Project Work :

Final ISA(Review 4) / ESA 2020

Project Title : Question Answering System
Project ID : PW20MHR04
Project Guide : D.r Mamatha H R
Project Team : Pragya Agrawal - 01FB16ECS258
Priyanka A - 01FB16ECS278
Ranjitha nayak - 01FB16ECS298

Problem Statement

- The goal of the project is to produce a question answering system that generates answers from the knowledge base or text paragraphs for the questions posed as input.
- The answer to a question is always a part of the context and a direct answer is expected in response to a submitted query, rather than a set of references that may contain the answers.
- The system has also been extended to answer questions on selected topic without an input context.

User Profile

- Search engines that give answers back from a pool of websites (can be considered as documents).
- Chatbots trained to answer questions posed on predefined topics.
- Institution centric question answering system.
- Automated customer service for issues.

Research Papers

- **Title** : Closed Domain Keyword based Question Answering System for Legal Documents of IPC Sections & Indian Laws
    - http://www.ijircce.com/upload/2015/june/77_%20Closed.pdf
- **Author -** Shubhangi C. Tirpude, Dr. A.S. Alvi
- **Year of publication -** 2015
- **Methodology -** Uses domain resource and concepts dictionary for extracting the keywords from the questions to generate and select the relevant answer for the Indian laws.
- **Dataset -** IPC sections and laws documents from various websites
- **Critics about the paper -** It returns a paragraph as an answer to the users' questions instead of returning exact answer.

Literature Survey

- **Title** : Question and Answering on SQuAD 2.0: BERT Is All You Need
  - https://web.stanford.edu/class/cs224n/reports/default/15812785.pdf
- **Author:** Sam Schwager, John Solitario
- **Year of publication -** 2019
- **Methodology-** Fine-tuning of the original BERT model by experimenting with learning rate, dropout, batch size, and training epochs. Weakening the loss function in an attempt to improve BERT. Several BERT models are built and certain models perform better on certain tasks.
- **Dataset -** SQuAD
- **Critics about the paper -** Leveraging an intelligent version of ensembling in the final model, which achieves an F1 score of 76.545 and an EM score of 73.609 on the test set.

Literature Survey

- **Title** : BERT for Question Answering on SQuAD 2.0
  - https://web.stanford.edu/class/cs224n/reports/default/15848021.pdf
- **Authors :** Yuwen Zhang, Zhaozhuo Xu
- **Year of publication :** 2019
- **Methodology :** By replacing the linear BERT output layer with an encoder-decoder architecture, authors successfully implemented the task-specific layers that can deal with the SQuAD 2.0 problems quite well
- **Dataset :** SQuAD 2.0
- **Critics about the paper :** Further improve the performance of the model by fine-tuning the parameters in each layer. There is still a gap in has-answer and no-answer accuracy which should be tried to compensate with ensemble on no-answer predictions.

Literature Survey

- **Title** : How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations
  - https://openreview.net/pdf?id=SygMXE2vAE
- **Author:** Anonymous
- **Year of publication -** 2019
- **Methodology -** The analysis focuses on models fine-tuned on the task of QA as a complex downstream task. Inspection on how QA models transform token vectors in order to find the correct answer. A set of general and QA-specific probing tasks are applied that reveal the information stored in each representation layer. The analysis of hidden state visualizations provides additional insights into BERT's reasoning process.
- **Dataset -** SQuAD, HotpotQA, bAbI
- **Critics about the paper -** The training analysis shows the varied transformation phases for correct and wrong predictions giving deeper insights into the layers of the model and the reasons for selection of a wrong candidate.

PES UNIVERSITY

Proposed Solution

- First step is Document Retrieval
    - to retrieve the correct document to look for answer given a question.
    - Use Latent Semantic Analysis for document retrieval.

- Second step is build a QA Model
    - Defining a new layer on top of existing BERT to fine tune it to make an improved question answering system.

The context for the built question answering system is “computer security”.

Why Your Solution is Better?

- BERT is a pre trained contextual model trained on a lot of wikipedia content and other articles.
- It generates a representation of each word that is based on the other words in the sentence, unlike the RNN and CNN models take in words sequentially, making it difficult for long term dependencies, with no context and makes difficult to achieve the power of parallel processing.
- Integrating BERT QA with IR makes it better and faster to answer questions.

Technologies / Methodologies

- BERT pre-trained model by Google
- NLP techniques
- Jupyter Notebook and Colaboratory
- Google Cloud Platform
- PyTorch
- Information retrieval techniques to find relevant documents given query.

