HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOL OF INFORMATION TECHNOLOGY AND COMMUNICATION

-----□□🕮□□-----

AI PROJECT REPORT

Project name: COVID-19 INTELLIGENT VIRTUAL ASSISTANT USING BIGBIRD

Supervisor Assoc. Prof: Phạm Văn Hải
Group number: Group 4
Student names:

Đinh Thị Kiều Trinh	**20194864**
Đặng Quốc Tú	**20194871**
Đinh Thế Kiệt	**20194783**

Hà Nội, tháng 1 năm 2021

Table of Contents

AI Project name:

COVID-19 Intelligent Virtual Assistant Using BIG BIRD

Student name: Đinh Thị Kiều Trinh, Đặng Quốc Tú, Đinh Thế Kiệt.

Class ICT Global, Hanoi University of Science and Technology, No1. Dai Co Viet st., Hanoi, Vietnam

ABSTRACT

People are demanding easy access to reliable information regarding the extremely potent and rapidly COVID-19 pandemic with precise, up-to-date virus information difficult to obtain. Especially in remote areas, it is becoming more difficult to consult a medical specialist when the immediate hit of the epidemic has occurred. We launched this project to ensure that everyone has quick access to any information relating to COVID-19 to help them manage their health concerns during this difficult time. We use the Big Bird model to attempt to propose applications in question answering about COVID-19. The Big Bird model is a sparse attention method that reduces quadratic dependency to linear dependency. The proposed model replaces the quadratic attention mechanism in the transformer with a mix of random attention, window attention, global attention to achieve a linear complexity requirement. As a result, experimental

results show that the proposed model can process longer sequences than traditional transformers like BERT and achieve better results in some NLP tasks.

Keywords

Artificial intelligence, chatbot, COVID-19, Big Bird model, question answering, NLP tasks, linear complexity.

1 Introduction

Several interesting attempts were aimed at alleviating the quadratic dependency of Transformers to answer the question and summarize documents on several different datasets, which can be broadly categorized into two directions.

- First line of work embraces the length limitation and develops method around it. Most prominently, SpanBERT, ORQA, REALM, RAG have achieved strong performance for different tasks. However, it is worth noting that these methods often require significant engineering efforts (like back prop through large-scale nearest neighbor search) and are hard to train.

- Second line of work questions if full attention is essential and has tried to come up with approaches that do not require full attention, thereby reducing the memory and computation requirements. Prominently, Dai et al., Sukhbaatar et al., Rae et al. have proposed auto-regressive models that work well for left-to-right language modeling but suffer in tasks that require bidirectional context. Recently, Longformer introduced a localized sliding window-based mask with few global masks to reduce computation and extend BERT to longer sequence-based tasks.

BIG BIRD comes to handle sequences of length up to 8x of what was previously possible using similar hardware. BIG BIRD is a Google Researchers's work that is closely related to and built on the work of Extended Transformers Construction. BIG BIRD is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model. Along the way, their theoretical analysis reveals some of the benefits of having O(1) global tokens (such as CLS), that attend to the entire sequence as part of the sparse attention mechanism. As a consequence of the capability to handle longer context, BIGBIRD drastically improves performance on various NLP tasks such as question answering and summarization.

In this paper, we apply the BIGBERD model to have a conversation with millions of pages of information from the WHO (World Health Organization) to quickly get answers conversationally about the COVID-19 Pandemic.

2 Preliminaties

2.1 BIG BIRD Architecture

In this section, we describe the BIGBIRD model using the *generalized attention mechanism* that is used in each layer of a transformer operating on an input sequence $X = (x1,...,xn)$ 2 R$n \rightarrow d$.

