Closed Domain Question Answering, Text Summarization and Automatic Slide Generation Using Natural Language Processing

Guide: Asst. Prof. M. M. Swami

Group Members
17CO026 - Vishakha Kokate
17CO028 - Shrayesha Kukade
17CO030 - Kunchal Kulkarni
17CO044 - Akanksha Pawar

TABLE OF CONTENT

PROBLEM STATEMENT

The pandemic has forced schools to shift to the online classes. So, turning teaching material into digital format at a short notice has been a challenge. There is a need for a system that can generate presentations of the provided document. Students are facing difficulties in interacting with faculties virtually, so there is a need for question answering system, which will help students with precise answers to their questions.Also a system that will summarize any document in a short time for any revision or practice.

ABSTRACT

In this pandemic, the teaching-learning process has changed a lot. Due to online classes, teachers are now burdened with creating teaching aids on large scale. E-learning is promoted due to lockdown. The proposed system will help learners to resolve their queries using the question answering system. The system will also provide summary of the learning document provided by the learner. The proposed system will generate ppt for reducing the effort of the teaching faculty, so that the teachers can concentrate more on teaching.

GOAL & OBJECTIVE

- Reduce the burden of preparing e-learning resources for teachers.
- Help students in self-study and effective learning.

MOTIVATION

- The great amount of availability of the information in web, difficulty in finding precise answers.
- Time consumed in preparing learning materials like PPTs.
- Promoting self study.
- Thus, for minimizing the time and efforts of teaching faculty as well as learners in learning and preparing for exams.

LITERATURE SURVEY

YEAR	PAPER	AUTHOR	TOPIC	DESCRIPTION
2019	**IEEE paper**	**J.N.Madhuri Ganesh Kumar.R**	**Extractive Text Summarization using sentence Ranking.**	**In this paper, extractive text summarization approach is used which uses statistical features for selecting useful informative sentences.**
2015	**IJSRCSETI**	**M. Sravanthi, C. Ravindranath Chowdary**	**SlidesGen: Automatic Generation of Presentation Slides for a Technical Paper Using Summarization**	**This paper proposed a system where LATEX document is converted into XML and then PPT. This system took input only in LATEX form.**

YEAR	PAPER	AUTHOR	TOPIC	DESCRIPTION
2016	**International Journal of Engineering Science & Research Technology**	**Sweta P. Lende, M. M. Raghuwanshi**	**Closed domain Question answering system using NLP techniques**	**System is pre trained for educational acts dataset. The dataset contains 583 sections. Dataset of 300 different questions is maintained to train the QA system. Out of 100 questions asked, 76 were answered. Out of which 68 were accurate. Precision is 0.89 Recall 0.68**
2016	**Stanford University Paper**	**Yuwen Zhang, Zhaozhuo Xu**	**BERT for Question Answering on SQUAD 2.0**	**This paper proposed a model by using bert and fine - tuned it with additional task-specific layers to improve its performance on Stanford Question Answering Dataset (SQuAD 2.0). as base model and achieved F1 score of 77.96 on the dev set and 77.827 on SQUAD 2.0 dataset.**

YEAR	PAPER	AUTHOR	TOPIC	DESCRIPTION
2016	**International Journal of Innovative Research in Science, Engineering and Technology**	**Ektaa Meshram , D. A. Phalke**	**Technique for Generating Automatic Slides on the basis of Paper Structure Analysis**	**Tokenization, stemming, removal of stop words such processes are performed on input document. Part of speech tagging and score is evaluated using NLP. Images are extracted using PDFbox and slides with graphical images are generated.**
2016	**2016 International Conference on Computing, Analytics and Security Trends (CAST)**	**Prachi Ganguly, Prachi Joshi**	**IPPTGen-Intelligent PPT Generator**	**This paper proposed a system where input file was first preprocessed and closeness of the document was determined using cosine similarity. After feature extraction and reprocessed data as merged and finally a summary was generated. After obtaining the summarized and relevant contents from the input file and processed documents, the presentation slides are generated.**