Dependencies

Software Requirements:

- **Technologies:** PyTorch,flask
- **Tools:** GPU by google colab ,anaconda navigator.
- **Domain:** Deep learning, Natural Language Processing
- **Browser:** Works on all available browsers

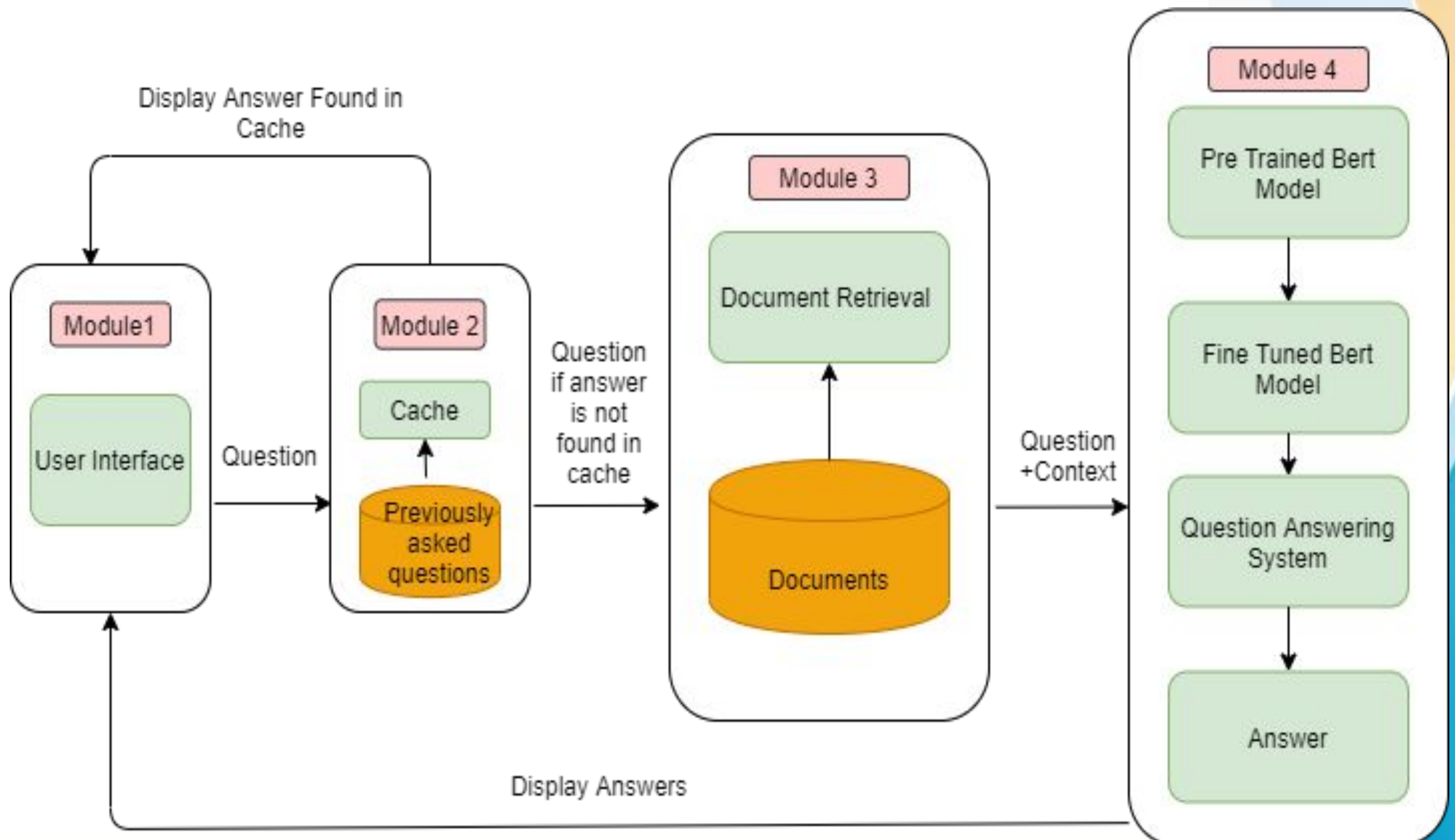
Hardware Requirement:

- Good Internet Connection.
- Enough disk space for document database.