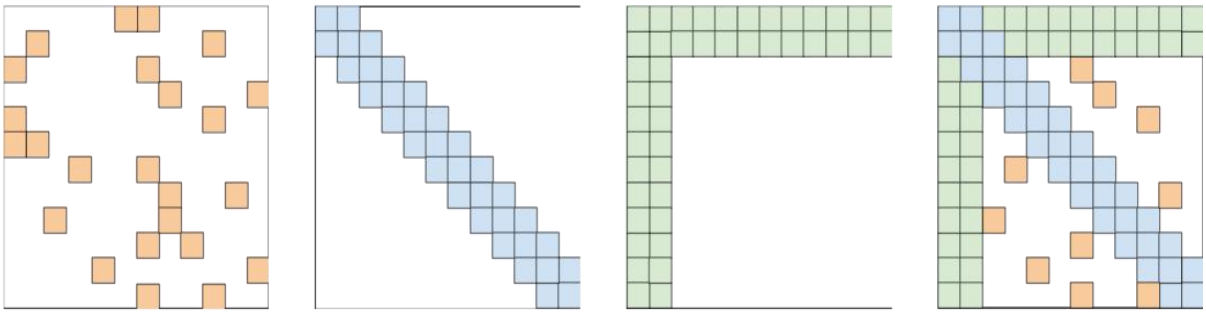


Figure 1: Building blocks of the attention mechanism used in BIG BIRD. The white color indicates the absence of attention. (a) random attention with r = 2, (b) sliding window attention with w = 3 (c) global attention with g = 2. (d) the combined BIGBIRD model.

First, we consider the simplest random graph construction, known as the Erdos-Rényi model, where each edge is independently chosen with a fixed probability. Thus, we propose sparse attention where each query attends over r random number of keys i.e. A(i, ·)=1 for r randomly chosen keys (see Fig. 1a).

Next, we begin with a sliding window on the nodes. Then a random subset (k%) of all connections is replaced with a random connection. The other (100 - k)% local connections are retained. However, deleting such random edges might be inefficient on modern hardware, so we retain it, which will not affect its properties. In summary, to capture these local structures in the context, in BIGBIRD, we define sliding window attention, so that during self-attention of width w, query at location i attends from i - w/2 to i + w/2 keys. In our notation, $A(i,i\ w/2 : i+w/2) = 1$ (see Fig. 1b).

The *final* piece of BIGBIRD is inspired by "global tokens" (tokens that attend to all tokens in the sequence and to whom all tokens attend (see Fig. 1c). These global tokens can be defined in two ways:

• BIGBIRD - ITC: In internal transformer construction (ITC), we make some existing tokens "global", which attend over the entire sequence. Concretely, we choose a subset G of indices (with g:= |G|), such that A(i, :) = 1 and A(:, i)=1 for all i ∈ G.

• BIGBIRD-ETC: In extended transformer construction (ETC), we include additional "global" tokens such as CLS. Concretely, we add g global tokens that attend to all existing tokens. In our notation, this corresponds to creating a new matrix B ∈ [0, 1](N+g)↠ (N+g) by adding g rows to matrix A, such that B(i, :) = 1, and B(:, i)=1 for all i ∈ {1, 2,...g}, and B(g + i, g + j) = A(i, j) ∀i, j ∈ {1,...,N}. This adds extra

location to store context and as we will see in the experiments improves performance.

The final attention mechanism for BIGBIRD (Fig. 1d) has all three of these properties: queries attend to r random keys, each query attends to $w/2$ tokens to the left of its location and $w/2$ to the right of its location and they contain g global tokens (The global tokens can be from existing tokens or extra added tokens).

2.2 Implementations

A. AIML:

To create our knowledge base for normal conversation, we have used AIML files to store the question and answers pair. When the user converses with our chatbot, the input is matched to patterns listed in AIML files and the corresponding answer is returned as a response. The sample AIML file structure is HELLO USERNAME Hello User!

B. Lemmatization and POS Tagging Using WordNet:

Information extraction from the input text was done by extracting keywords. For example, “What is the current placement scenario?” contains “current”, “placement” and “scenario” as the keywords. Appropriate Lemmas of the keywords were found using Lemmatization and POS tagging, to group together the different inflected forms of the words. For example, requiring,

