Development and Comparative Analysis of Generative Chatbots Using BERT and BART Architectures Trained on the Stanford Question Answering Dataset

Trevor McGirr, Eyoha Mengistu, Reed Oken

University of San Diego, Shiley-Marcos School of Engineering

Masters in Applied Artificial Intelligence

Natural Language Processing (AAI-520-02)

Professor Roozbeh Sadeghian

October 23, 2023

Abstract

This project focused on designing and implementing a generative-based chatbot capable of conducting question and answer style conversations, adapting to different contexts, and handling a variety of topics. For this purpose, the Stanford Question Answering Dataset (SQuAD) was used and two state-of-the-art architectures, BERT (Bidirectional Encoder Representations from Transformers) and BART (Bidirectional and Auto-Regressive Transformers), were fine tuned on the dataset. Results showed that the BERT model, when fine-tuned on the SQuAD and coupled with Pinecone as a vector database, outperformed the BART model in terms of the evaluation metrics used, including BLEU and ROUGE scores. Consequently, the final chatbot was implemented using the BERT model.

Introduction

Chatbots have revolutionized the way we interact with technology, providing a more natural and interactive means of communication. With the integration of artificial intelligence, chatbots have evolved into intelligent conversational agents capable of understanding and responding to user queries. The goal of this project was to build a chatbot that could handle question and answer style prompts that are adapted to context (provided by the user), and handle a variety of topics, focusing primarily on the fine-tuned data set. The SQuAD is a popular dataset in the natural language processing community, consisting of questions posed by crowdworkers on a set of Wikipedia articles, with the answer to every question being a segment of text from the corresponding reading passage. By leveraging this dataset, a chatbot capable of providing accurate and relevant answers to user queries was developed.

Data Cleaning/Preparation

SQuAD was imported and preprocessed to format it correctly for training. This involved extracting question-answer pairs and their corresponding contexts, checking for any missing values, and handling any data cleaning necessary to make the dataset suitable for training and fine tuning of the Transformer models. Data was then transformed into a format that could be used to train the BERT and BART models. The preprocessing steps were crucial in ensuring that the dataset was free from any inconsistencies or errors which could affect the performance of the models. Additionally, the preprocessing steps also helped to structure the data in a way that facilitated the training process, ultimately leading to better model performance.

Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to gain a better understanding of the characteristics of the dataset. This included analyzing the distribution of question lengths, answer lengths, and context lengths (Appendix D, E, F). Word clouds were generated for questions, answers, and contexts to visualize the most common words and identify key topics and themes (Appendix G, H, I). The results of the EDA provided valuable insights that informed the model training and evaluation process. For instance, the analysis of question and answer lengths helped to determine the appropriate input size for the models, while the word clouds helped to identify common themes and topics that could be used to evaluate the chatbot's performance.

Model Selection

Two state-of-the-art models were considered for the project: BERT (Bidirectional Encoder Representations from Transformers) and BART (Bidirectional and Auto-Regressive

Transformers). Both models represent the cutting edge in natural language processing and have been successful in various tasks. Additionally, both models are based on the Transformer architecture first proposed by Vaswani et al. in June 2017, which has revolutionized the field of natural language processing (NLP).

BERT is a transformer-based model that has revolutionized the field of natural language processing with its powerful representation learning capabilities. It has demonstrated superior performance in various natural language understanding tasks such as sentiment analysis, text summarization, and question-answering. BERT's architecture allows it to capture the context of the input text and generate responses that are highly relevant to the query. The decision to use BERT was based on its proven effectiveness in similar tasks and its ability to handle multi-turn conversations, which was a key requirement for our chatbot.

On the other hand, BART is a state-of-the-art model specifically designed for sequence-to-sequence tasks such as text generation and question-answering. BART's architecture combines the benefits of autoregressive models and denoising autoencoders, making it an ideal choice for generating coherent and contextually relevant responses. The decision to experiment with BART was based on its potential to generate high-quality responses for the chatbot, given its success in similar text generation tasks.

