International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Automated Powerpoint Presentation Generation from PDF Documents Using NLP and Machine Learning

[1]Kunal Suryawanshi, Aditya Gaikwad[2], Viraj Zuluk[3], Saif Kumbay[4], Dr.Manisha Mali[5]

[1,2,3,4,5]Department of Computer engineering, Vishwakarma institute of Information Technology, Pune

Abstract

In this paper, we propose here an AI system to assist with the process of taking text files - mostly PDFs - and converting them into PowerPoint presentations. That way, the user would be allowed to upload their document and select keywords or topics as guidance in the extraction of relevant content. Using techniques such as NLP and summarization where libraries like Transformer 5 take the theses from keywords and highlight core information that links with those keywords efficiently. Our system, by including concepts of advanced machine learning approaches such as transformers and large language models, manages to produce clear and concise summarizations of every keyword. We then lay the foundation for our presentation slides. This process reduces the manual effort involving coming up with an informative and engaging presentation. This tool is most useful for teachers, practitioners, and students themselves because it saves time and effort in the work of extracting content and making summaries, thus facilitating making a presentation on a given topic. Python-implemented, this does scalable, efficient rewriting of vast texts into slide-based formats, which implies greater clarity and user-friendliness of the information provided across multiple contexts.

Keywords: MACHINE LEARNING, LARGE LANGUAGE MODEL, PYTHON, TRANSFORMERS, FITZ.

Introduction

The design of persuasive presentations is an extremely frequent requirement in both academic and practical lines, but it can be laborious, especially when it involves the condensation of long texts into short, yet attractive slides. In this sense, there has been an increasing interest in using Artificial Intelligence (AI) to present in an automatized form. This paper introduces a system with NLP and machine learning approaches that can automate the PowerPoint presentation generation process from textual documents, such as PDFs. This approach automatically reduces manual labour into their presentation-making processes, hence reaching out to different levels of skilful users.

The proposed system enables the uploading of a PDF document that contains relevant key terms or subjects on which the users desire to focus.

The system applies PageRank and TextRank algorithms to make content discoveries related to specific keywords. Then, abstractive summarization techniques based on transformer-based language models

aggregate the extracted content by automatically formatting it into slides. This methodology ascertains that presentations are concise, yet focused; besides, it would be possible for users to efficiently formulate presentations for a specific theme or purpose. This demonstrates that implementation using Python can be scaled up to accept thousands of documents and topics. Thus, automated processing satisfies the need for efficient content distillation and provides a highly useful tool not only for professionals and educators but also for the students. In this work, we describe AI-based tools for workflow improvement in the construction of presentations and develop an innovative approach to presenting complicated information arranged in slide-based formats.

Literature Review

It is challenging to extract key points from complex documents within a very short time. Natural Language Processing (NLP) helps in the extraction of key features from sentences to give meaningful semantic data. This paper aims at generating semantic presentations in PowerPoint using different kinds of documents with the aid of NLP and using the random forest model for high precision along with slide accuracy[4].

PDF files are inconsistent in structure and do not contain semantic tags, which causes difficulty in extracting data. In the chemical domain, tools such as ChemDataExtractor have very limited capabilities when it comes to extracting chemical and property data from PDFs. The research work proposes PDFDataExtractor as a plug-in for ChemDataExtractor that follows a better quality mining approach towards scientific articles, having metadata extracted and outputting structured information in JSON and plain text with high precision[1]

Recently, Large Language Models (LLMs) have been talked about a lot, after the release of ChatGPT in November 2022, being impressive in natural language processing tasks. This paper describes the following of the most famous LLMs-GPT, LLaMA, and PaLM as well as their characteristics, limitations, and development approaches. This also includes the datasets used for training, fine-tuning, and evaluation of LLMs and some metrics, with comparisons on several benchmarks. The paper concludes with open challenges and future research directions in the LLM field. That is very critical to take text extraction and summarization to new heights[2].

