Learning Based Slide Generator

Kevin Shaj
Department of Computer Science and Engineering,
St. Thomas College of Engineering and Technology
Chengannur, Alappuzha, Kerala.

Sara Susan John
Department of Computer Science and Engineering,
St. Thomas College of Engineering and Technology.
Chengannur, Alappuzha, Kerala.

Roshney Philip
Department of Computer Science and Engineering,
St. Thomas College of Engineering and Technology.
Chengannur, Alappuzha, Kerala.

Prof. Anish George
Department of Computer Science and Engineering,
St. Thomas College of Engineering and Technology.
Chengannur, Alappuzha, Kerala.

***Abstract*—An automated slide generator will help to save time, effort and subsequently cost. At present, tools such as Microsoft PowerPoint and Open Office assist researchers in providing an outline and theme for the slides but do not help researchers to select contents of slides. An academic presentation is a sort of advertisement for the paper than an attempt to present all the information in the paper. In the PowerPoint presentation, it is an acceptable idea to find a picture that describes the aim of your research. Visuals are considered very effective tools for engaging the audience and maintaining their interest in conveying an important point or thought. This project proposes an automated system which generates presentation slides from research papers. The proposed system accepts research papers in PDF format as input and helps to generate the corresponding presentation slides. Papers and slides are learned and trained by Bidirectional Encoder Representations from Transformers (BERT) model. The research papers are summarized using the Google BERT algorithm which is a custom module that was released by Google. The sentence important scores are predicted by the pre-trained model of BERT. Text from the papers is extracted using Python's *unpdfer* tool. The generated summary is used by BERT for making slides. As the presentation slides are of vital importance in a person's career, a significant amount of time and effort is spent on its preparation.**

Keywords*—Automated slide generator, Python, BERT, NLTK toolkit, unpdfer tool, text extraction

I. INTRODUCTION

Presentation slides have always been the most popular, creative and simplistic visual aid to deliver information. Slides provide an easy way of collaborating knowledge among different groups of people. Researchers, educators and learners use slides to explain their lectures, works or academic projects briefly and conveniently in most conferences and meetings. Since the past few years, software tools such as PowerPoint in Windows and Open office in Linux are available which are only used for designing, formatting and aligning the slides, but not in its content preparation. Recently, Google Slide was developed by Google for creating and formatting slide in open source but not in content. The process of slide generation becomes cumbersome when the users are required to take down critical points from the academic papers to include them in the slides while preparing them. But this project proposes a methodology for generating presentation slides that include important ideas from a research paper.

The proposed automated slide generation system will assist to optimize time and value for money. Our system will help in providing the precomputed presentation slides, which contain the eminent points of the given research paper. The critical content will be highlighted as bullet points in the slides. The system feeds in research papers in PDF format and generate the corresponding presentation slides based on the input papers. The BERT algorithm is used to summarize these research papers. The sentence important scores are predicted by BERT pre trained model, from which summary of each paper is generated. Slides are created from the generated summary. This work considers only the text elements present in the papers. The system uses Python's NLTK toolkit for sentence ranking.

II. RELATED WORKS

Yue Hu and Xiaojun Wan [1] proposed "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers" in 2015. They suggest an innovative system called PPSGen to investigate the daunting task of automatically generating presentation slides from the academic papers. These slides can be used by the presenters to prepare official slides in less time. Regression methods are applied to determine the value of sentences in academic papers. Highly organized slides with good or solid structure is devised using the integer linear programming (ILP) method by choosing key phrases and sentences. The enhanced slides are created from the submitted PPSGen system.

In 2016, Ektaa Meshram and D. A. Phalke [2] suggested a "Technique for Generating Automatic Slides on the basis of Paper Structure Analysis". Most areas in all fields use slide presentations in an easy and aesthetically pleasing format to collaborate information with all concerned parties. Tools such as Microsoft PowerPoint, Open Office, and Apple Pages are used to conventionally prepare slide presentations. This resulted in an exhaustive process of preparation with a chance for failure. The biggest challenge would be missing out important information from research journals and conference papers. To overcome this trial, an intelligent tool was developed to create slides with less human error. Their proposal draws the graphical elements as well as text from a paper which was a failure in many other existing automatic tools.

III. PROPOSED SYSTEM

The proposed system is meant to provide efficient means to simplify slide generation for a wide variety of users. We use Python’s NLTK toolkit to extract the text elements and convert it into a preprocessed form, as an input to BERT. BERT model is used to summarize the text into required number of presentation slides. The input obtained from the user is the research papers in the form of portable document format (pdf) file. Additionally, the required number of slides is also obtained.