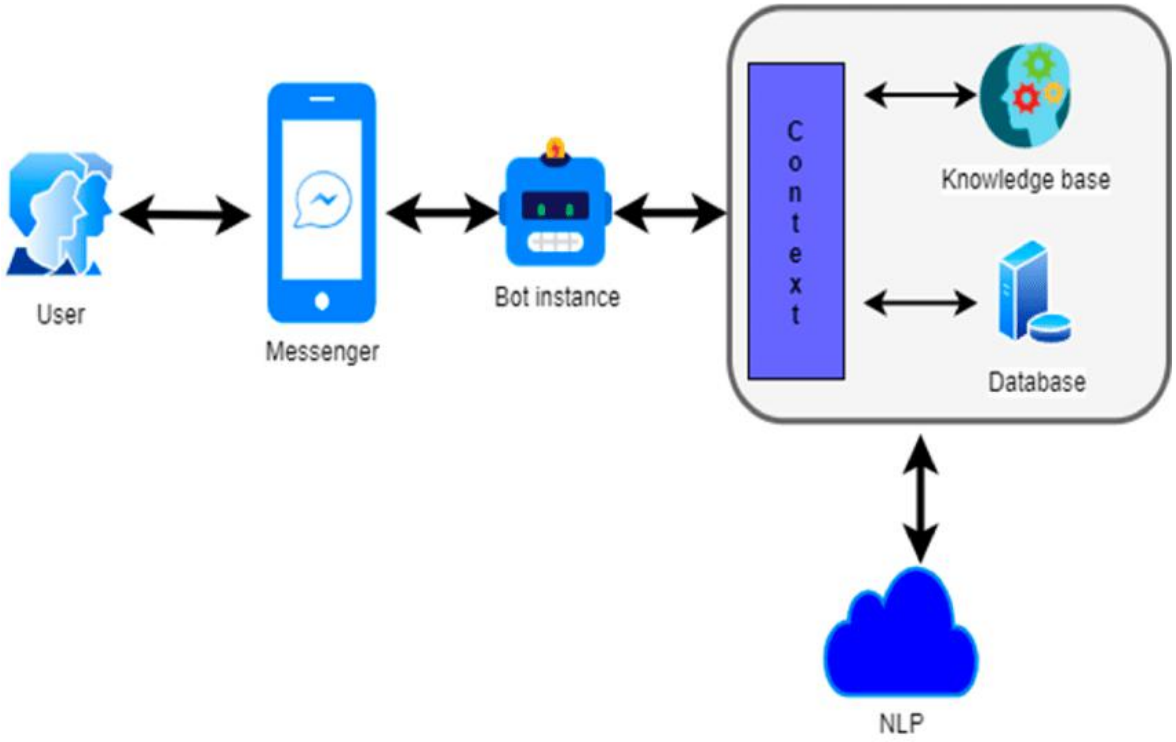
require, and required should map to require. WordNet from Python's "nltk" package was used for this purpose.

C. Semantic Sentence Similarity:

There are various combinations in which users can input the same query. For example, Q1: What is the notice regarding PG courses re-registration? Q2: Tell me about re-registration in PG courses in our college.

3 The Proposed Model

The proposed model is used to extract information from a paragraph and then answer specific questions.



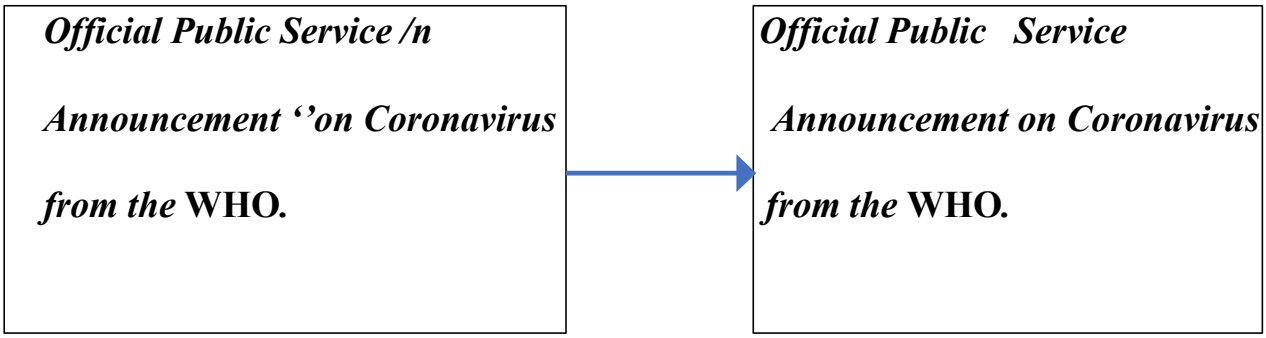
A. Collect Database

WHO is the United Nations agency that connects nations, partners, and people to promote health, keep the world safe and serve the vulnerable – so everyone, everywhere can attain the highest level of health. Information is extracted from the WHO website and be preprocessed to complete the database.