INTRODUCTION

THE PROPOSED SYSTEM CONSISTS OF 3 MODULES :-

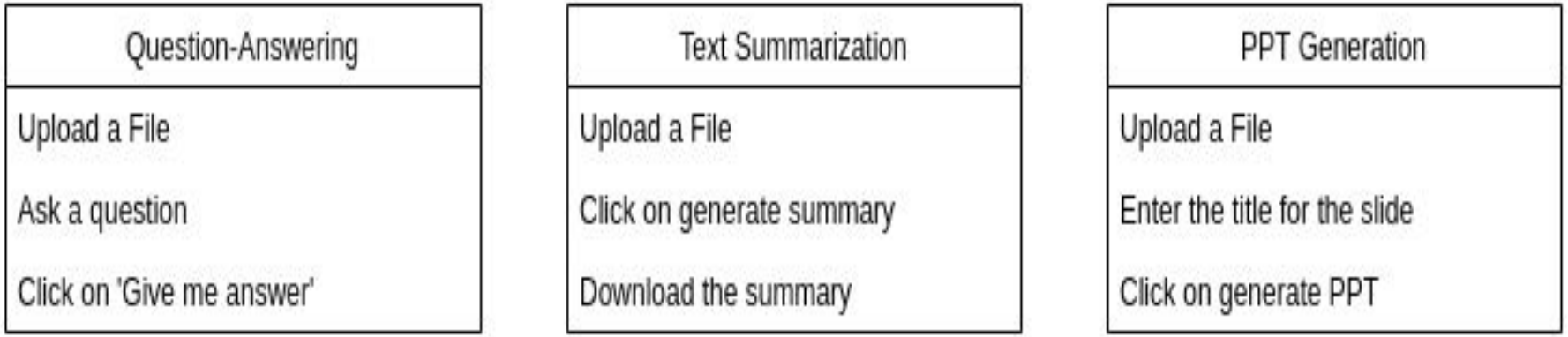
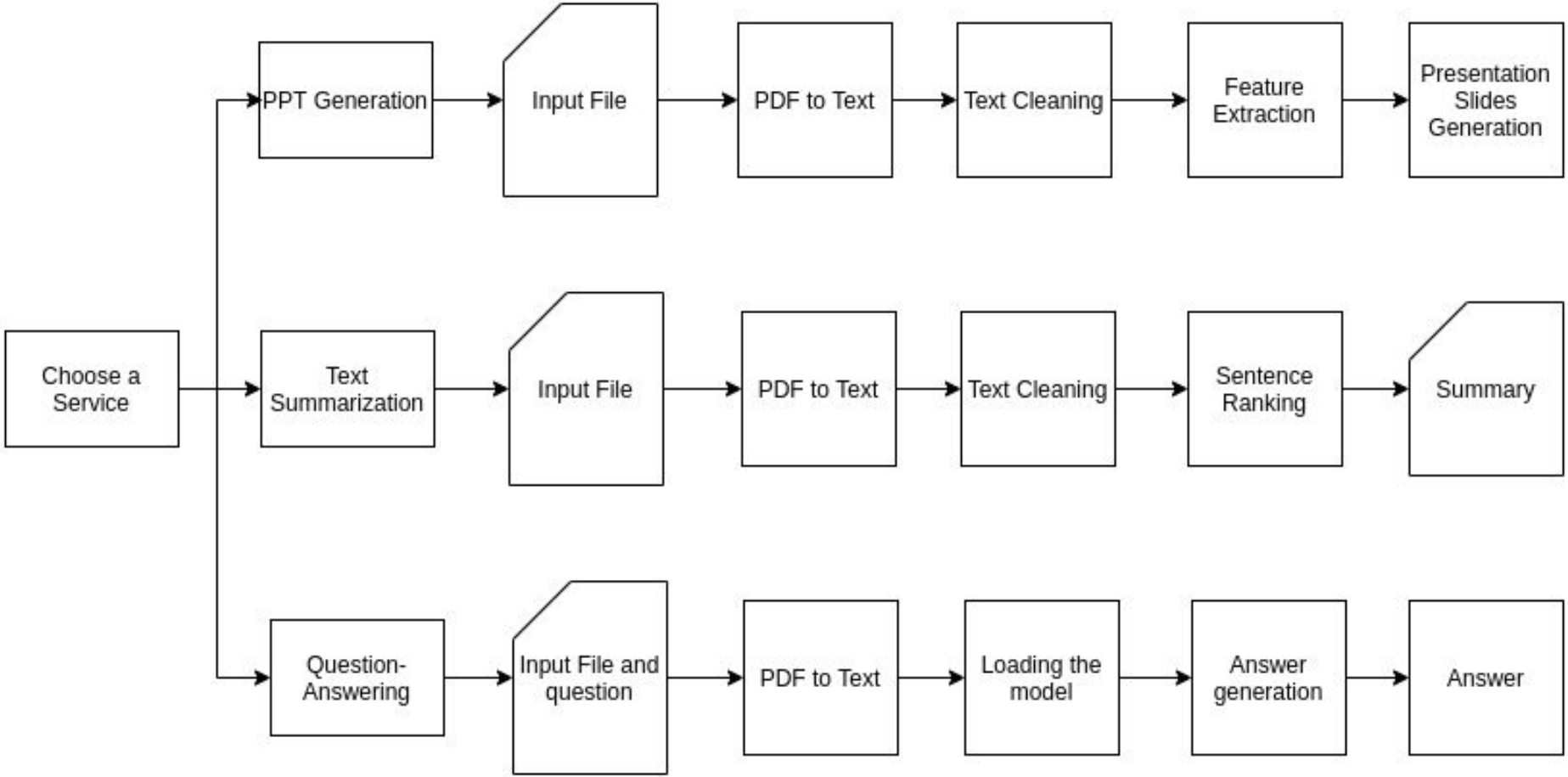