Both models were fine-tuned on SQuAD, with BERT being coupled with Pinecone as the vector database to enhance its performance further. Pinecone, a vector database, provides a convenient and efficient way to store and retrieve high-dimensional vectors, which is essential for capturing the semantic relationships between words and phrases. The integration of Pinecone with BERT aimed to leverage these semantic relationships to generate more accurate and relevant responses.

Model Analysis

The BERT model achieved impressive results when fine-tuned on the SQuAD and coupled with Pinecone as the vector database. The BLEU scores ranged from 0.107 to 1.030, and the eval loss consistently decreased from -5.30 to -7.67, indicating that the model was learning effectively and generating high-quality responses. These results highlight the effectiveness of BERT in capturing the context and generating relevant responses, which is crucial for a question-answering chatbot. The integration of Pinecone further enhanced the model's performance by leveraging the semantic relationships between words and phrases, leading to more accurate and contextually relevant responses.

In comparison, the BART model faced some challenges during the training process, which ultimately affected its performance. The average BLEU score was 0.0001, and the average training and validation losses were 1.81 and 2.88, respectively. Despite various fine-tuning approaches taken to improve the BART model's performance, such as adjusting the learning rate, batch size, and number of training epochs, the results were not as satisfactory as those achieved by the BERT model. Several other advanced approaches were also tried, including; Early Stopping which monitors the validation loss and stop training once it stops decreasing (and starts increasing),

Regularization, which adds regularization terms to the loss function to penalize large weights. Implement Dropout to add dropout layers to the model to randomly zero out some of the weights during training, which can prevent overfitting. These challenges highlight the limitations of BART in handling the complexity of SQuAD and the requirements of the chatbot.

The decision to implement the final chatbot using the BERT model was based on its superior performance and the potential benefits of using Pinecone as the vector database. However, it is important to note that the BERT model is not without its challenges. Fine-tuning BERT requires careful consideration of various hyperparameters such as learning rate, batch size, and number of training epochs, which can significantly affect the model's performance. Moreover, BERT's architecture is complex and requires substantial computational resources, which can be a limitation for some applications. Despite these challenges, the benefits of using BERT, such as its powerful representation learning capabilities and proven effectiveness in natural language processing tasks, outweigh the risks. The analysis of the BERT and BART models revealed distinct differences in their performance and the challenges faced during fine-tuning.

Conclusion

In conclusion, the BERT model, when fine-tuned on the SQuAD and coupled with Pinecone as the vector database, outperformed the BART model in terms of the evaluation metrics used. Consequently, the final chatbot was implemented using the BERT model. This report detailed the challenges faced, solutions implemented, model architecture, evaluation and results. Looking ahead, future research and testing should explore integrating advanced NLP techniques such as sentiment analysis, entity recognition, and language translation to further augment and fine tune the chatbot's functionality. The integration of real-time learning capabilities, where the chatbot learns and evolves from user interactions, is another exciting possibility that could be explored in future iterations of this project.

References

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1606.05250

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. *In Advances in Neural Information Processing Systems* (pp. 30-80).