B. Context Identification

Step 1: Clean raw text.

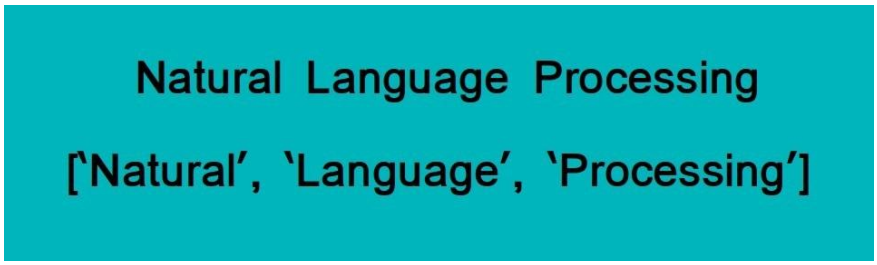
Raw data might contain irrelevant or meaningless data termed as noise which can significantly affect various data analyses such as HTML tags <h1>, <h2>. Cleaning text better our performance.



Step 2: Pre train model and initialize tokens.

NLP pre-trained models are useful for NLP tasks like translating text, predicting missing parts of a sentence, or even generating new sentences.

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units is called a token.



Step 3: Embed tokens, paragraphs, questions into our model.

After preprocessing our raw data to boost our model's performance, we embed tokens, paragraphs, questions into our model.

4 Experiments and results

4.1 Dataset

We present TriviaQA, a challenging reading comprehension dataset containing over 650K question-answer-evidence triples. TriviaQA includes 95K question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average, that provide high-quality distant supervision for answering the questions.

Dataset	Large scale	Freeform Answer	Well formed	Independent of Evidence	Varied Evidence
TriviaQA	✓	✓	✓	✓	✓
SQuAD (Rajpurkar et al., 2016)	✓	✓	✓	✗	✗
MS Marco (Nguyen et al., 2016)	✓	✓	✗	✓	✓
NewsQA(Trischler et al., 2016)	✓	✓	✓	✗*	✗
WikiQA (Yang et al., 2016)	✗	✗	✗	✓	✗
TREC (Voorhees and Tice, 2000)	✗	✓	✓	✓	✓

*Table 1: Comparison of TriviaQA with existing QA datasets. Our dataset is unique in that it is naturally occurring, well-formed questions collected independently of the evidence. *NewsQA uses evidence articles indirectly by using only article summaries.*

- **Example**:

Question: American Callan Pinckney's eponymously named system became a best-selling (the 1980s-2000s) book/video franchise in what genre?

Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other fitness video in the US.

→ *Answer: Fitness*

- Dataset Collection:

A large dataset was collected to support the reading comprehension task. First, they gathered question-answer pairs from 14 trivia and quiz-league websites. They removed questions with less than four tokens since these were generally either too simple or too vague.

