NORTH SOUTH UNIVERSITY

Ask Anything About NSU:

A RAG Model Based Question Answering Platform Based On North South University Knowledge-Base

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL AND COMPUTER ENGINEERING
OF NORTH SOUTH UNIVERSITY
IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

CSE499B, SUMMER 2022
SENIOR DESIGN PROJECT

Declaration

It is hereby acknowledged that,

- No illegitimate procedure has been practised during the preparation of this document.
- This document does not contain any previously published content without proper citation.
- This document represents our own accomplishment while being undergraduate students in the North South University.

Sincerely,

Shahriar Khan
1821133042

Ahmed Al Jawad
1821526042

Sabbir Ahmed Sozol
1821926042

Hasibur Rahman Ridoy
1812046042

Approval

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Dr. Mohammad Ashrafuzzaman Khan
Associate Professor
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation.

Dr. Rajesh Palit
Professor & Chair
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh

Table of contents

List Of Figures:

List Of Tables:

Abstract

Training computers to comprehend human writing and speaking is referred to as natural language processing. Teaching computers for knowledge intensive tasks is still a challenge for AI researchers.There has been significant success in question-answering tasks for models that use parametric knowledge. Since large pretrained models store factual knowledge in their parameters, their ability to manipulate and access precise information is very limited. Retrieval-Augmented Generation (RAG) performs better in these cases since the model uses both parametric memory and non-parametric memory. The use of non-parametric memory allows the model to use updated documents for context. A seq2seq generator then generates an answer using the question and context. Using the North South University dataset as an external knowledge base for the RAG model, we simulated various scenarios and extrapolated the outcomes. Our research finds the performance and limitations of RAG model for a specific domain through these outcomes.

CHAPTER 1
Introduction

Introduction

Humans are social creatures and to function normally, they must communicate with one another. Speech is the best form of communication. There is a lot of information to remember relating to a university. Whether it is looking for emails, course-related data, or the location of a faculty's office space, it takes a lot of effort to find them.To ease this task of communication, researchers have trained computers to comprehend human writing and speaking that is referred to as natural language processing(NLP). One of the most challenging NLP problems is Question-Answering, but lately there has been a significant breakthrough in NLP. A user can ask a query in natural language to the system and get an instant answer. In today's world, Question-Answering systems are found in websites, these systems are capable of responding to little information. For more challenging queries they can not perform properly. To find the solution to these problems researchers have introduced the Retrieval-Augmented Generation(RAG) model. This external knowledge base RAG model is performing very well in case of question answering. By analysing its performance and attempting to identify the hurdles that prevent the model from performing at its best. North South University dataset is used as an external knowledge base for the RAG model, it helped to simulate various scenarios and extrapolated the outcomes. The research finds the performance and limitations of RAG model for a specific domain through these outcomes. If anyone could overcome the limitation then it will be very helpful for mankind and it will automate the question answering or information retrieving process to the next level.

1.1 Question-Answering platform

Information is a very crucial element in everyday life. To ease the access of information various tools are used. Chatbots or question answering platforms are one such tool that can help answer a query without the direct contact of humans on the other end. With the help of such a platform basic information can be retrieved saving time. This type of platform is proposed to use in the project to get answers for North South University related questions. The university has a website rich in information. The chatbot can answer the queries of the user using the information from the website. The question-answering platform can be used to get answers for the queries that can consist of basic information regarding courses, fees, faculty, programs and other facilities provided by the university.

1.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a model developed to combine the power of parametric knowledge with the flexibility of non-parametric external knowledge base [1]. The working mechanism of the RAG model might make it look like a standard seq2seq model. But the use of an intermediary retriever differentiates the performance of the model. The model uses an external knowledge base(kb) that works as the non-parametric memory. The kb is accessed using a retriever which is pre-trained [2]. The retriever uses a similarity search to find the documents relevant to the query [2]. Then the seq2seq model generates a prediction[1]. Thus the combination of parametric and non-parametric memory helps the RAG model to generate a prediction that is much more accurate than other models.

1.3 Purpose Of the Project

Facebook introduced a powerful NLP model and claimed that it provides state of the art performance for retrieving unstructured text documents. Among all the NLP tasks we have focused on the open-domain QA area. We've made an attempt to study its boundaries. In this project, for research purposes we have collected valuable information about North South University to implement this as an external knowledge base for the RAG. Through the research we were aimed to elaborate the current obstacles for retrieving domain specific information and propose a few solutions to the problems.If we can solve the problems then RAG can be used to retrieve exact information very easily. In future everyone will keep in mind the limitation and will surely make the best use of this model.Currently no other NLP model is able to provide accurate results like RAG.so it will be beneficial for people who will implement RAG in their question answering system.

1.4 Goals Of The Project

- determine the title's dependence.
- capability of the model for finding answers.
- find the constraints.
- Measure of information retrieval accuracy.

CHAPTER 2
RELATED WORKS

Related Works

In recent years, open-domain textual question answering (QA), which aims to respond to inquiries from vast data sources like Wikipedia or the web, has attracted the attention of a lot of researchers. Deep learning techniques, particularly those for reading comprehension and information retrieval using neural networks, have made significant improvements recently, which has allowed models to continuously update their state-of-the-art performances. Hence, this has led to further advancements in open-domain textual QA.