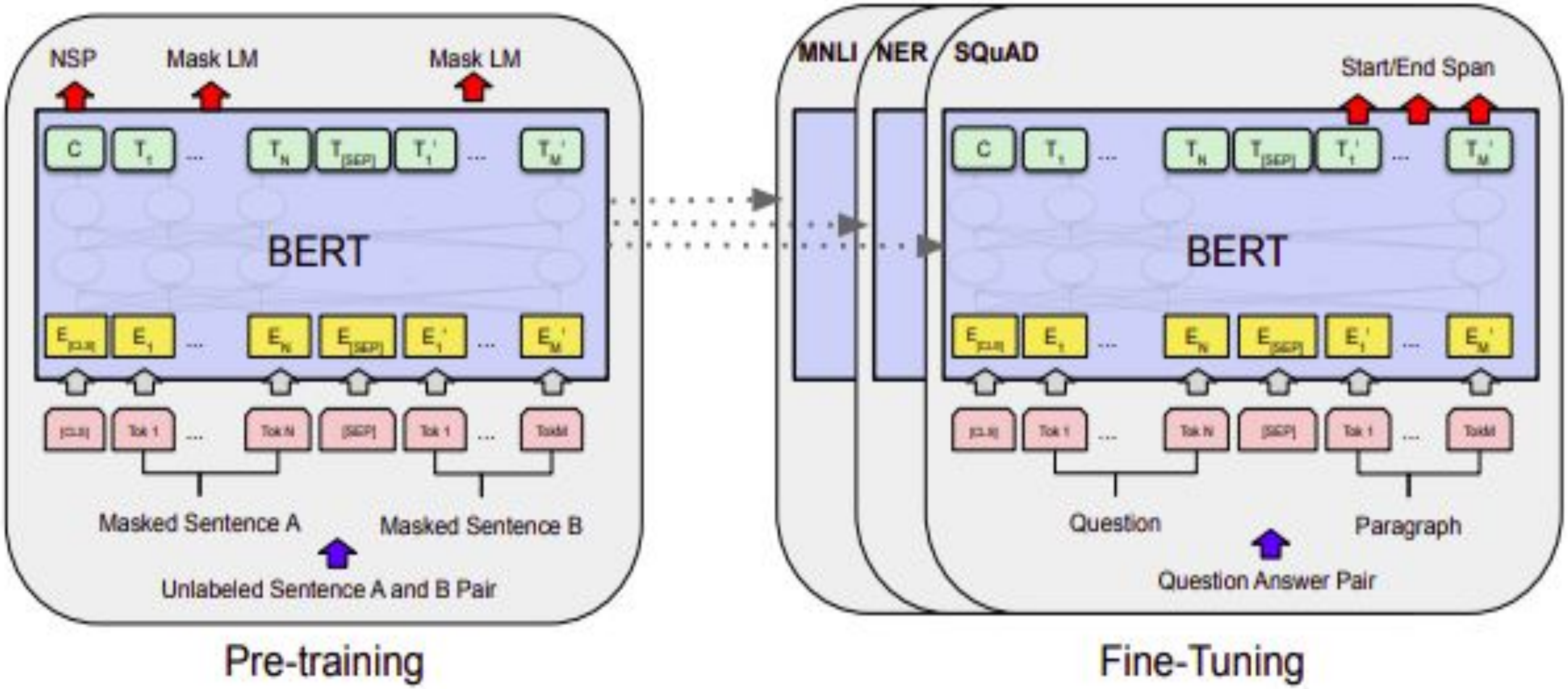
Risks

- Any question which is out of the domain will result in an unexpected answer
- The answers can arbitrary span within chosen document rather than a limited set of multiple choices
- The performance of the model might worsen when there is syntactic divergence between the question and the sentence containing the answer

System Architecture



Defining a new layer on top of existing BERT to fine tune it.



Modules

- Module 1 : User Interface
  - User can type in the question and submit to retrieve the answer.
  - Final answer and the context from which answer was picked will be displayed.

- Module 2: Cache
  - Stores the previously asked questions’ answers and context into database
  - Returns the answer to the UI when such questions are asked again saving the processing time.

Modules

- Module 3 : Document Retrieval
  - Latent Semantic Indexing.
  - Also tried TF-IDF followed by cosine similarity.

- Module 4 : QA model
  - Pretrained BERT model.
  - Fine tuning it to make a QA model using SQUAD.

Design Constraints, Assumptions & Dependencies

Design Constraints

- Training on domain specific corpus yields better performance than fine tuning BERT.
- BERT has a limitation on the sentence length.
- The higher the maximum length of the sentence, the more GPU memory or CPU time will be required
- Giving context along with the question limits the generic behavior of the system

Design Constraints, Assumptions & Dependencies

Assumptions

- The length of the sentence is limited to match with the memory available
- Fine tuning the BERT model is done with the help of domain specific data

Dependencies

- Context is required with the question which should contain the answer to the question.

