IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

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

Vishakha Kokate[1], Shrayesha Kukade[2], Kunchal Kulkarni[3], Akanksha Pawar[4], M. M. Swami[5]
Students, Department of Computer Engineering[1,2,3,4]
Assistant Professor, Department of Computer Engineering[2]
AISSMS College of Engineering, Pune, India

Abstract: *In Covid pandemic, the teaching-learning process has changed a lot. Due to the online classes, teachers are now burdened with creating teaching aids on large scale. E-learning is promoted due to the lockdown. The great amount of availability of the information in web causes difficulty in finding precise answers. Searching precise answer can be a time-consuming task. The paper aims at an intelligent system that will take a PDF file as an input and gain knowledge from the given file. The system will try to answer the questions posed by the user and summarize the given text. Presentation plays a vital role in e-learning and has revolutionized the concept of providing knowledge at all levels. But manually preparing presentation consume more time. So, as a solution to this problem we are proposing a system which will help in automatic generation of the slides from the given document and save time spent in preparing slides. Therefore, proposing system will help learners to resolve their queries using the question answering system and provide summary of the learning document provided by the learner. The system will also be able to generate PPT for reducing the effort of the teaching faculty, so that the teachers can concentrate more on teaching and also motivate the lazy learner or less motivated learner to self-study and prepare well for their exam.*

Keywords: Transformers, Information retrieval, Automatic text summarization, extractive summarization, sentence extraction, term weight, Natural Language Processing, Feature Extraction.

I. INTRODUCTION

There are many search engines available. Given some keywords, instead of giving direct answers they only return the documents that contain the keywords. Closed Domain Question answering system provides a precise answer to the questions under a definite domain as opposed to the search engines. The aim of the system is to present short and precise answer to the user query. User will ask a question and the system will retrieve the most accurate answer. Automatic Text summarization is the technique to identify the most useful and necessary information in a text. The approach being used is Extractive text summarization. In this paper, a novel statistical method to perform an extractive text summarization on single document is demonstrated. The method extraction of sentences, which gives the idea of the input text in a short form, is presented. Sentences are ranked by assigning weights and they are ranked based on their weights. Highly ranked sentences are extracted from the input document so it extracts important sentences which directs to a high-quality summary of the input document.

The content displayed on the slides is the summarized text of the information in the PDF. But the summarized text is to be displayed in a proper sequence to make audience understand presenter's ideas. The main objective of our system is to generate presentation slides automatically by summarizing the text using data processing methods, ranking the sentence based on the importance and finally generating the well-structured slides with other graphics.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

II. LITERATURE SURVEY

A survey of various methods getting used for question answering, text summarization, PDF generation systems with the purpose they were used for was carried. Here we present a quick survey of few of them.

2.1 Deep Learning Approaches for Question Answering System [1]

In 2018, Sharma and Gupta proposed "Deep Learning Approaches for Question Answering System". In this paper, they proposed a neural network-based framework for general question answering tasks that are trained using raw input-question-answer triplets. Dynamic Memory Networks Algorithm gave better performance compared to other algorithms like POS Tagging with tf-idf, LSTM Baseline and Memory Networks.

2.2 Implementation of Question Answering Retrieval System in Natural Language Processing [2]

In 2017, Vishwakarma and Bhatt proposed "Implementation of Question Answering System in Natural Language Processing". They developed a Question Answering system that answers simple factoid, WH-questions by using a technique called Semantic Role Labelling. This framework is given for restricted domain, it also handles the issues of word sense disambiguation. But the answer to the question is given with the help of search engine.

2.3 Question Answering System on Education Acts using NLP Techniques [3]

In 2016, Lende and Raghuwanshi proposed "Question Answering System on Education Acts using NLP Techniques". They proposed a closed domain Question Answering system in which they pre-processed the corpus and stored the extracted keyword in the index term dictionary, then the keywords extracted from the question is matched with the indexed dictionary. For finding the relevant keyword as an answer they used Jaccard similarity.