2.1 RAG Architecture

The Facebook AI Researchers conducted a study on Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. For language generation, they introduce the RAG Model, which combines pre-trained parametric and non-parametric memory. They use a pre-trained seq2seq model for parametric memory, and a dense vector index of Wikipedia for non-parametric memory, which is accessed with a pre-trained neural retriever.They compare two RAG formulations, one of which uses the same retrieved texts throughout the produced sequence, and another that can utilise various passages for each token. They fine-tune and assess the model on a variety of knowledge-intensive NLP tasks, as well as on three open domain QA tasks, where it outperforms parametric seq2seq models and task-specific retrieve-and-extract architectures. RAG models produce more precise, diversified, and factual language than a state-of-the-art parametric-only seq2seq baseline for language production tasks.In their experiments they found that RAG scores are within 4.3 percent of state-of-the-art models, which are complicated pipeline systems with domain-specific architecture and intensive engineering that are trained utilising intermediate retrieval supervision, which RAG does not require.

2.2 Realm Pre-Trained Architecture

The Realm(Retrieval-Augmented Language Model Pre-Training) architecture is a natural language processing model. The Realm architecture's main function is the Pre-training corpus. When the related question is asked the pre -training corpus “MASK” the unlabeled the question(text). At this time the model retrieves the document that matches with the questions with the help of vector function. Then the natural knowledge retrieves the textual knowledge from the right pointed document on the scores of the vector function.The knowledge-augmented encoder then compares the masked query and the recovered document, and the masked portion of the unlabeled text is replaced with the correct response from the obtained document. The generator then provides the correct response.

2.3 Proposed solution

Our approach is based upon the RAG architecture. The RAG architecture is a short form base question-answering model that was fine tuned on the Wikipedia website. The problem is finding out the faculty's email, office room number, phone number, or a description about a course, fees about the course, department info. All of these have to be found by manual findings. Our solution is that we have a dataset on the entire north south university and with the RAG architecture's question-answering model we can be able to provide the students of the north south university with all this information that is sometimes frustrating to search for and time consuming.

The infrastructure that we will be using for our auto-reply machine learning model is hay-stack, gradio for user interface and hugging-face for a different scalability for the project. The hay-stack pipeline is much faster and the documentation is very easy to read. We will be focusing our model building

around hay-stacks for its easy to use and scalability. We also want to venture around another pipeline which is the hugging-face. We create our auto reply bot in both of these pipelines and we will have 2 perspectives. The gradio user interface will be used for our machine learning model for fast testing and easy to have an interface for better understanding. Our goal is to venture through these pipelines and find out which one is working better for our suit.

CHAPTER 3
BACKGROUND & DESIGN OF THE SYSTEM

BACKGROUND & DESIGN OF THE SYSTEM

3.1 Retrieval Augmented Generation

Retrieval Augmented Generation is a relatively new NLP architecture which is capable of leveraging external documents to augment its knowledge and produce state-of-the-art results. RAG consists of an input encoder, an output decoder and an intermediate block consisting of a dense passage retriever, which allows it to leverage an external knowledge base. In our work, we have used the Haystack framework to implement the RAG architecture [4]. Haystack consists of models from both the Hugging Face and FARM model repositories. We have experimented with a few question answering generators including – the RAG seq2seq generators (facebook-token-nq and facebook-sequence-nq), Alberta, and a BERT that was fine-tuned for long-form question answering. Future work is set to slice GPT and integrate with the system.

3.2 Analysis of the design principles

RAG has an intermediary step that elevates its performance compared to other seq2seq based Question-Answering platforms. RAG consists of two main sections - the first section consisting of a set of encoders along with the DPR, and the latter section consisting of a seq2seq generator. The first section consists of input query encoder which encodes the query given as an input, as well as a document encoder that encodes external documents. It also consists of the Dense Passage Retriever which is responsible for performing the Maximum Inner Product Search(MIPS) and retrieving the top-K passages with the highest scores. For quick MIPS, DPR makes advantage of Facebook AI's Similarity Search.

