2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-3701-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/Confluence52989.2022.9734207

Context-Based Question Answering System with Suggested Questions

1st Vijay Kumari
Department of Computer Science and Information Systems
Birla Institute of Technology and Science
Pilani, Rajasthan, INDIA
p20190065@pilani.bits-pilani.ac.in

2nd Srishti Keshari
Department of Computer Science and Information Systems
Birla Institute of Technology and Science
Pilani, Rajasthan, INDIA
h20200266@pilani.bits-pilani.ac.in

3rd Yashvardhan Sharma
Department of Computer Science and Information Systems
Birla Institute of Technology and Science
Pilani, Rajasthan, INDIA
yash@pilani.bits-pilani.ac.in

4th Lavika Goel
Department of computer science and engineering
Malaviya National Institute of Technology
Jaipur, Rajasthan, India
lavika.cse@mnit.ac.in

***Abstract*—Question Answering and Question Generation are well-researched problems in the field of Natural Language Processing and Information Retrieval. This paper aims to demonstrate the use of novel transformer-based models like BERT, AlBERT, and DistilBERT for Question Answering System and the t5 model for Question Generation. The Question Generation task is integrated with the Question Answering System to suggest relevant questions from the input context using the transfer learning-based model. The question generation model generates questions from the context input by the user and uses different models like DistilBERT, RoBERTa for getting answers from the context. Suggested questions are ranked using BM25 scores to show the most relevant question-answer pairs on the top. The input context can be given as PDF or image(extract texts from image).**

***Index Terms*—QAS, Question Generation, BERT, Language Modeling, t5,BM25**

I. Introduction

Question Answering System(QAS) and Question Generation Systems are the two most challenging domains of Natural Language Processing and Information Retrieval. The Question Answering System aims to generate answers to the user's natural language questions from the given context or knowledge base. The answers can be generated from open as well as closed domains. Open-domain systems can answer questions in human-understandable language from the large knowledge base; in contrast, the answers in a closed domain are restricted to a particular context. Generating answers to the questions from specific contexts is another problem as any question can be interpreted in different ways and can have different answers depending on the context it refers [1]. Question Answering Systems has become an essential tool because of its many applications like personal assistants, social media, machine reading comprehension, etc.

Another domain that has received tremendous interest in recent years is the Question Generation Task which involves generating a question from a given context and an answer phase. It can be useful in generating suggested questions from a given context to understand the context better. Question Generation System can be used in many applications like dataset creation and generating relevant questions in the education domain. For example, relevant questions from a book chapter can be generated to test the understanding of the topic.

This work combined both tasks to make the search easier for the user and give the exact answer to the user query. The question generation task is used to generate the questions from the context; these questions are suggested to the user along with the answers to the asked question.

QA Systems have traditionally been using NLP techniques like parsing, parts of speech tagging, etc.; these techniques can only comprehend the natural language context and the questions syntactically. However, semantic aspects of the natural language features play a significant role in understanding and correctly comprehending the language. Earlier, context-free models such as word2vec and GloVe were used for language modeling; with the evolution of deep learning models, many better language modeling techniques have evolved by including combinations of both left to right and right to left contextual models using RNNs, CNNs, LSTMs.

The idea of transfer learning helped overcome the shortage of task-specific annotated datasets by dividing the task into two stages of pre-training and fine-tuning. With the introduction of transformer framework and attention mechanism, BERT(Bidirectional Encoder Representations from Transformers) was introduced, the first fully and deeply bidirectional

Department of Science Technology,New Delhi

978-1-6654-3701-1/22/$31.00 ©2022 IEEE

language representation model [2] and gave the state of the art performance in many of the NLP tasks. Other versions of the BERT like DistilBERT were later introduced, which compressed the size and increased the model's speed while retrieving almost 95 percent of the performance of BERT[3]. Similarly, Question Generation models mainly relied on Recurrent Neural Networks(RNNs) like LSTMs and Gated Recurrent Units which only consider sentence-level information as they are not suitable for long inputs; hence, a better approach needs to be followed to consider the paragraph level information. The use of transfer learning models such as t5 can improve the performance of Question Generation [4]

Due to the lack of technical skills and extensive human effort, not every user can write a query using a query language. It increased the demand for systems capable of working with natural language. To reduce the burden of searching different websites and finding the required information, Question answering systems aim to provide the correct result directly. It is pretty helpful for the user to get a list of suggested questions while answering the specific question from the context. The user wants the most relevant information in the least possible time. The recently proposed techniques in deep learning open up new opportunities for improving the results of all NLP tasks, including Question Answering and Question Generation.