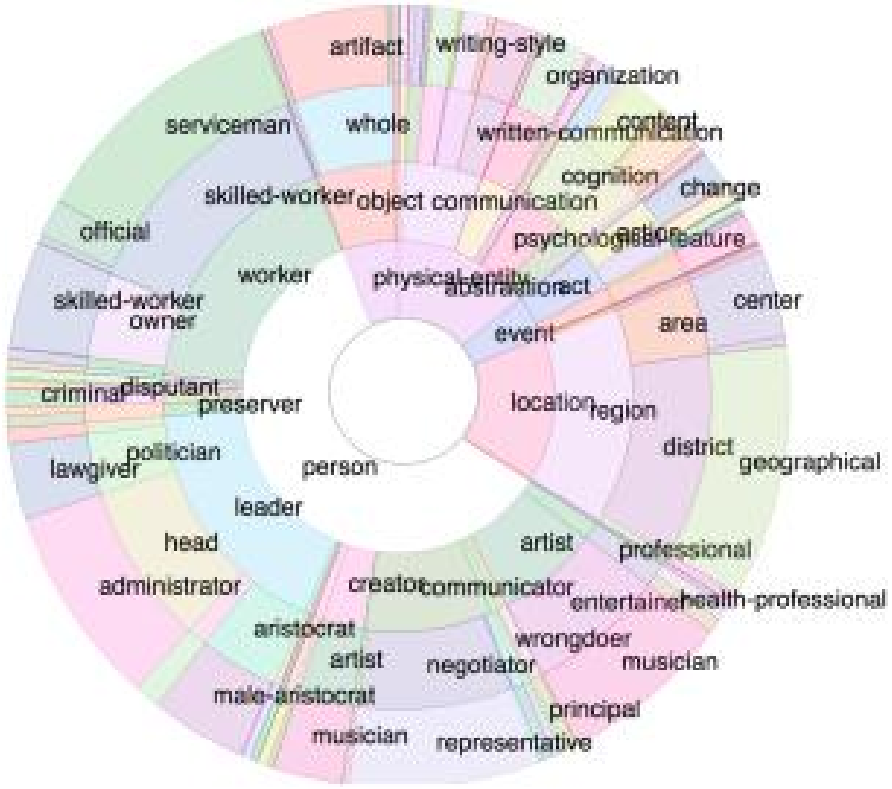


Figure 1: Distribution of hierarchical WordNet synsets for entities appearing in the answer. The arc length is proportional to the number of questions containing that category.

4.2 Evaluation:

The exact match ratio is a very strict measure of the model performance. It increases only when the model correctly identifies every possible label that an example has, without any false positive.

$$\text{Exact match ratio} = \frac{\text{Number of examples with exact label match}}{\text{Total number of examples}}$$

The max sequence length of the big bird model is 4096 tokens. We have now preprocessed the complete dataset TriviaQA and a shorter version short_validation_dataset that only consists of data samples < 4096 tokens.

```
print("Exact Match (EM): {:.2f}".format(100 * sum(results_short['match'])/len(results_short)))
```

```
Exact Match (EM): 81.29
```

- Exact match in short_validation_dataset: 81.29

The whole dataset shows a degradation in exact match, which is expected since the added data samples all have a longer context than the model can handle. It is still a very good score though.

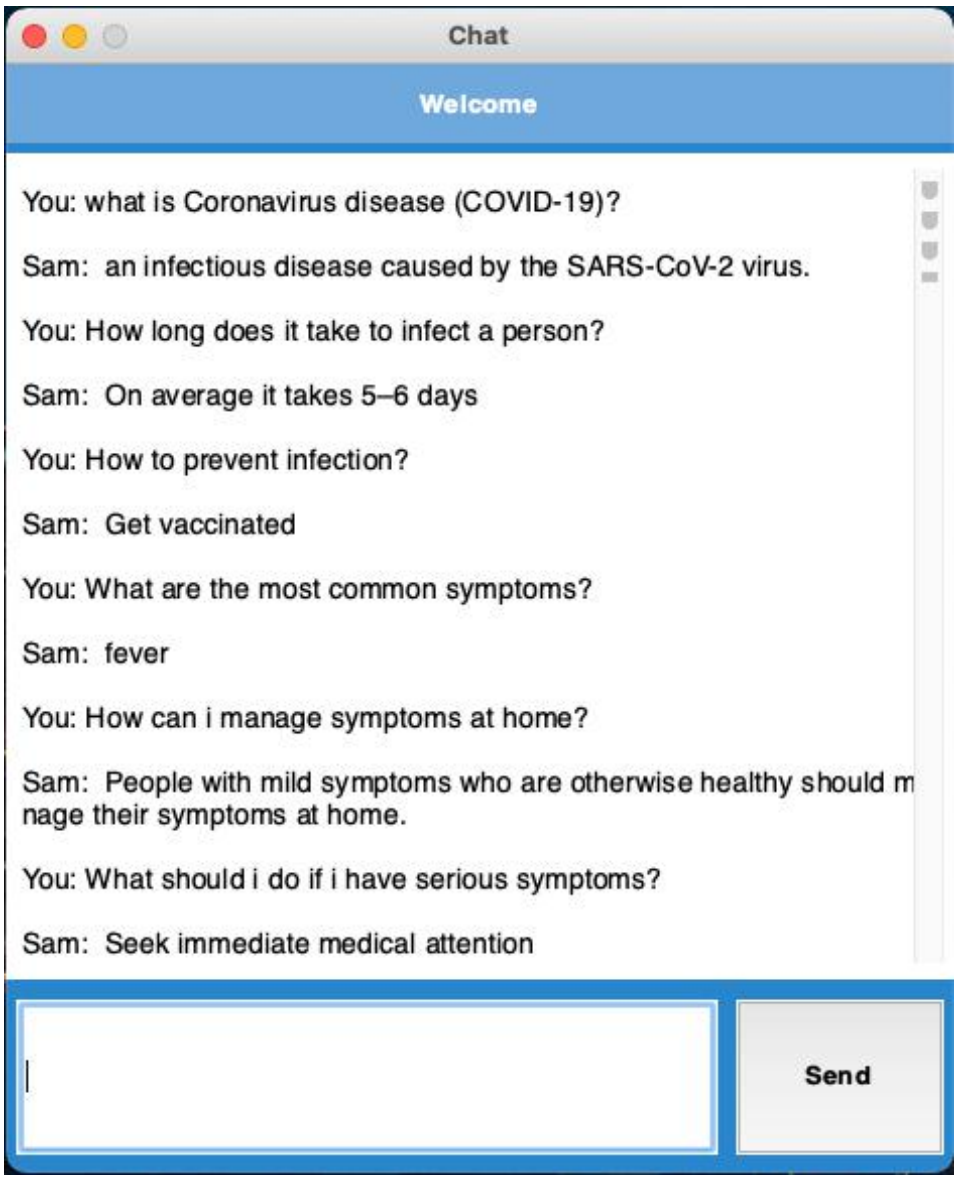
```
print("Exact Match (EM): {:.2f}".format(100 * sum(results['match'])/len(results)))
```

```
Exact Match (EM): 65.78
```