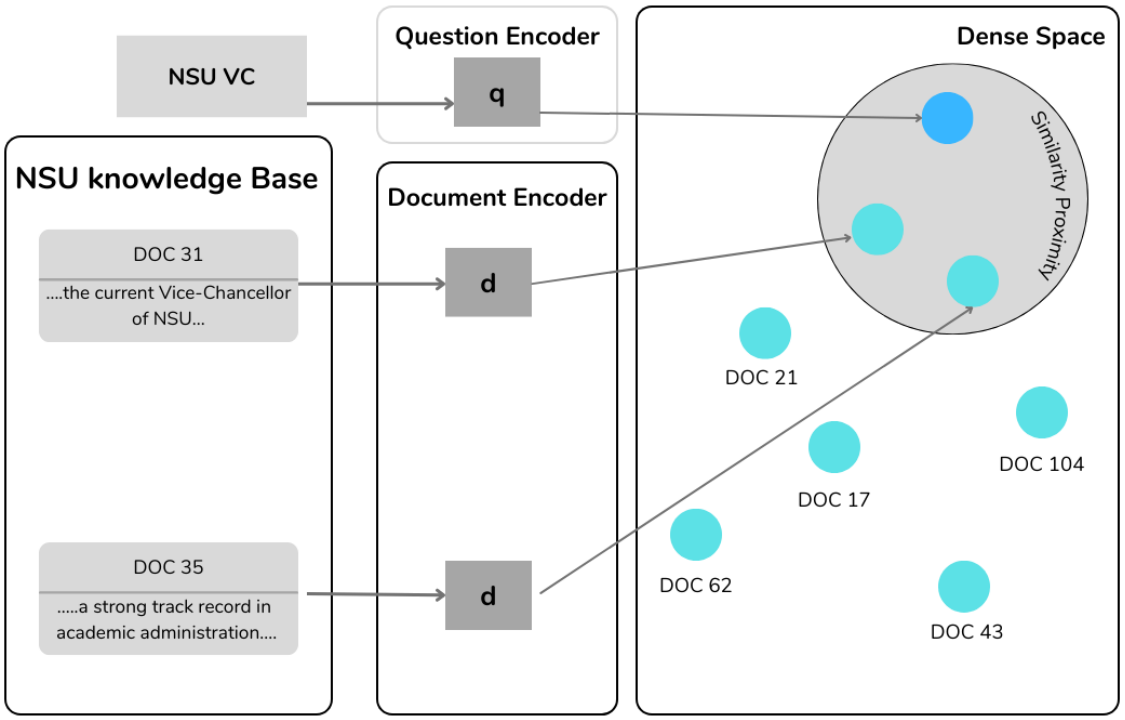

Figure : Dense Space Search of RAG model

Before feeding the question in a seq2seq model, RAG uses the document encoder to encode the documents or passages from the external knowledge base and store them in dense space. Upon being given a query, RAG then searches for a set of support documents from the dense space. These support documents are then concatenated and as a latent variable to generate the final output. Then the question and context is fed in a seq2seq model that produces the actual output.

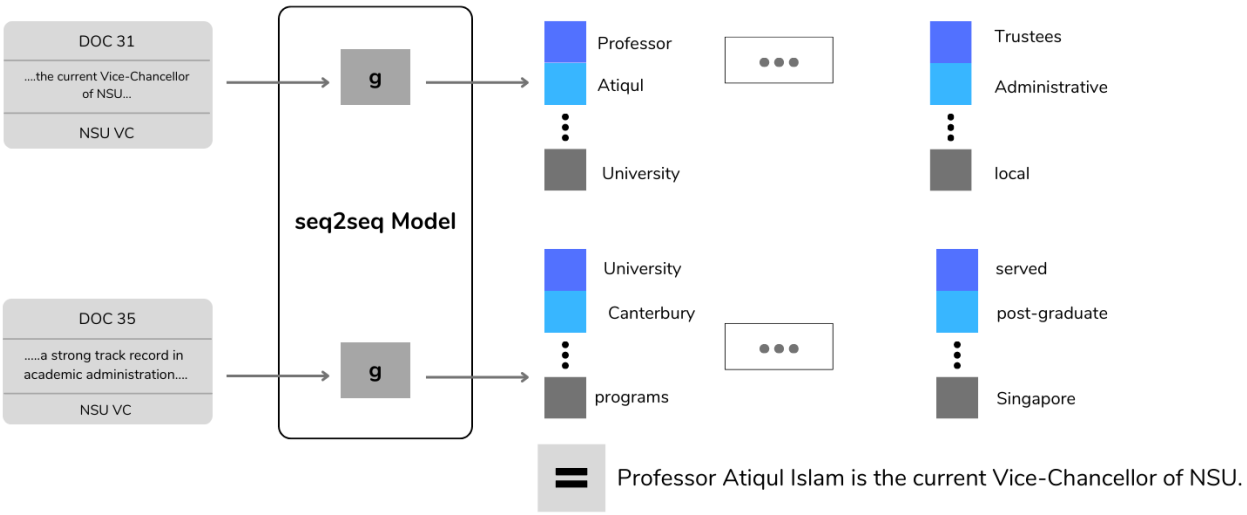

Figure : Response generation from seq2seq generator

3.3 Usability

Once the system is running, the user is presented with an interface, to interact with the question-answering system. The user interface of our system is significantly simple in design as well as easy to use. We just need to feed the knowledge base to the model and embed it as this will save time for there is training. This interface is user-friendly, it's quite confirmatory to the user to act according to their needs. This system is able to show error messages to the users and point them where it was made and helps them to fix it. If an error occurs, the UI points out the error in a pleasant format to the user. Due to the low consumption of power and the ability to run on CPU-only without suffering from too much delay, it stands as a very effective method of delivering answers to user queries.The functionality of the interface is quite effective and fast, the answers to the user’s queries are executed in approximately from a few milliseconds to a few seconds. It takes only a single click to switch between two desired generated outputs - either the RAG seq2seq generator or the GPT-3. Answers related to questions are given in a purposeful way as they are related. The DPR-pointed passage is also displayed to the user along with the generated output, if the option for viewing the passage is selected in the UI prompt radio button. It's designed to the edges with easy to read text, also easy to locate tools like navigation function and search for question-answering.

3.4 Manufacturability

This project was quite feasible and cost-effective. The necessary products that were needed for making the question-answering model were free and cheap of cost. For this project's planning and refining product prior to full-scale planning the cost estimation is overall quite less. Products which were used for building up this project include Google colab, RAG model, GPT-3, gradio(API), Open AI(haystack library). All of the mentioned products were free

and few were very cheap. The environment for building the project was very calm and made sure that to minimise dangers, lessen mishaps, and promote adherence to good manufacturing methods. Health measures were taken very seriously while creating the project, mental stress or any other physical problems were taken into consideration and adequate solutions were given. Overall safety, health and environment management were taken care of with greater adherence. This project was completely created in Google collab which is a cloud based system. It was free of any charge, all the components were installed in the cloud system and were used for fabricating the project. For this reason we didn’t need to have on premise hardware. For maintenance of this project the products that are needed are free and ready to use any time.