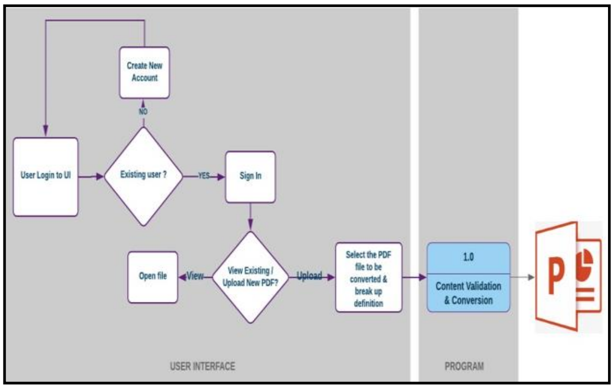

Create New Account
NO
User Login to UI
Existing user ?
YES
Sign In
Open file
View
View Existing / Upload New PDF?
Upload
Select the PDF file to be converted & break up definition
1.0
Content Validation & Conversion
USER INTERFACE
PROGRAM

Fig 1: Process Flow

A. *System Design*

The system architecture of the proposed system is as follows:

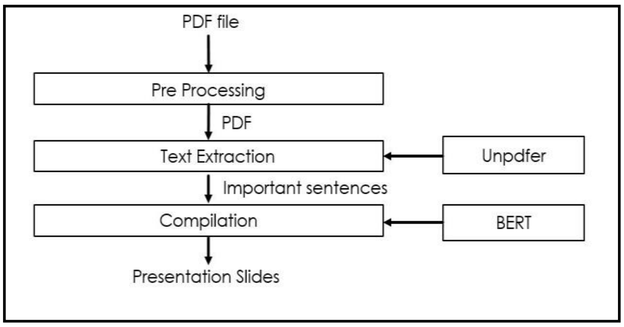


Fig 2: System Architecture

The system mainly consists of three modules namely:

- Pre-processing
- Text Extraction
- Compilation

a) Pre-processing

Considering the user interface is in the form of a web application, we can explain the pre-processing stage where user provides the input parameters. The UI is a php page which first takes the user to a login page. Here, the user is required to make an account. It is done by giving user details such as name, email id, phone number and a password. If the user already has an existing account, he/she can simply log in by typing in their credentials. In case the user has forgotten the password, it can be recovered using the respective phone number. Once the user authentication is done, next the input file should be uploaded. The interface does not provide any academic papers and should essentially be provided by the user. The following option is to input the number of slides that the user wants to distribute the pdf into.

b) Text Extraction

The next module is where text extraction occurs [Fig 3]. The main aim of NLTK toolkit is to derive the importance of each word present in the text elements. Python has various summarizing tools. For the purpose of text extraction we first use ‘unpdfer’. It is a package that takes a pdf document and produces a blob of text and some additional information related to the document. After that the tokens are extracted using tokenize function call which means the words are separated. This is followed by stop word removal which is to remove stop word such as the, an, a, have, has, and so on. Next step is POS (part of speech) tagging where the words are marked according to the corresponding parts of speech such as noun, adjective, verb and so on. Following that lemmatization is done in order to remove inflectional endings and return it to the root words.

Vectorization is the process of converting text and related information in a number format. It returns vector numbers ranging from a certain interval. The words with largest vectors are considered important. The similarity index function checks if any sentences are repeated and discards the rest after keeping just one copy.

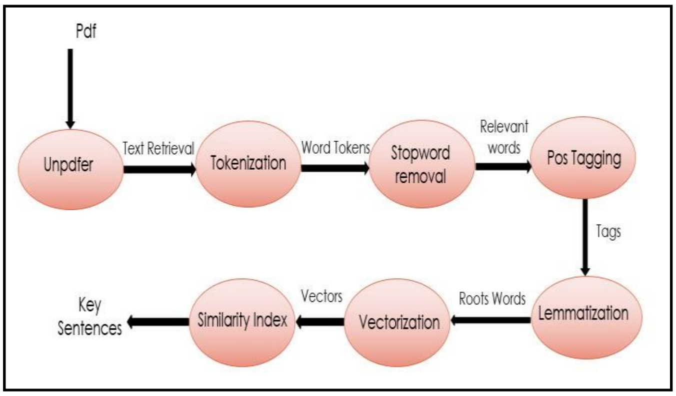


Fig 3: Decomposition of Text Extraction

c) Compilation

The final module is compilation which observes the content and each sentence importance and deduces the final output. The BERT model is responsible for the compilation. It is a readily available natural language processing model. It is trained to obtain presentation slides from the key sentences. The key sentences are evaluated and narrowed down in order to fit the specified number of slides. It considers the vectors and arranges the important sentences. The slide, even though it is a draft, maintains the structural format of PowerPoint slides, for example, bullet points. Therefore, the draft slides are obtained.