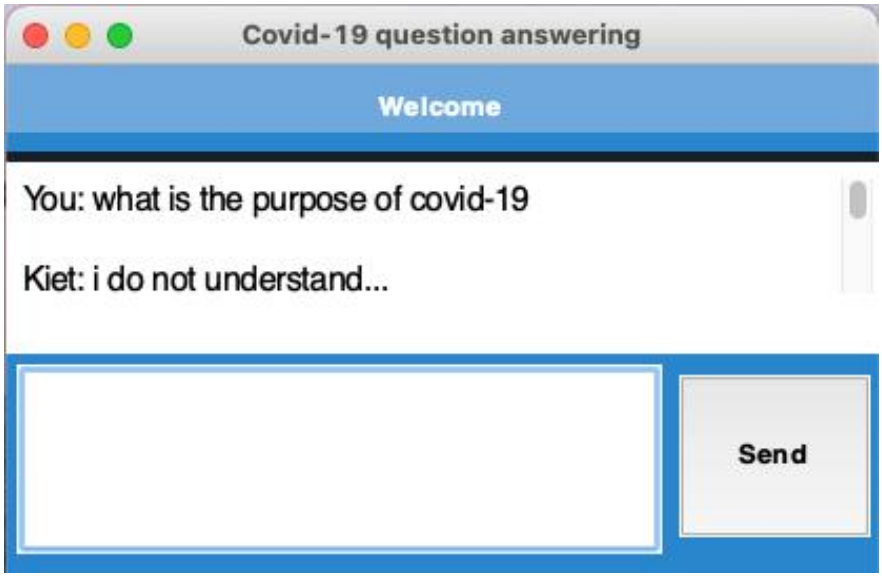
- Exact match in short_validation_dataset: 65.78

4.3 Result

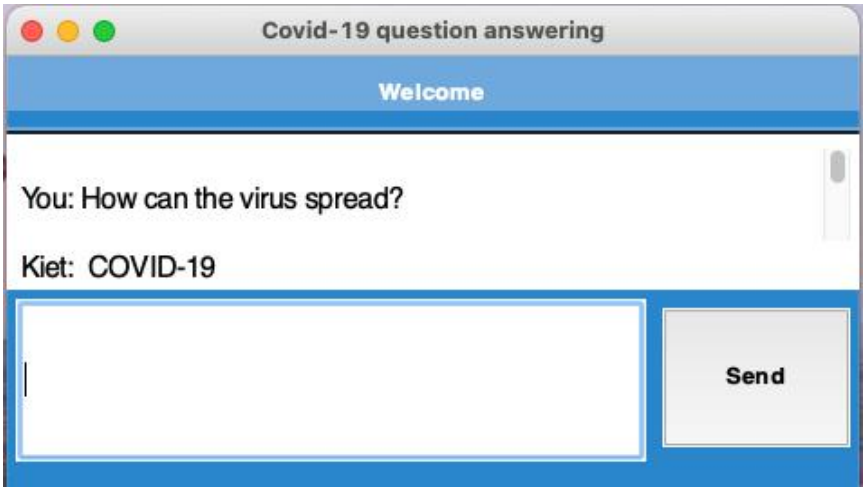
The big bird model extracts information from the excerpt and returns the answer. It returns “I don’t understand” when the question is meaningless and sometimes wrong answers when it misunderstands questions.