3.5 Sustainability

This project's support for the long run is quite feasible. RAG architecture is META’s new NLP models and its algorithms are very energy-efficient as these big tech companies are moving towards energy saving systems. This project's system requires less electricity as it has no external hardware for running except the computer browser that it's running on. The models that were used for making this project are RAG, GPT-3, gradio, google colab, haystack and hugging face. All of these require less money and hardware capacity requirement is as little as possible. Monitoring for this project is attainable for the long seeable future. The project was also hosted in a 100% renewable energy powered server. If we want to upgrade the project's dataset we just only need to change its knowledge base and embed it into the model. There is no dataset changing, training or testing required for the model. This is cost-effective as there needs to be less work to be done and less time to consume for maintaining and upgrading the system. Also this project's interaction with the economy is very feasible, a small amount of money is required for sustaining the project as all of the models are cheap.

Chapter 4

Dataset

Dataset

The dataset used as the external knowledge base has been generated by using web scraping techniques. The python library *BeautifulSoup* was used to scrape HTML documents from the North South University official website. A total of 676 web pages from the website have been used to create the dataset. It has been formatted into SQuAD format [3], with a title and its corresponding text column. We have collected dataset in three different labelling format,

- Manually Titled(Human)
- Auto-Generated Title(AI)
- Default Titled (Given by North South University)

Firstly we used default titles to evaluate the results and then for research purposes we prepared another two dataset with auto labelled title and manually labelled title. Using these three dataset we evaluated the differences and performance of the model for information retrieving.

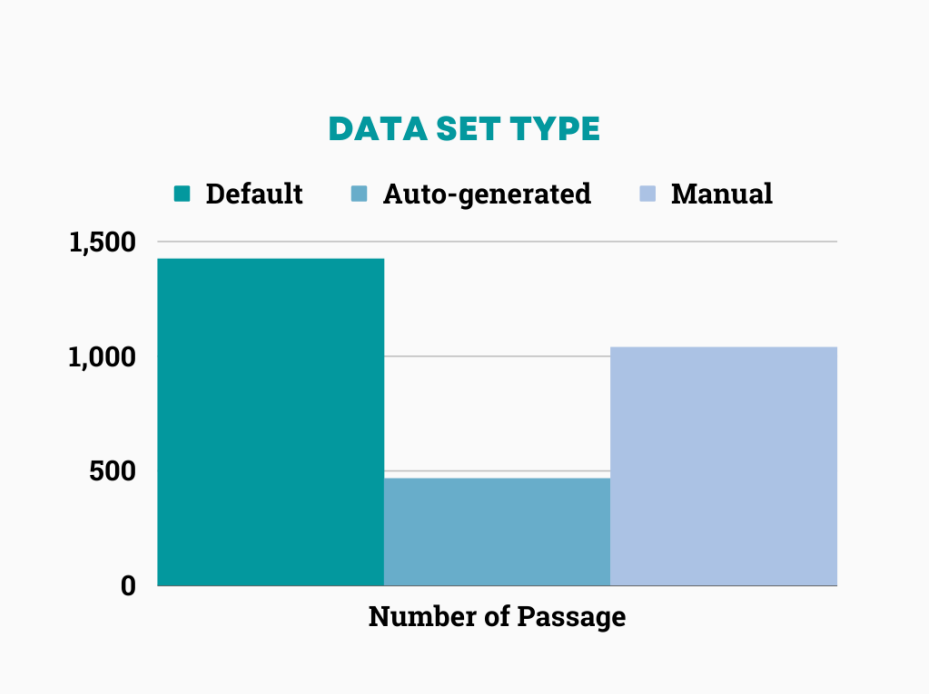

Figure: Passage number for different datasets

Default Labelled Dataset : The first kind of dataset used in this research is the default dataset that is collected using scraping of the North South University website. For the usage of external knowledge, we need a dataset that contains two columns namely title and text. The text is the passage that RAG will use as context and the title is the keyword for passage. In the default labelled dataset, our title was manually retrieved from the code of the website. These titles were not always accurate corresponding to the content of the webpage. Moreover, when a page was updated the title was not updated and for that reason we can see passages with titles that do not make sense. The use of the same code sometimes creates the same title for web pages that contain different information.

Auto-generated Labelled Dataset: Due to irrelevant results from default labelled knowledge base we experimented with the performance of our model using auto-generated labelled dataset. For this, we used a T5 base model that has been trained on a collection of 500k articles with headings. It can produce a one-line heading suitable for the given article. With the help of an inference API of hugging face, we generated a title for every passage of the default dataset. By this, we get a title for all the passages. The accuracy of the title depends on the accuracy of the headline generation of the t5 model. The titles were not fixed or updated post generation. As a result, the titles are completely independent of human interpretation. The titles and corresponding passages are then converted into a CSV file format to use as the knowledge base for our model.