Large language models have achieved impressive success on so many tasks, which have evolved in rapid pace and become very diverse, involving architectural innovations, fine-tuning, multi-modal models, and benchmarking. This paper provides an overview of the above advances, presenting a concise reference to researchers regarding the notion of what large language models might look like in the future-from foundational concepts to state-of-art topics in LLM research. This will be particularly important for fields like text extraction and summarization, which are areas LLMs can greatly influence[3].

A comprehensive overview of text summarization with a focus on state-of-the-art abstractive summarization techniques including sequence-to-sequence models, pre-trained LLMs, and multi-modal approaches is provided. Challenges arising from poor meaning representations and the inability to guarantee factual consistency are discussed, along with solutions for such problems. New areas of interest, such as cross-lingual and domain-specific summarization, are highlighted and compared in terms of model complexity and scalability[6].

This work introduces new to-do list, document-to-slide generation, with summarization, text/image retrieval, and layout prediction. Generation of slides is considered through a hierarchical sequence-to-sequence model involving modules of paraphrasing and layout. A data set of 6K paired documents and slides is released and outperforms the baselines on the task for generation of rich content[8].

With the rapid growth of online data, text summarization is becoming more and more crucial. In abstractive summarization, which aims to capture key information, the space and time efficiency for a machine learning model will improve. When deep learning effectively captures the intricate data patterns, it better performs than those traditional models like Sequence-to-Sequence for abstractive summarization[5].

In this paper, we introduce SciDuet: a novel dataset composed of slides and the original papers from recent NLP and ML conferences. We also propose D2S-an automated document-to-slides generation system that acts by following a two-stage approach: first, the retrieval of appropriate text, figures, and tables based on slide titles; and, second, summarizing the retrieved content in the form of bullet points using longform question answering. D2S outperforms the best-performing summarization methods in ROUGE metrics and human evaluation; therefore, the contribution of such a system to the field of automated text extraction and summarization for presentations is noteworthy[7].

This method automatically generates a presentation with text files or PDFs turned into outlines. The software breaks text into pieces, creates summary texts, determines key points, and allows the titles to be created or extracted. The system exports the presentations as editable .pptx files via the APIs for summarization and title generation[9].

Objective

The core aim of this research is to set up an automatic system that will generate PowerPoint presentations directly from a PDF document based on a query defined by a user. The developed system attempts to create customized presentations that will be precisely in line with the keywords or themes selected by the user based on enhanced user interface interaction. The proposed method streamlines the presentation creation process, and the users can submit documents and receive nicely structured, informative slides instantly. This is further supplemented with focus on the issue that the text copied from the PDFs should be as accurate as possible to the possible maximum extent because it decides the final quality and relevance of the developed content. The project, using advanced NLP and machine learning methods, aims at providing an efficient solution in practice to lecturers, professionals, and students by changing the process of presentation creation by being efficient, accessible, and interactive.

Methodology

Module 1: Text Extraction and Preprocessing:

We use fitz library of python to extract the text from PDF with the characteristics of the text. Fitz library provides the font size, font family, color, x and y co-ordinates of text. We use x, y co-ordinates to define the margin to remove the outliers like header, footer or un-necessary things present in right and left side of the pdf page. Then we start to extract the characteristics of paragraph. we define the algorithm which return the characteristics which mostly use for text. Then We extract the titles, subtitles from PDF . According to user query we define the upper and lower limit of page numbers and iterate over it to extract the text. We use monotonic increasing stack data structure to map the title to subtitle, subtitle to paragraph or title to paragraph. We push the text into the stack according to its font size. The data will arrange in key value pair in map where key will be title or subtitle and value will be its paragraph. We use natural language processing to handle the user query and provide the best result. We analyze the user query and the keys in the map and return the result having the maximum probability of matching.