B. System Requirements

For the system to be set up, it has the following requirements;

a) HARDWARE REQUIREMENTS

- Processor – Intel i3 (min)
- RAM – 6 GB (min)
- System Type – 64-bit Operating System
- Keyboard – Standard Windows Keyboard

b) SOFTWARE REQUIREMENTS

- Operating System – Windows
- Programming Language – Python
- Database – MySQL
- Tools – BERT, Unpdfer

PYTHON

Python is an object-oriented, high-level programming language having dynamic semantics. Python's easy to learn syntax emphasizes on readability and therefore, reduces the cost of maintenance. Python supports modules and packages, which encourages modularity and reusability of code.

MYSQL

MySQL is a well-known Open Source SQL database management system. It is an Oracle-backed open source relational database management system (RDBMD) based on Structured Query Language (SQL).

BERT

Bert is a Python library that enables us to deploy pre-trained BERT models in our machine. This Bert library is used for summarizing the input text that is extracted using the *unpdfer*.

UNPDFER

Unpdfer takes in a pdf document and returns the text from within it along with additional information about the document. It has SCRUB and VERBOSE flags.

IV. RESULTS

Learning based slide generator is a system that is used to automatically generate summarized content of a given PDF. Here, PDF of a scientific paper is uploaded to the system. The system then converts this PDF file to text file to run the summarization. The text summarizer then applies different functions to keep only the relevant words. The text is then send to BERT system for the final summarization. Hence, we get a summarized content of a long scientific paper which can be used in Presentation Slides.

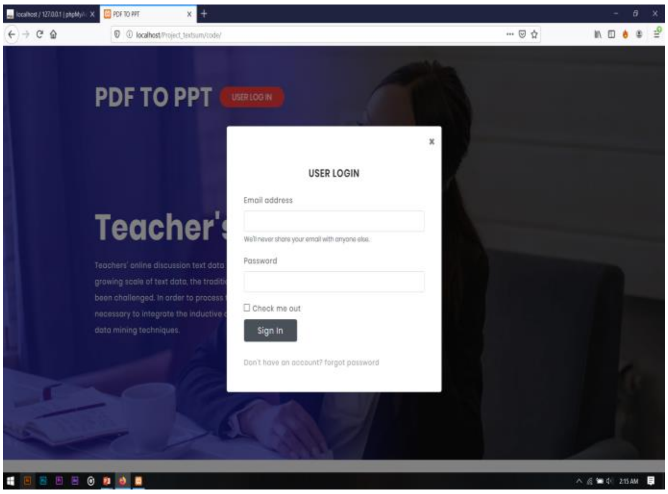

localhost / 127.0.0.1 | phpMy
PDF TO PPT
localhost/Project_textsum/code/
PDF TO PPT
USER LOG IN
Teacher's
Teachers' online discussion text data
growing scale of text data, the traditi
been challenged. In order to process
necessary to integrate the inductive
data mining techniques.
USER LOGIN
Email address
We'll never share your email with anyone else.
Password
☐ Check me out
Sign In
Don't have an account? forgot password

Fig 4: User Login Page

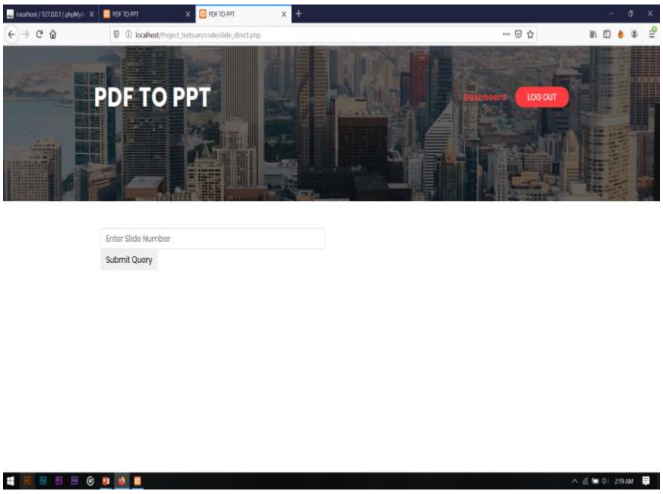
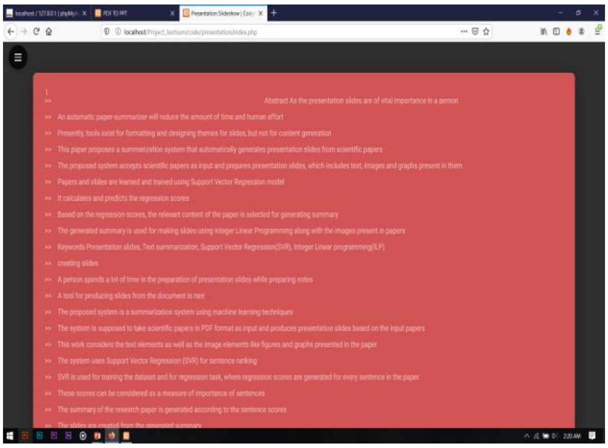

localhost / 127.0.0.1 | phpMy
PDF TO PPT
PDF TO PPT
localhost/Project_textsum/code/slide_direct.php
PDF TO PPT
Dashboard
LOG OUT
Enter Slide Number
Submit Query