Manually Labelled dataset: After exploration of default and auto-labelling we followed another approach for more accurate retrieval of the passage. We experimented with the model and tried to investigate how it works with manually labelled dataset as a knowledge base.It was not convenient to label manually. Instead of web scraping the website data directly into a CSV(comma separated values) file format, it was individually scraped so that

the content of the web pages is converted into a text file format. To manually label the data, we summarised each text file to get the gist of the content. Afterwards, an appropriate file name was designated and the text file was renamed. After manually labelling the file name of every text file, we split them into passages of 100 tokens. And for each passage, the title was taken from its corresponding file name. By this, a labelled dataset was created that contains the best possible title corresponding to the passages. We had to go through every collected passage by reading them and from the understandings and the brief idea of what the passage is about, we have decided a title for the passage. Though Manually Labelled dataset provide better results but there are multiple constraints can be seen to apply manually labelled methods which make the model less efficient. Noticeable constraints are,

- Inconsistency in data entry, room for errors, miskeying information.
- Large ongoing staff training cost.
- Labelling quality is dependent on good individuals.
- Time consuming.
- Lack of security.
- Duplication of data entry.

CHAPTER 5
Implementation of the system

Implementation of the System

The paper introduces steps to implement the RAG based question answering platform. The project is divided into various steps to complete the objective one by one.

The first step was to identify the scope of the project. In order to identify the scope and usability of the model in the project, previous works and papers related to the model were studied and the key themes were marked. Once the key points were found, the next step was to collect data to create a dataset.

The next step in this project was quite challenging. This step consists of collecting passages regarding North South University from the website. Using web scraping the passages were collected. In order to get precise information, the default title of the passages were not enough. As a result, two different approaches were taken in order to create a correct title for the passage. One of them was to get a title from an NLP model e.g. title generators and another was to read every passage manually and give a just title accordingly.

In the following step, we tried the RAG model from hugging face and built a function to try out Facebook AI Similarity Search (FAISS) and how it helps RAG's dense passage retriever (DPR) to get the similarity of the passages/documents and questions. Then using the RAG model from hugging face we created embeddings for our passages and performed answer generation from query. But the response was slow.

Next, we tried to find a faster way to get the answer. Using an open source framework namely Haystack, we found a faster response time for RAG models. Using different custom functions to generate a clean answer, a question answering platform was created from this that generates using the generator of RAG.

But the response from the RAG generator was sometimes underwhelming and less human like. For this, various other generator models were explored and Generative Pre-trained Transformer (GPT) was decided as the alternative. GPT is a state of the art seq2seq generator that produces human-like answers.

To get the answer from GPT, we need to pass the context of the question. So, in this step we tried various approaches to get the best passage from DPR and pass it as context for GPT.

The new GPT generator was integrated with the RAG model. Lastly, the whole system was deployed and a user interface was created to interact with the system.

CHAPTER 6
Process of the Development

Process of the Development

At the initial phase of the development, several studies were carried out on recent approaches for question answering systems. The research paper for Retrieval Augmented Generation as well as Dense Passage Retrieval for Open Domain Question-Answering were studied to grasp the underlying architecture of how these models work. Next, we carried out some research on SQuAD datasets and brainstormed the process for the creation of the university dataset, and in due course, performed scraping on the official North South University website. After performing text-preprocessing and word chunking, the initial dataset was prepared, with the titles kept the same as that of their corresponding HTML title tags. In the next phase, the design for our system

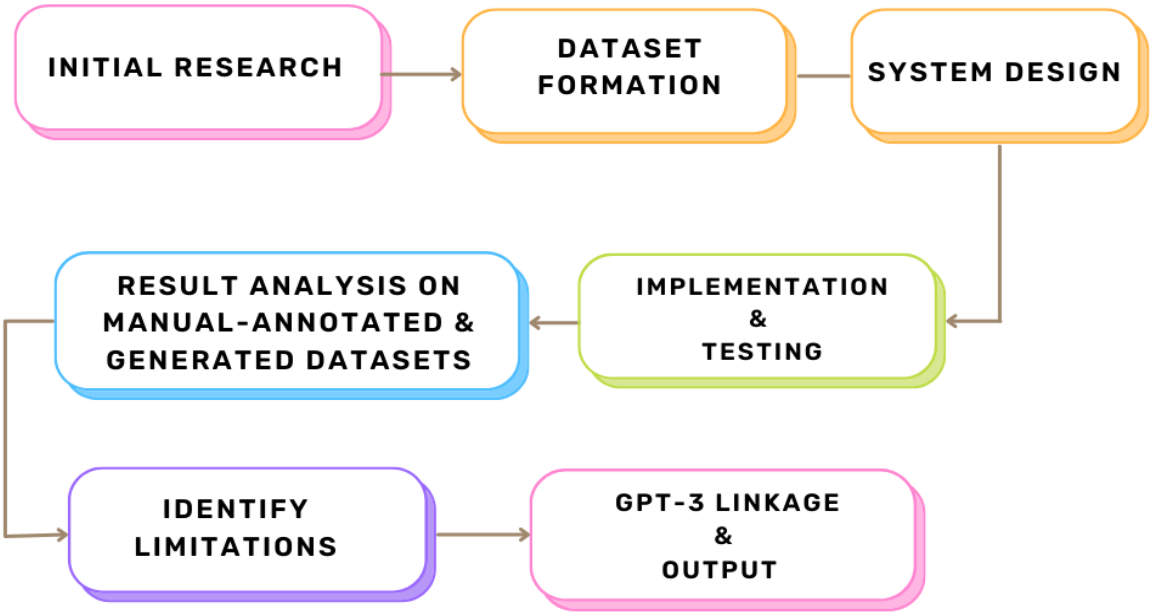

Figure: workflow of the development process

was finalised and Haystack framework was utilised to lay the foundation for the RAG architecture. With our initial dataset, exploratory results were analysed to understand capabilities and constraints. Subsequently due to title-dependency, another two datasets were formed, one with manual hand-crafted title annotations based on the context of the passage, and the other was generated with a title-generating deep learning model - a T5 based model which was trained to perform title generation on newspaper articles.