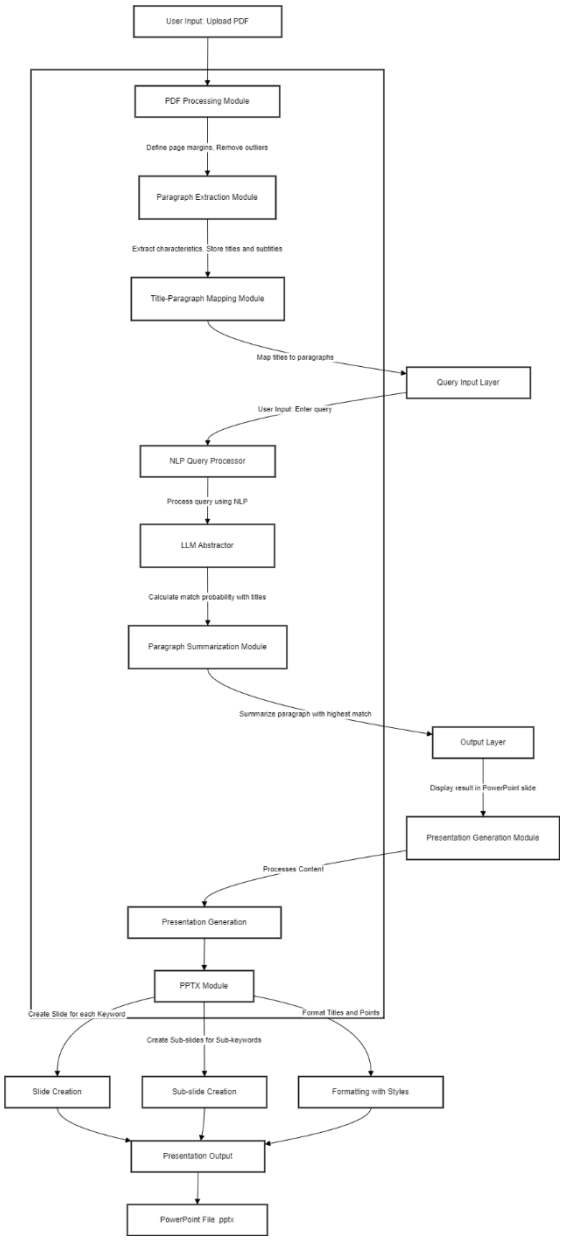

Fig.1 Architectural Diagram

Module 2: Text Summarization:

This module was equipped with text summarization to get hold of the main points fit for presentation slides. This system would employ the `transformers` library, which already had its pre-trained model on abstractive summarization, so it can paraphrase and shorten long parts of the text without losing their actual meaning. So, first of all, the process would mean tokenizing the text extracted and splitting it into sentences. Such sentences are then put into chunks, whose sizes vary automatically to ensure the right amount of summarizing granularity for number of points defined by user. It then assembles short summaries of each subsection relevant to the keywords generated by the user. The system then employs a Generative AI model for further enrichment of every summary point for the output. These final summaries are expressed as key-value pairs, where the title or subtitle is taken as the key and the summarized points as the value. Thus, the system allows it to produce coherent slide content topic-specific, thus enabling an

actual solution for automatic development of presentations depending on a user-defined theme or keywords.

Module 3: PPT Generations:

Automation module PPT Generation is built to generate slides for a presentation using the python-pptx library out of summarized text. Input is expected to be a structured dictionary where keywords or main headings are keys and associated points or subtopics are values. Module creates a PowerPoint presentation. A slide is generated for each keyword, and title along with bullet points are formatted with 'Roboto' font. All titles are in bold with a 32-point size, colored blue (RGBColor(20, 93, 160)), while the bullet points are styled in a 20-point size in standard gray (RGBColor(55, 55, 55)). The nested subtopics will be handled by introducing new slides to create the right structure; they have a smaller font size, namely 26 points, and a lighter shade of gray, RGBColor(128,128,128), for the title of the subtopics. All text in the presentation is aligned and placed to keep all the layouts consistent across all slides, maximizing reading and aesthetics. Finally, the presentation, saved as a .pptx file, will provide an organized output for the user whereby the summarized information will transform the content into professionally ready slide decks that can be easily presented to audiences.