The chatbot returns answers.



The chatbot return "I don't understand" when the question is meaningless.



The chatbot returns wrong answer.

5 Conclusion

In this study, a solution has been proposed against the problem of lacking precise information that arises during the COVID-19 epidemic process, with a question answering system where people can quickly and easily get correct answers to their questions about the disease responding to the information of the WHO. The investigation in this paper presented the BIG BIRD model, a sparse attention mechanism that is linear in the number of tokens to have a conversation with millions of pages of information from WHO (World Health Organization) to quickly get answers conversationally. Looking at the results of the research, it has been seen that the questions were answered correctly to a large extent. Considering the rapid increase in the number of question answering systems and the widespread use of question-answer systems today, it is clear that more work needs to be done in this field. However, the model can play an important role in reducing the burden of the health system in critical periods such as the epidemic process we live in these days.

6 Future Enhancements

We can include voice-based queries. The users will have to give voice input and the system will give the text output and voice response. Also, after successful execution of chatbot in COVID-19 pandemic domain, we can

implement it in other domains like medical, forensic, sports, etc. It will be beneficial in all the fields as without spending much time, we are accessing the relevant information and that too without any sorting.

7 References

[1] A. Abboud, V. V. Williams, and O. Weimann. Consequences of faster alignment of sequences. In *International Colloquium on Automata, Languages, and Programming*, pages 39–51. Springer, 2014.

[2] A. Abboud, A. Backurs, and V. V. Williams. Tight hardness results for lcs and other sequence similarity measures. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 59–78. IEEE, 2015.

[3] J. Abreu, L. Fred, D. Macêdo, and C. Zanchettin. Hierarchical attentional hybrid neural networks for document classification. In *International Conference on Artificial Neural Networks*, pages 396–402. Springer, 2019.

[4] J. Ainslie, S. Ontanon, C. Alberti, P. Pham, A. Ravula, and S. Sanghai. Etc: Encoding long and structured data in transformers. *arXiv preprint arXiv:2004.08483*, 2020.

[5] C. Alberti, K. Lee, and M. Collins. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, 2019.

[6] J. Alt, R. Ducatez, and A. Knowles. Extremal eigenvalues of critical erd\h {o} sr\'enyi graphs. *arXiv preprint arXiv:1905.03243*, 2019.

[7] A. Backurs and P. Indyk. Edit distance cannot be computed in strongly subquadratic time (unless seth is false). In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 51–58, 2015.

[8] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[9] F. Benaych-Georges, C. Bordenave, A. Knowles, et al. Largest eigenvalues of sparse inhomogeneous erdos–rényi graphs.″ *Annals of Probability*, 47(3):1653–1676, 2019.

[10] F. Benaych-Georges, C. Bordenave, A. Knowles, et al. Spectral radii of sparse random matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 2141–2161. Institut Henri Poincaré, 2020.

[11] R. Bharanikumar, K. A. R. Premkumar, and A. Palaniappan. Promoterpredict:

sequence-based modelling of escherichia coli 70 promoter strength yields logarithmic dependence between promoter strength and sequence. *PeerJ*, 6:e5862, 2018.

[12] S. Buldyrev, A. Goldberger, S. Havlin, R. Mantegna, M. Matsa, C.-K. Peng, M. Simons, and H. Stanley. Long-range correlation properties of coding and noncoding dna sequences: Genbank analysis. *Physical Review E*, 51(5):5084, 1995.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledgeintensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

[14] Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Shreya Bisen, Vasundhara Rathod. Implementation of a Chatbot System using AI and NLP, pages 27-28.