A set of 100 discrete query questions were prepared to serve as an evaluation metric for the question-answering capability. As we used pre-trained RAG models without any prior fine-tuning, the resulting answers from the RAG generator were short answers, as they were trained on a short-answer version of Google's Natural Questions dataset. To combat this issue and to generate a more smooth, generated response, we decided to connect GPT-3 with RAG. Using OpenAI's APIs, GPT-3, specifically Da-Vinci, was paired with our system. DPR embeddings were converted to textual format and passed directly to Da-Vinci as a string, to perform reading comprehension, and the results are fetched and marked as the final output of our entire question-answering system.

CHAPTER 7

Economical,Social,Political and Health Impact

Economical, Social, Political and Health Impact

Question-Answering in Natural Language Processing is an important area of interaction between Human and Computer because they provide a unique way for the people to get an instant answer to their query. Question-Answering platforms are potentially powerful tools to provide relevant information at any time to customers, that is the reason for building them frequently. It is easier to ask about something to the computer than any human being who is supposed to be the representative of that particular service. On the other hand, this technology of Question-Answering may be developed broadly for the purpose of Economical, Social, Political and Health consideration in upcoming days.

7.1 Economical Impact:

Question-Answering represents chances for good economical influence. Quality assurance may make essential services more accessible, available, and inexpensive. All questions that were previously asked in person to a representative may now be answered through the platform, lowering the expense of transportation as well as individuals' time. This project is based on a specific University Data, so students, parents, and individuals who are interested and have inquiries but are unable to come and acquire the facts in person may get trustworthy results to their inquiries.

At the same time it is beneficial for the institution or corporation because they can reduce their reliance on representatives and make it more accessible to individuals. Also, It will be financially beneficial to the company because hiring someone is more expensive than developing a Question-Answering software.

7.2 Social Impact:

This is a technology that has the ability to enhance the frontline feedback, services and social issues. It will affect many aspects of our existence. Gathering information used to be a challenging chore, but it is becoming easier by the day. This initiative will make the work easier than ever before, therefore improving the way knowledge is retrieved in society. If a platform like this one is applied anywhere people generally seek services, everyone in society will be better educated on the requirements and complexities of that specific service, which would eventually decrease stress. Artificial intelligence has always been only focused on enhancing social interactions.

As we concentrated on the North South University information in this study, a huge number of students benefited from the knowledge. If it is to be used in other services, it simply has to alter the knowledge-base and it is ready to act as a Question-Answering Platform for that specific service.

7.3 Political Impact:

This project has been looked up to from the neutral point of view. There has been no sort of hampering with the data that had been collected. From the start of the project this had been made sure. Data that was collected from the organisation that we intend to make our question-answering models knowledge base was not hindered in any sort of way. A neutral point of consideration was made sure from the start. The scraped data was handled with prior caution and was checked several times from an unbiased position. Nonetheless if the data that was collected from the organisation already had a political issue there is no means from our project point of view as we are not the rightful owner of the organisations data. The question-answering model is built in a way that it provides the correct answer and also gives answers from a neutral point of view.

7.4 Health Impact:

Mental satisfaction is requisite for everyone's lives.This projects can benefit its users by providing them with correct information about North South University. A university student or their parents or any person needs information about a matter of their interest related to the university. They don't need to contact the North South University for that matter; they can just ask our question-answer model for the right details. As a matter of fact these will help the students or their related ones figuring out the facts easily and not waiting several minutes for the right details.

Also another issue is reduced by a far means that is the reduction of travel stress. Nowadays travelling around the city costs much and the level traffic jams in the roads are getting worse each day, not to mention the weather also. Any queries that need could be done by at home our question-answering model can provide the user with useful details.

CHAPTER 8
Environmental consideration and sustainability

Environmental Consideration & Sustainability

As this project is a software based product there are few environmental considerations to think about. There is no need for training for the model as we have used pre-trained models. Only the knowledge base which is collected from the organisation needs to be embedded in the model. The question-answering model is only required to run once, and after the document indexing is done, it is ready for answering queries. As a result the net power consumption is significantly less. This project can be instantiated with CPU only for lesser power consumption, as GPU runtime requires more energy than CPU runtime. Furthermore, since this entire system works online, users do not need to travel to their organisation just to gain information regarding their queries, as a result, the cost of travelling and fuel consumption is saved. It also prevents additional exhaustion of human time and effort. Using pretrained models also saves training costs. The products that were used for making the question-answering model are very cheap. This project's models included RAG, GPT-3, Gradio, Google Colab, Haystack, and Hugging Face. All of these are less expensive except for GPT-3 integration, and require the least amount of gear. For the foreseeable future, monitoring for this project is feasible. The project was also housed on a server that was entirely run on renewable energy. The project's dataset can be upgraded by simply altering the knowledge base and embedding it into the model. Because less time and labour must be invested in maintaining and updating the system, this is cost-effective. The project's relationship to the economy is also highly realistic; because all of the models are inexpensive, only a small sum of money is needed to support it. Making the sustainability of this project very feasible in the future.