This work aims to study the practical applications of Question Answering and Question Generation by implementing both models together. The system aims to take the context and natural language questions from the user and give the answers present in the context by the user and give relevant links from the other domain QAS if the answer to a particular question is not found in the context given input by the user. Our model also aims to suggest other questions formed from the input context and their answers to get a comprehensive set of relevant information present in the context.

The paper is organized as follows: Section II contains the survey of related work. Section III gives brief thought regarding the working of the framework. Section IV explains the dataset used for the task. Section V explains Experiments and Results followed by the conclusion and future work section.

II. Related Work

Context-free models word2vec [5] and GloVe [6] were widely used in the past and could not entirely capture the input's context. As many developments in deep learning were made, the context in both left to right and right to left directions could be captured. However, the drawbacks of the RNNs having vanishing descents made capturing the context of large inputs challenging [7].The classification of the transfer learning models are given in figure 1.

The Introduction of a transformer-based framework and the attention mechanism solved this problem to a large extent [8]. The transfer learning technique divided the language modeling into two tasks - pre-training and finetuning. The BERT model was introduced using these novel techniques, which uses a series of encoder blocks in the transformer architecture and is trained in two phases - pre-training and finetuning [2]. The pre-training of the BERT involves two steps - Masked Language Modelling(MLM) and NSP(Next Sequence Prediction).During the MLM steps, around 15 percent of the words in the dataset are masked. The model tries to learn the contextual representation of each word by predicting these masks. The NSP task predicts whether B occurs after A in the corpus by giving A and B sentences. During training, 50percent of the time, the second sentence is the actual next sentence in the original document, while in the other 50 percent, the random sentence is chosen from the corpus. During the fine-tuning phase, the model is trained for specific tasks such as Question Answering. For the Question Answering task, the model returns a span with the start and end of the answer to the given question.

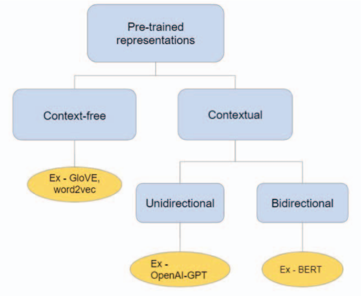


Fig. 1. Classification of Transfer Learning models

Many more models like Transformer-ELMo, XLNET, RoBERTa, MegatronLM have been introduced after BERT. These models perform very well on many tasks but require training on huge datasets, and their size increases exponentially, requiring training on huge datasets. For instance, the MegatronLM model by NVIDIA has 8.3 billion parameters. The RoBERTa model was trained on 160 GB of text [9]. Due to such large size of the models, the training, inference time, and computation resources required increase. Hence, smaller distilled versions of these models, such as DistilBERT and tinyBERT, come with a significantly smaller size than these models but give comparable performance by using the technique of Distillation [3]. For example, the DistilBERT has 66million parameters compared to 340 million parameters in the BERT Large model, but it retained around 95 percent of the performance in most NLP Tasks. Table 1 shows a comparison of different models introduced since the introduction of BERT. It can be seen that DistilBERT has the smallest size compared to other models. Hence, DistilBERT is a faster, cheaper, accurate model for language modeling.Our model will generate answers using either BERT [2], AlBERT [10], or DistilBERT[3] models.

III. Proposed Techniques And Algorithm

This section will explain the techniques used for developing the model. A brief overview of the proposed model is given in figure 2.