2.4 Automatic Slide Generation for Scientific Papers [4]

In 2019, Sefid, Mitra, Lee Giles and Wu proposed "Automatic Slide Generation for Scientific Papers". They proposed a model which used CNN neural network for sentence ranking and ILP as sentence selector. Their model achieved best Fscores in terms of ROUGE-2, ROUGE-L, and ROUGE-W compared with baselines like AvgTFIDF, TextRank and MEAD.

2.5 Enhancing Automatic PPTGeneration Technique through NLP for Textual Data[5]

In 2017, Belote, Bidwai, Jadhav, Kapadnis and Sharma proposed "Enhancing Automatic PPTGeneration Technique through NLP for Textual Data". They used Fuzzy Classification method for scoring the sentences. They also utilized fuzzy inference engine to remove adjust conclusion from estimated information and Fuzzy restrictive proclamation to decide the vital sentences.

2.6 IPPTGen-Intelligent PPT Generator [6]

In 2016, Ganguly and Joshi proposed "IPPTGen-Intelligent PPT Generator". They used TFIDF for feature Extraction, clustering for similarity between the feature vectors but the relevant sentences were extracted and merged into two phases- one used cosine similarity approach and the other used position Score Algorithm, and then slides were generated.

2.7 Extractive Summarization [7]

In 1958, most of the earlier works are done on the single document mainly focusing on technical document. (Luhn, 1958), he did his research on the extractive summarization. In his research he extracted sentence by calculating word frequency and phrase frequency that gives the useful measure of its significance.

In 1958, Baxendale has done his research at IBM on Extractive summarization. He extracted important sentence by using the position of text. The author has tested 200 paragraphs towards his goal to find that in 85% of the sentences which author has taken first topic which is main topic sentence and the last sentence came 7%. The most accurate

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

sentence would be selected from these two sentences. In 1969, Edmundson has done research on extracted summarization in this he extracted important sentence by using two features position and word frequency importance were taken from the previous works. The author has added two they are: presence of cue words, and the skeleton of the document.

III. Our Proposed Work

3.1 Overview

There are many different Question Answering System as well as web available for online learning. But great amount of availability of the information on web cause difficulty in finding precise answers and is a time-consuming procedure. Open domain systems are less accurate in providing answers compared to closed domain systems. In this paper, we investigate the problem of digital learning in existing system for learners and we are proposing a design to support e-learning.

Turning teaching material into digital format at a short notice has been a challenge. Therefore, there is a need for a system that can generate presentations after summarizing the provided document. Learners face 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. The overall system design consists of following modules:

1. Question answering,
2. Text summarization and
3. Automatic PPT generation on a given input file.

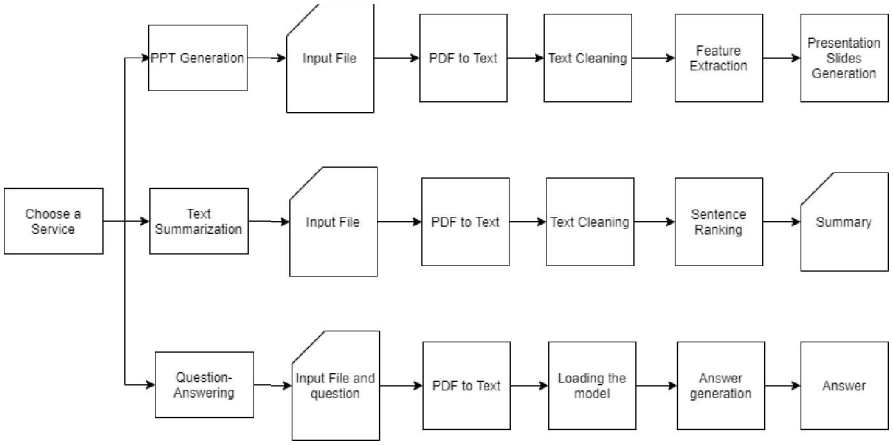

Figure 1: System architecture

3.2 Module Design

3.2.1 Question Answering

We use the hugging face transformers model to generate answers from the PDF provided. We provide text extracted from the PDF and question asked by the user as an input to the model and get answer as the output. First the user uploads the file in PDF format, then that file is converted to text. The converted text is used to search for the answer to the question asked. The answer is then displayed to the user. Extractive Question Answering is the task of extracting an answer from a text given a question.