Test Strategy

1. **Unit Testing:** It was done when a new feature, document retrieval, was added to the system. This was done before this unit was added to the system.
2. **Component Testing:** It was done to check the interaction between the document retrieval and the UI module, and also between module 2 and module 3.
3. **System Testing:** It was done to check the complete working of the system.

These strategies are used in order to follow a systemized testing pattern and not a have a big bang approach to it. Directly performing system testing may result in failures which could have been solved by just unit or component testing.

Test Cases

- Testcase 1 : Check interaction between UI and Document retrieval
- Testcase 2 : Check if proper document is picked given a query.
    - Tested with a lot of queries and it picks the right document 85 - 90% of the time.
    - wrote a python script which automatically does this.
- Testcase 3 : Check if the interaction between document retrieval and model.
- Testcase 4 : Check if expected result is returned.

Implementation Details

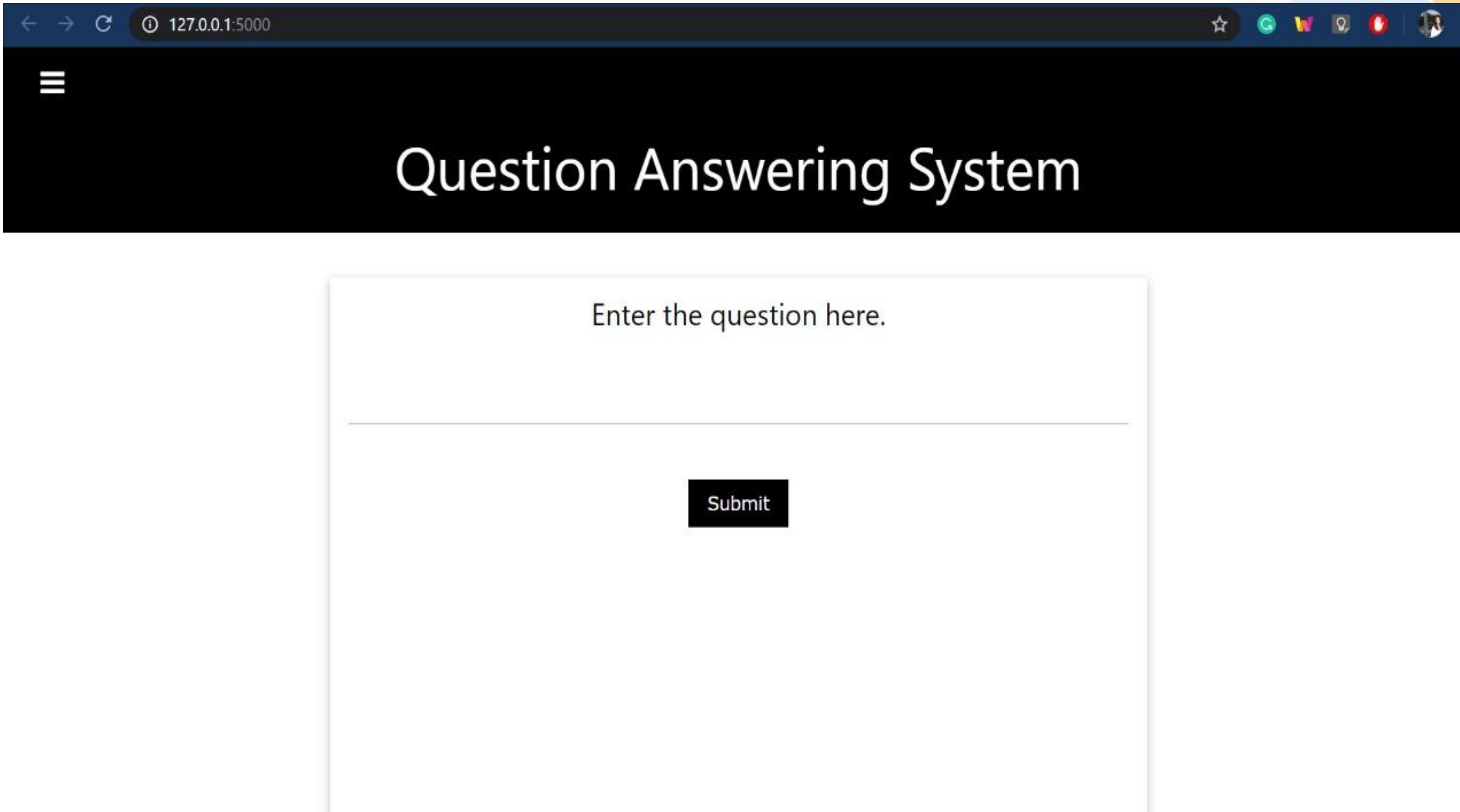
Module Name	Technologies Used	Lines of code	Algorithm used
User Interface	HTML, Javascript, Python Flask.	100	-
Document Retrieval	Python	180	LSI, TF-IDF, cosine similarity. 85 - 90% accuracy.
QA model	Pre trained BERT model in python	150	BERT QA provided functions.