TABLE I
COMPARISON OF SIZE OF DIFFERENT MODELS USED FOR LANGUAGE MODELLING

Name of the Model	Organization	Number of Parameter
BERT-Large	Google AI	340 million
Tranformer-ELMo	AI2	465 million
GPT-2	Open AI	1500 million
MT-DNN	Microsoft	330 million
XLNET	Google and Carnegie Mellon Univ.	340 million
XLM	Facebook	665 million
Grover-Mega	University of Washington	1500 million
RoBERTa	Facebook	355 million
MegatronLM	Nvidia	8300 million
DistilBERT	Hugging Face	66 million

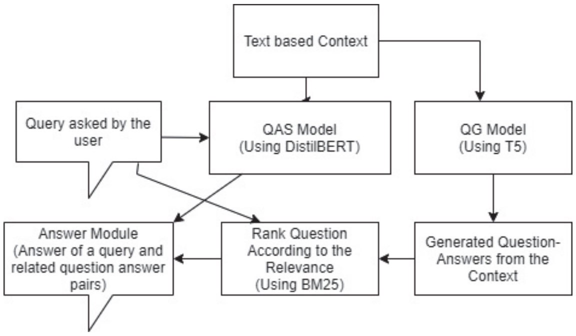

Text based Context
Query asked by the user
QAS Model (Using DistilBERT)
QG Model (Using T5)
Answer Module (Answer of a query and related question answer pairs)
Rank Question According to the Relevance (Using BM25)
Generated Question-Answers from the Context

Fig. 2. Block diagram of the proposed model

A. *Question Answering System*

As the transfer learning technique is becoming more prevalent in Natural Language Processing due to its novelty and promising performance, it is challenging to operate these models in one the edge or constrained computation training or inference budgets. Lowering the latency of these models in production is a challenge. Hence, compressing these models to be suitable for environments with lower computation capabilities and decreasing the inference time of the models is an area that has attracted interest recently. One such technique to lower the model's size and hence the inference time is called the distillation of knowledge. Some other techniques which have been used before are quantization, weight pruning, etc.

DistilBERT is one such model which employs the technique of knowledge distillation from the bigger BERT model. It is also called the teacher-student paradigm, where a smaller model uses the knowledge of the more extensive and deep model to produce similar results.
Here, the DistilBERT uses the logits produced from BERT, but these are not used as they are; instead, the DistilBERT uses the dark knowledge from the BERT. Dark Knowledge can be referred to as the knowledge that the BERT model discards while taking the word with maximum probability.

For example, the encoder blocks from the DistilBERT model produce the probability of the words that can be used in the place of masked words. The vector of the probabilities is then one-hot encoded, and the word with maximum probability is set to one while others are set to 0. However, the other probabilities which are near to the highest probability can also be helpful to determine other possible words that can be present in place of the masked word in the sentence.

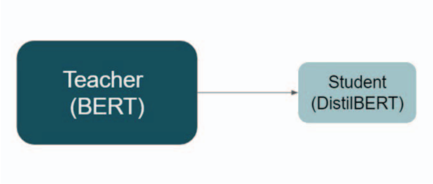


Fig. 3. Relationship between BERT and DistilBERT

Let us consider a sentence - ”New Delhi is a city in India.” Suppose ”city” is masked in the sentence. The masked sentence becomes - New Delhi <mask>in India. Here, different probabilities of the possible words will be generated, and other possible words like ”place” or ”territory” will make more sense as compared to other words, which will have a small probability$(\tilde{0})$ like ”computer.”

The DistilBERT takes this dark knowledge from the BERT models to train, which is done by taking the vector of predicted probabilities of the BERT and using as ”t” in the loss function shown in equation 1. In the equation, t is the logits from the teacher, and s is the logits of the student. Rather than training with a cross-entropy over the hard targets (one-hot encoding of the gold class), we transfer the knowledge from the teacher to the student with a cross-entropy over the soft targets. This loss is a richer training signal since a single example enforces more constraint than a hard target.

$$L = -\sum_i t_i * log(s_i) \tag{1}$$

There is another challenge that needs to be considered while distilling the dark knowledge from the teacher model is that the probabilities which are close to the highest predicted probability should only be taken into account and not all the probabilities that are tending to zero so that irrelevant knowledge is not passed on to the student model. For this, the calculation of the softmax layer outputs is changed by introducing a softmax temperature [11]. The modified equation for the softmax is shown in equation 2.

$$p_i = \exp(z_i/T)/\sum_j \exp(z_j/T) \tag{2}$$

The ”T” in the equation is the softmax temperature, and modifying the value of T changes the distribution of the softmax layer outputs. If T = 0, then the distribution tends to form a distribution similar to the one-hot encoded output where the highest probability word is given a value of 1 while others are given 0. The same temperature value is applied on both the teacher and the student while training; this reveals more training signals for each training example. This value is set to 1 at inference time to rev=cover the standard softmax value. In this way, the dark knowledge helps in training the student model.