Transformers provide general-purpose architectures (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet…) for Natural Language Processing.

3.2.1 Text Summarization

In this proposed approach, we are using extractive method to get summary of given input. We are taking input as text file .txt.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

a. Firstly, the file which is given as input is tokenized in order to get tokens of the terms.
b. The stop words are removed from the text after tokenization. The words which remain are considered as key words.
c. The key words are taken as an input for that we are attaching a part of tag to each key word. After adding the parts-of-speech tag to tokens or terms each individual weight is assigned to the tokens. The term weight is calculated as follows:

$$W_t = \frac{frquency\ of\ term}{total\ no.of\ terms\ in\ document}$$

d. Now maximum weight of the token is considered after finding maximum weight. The weighted frequency of the document is calculated as follows:

$$W_{tf} = \frac{frequency\ of\ a\ term}{maximum\ frequency\ of\ the\ term}$$

e. In this step, the frequencies are connecting in place of corresponding words in sentence and sum of it is found. The ranks are found based on the weighted frequency. The sentences are sorted based on their weighted frequency ranks like highest rank to lowest. The sentences are arranged in descending order.
f. Finally, summarizer will extract sentences which are ranked highest from the document and the sentences which are extracted accordingly.

3.2.3 Automatic PPT generation

Traditionally, there are many applications for converting PDF to text but there is no text summarization and also the presentation generated is not in proper format. The proposed system automatically generates the slides by using the information from the given document.

To generate the presentation slides automatically from the given PDF, the PDF format is first converted to text format and then data is pre-processed. Data Pre-processing includes tokenization, removal of special symbols and stop words, identifying the stemming substrings and replacing the substring with desired string and then finally concatenating the strings. Next, feature extraction is done by calculating the repeated occurrences of non-stop words.

IV. IMPLEMENTATION

1. Taking a file as an input.

A. file with an extension, pdf is taken as an input for all the three modules.

```php
<?php
set_time_limit(500);
if(isset($_POST['upload'])){
  $file = $_FILES['file'];
  $fileName = $_FILES['file']['name'];
  $fileTmpLoc = $_FILES['file']['tmp_name'];
  $fileSize = $_FILES['file']['size'];
  $fileType = $_FILES['file']['type'];
  $fileError = $_FILES['file']['error'];
                          @var mixed $fileName
  $fileExt = explode('.',$fileName);
  $fileActualExt = strtolower(end($fileExt));
```

Snippet 1. File input

2. PDF file converted to a Txt file

In case of the file being a PDF file, is converted into a .txt file for extraction of the text in that file.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

```
def convert_pdf_to_txt(self, path):
    print("in p2d")
    rsrcmgr = PDFResourceManager()
    retstr = io.StringIO()
    codec = 'utf-8'
    laparams = LAParams()
    device = TextConverter(rsrcmgr, retstr, codec=codec, laparams=laparams)
    fp = open(path, 'rb')
    interpreter = PDFPageInterpreter(rsrcmgr, device)
    password = ""
    maxpages = 0
    caching = True
    pagenos = set()

    for page in PDFPage.get_pages(fp, pagenos, maxpages=maxpages,
                                  password=password,
                                  caching=caching,
```

Snippet 2. PDF to text

A. Question-Answering Portal

1. Loading the transformers model

The transformers model which is used for generating answers is loaded.

```
from transformers import pipeline

def load_qa_model():
    model = pipeline("question-answering")
    return model
```