CHAPTER 9

Ethical & Professional Responsibility

Ethical & Professional Responsibility

The textual material was obtained from the website of North South University. Because the method of collection was manual labelling, it is important to remember that competent humans should be used to develop these annotations unless it would operate differently. Misuse of information or irrelevant labelling must be avoided, otherwise people may be misled by disinformation. It is a critical component of the project. These things should be done carefully, and the labelling person should execute his or her job extremely properly unless he or she may hurt the individuals who will benefit.

During the datasets creation process, it was ensured that only one individual performed this work so that the quality of labelling remained consistent. It was meticulously maintained and double-checked that no information was misleading or untrue. The data was not tampered with, and this information was not obtained from any other third-party website save the North South University website. We were truthful, and the restrictions were acknowledged based on the outcomes.

CHAPTER 10

Tools and Technologies Used

Tools and Technologies used

Python, Gradio (for interface and deployment of the project), and Google Collab would be used to implement the project. Making the system a web application ensures an easier solution for the large range of devices because the system will be created using Gradio api silent-server. To use this service, only a computer or other device with an active internet connection and a web browser would be needed. To implement the client side of the application, Gradio would be used. To offer a seamless experience across most devices, the user interface will be responsive. Other tools and technologies that were used are given below.

- For the preparation of the dataset, BeautifulSoup was used as the primary tool used for data scraping.
- The development was carried out on Google Colab for better computational resources.
- For the implementation of the RAG architecture, we have used the Haystack framework. Haystack allowed us to rapidly build and perform experiments due to its available pipelines. Each of the RAG components were designed with Haystack.
- All the pre-trained models used in our setup were taken from the Hugging Face model repository and FARM repository.
- FAISS was used for performing dataset-indexing for the external knowledge base.
- For the generation with GPT-3, OpenAI's APIs were used.
- Lastly, for comparison and result visualisation between RAG generator and GPT-3 based generation.

By making such a reliable system available at a low cost, many organisations would be encouraged to use it, which would not only secure a sustainable

environment but also benefit many people who are in need of accurate information on the information they are looking for for their linked issue.

CHAPTER 11

Result Analysis

Result Analysis

Due to the constraints of title dependency, we used manually labelled dataset as knowledge base and found interesting outputs for different scenarios. The RAG performs better when the question is precise and has “clues” for the model to “cue” the candidate documents from dense space. Also we found that, since the model is pre-trained on wikipedia dataset, the output accuracy also depends on the format of the passages/documents.

RAG performs well when we can minimise these constraints. In our research using manually labelled dataset, we used two different generators: the default RAG generator which consists of BART model, and the state of the art GPT model. For both generators, the passage retriever is the RAG’s dense passage retriever (DPR).

We compared the results of 100 queries for both generators along with the DPR. While the RAG’s generator performs well, the GPT clearly outperforms with the same passage as context. GPT also generates more human-like answers and provides long form answers. Although, the outputs from GPT may vary for different temperature parameters which controls the randomness and creativity of the output. The performance for the generators and DPR is represented in the bar chart.

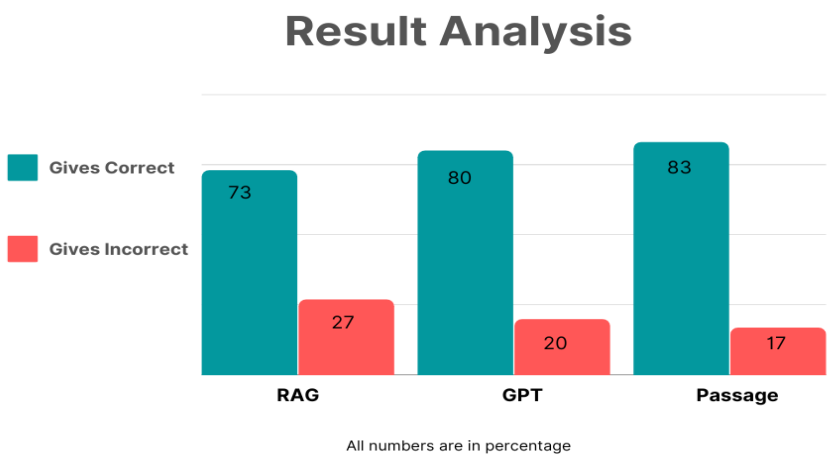

Figure: Bar graph of query response

11.1 Difference of result for different datasets

Our default labelled dataset was very weak in the context of association between title and text. The use of this dataset as knowledge base for our RAG model enabled us to gain the insight of the importance of title for accurate passage retrieval from the dense space.

The model performs well for auto-generated dataset since the titles were accurately generated for most of the passages. But, for passages that contained a broad range of information, the headline generator model generated a title that covers some information ignoring the others. As a result, for questions regarding those missing information, RAG tends to give non-factual answers from its parametric memory. This performance can be improved by using models for title generation that provides more accurate titles based on the semantic meaning.

Surprisingly among the other two approaches , manually annotated dataset produces the best result using the RAG model.For example against the same question of previous approaches, when manually labelled dataset is used as knowledge base, the model provides better and accurate answers.The correct results from this approach bolsters the fact that an accurate title can retrieve relevant passages precisely in the RAG model. The use of a correct title can increase the chance of using the correct context that can produce more accurate and factual answers.