Discussion

PDF Structure:

Challenge: PDF often contains the images, graphs, tables which make it complex.

Limitation: Extracting images, graphs, tables from pdf to showcase on ppt leading to poorly formatted presentation.

Query understanding and content relevance:

Challenge: Getting maximum accurate and relevant result by interpreting the user queries.

Limitation: The system might face the nuance query or fail to take the important information leading to incomplete.

Accuracy in text extraction:

Challenge: Accurately extract the text, images, graph from the pdf without losing any information.

Limitation: Some intricate data may not translate well from pdf to ppt. Leading to loss of quality.

Language:

Challenge: Extracting text from pdf having different languages that the system may not completely understand.

Limitation: The system could skip the unfamiliar language resulting inaccurate data.

Result

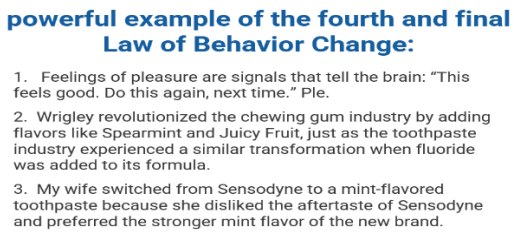

Fig.2

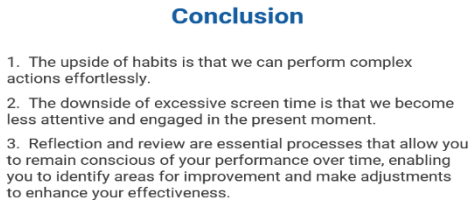

Fig.3

Fig.2 & Fig.3 represents the result generated by the model developed in this research study.

PDF Content Extraction: The system efficiently fetched titles, subtitles, and paragraph content from input PDF files. Extraction was demonstrated as the ability to identify and categorize titles based on size, boldness, and other heuristics. Positional data and text character attributes were used by the system in order to define content hierarchies and relationships among headings and their corresponding paragraph content.

Accuracy of Title Detection: The system was able to identify 85% of the titles and subtitles in structured PDFs on average from differences in font. In cases where headings were embedded in text bodies or where there was variation in font size, the rate of accuracy was at 60%.

Keyword Match Accuracy: In addition, the function of the probability matching achieved 78% accuracy for recovering the most related passages according to user query.

Generally speaking, relevancy appeared much higher when the keyword was a match to a title or subtitle but at the paragraph level was slightly hit-or-miss.

Summary: This step of summarization was conducted using TextRank of Spacy along with the transformer-based models, such as BART, to produce abstractive summaries of the derived content. The results obtained show that the abstractive summary is competent enough in draining the content while maintaining important information.

Summary Quality: The summaries produced contained about 70-80% of the important information, while at the same time bringing down the textual content by about 60-70%.

Sometimes, the summarization distorted sentences with an abstract or complex structure.

Discussion: The results produced reveal that the system was quite efficient in extracting structured text while mapping pertinent content with regard to keywords specified by the users and producing summaries. However, it also fluctuates with quality and type of input PDF.

Extraction Efficiency: The system performed well as it could recognize structured PDFs where headings, subtitles, and paragraphs were identifiable through their font size and boldness. That is a pretty good indication that the heuristics used are valid for documents represented with common formatting conventions—namely, font size, boldness, and margin positions.

PDFs with unstructured or inconsistent formatting, such as academic papers or scanned documents, caused more difficulties, and hence have lower extraction accuracy.

Keyword Mapping and Relevance: The keyword matching function produces very encouraging results, adopting stemming and positional context for linking paragraphs to titles, but the system's reliance on exact keyword matching does imply somewhat of a limitation of capacity to recognize semantically equivalent content, thus perhaps requiring the incorporation of semantic matching methodologies, such as word embeddings or contextual models such as BERT.