B. Suggested Questions

In the Question Answering System, it can be helpful to the user if the user gets a set of relevant questions and their answers from the context along with the questions that were asked. It can make the search easier for the user and helpful if a user cannot frame a query; he/she can directly go to the generated questions list to see the answers. For question generation, two types of approach can be followed-

- Answer Aware Question Generation – the question is generated based on answers.
- End to End Question Generation – The questions are self-generated without any already known answers.

We used the answer-aware question generation technique in our work because it will be better if the user gets answers along with the question when a suggested question is given. If end-to-end generation were used, the answers to the generated question would again be found from the context.

Google's t5, which is a transfer learning model, is used for generating the questions. The t5 model is a simple encode decoder model, and it converts the problems into the text-to-text format, making it different from other popular transfer learning models such as BERT, which converts the natural language into word embeddings.

For answer-aware question generation using t5, the positions of the answer need to be extracted first. The answers from a given context are generated by training the t5 model to detect answer-like spans from a given context. The answers in a sentence are highlighted with <sep>tags wherever an answer is present. The set for the current context is ended by using a </s>tag.

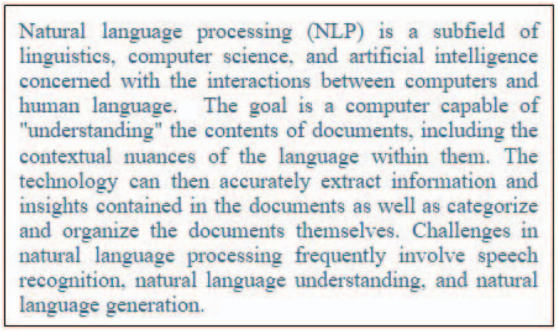
For example - Java was created by Sun Microsystems in 1990s Target text – Sun Microsystems <sep >1990s<sep ></s>

In this way, the t5 is trained to extract answer spans from the given context.

Since the t5 model is now trained to extract the answer spans from the context, the input for the answer-aware question generation can be made using the above model. The context input by the user is preprocessed before giving as input for question generation. The context is divided into sentences, and the possible answer spans are generated by the model described above. The answers in each sentence are then surrounded with a <h1>tag so that the question generation model is aware of the answer in the given sentence and can generate the possible questions while it is aware of the answers.

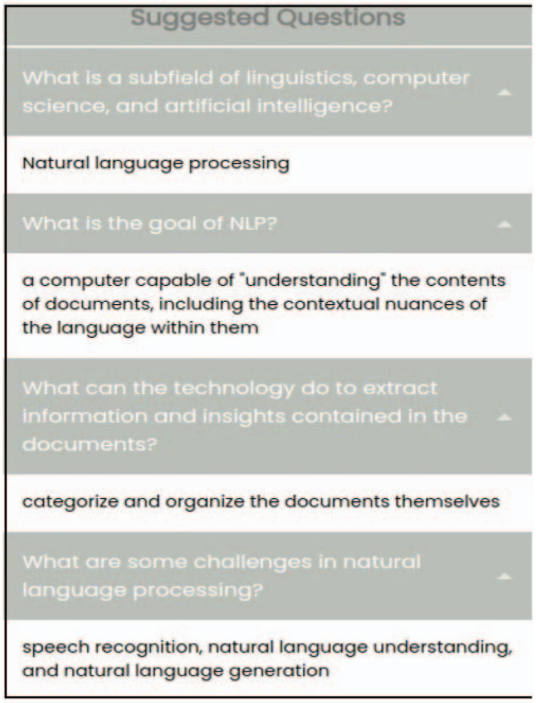
For example – Context - 42 is the answer to life, universe and everything. Preprocessed - <h1>42 <h2>is the answer to life universe and everything.

The Question is then generated by the t5 model from the preprocessed sentence with the answer token. Final result – What is the answer to life, universe and everything. Examples of the questions generated from the context shown in figure 4 by the system are shown in Figure 5.



Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Fig. 4. The context is given as input for Question Generation



Suggested Questions

What is a subfield of linguistics, computer science, and artificial intelligence?