Snippet 3. Loading the pipeline

2. Converting PDF file to text:

```
def text_extract(file_path):
    article=""
    pdfreader = PdfFileReader(file_path)
    count = pdfreader.numPages
    for i in range(count):
        page = pdfreader.getPage(i)
        article += page.extractText()
    return article
```

Snippet 4. PDF to text

3. Searching for answers in the text using transformer model:

```
qa = load_qa_model()
q = lines
sentence = content
answers = qa(question=q, context=sentence)
ans = answers['answer']
```

Snippet 5. Getting answer

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

B. Text-Summarization Portal

1. Text-Ranking Algorithm

Using the text ranking algorithm to generate summary

```
for sent in sentence_tokens:
    for word in sent:
        if word.text.lower() in word_frequencies.keys():
            if sent not in sentence_scores.keys():
                sentence_scores[sent]=word_frequencies[word.text.lower()]
            else:
                sentence_scores[sent] += word_frequencies[word.text.lower()]

select_length = int(len(sentence_tokens)*0.4)
summary = nlargest(select_length,sentence_scores,key=sentence_scores.get)
final_summary= [word.text for word in summary]
summary=' '.join(final_summary)
return summary
```

Snippet 6. Text ranking

C. PPT-Generation

1. Feature Extraction

Feature extraction is done to generate important sentences for the ppt.

```
token_list = word_tokenize(sentence.lower())
featureExtractor = FeatureExtractor()
stop_words_per = featureExtractor.getStopWordsPer(token_list)
num_nouns = featureExtractor.getNumNounPhrases(token_list)
num_verbs = featureExtractor.getNumVerbPhrases(token_list)
overlapping_word_count = featureExtractor.getNumOverlappingWords(sentence.lower(), title.lower())
sentence_pos = featureExtractor.getSentencePosition(sentence.lower(), sub_sentence.lower())
avg_sent_len = featureExtractor.getAvgSentenceLength(sentence)
print(stop_words_per, num_nouns, num_verbs, overlapping_word_count, avg_sent_len)
```

(variable) token_list: Any

Snippet 7. Feature extraction

2. Slide Generation

After feature extraction, slides are generated using python pptx library.

```
def create_presentation(self, output_file, title='', sub_title='', contents=[]):
    self.create_title_slide(title, sub_title)
    for i in range(0, len(contents), 5):
        bullets = contents[i:i + 5]
        bullets.insert(0, "")
        bullet_title = self.get_bullet_title(bullets)
        if bullet_title:
            bullet_title = bullet_title.title()
        bullets = self.get_cleaned_bullets(bullets)
        self.add_bullet_slide(bullet_title, bullets)
        print("sg object from")
        #self.set_logo((i / 5) + 1, logo)
        #self.set_footer((i / 5) + 1, footer)
    self.prs.save(output_file + '.pptx')
```

