in team issue solving. Information is retrieved using IR S YSTEM from the accessible resources. This software gives users access to books, papers, journals, and other resources. Here, SERVER LOGIC executes all of the necessary mechanisms needed to respond to the user's query. This is where the entire query is processed. The system that will use the query and deliver an answer that is pertinent to it is called QA S YS TEM. This bot's architecture closely resembles the architecture outlined in the book. The QA System's primary modules are: Processing of questions: The bot determines the type of query in this stage and answers it. When retrieving a passage, a question vector and a passage vector are generated using the TF-IDF feature. The cosine similarity between the two vectors is then calculated, and the top three closely related passages are returned. By eliminating Stop Words and utilizing Porter Stemmer, this step has undergone further development. Sentence Retrieval: Following the passage's retrieval, sentences are tokenized, and the n-gram similarity between the question and the sentence is calculated. finding the most pertinent sentences as a result. Processing of answers: Depending on the anticipated answer type, the processing of the answers identifies a specific entity using a named-entity recognition technique and a voice tagging technique. Text Summarization: If a question is defined or a named entity cannot be determined from the question, the bot summarizes the text using n-gram tilting.

VII. CONCLUSION

In recent years, the promising concept of pre-trained language models has attracted the attention of researchers. This is an emerging model that aims to create better contextualized word representations for conversational systems to better understand context. This paper is an attempt to investigate recent trends introduced in language models and their application to dialog systems. We focused our discussion on question-and-answer systems because they were most affected by these advances. Next, we briefly highlight the progress in the remaining two categories. Although these pre-trained models have the potential to solve most of the limitations posed by the previous methods, there are still some open problems that need to be addressed. The question answer system deals with a variety of articles and a variety of questions.

VIII. REFRENCES

[1] Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - 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

[2] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large- Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 5016–5026. https://doi.org/10.18653/v1/d18-1547

[3] 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

[4] 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

[5] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). 4171– 4186.

[6] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Con- versational AI. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July08-12-2018.1371–1374.https://doi.org/10.1145/3209978.3210183

[7] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data Learning of New Tasks. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008. 646–651.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019). arXiv:1907.11692

[9] Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension. In Proceedings of the First Workshop on NLP for Conversational AI. Association for Computational Linguistics, Florence, Italy, 11–17. https://doi.org/10.18653/v1/w19-4102

[10] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. 2009. Zero-shot Learning with Semantic Output Codes. In Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Informa- tion Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada. 1410–1418.

[11] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Rep- resentations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). 2227–2237. https://doi.org/10.18653/v1/n18-1202

[12] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019. 1133–1136. https://doi.org/10.1145/3331184.333134