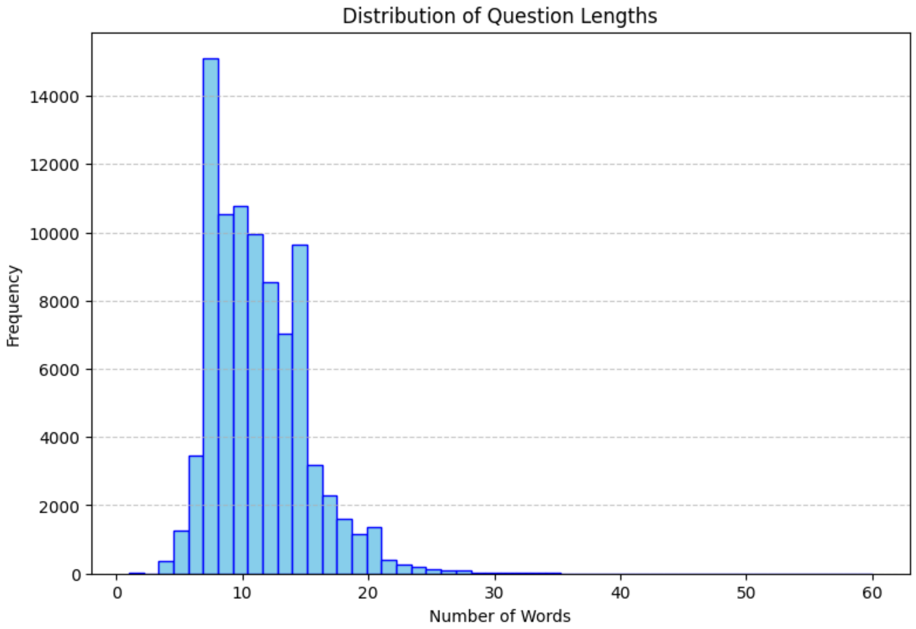
Appendix

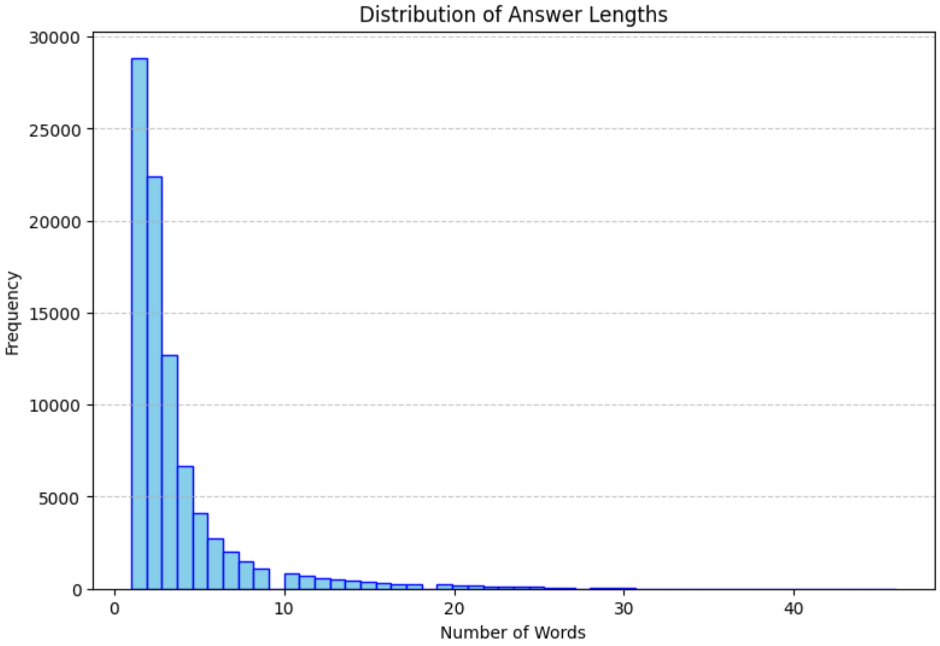
A. Github Repo: https://github.com/tmcgirr/SQUAD-BERT-chatbot-AAI/