Snippet 8. Slide generation

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

V. RESULTS

5.1 Question Answering

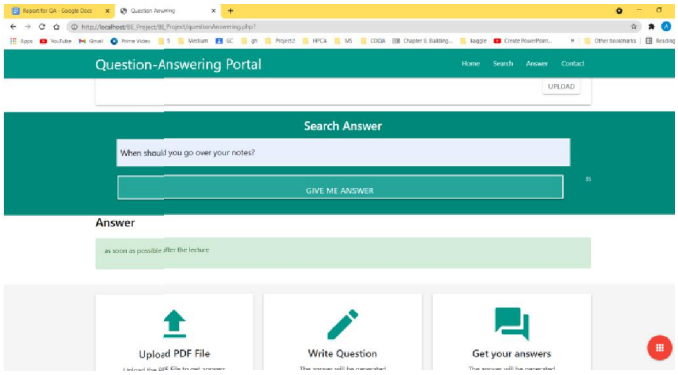

Figure 2: Question answering

5.2 Generated Summary

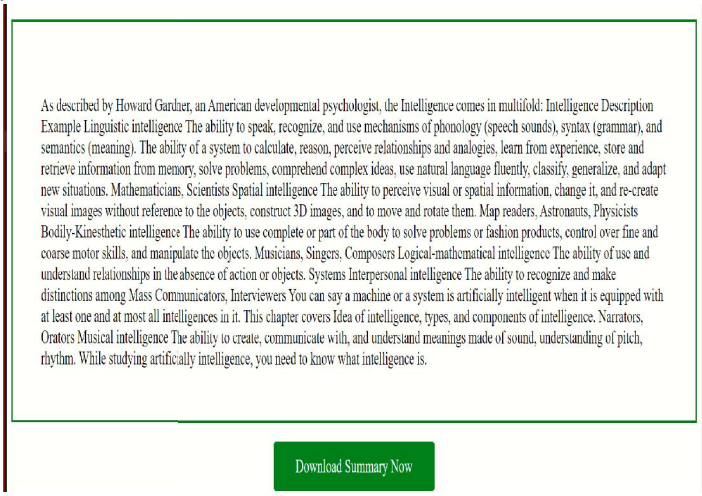

Figure 3: Summarized Text

5.3 PPT Generation

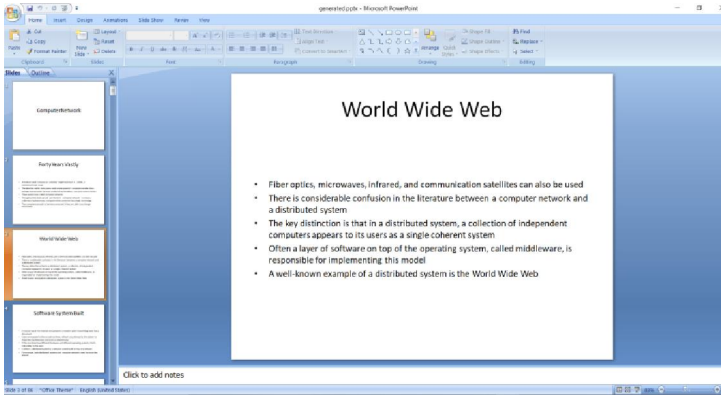

Figure 4: Generated PPT

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 6, Issue 1, June 2021
Impact Factor: 4.819

VI. CONCLUSION

In this paper, we have discussed closed domain Question answering system using the Hugging face transformers, Text summarization using Extractive summarization and Slide generation using feature extraction. The system can be enhanced by automatically putting the images related to the text in proper slides without any human interference.

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

REFERENCES

[1] Sharma and Gupta, "Deep Learning Approaches for Question Answering System", Elsevier Ltd., ScienceDirect, ICCIDS 2018

[2] Jaylalita Vishwakarma and Prof. Mayant Bhatt, International Journal for Rapid Research in Engineering Technology & Applied Science Vol 3 Issue 11 November 2017 ISSN(Online): 2455-4723, Paper ID: 2017/IJRRETAS/12/2017/34611, 2017

[3] Sweta P. Lende and Dr. M. M. Raghuwanshi, "Question Answering System on Education Acts using NLP Techniques", IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'16), 2016

[4] Athar Sefid, Jian Wu, Prasenjit Mitra and C. Lee Giles, "Automatic Slide Generation for Scientific Papers", SciKnow'19, November 19-22, 2019

[5] Pooja Belote, Sonali Bidwai, Snehal Jadhav, Pradnya Kapadnis and Nakul Sharma, "Enhancing Automatic PPTGeneration Technique through NLP for Textual Data", IJARCCE, Vol. 6, Issue 3, March 2017.

[6] Priya Ganguly and Dr. Prachi M. Joshi, "IPPTGen-Intelligent PPT Generator", International Conference on Computing, Analytics and Security Trends (CAST), 2016