Fig No 1:Proposed System


Choose a Service
PPT Generation
Input File
PDF to Text
Text Cleaning
Feature Extraction
Presentation Slides Generation
Text Summarization
Input File
PDF to Text
Text Cleaning
Sentence Ranking
Summary
Question-Answering
Input File and question
PDF to Text
Loading the model
Answer generation
Answer

TEXT SUMMARIZATION

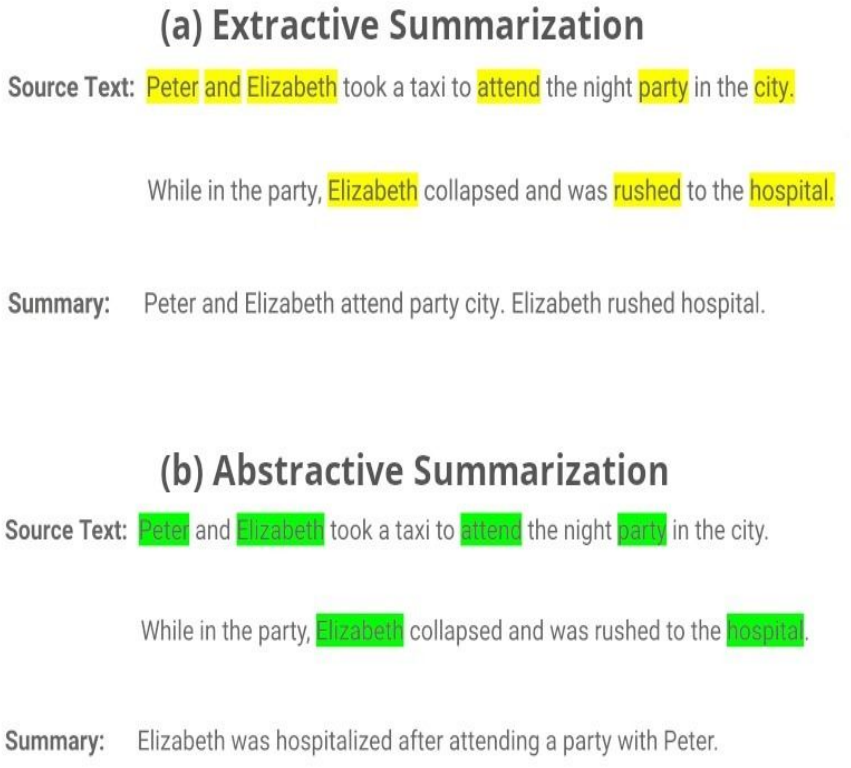
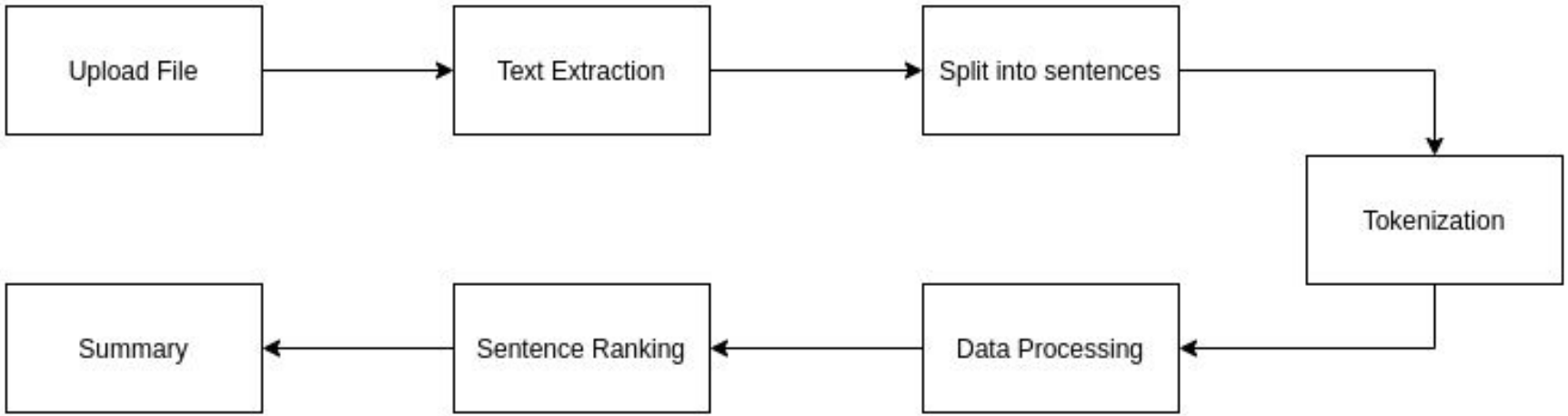


Fig No 5:Difference between Extractive and abstractive Summarization

- Text summarization refers to the technique of shortening long pieces of text.
- The intention is to create a coherent and fluent summary having only the main points outlined in the document.
- Automatic text summarization is a common problem in machine learning and natural language processing (NLP).
- There are two main types of how to summarize text in NLP:

https://devopedia.org/text-summarization

TEXT-SUMMARIZATION PROCESS



QUESTION ANSWERING SYSTEM

TYPES OF QUESTION ANSWERING SYSTEM

1. **Open domain (ODQA)**

   Deals with questions about anything

   e.g. DrQA by facebook, Google

2. **Closed Domain (CDQA)**

   Deals with questions about specific domain.

   We can train the system on specific topic or upload a file and search answers.