B. StreamLit App: https://squad-bert-chatbot-aai.streamlit.app/

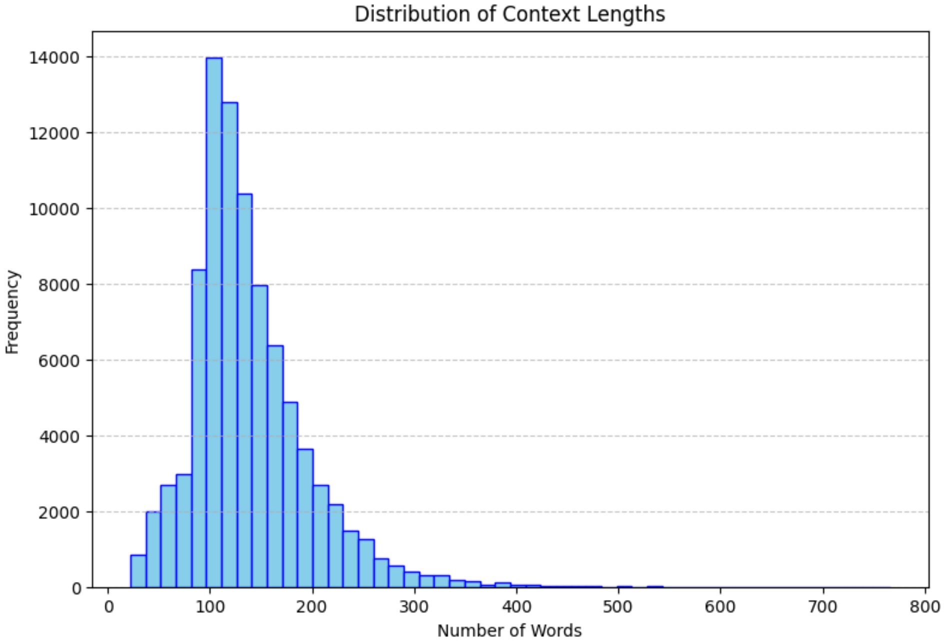
C. Hugging Face deployment: https://huggingface.co/tmcgirr/BERT-squad-chatbot-AAI/

D.

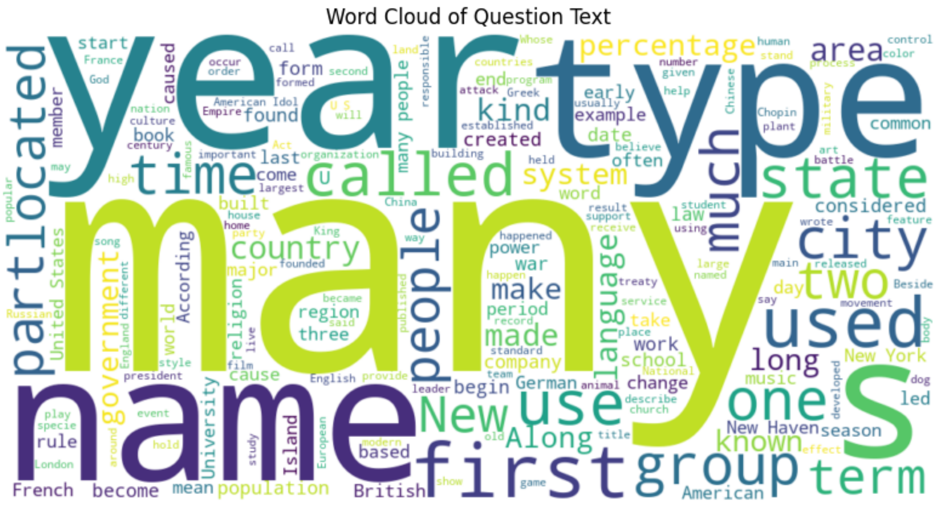



Distribution of Answer Lengths
Frequency
Number of Words
0
5000
10000
15000
20000
25000
30000
0
10
20
30
40