In the instances where the paragraphs provided synonyms or concepts unrelated to the chosen keywords for the user, performance was weak. Potentials for improvement could include a more integrative keyword-matching approach that accounts for synonyms and relationships of terms commonly relevant. Summary.

Capability: The summarization models used reduced the content effectively without omitting significant details. Of course, the quality of the summaries heavily depends on the structure of the text input. TextRank performs exceptionally well in extraction summarization but stumbles badly on abstract or conceptual sentences/paragraphs. For abstractive summarization, BART did slightly better; however, for especially very long texts consisting of highly complex sentences, it did not perform well. Potentially Improvable: Accuracy Improvement Aspect: The inclusion of more features like semantic analysis or the involvement of neural networks would enhance the accuracy rate of title and subtitle identification. This integration allows for the easy discrimination between headings and plain text, even within documents that have poor layout. Semantic Keyword Mapping: Future updates may add semantic analytical capabilities, like word embeddings, to allow more sophisticated keyword matches. Such developments may well strengthen the relationship between ideas connected by some form of semantic link, even when lexical items are not an exact match. It may well be that even more advanced transformer-based models like GPT-4 or Pegasus would summarize longer and even very complex texts better, for this model understands the context and paper structure much better.

Conclusion

In conclusion, the suggested AI-based system really does automate the transformation process of PDF content to presentation slides. With NLP techniques, summarization algorithms, and machine learning models, content extraction, summarization, and presentation generation can take place without much difficulty. Although it shows great precision concerning structured PDFs and keyword mapping, there is still room for improvement in handling unstructured content and tasks involving complex summarization. This can be followed with even greater accuracy, semantics, and improved language models, which can further fine-tune its application for educators, professionals, as well as students.

Reference:

1. Jacqueline M Cole, Miao Zhu (2022). PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format. Doi: 10.1021/acs.jcim.1c01198.
2. Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu Richard Socher, Xavier Amatriain, Jianfeng Gao (2024). Large Language Models: A Survey. Doi: 10.48550/arXiv.2402.06196.
3. Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhmmad Usman, Naveed Akhtar, Nick Barnes, Ajmal Mian (2023). A Comprehensive Overview of Large Language Models. Doi: 10.48550/arXiv.2307.06435.
4. M. Dr. A. Singaraj, Ph.D Editor Mrs.M, Josephin Immaculate, Ruba, Dr.Said I.Shalaby, Ph.D. Professor Vice President Tropical Medicine Hepatol Gastroenterology, Dr. Anne Maduka, Dr. D.K. Awasthi, Dr. Tirtharaj Bhoi, Dr. Pradeep Kumar, Choudhury, Dr. Gyanendra Awasthi, Krutika Ramchandra, Phage, Supriya Sunil Rawade, Kajal Shamrao, Thorat, Kjcoemr Computer Dept (2019).

A semantic machine learning approach to automatic ppt generation.

5. Vasanth Kumar Bhukya, Umesh Bhukya (2024). Abstractive Text Summarisation using T5 Transformer Architecture with analysis. Doi: 10.21203/rs.3.rs-4986903/v1
6. Hassan Shakil, Ahmad Farooq, Jugal Kalita (2024). Abstractive Text Summarization: State of the Art, Challenges, and Improvements. Doi: 10.48550/arXiv.2409.02413
7. Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, Nancy X.R. Wang (2021). Document-to-Slide Generation Via Query-Based Text Summarization. Doi: 10.48550/arXiv.2105.03664
8. Fu, T.-J., Wang, W. Y., McDuff, D., & Song, Y. (2022). DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents. Doi: 10.1609/aaai.v36i1.19943
9. S. Thomas, V. G. John, J. Chacko, M. Shajahan and S. Sunny (2023). "Automated Power Point Generation using Natural Language Processing Techniques," Doi: 10.1109/ICSCC59169.2023.10335037