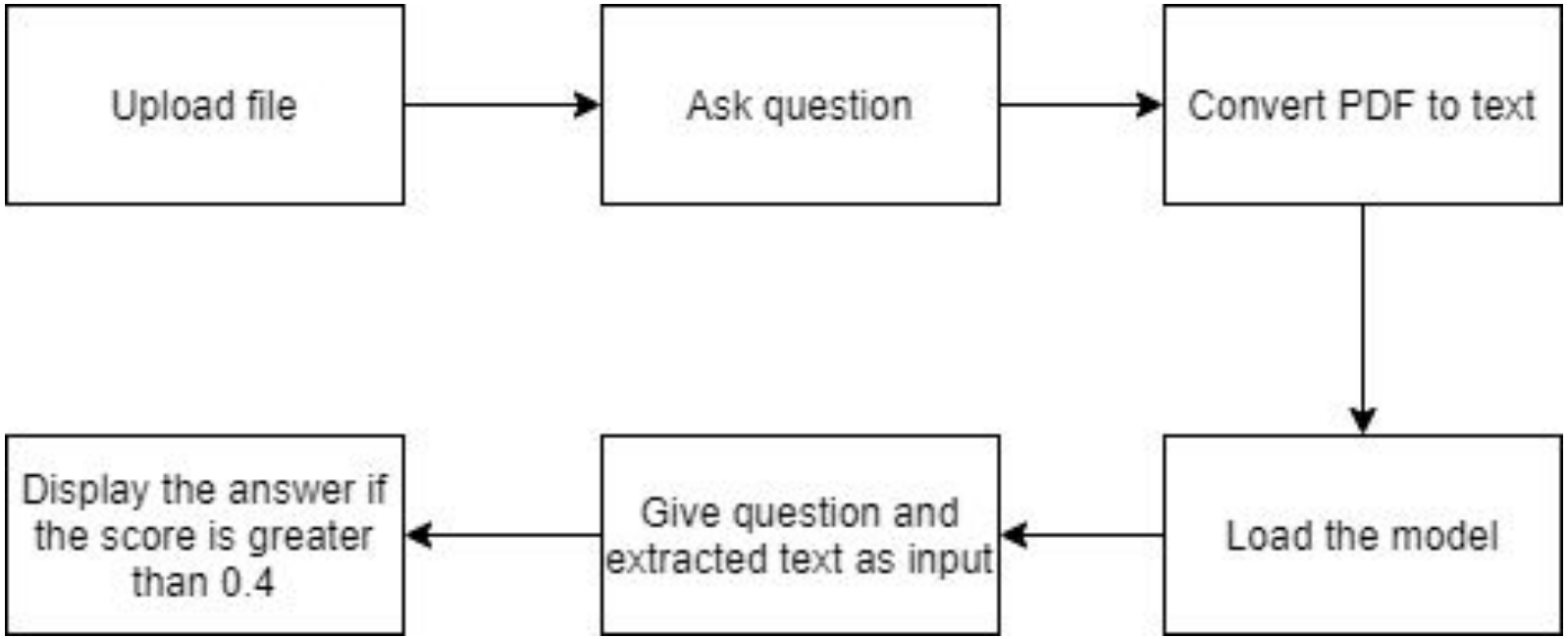
Steps for Question Answering

1. Install Huggingface transformers library
2. Import library
3. Build the question answering pipeline
4. Define the context and question to ask
5. Perform question answering

Here we have used DistilBERT model instead of BERT.

It is smaller, faster than BERT, which is trained on the same data as BERT.

DistilBERT retains 97% performance of the BERT with 40% fewer parameters.[8]


Upload file
Ask question
Convert PDF to text
Load the model
Give question and extracted text as input
Display the answer if the score is greater than 0.4