Fig 5: Option to define number of slides into which content should be distributed


localhost / 127.0.0.1 | phpMy
PDF TO PPT
Presentation Slideshow | Cod
localhost/Project_textsum/code/presentation/index.php

1

>> Abstract As the presentation slides are of vital importance in a person

>> An automatic paper-summarizer will reduce the amount of time and human effort

>> Presently, tools exist for formatting and designing themes for slides, but not for content generation

>> This paper proposes a summarization system that automatically generates presentation slides from scientific papers

>> The proposed system accepts scientific papers as input and prepares presentation slides, which includes text, images and graphs present in them

>> Papers and slides are learned and trained using Support Vector Regression model

>> It calculates and predicts the regression scores

>> Based on the regression scores, the relevant content of the paper is selected for generating summary

>> The generated summary is used for making slides using Integer Linear Programming along with the images present in papers

>> Keywords Presentation slides, Text summarization, Support Vector Regression(SVR), Integer Linear programming(ILP)

>> creating slides

>> A person spends a lot of time in the preparation of presentation slides while preparing notes

>> A tool for producing slides from the document is rare

>> The proposed system is a summarization system using machine learning techniques

>> The system is supposed to take scientific papers in PDF format as input and produces presentation slides based on the input papers

>> This work considers the text elements as well as the image elements like figures and graphs presented in the paper

>> The system uses Support Vector Regression (SVR) for sentence ranking

>> SVR is used for training the dataset and for regression task, where regression scores are generated for every sentence in the paper

>> These scores can be considered as a measure of importance of sentences

>> The summary of the research paper is generated according to the sentence scores

>> The slides are created from the generated summary

Fig 6: Output Screen

V. CONCLUSION

Instead of spending hours for the preparation of presentation slides, contents will be generated with the click of a button. This project explains a system for the automated generation of slides from research papers. The proposed system generates slides of scientific papers. The system will ensure that only the important sentences would be included in the respective slides. BERT library is used for sentence scoring for importance calculation and NLP method for generation of slides.

A. Advantages

- Can obtain maximum accuracy on the text summarizing.
- Innovative and technologically advanced tools are used.
- Easy to operate.
- Saves a lot of time.

B. Disadvantages

- The animation and presentation perfection cannot be achieved as compared to manually drafter slides.
- The proposed system cannot work efficiently if the hardware and software requirements are not met correctly.
- Proper network should be available for uninterrupted service.

VI. FUTURE SCOPE

The proposed system can be extended in future for slide generation of any kind of documents or pasting a link of a particular website and summarize its content. We could also include a search engine within the application for obtaining the input documents based on keywords or commonly used phrases. We can also modify it in order to include tables, graphs and figures.

REFERENCES

[1] PPSGen: Learning-Based Presentation Slides Generation for Academic Papers, Yue Hu and Xiaojun Wan IEEE Transactions on Knowledge and Data Engneering, Vol. 27, No. 4, April 2015.

[2] Autade Dhanshri P, Prof. Raut S.Y, "SLIDEGen: Approach to automatic Slides Generation in International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 01— Jan-2016.

[3] Abu-Jbara and D. Radev, "Coherent citationbased summarization of scientific papers", in Proc. 49th Annu. Meeting Assoc.Comput. Linguistics: Human Lang. Technol.-Volume 1, 2011, pp. 500509.

[4] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "SlidesGen: Automatic generation of presentation slides for a technical paper using summarization", in Proc. 22nd Int. Flairs Conf,2009, pp. 284289.

[5] M.Y. Kan, "SlideSeer: A digital library of aligned document and presentation pairs, in Proc.7th ACM/IEEE-cs Joint Conf. Digit. Libraries", Jun. 2006, pp. 8190.

[6] https://www.searchenginejournal.com/bert-explained-what-you-need-to-know-about-googles-new-algorithm/337247/#close

[7] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[8] S. Syamili, Anish Abraham, "Presentation slides generation from scientific papers using support vector regression" in International conference on Inventive Communication and Computational Technologies (ICICCT), July 2017