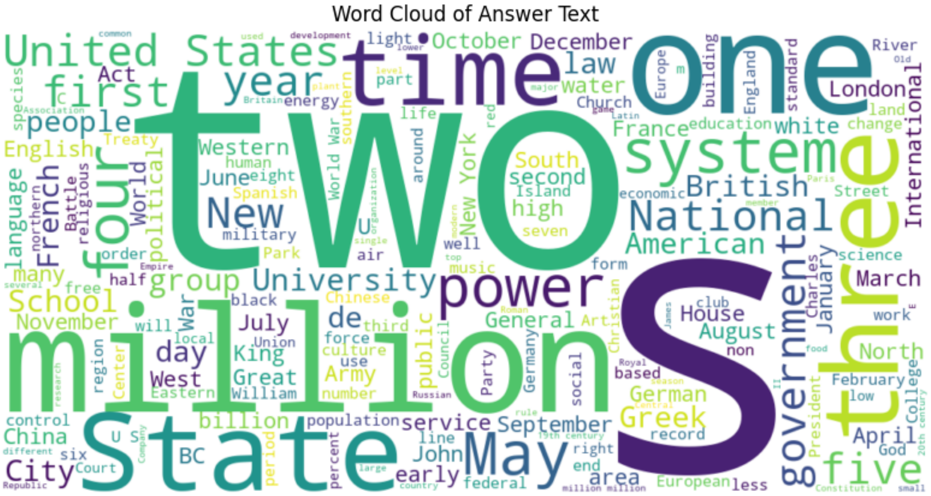
E.


Distribution of Context Lengths
Frequency
Number of Words
0
2000
4000
6000
8000
10000
12000
14000
0
100
200
300
400
500
600
700
800

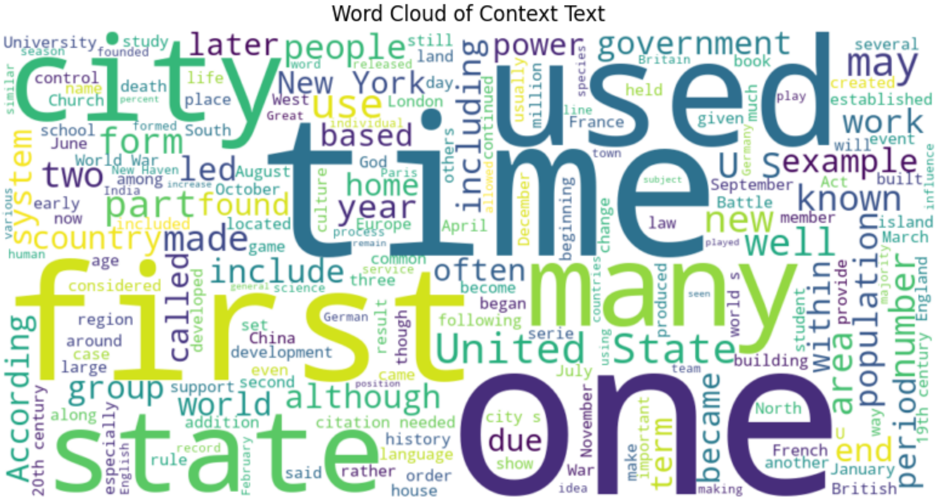
F.


Word Cloud of Question Text
many
year
type
name
s
located
part
time
called
state
city
used
two
first
group
term
one
use
New
people
much
language
government
country
area
percentage
kind
system
made
make
long
known
United States
University
Island
New Haven
season
start
example
early
created
date
power
war
period
region
three
company
school
change
begin
German
based
British
American
French
become
mean
population
rule
religion
According
world
member
caused
book
last
come
found
form
common
considered
law
day
New York
led
music

G.


Word Cloud of Answer Text
two
s
one
million
State
time
three
United States
first
year
power
government
May
system
National
New
four
five
people
American
British
University
group
School
English
French
language
political
World
South
second
high
Western
June
day
July
King
West
Great
William
Army
billion
population
service
September
Greek
German
General
House
August
North
January
February
April
March
October
December
November
law
water
London
Church
France
white
International
City
China
BC
John
early
federal
area
European
less
record
power
Act
Treaty
de
War
Party
social

H.


Word Cloud of Context Text
time
used
one
first
many
city
state
United State
people
later
power
government
may
New York
use
form
based
home
year
part
found
country
made
include
called
including
often
two
led
system
According
group
world
although
due
became
term
end
number
period
population
within
area
known
example
U S
new
well
work
established
several
book
school
June
World War
August
October
region
around
large
state
history
language
order
house
French
another
January
British
North

I.