PPT GENERATION

PROCESS OF PPT GENERATION

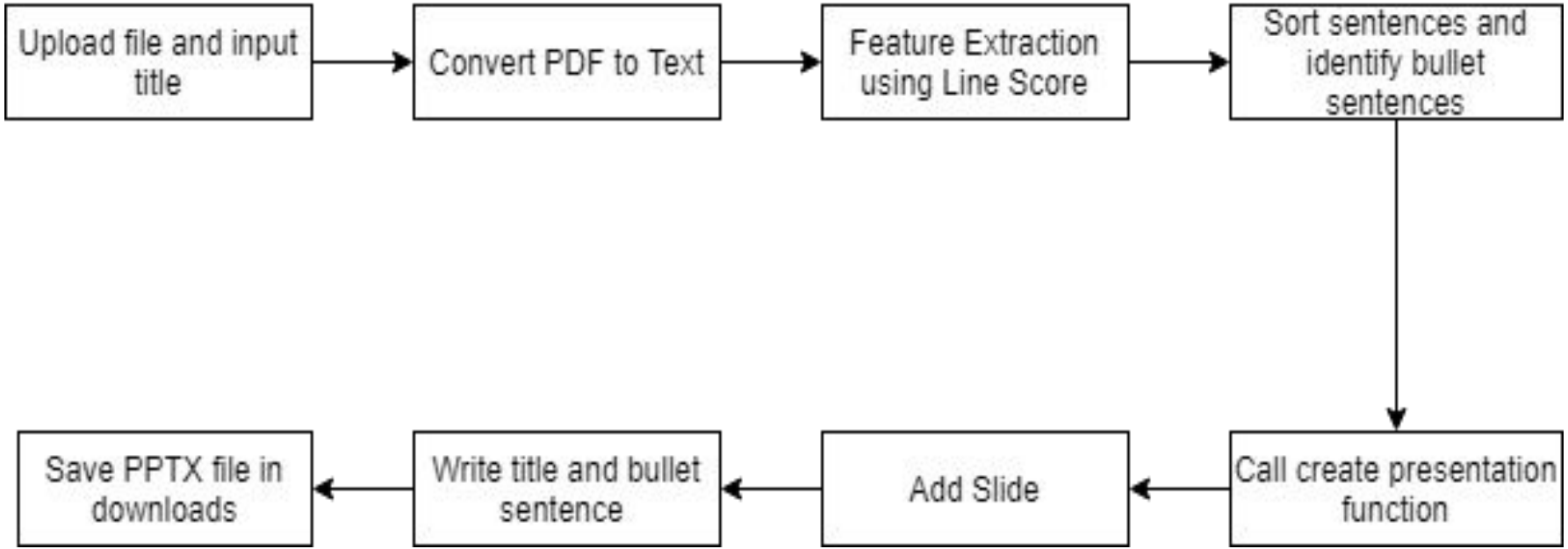


Fig No 8:PPT Generation Process

CONCLUSION

The proposed system helps students to improve their learning capability and promotes self-study and e-learning by answering the question with maximum possible accuracy and provides high quality text summarization for given document. The proposed system also reduces the efforts of teaching faculty by providing automatic presentation slides from the document and saves course delivery time of the facilitator.

REFERENCES

[1] André Macedo Farias: https://towardsdatascience.com/how-to-create-your-own-question-answering-system-2eff8a

[2] Ektaa Meshram , D. A. Phalke,"Technique for Generating Automatic Slides on the basis of Paper Structure Analysis" in IRJET Vol. 5, Issue 6, June 2016.

[3] J.N.Madhuri,Ganesh Kumar.R,"Extractive Text Summarization Using Sentence Ranking" in 2019 International Conference on Data Science and Communication,29th August 2019.

[4] Priya Ganguly, Prachi M. Joshi,"IPPT-Gen Intelligent PPT generator" in 2016 International Conference on Computing, Analytics and Security Trends (CAST) 19th December 2016.

[5] Yuwen Zhang,Zhaozhuo Xu,"BERT for Question Answering on SQuAD 2.0", Stanford University Research Paper, 2019.

[6] Yashu Seth,"Bert explained-A list of frequently asked questions", Blog on data science, Machine Learning and Artificial Intelligence, June 12, 2019.

[7] Ashish Vaswani, et al. "Attention Is All You Need"

[8] https://analyticsindiamag.com/python-guide-to-huggingface-distilbert-smaller-faster-cheaper-distilled-bert/