Project Completion

Project Report Status – Ready

Project Demo – Completed till Review 3

Project Results - UI



Project Results - Query Submission

Question Answering System

Enter the question here.

What does ACL stand for?

Submit

Or try some previously asked questions.

What can be caused by neglecting a computer security system?

Submit

PES UNIVERSITY

Project Results - Document Retrieval

- Document retrieval - Picks the closest matched document when query is given by the user.
    - Question: What is the protection of information systems?
    - Document picked : **Computer security, also known as cybersecurity or IT security, is the protection of information systems** from theft or damage to the hardware, the software, and to the information on them, as well as from disruption or misdirection of the services they provide. It includes controlling physical access to the hardware, as well as protecting against harm that may come via network access, data and code injection, and due to malpractice by operators, whether intentional, accidental, or due to them being tricked into deviating from secure procedures.

Project Results - QA Model

- QA model given query and document picked , it parses through the document to find the answer. Answer should be in the context.

- **Question** : What is the protection of information systems?
- **Document picked : Computer security, also known as cybersecurity or IT security, is the protection of information systems** from theft or damage to the hardware, the software, and to the information on them, as well as from disruption or misdirection of the services they provide. It includes controlling physical access to the hardware, as well as protecting against harm that may come via network access, data and code injection, and due to malpractice by operators, whether intentional, accidental, or due to them being tricked into deviating from secure procedures.
- **Answer** : security

PES UNIVERSITY

Project Results - Previously Asked Questions

Question Answering System

Enter the question here.

what does ACL stand for ?

Submit

Or try some previously asked questions.

what does ACL stand for ?

- what does ACL stand for ?
- What can be caused by neglecting a computer security system?
- What does CCIRC stand for?
- Firewalls and exit procedures are considered what?
- What is computer security also known as?

Project Results - Answer Returned

Question Answering System

The answer is

access control lists

It is chosen from the context

Within computer systems, two of many security models capable of enforcing privilege separation are access control lists (ACLs) and capability-based security. Using ACLs to confine programs has been proven to be insecure in many situations, such as if the host computer can be tricked into indirectly allowing restricted file access, an issue known as the confused deputy problem. It has also been shown that the promise of ACLs of giving access to an object to only one person can never be guaranteed in practice. Both of these problems are resolved by capabilities. This does not mean practical flaws exist in all ACL-based systems, but only that the designers of certain utilities must take responsibility to ensure that they do not introduce flaws.[citation needed]

Planned Effort Vs Actual Effort

Phases	Planned Effort	Actual Effort	Reason for Deviation
Literature Survey	3 weeks	3 weeks	-
Data Collection	1 week	2 weeks	More time was spent analyzing the data and trying to find domain-specific dataset.
Modelling and Production	1 ½ months	2 months	Encountered various errors while installing the required libraries, network errors during training of the model
Analysis and Testing	4 days	2 days	-
Report Writing	1 week	4 days	-

Lessons Learnt

- Importance of analyzing and cleaning the data.
- How Attention mechanism is better than basic RNN model.
- BERT model's functioning and its various applications in the space of Natural Language Processing.
- Understanding various algorithms for document retrieval with best match with a given question and deciding on a best one.
- Understanding and implementing the prediction algorithm.

References

- BERT Research - Ep. 1 - Key Concepts & Sources, Chris McCormick, (2019, Nov 11)
  - https://mccormickml.com/2019/11/11/bert-research-ep-1-key-concepts-and-sources/
- Question Answering with BERT, XLNET, XLM, and DistilBERT using Simple Transformers , Thilina Rajapakse ,(2019 , Nov 8)
  - https://towardsdatascience.com/question-answering-with-bert-xlnet-xlm-and-distilbert-using-simple-transformers-4d8785ee762a
- Extending Google-BERT as Question and Answering model and Chatbot , Naga Kiran , (2019 , Aug 8)
  - https://medium.com/datadriveninvestor/extending-google-bert-as-question-and-answering-model-and-chatbot-e3e7b47b721a
- BERT Word Embeddings Tutorial , Chris McCormick and Nick Ryan , (14 May 2019)
  - https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/#2-input-formatting

Thank You