Question	Answer(RAG)		
	Default	Auto-generated	Manual
Who is the chair of the ECE department at North South University?	Dr. Abdur Rob Khan	Dr. Mohammad Rezaul Bari	Dr. Mohammad Rezaul Bari
How long is the admission test in North South University?	About three hours	Two and a half hour	Two and a half hour
When was NSU established?	January 17, 2017	1992	February 1993

Table : Response of query for different datasets

11.2 Analysis on the limitations Of RAG model

On the contrary, our study identified a few limits of RAG design and discovered that if these restrictions are avoided, RAG can outperform any other model now available. If answers to the limitations are to be discovered, the questions' state of the art outcomes should be expected. We made every effort to uncover answers, and in our experiment, we used three datasets techniques to do so.

Dependency on “Title”:

From our experiment with three different labelled dataset (default, auto-generated, manual), we find different results. This majority of the variance in result occurs due to the accuracy of titles. And this shows the dependency of the model on titles. The model performs well when the title is relevant to the content of passages. If the title is not relevant or lacks information regarding the passage we see a non-factual answer. It is due to the fact that, when title is not accurate then DPR does not retrieve the correct candidate passage. This is an inconvenience for domain specific tasks. Since, the title needs to be fixed according to the passage, if done manually then the accuracy of the title depends on the domain specific knowledge of the person. If we use a deep learning model to generate titles for the passages, it can produce some sort of accurate title. But the accuracy will largely depend on the training of that model. If the model is not pre-trained or fine-tuned on specific domain specific data then it might not be able to generate accurate title for every passage. This can hamper the performance of the RAG model since RAG depends on the title for better retrieval of passage.

In this research, we found that the manually labelled dataset, when used as a knowledge base, is best for accurate answers. The manually labelled dataset requires domain specific knowledge. In our case, since the dataset is scraped from our university website we have knowledge regarding the content of the passages. Although the auto-generated dataset that contains title generated by a T5 model does also provide accurate retrieval of the passage, when faced with specific semantic meaning of the passage contents, manually labelled dataset clearly outperforms when used as knowledge base for RAG.

When we extrapolate the results from the three kinds of datasets used as a knowledge base, it is clear the retrieval of passages in the RAG model is dependent on the accuracy of title.

Precise query for accurate answers:

The RAG model uses a question encoder that uses the dense passage space to retrieve the relevant documents. The similarity search between encoded question and encoded document determines the relevant document that can be used as context for our question. Through our research we found that the document retrieval works better when the question is more rigorous. The little details can improve the quality of encoding and give improved performance regarding document retrieval. From the following example we can see the difference of result for a question with slight difference.The difference in this result shows the importance of questions with precise words. Even little relevant words that can bolster the question's semantic meaning and so can bolster performance of the document retrieval process. The precise question also plays a key part for the answer generation of our seq2seq generator. If we look at another example we can see the evidence more clearly.

Question	Answer (RAG)
Where is North South University?	Bangladesh.
Where is North South University located in dhaka?	Bashundhara.

Table: Response of query

Here the passage retrieved by the Dense Passage Retriever(DPR) is the same in both cases. However, the answer differs due to the question encoding. The word choice of the question creates a variety of answer generation for the RAG model. It is due to the fact that the seq2seq model of our RAG gets the encoded question and document as prompt for generation. In this example, even though the passage pointed out by DPR is the same, the word choice is different. So the choice of words for a question is a clear factor for accurate answers.

For domain specific question-answering this precise word play can be proved as a limitation for this model. Since we can assume every individual using the question-answering platform won't have domain specific knowledge and cannot use precise words or phrases every time. As a result, the model might give incorrect answers. Even though the answer might exist in the knowledge base of the model, the choice of words might enhance or reduce the performance of the model.

CHAPTER 12

Conclusion

Conclusion

Using different domain specific knowledge bases, we can create a question answering platform using RAG. The ability to retrieve candidate documents or passages to be used as context makes this model superior to other chatbots or question answering platforms. Even if we need to change or replace the dataset, we don't need to train the model again. We can just swap the knowledge base with an updated data set and get factual updated answers from the same model. And with a state of the art generator like GPT, the RAG model can give more human-like precise updated factual answers in a really short time.

In the future we want to explore the performance and limitations of the RAG model upon fine-tuning. This will need a SQuad to original DPR format conversion of the dataset for fine tuning. Also we are interested in comparing the performance of RAG on different domain specific knowledge bases where we might not need much pre-processing for the dataset.

References

1. *Retrieval augmented generation: Streamlining the creation of Intelligent Natural Language Processing Models*. Meta AI. (n.d.). Retrieved April 24, 2022, from https://ai.facebook.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/
2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.
3. Rajpurkar, Pranav, et al. "Know What You Don't Know: Unanswerable Questions for SQuAD." ArXiv:1806.03822 [Cs], June 2018. arXiv.org, http://arxiv.org/abs/1806.03822.
4. Pietsch, M., Soni, T., Chan, B., Möller, T., & Kostić, B. (2020). Haystack (Version 0.5.0). GitHub. https://github.com/deepset-ai/haystack/
5. Kelvin Guu , Kenton Lee , Zora Tung , Panupong Pasupat , Ming-Wei Chang , guu20a.pdf (mlr.press)