Natural language processing

What is the goal of NLP?

a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them

What can the technology do to extract information and insights contained in the documents?

categorize and organize the documents themselves

What are some challenges in natural language processing?

speech recognition, natural language understanding, and natural language generation

Fig. 5. Question Generated from the given context

C. BM25

After question generation, scores of the generated questions are calculated with the questions input by the user, and the generated questions are ranked according to their relevance scores. We used the BM25 [12] to calculate the similarity score between the user query and the generated questions. The question with a higher score will be shown on the top; hence, the user will get the most relevant questions to their query.

IV. Datasets Used

We had used the SQuAD v1 dataset for training both question answering and question generation models. It is a reading comprehension dataset built on the questions posed by crowd workers based on Wikipedia texts. The answer to every question is a segment of text or span from the corresponding passage. It has about 100,000 questions that the crowd workers have written upon reading the given context [13]. BERT and DistillBERT are trained on the SQuAD V1 dataset; instead, the AlBERT model is trained on the SQuAD 2.0 dataset, containing the unanswerable questions.

V. Experiments and Results

In this section, we evaluate the performance of the developed model. Here, we present the evaluation matrices used for the task and then the evaluation results on different question context pairs created to explore the model performs.

A. Evaluation matrices

To evaluate the model performance, we had used f1-score and EM(exact match) for the question answering task and BLEU(bilingual evaluation understudy), ROUGE(Recall-Oriented Understudy for Gisting Evaluation) for the question generation, and METEOR(Metric for Evaluation of Translation with Explicit Ordering).

BLEU measures the closeness of the model-generated questions to the asked query, taking question length, word choice, and word order into consideration [14].ROUGE-L metric for evaluation was initially developed to evaluate the effectiveness of the text summarizing by comparing overlapping n-grams of a model's-generated summary versus a human-generated summary. We used generated questions from the context and query asked by the user. The matching is done by finding the longest common sequence in the sentences[15]. METEOR is based on first finding the harmonic mean of unigram precision and recall by giving recall a higher weight than precision[16]. It gauges the similarity between two texts coordinating with the main text to the second gathering of text. For this situation, the meteor measure will be utilized to find the comparability between the query and the generated answer by calculating the weighted F-score[17].

B. Evaluation Results

We tested the model with different unseen contexts and questions, and it performed satisfactorily. The exact match and f1 score on the SQuAD dataset are shown in Table 2. As we can see from the table, the BERT model performs well compared to its other variants.

The question generation model generated semantically similar and relevant questions for the same contexts. Since the question generation task was carried out using text to text (t5) model, the questions generated need to be evaluated based on the understandability of the questions compared to natural language. The metrics suitable for this task are BLEU, METEOR, and ROUGE scores and hence are used to evaluate the question generation (suggested questions) task. The results are evaluated on SQuAD 1.0 dataset. The nlg – eval package is used for calculating the metrics. The systems generated understandable questions in natural language for the context entered.

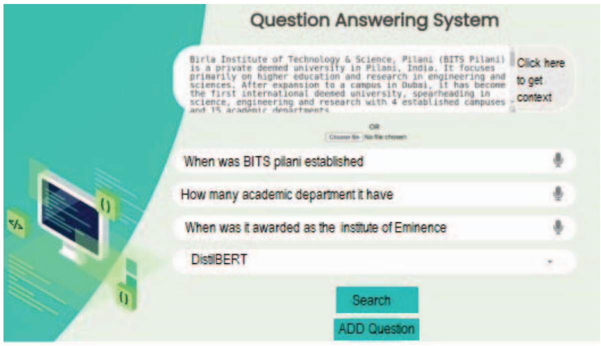
TABLE II
EM/F1 Score of the Models Used

Model Name	f1-Score	Exact Match
BERT	88.5	81.2
DistillBERT	86.692	78.845
ALBERT	87.462	84.418

TABLE III
Evaluation of Question Generation Task

Model	BLUE-4	METEOR	ROUGE-L
t5	18.5921	24.9915	40.1886

The result of the evaluation for the question generation task is shown in Table 3. An example of the Question Answering System with Suggested Questions developed is shown in Figure 6 and 7.


Question Answering System

Birla Institute of Technology & Science, Pilani (BITS Pilani) is a private deemed university in Pilani, India. It focuses primarily on higher education and research in engineering and sciences. After expansion to a campus in Dubai, it has become the first international deemed university, spearheading in science, engineering and research with 4 established campuses and 15 academic departments

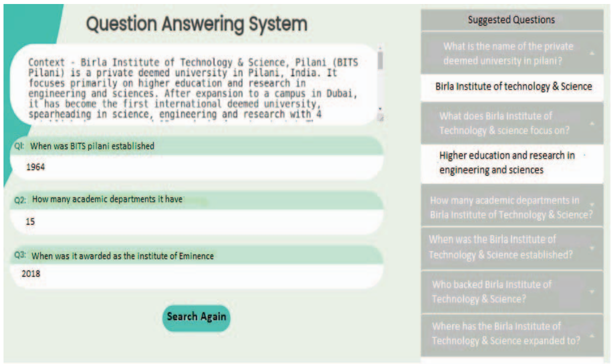
When was BITS pilani established

How many academic department it have

When was it awarded as the institute of Eminence

DistilBERT

Fig. 6. Giving context and Questions as input


Question Answering System

Context - Birla Institute of Technology & Science, Pilani (BITS Pilani) is a private deemed university in Pilani, India. It focuses primarily on higher education and research in engineering and sciences. After expansion to a campus in Dubai, it has become the first international deemed university, spearheading in science, engineering and research with 4

Ql: When was BITS pilani established

1964

Q2: How many academic departments it have

15

Q3: When was it awarded as the institute of Eminence

2018

Suggested Questions

What is the name of the private deemed university in pilani?

Birla Institute of technology & Science

What does Birla Institute of Technology & science focus on?

Higher education and research in engineering and sciences

How many academic departments in Birla Institute of Technology & Science?

When was the Birla Institute of Technology & Science established?

Who backed Birla Institute of Technology & Science?

Where has the Birla Institute of Technology & Science expanded to?

Fig. 7. Output with answers to the questions asked,suggested questions and their answers

VI. Web Application

A web interface is developed to provide the interface for asking the query. The User Interface is shown in Figure 4. The web application is developed using the Django framework in the backend and HTML, CSS, and JavaScript in the frontend. The application can take the context as input by typing text or using a PDF file. The context can also be input using an image. The text in the image is extracted using the integrated OCR model and considered as the context for answering the questions. If a user wants to ask questions about a particular topic but does not have any specific context, then the web is scraped to get the user's relevant information about the topic input. The questions can be input by typing or by voice. If the question is unanswerable by the system, then the relevant links are scraped from the web, and the most relevant ones are displayed to the user. The final output is the answered question from the context and the suggested question from the context.

VII. Conclusion and Future Work

With the already existing close domain context-based question answering system, we had integrated the Suggested Questions feature, which generates questions from the context input by the user and uses different models like DistilBERT and RoBERTa to get answers from the context. We had ranked the suggested questions based on BM25 scores to find their relevance with the asked query. The context can be given in the form of PDF or image(extract texts from image). Our work is still in the progress state, as the accuracy of the integrated question generation model can be improved. Further, getting answers from pictorial representations such as diagrams and charts can be explored, which can be very useful for getting complex answers by just looking at the pictorial representations.

Acknowledgment

The authors would like to convey their sincere thanks to the Department of Science and Technology (ICPS Division), New Delhi, India, for providing financial assistance under the Data Science (DS) Research of Interdisciplinary Cyber-Physical Systems (ICPS) Program [DST/ICPS/CLUSTER/Data Science/2018/Proposal-16:(T-856)] at the department of computer science, Birla Institute of Technology and Science, Pilani, India. The authors are also thankful to the authorities of Birla Institute of Technology and Science, Pilani, to provide basic infrastructure facilities during the preparation of the paper.

References

[1] A. Agarwal, V. Kumari, Y. Sharma, and L. Goel, "Ranking based question answering system with a web and mobile application," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 52–58.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[4] K. Grover, K. Kaur, K. Tiwari, P. Kumar *et al.*, "Deep learning based question generation using t5 transformer," in *International Advanced Computing Conference*. Springer, 2020, pp. 243–255.

[5] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[6] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[7] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[12] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[13] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[15] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[16] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[17] M. Zihayat and R. Etwaroo, "A non-factoid question answering system for prior art search," *Expert Systems with Applications*, vol. 177, p. 114910, 2021.