ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Technique for Generating Automatic Slides on the basis of Paper Structure Analysis

Ektaa Meshram [1], D. A. Phalke [2]

P.G. Student, Department of Computer Engineering, D.Y Patil College of Engineering, Akurdi, Pune, India[1]

Associate Professor, Department of Computer Engineering, D.Y Patil College of Engineering, Akurdi, Pune, India[2]

ABSTRACT: Slide presentations play an important role in most fields for sharing information in an easy-to-present and visually-appealing format. The slides for such presentations are traditionally prepared using various tools, including Microsoft PowerPoint, Open Office and Apple Pages. However, this traditional way of preparing slides is labor- intensive in nature and leaves scope for human-errors. Also, in case of research articles and lengthy discussion papers, there is a significant chance of some vital information being missed out. These drawbacks of the traditional way of manually preparing slides lead to a need for an intelligent tool, which is capable of automatically generating slides with minimum human interference. The existing automatic tools fail to fetch the graphical element from a given input paper and are capable of generating only textual drafts of the presentation. In this paper, we are suggesting an automatic slide generation tool, which fetches the graphical elements as well as text from a paper. The proposed system finds relevant images from each page and stores them in the map data. After that it checks the label of each image and adds images into the presentation according to the label of image. An evaluation result shows that the system is more reliable than the existing system.

KEYWORDS: *Abstracting methods, NLP, Text Mining, ILP, Slide Generation.*

I. INTRODUCTION

Presentation slides are an effective medium to convey and share information and deliver key-messages across the audience at professional as well as educational meetings. The research-presenter makes use of slides to share information in an orderly and lucid format [7][10]. The research-presenter has numerous programming tools to assist him in setting up the slides, including Microsoft Power-Point, Open Office and Apple Pages. Such tools help researchers in setting up the theme and outline of the presentation; however they do not help researchers in selecting the content for the slides. The traditional tools thus require a lot of investment, both in terms of time and efforts, from the researchers. In this work, a strategy is proposed for making presentation slides along with graphical elements for academic papers. The primary aim is to generate draft slides for the research-presenters so as to reduce their time and efforts in setting up presentation slides. Most academic and research papers have a relatively consistent structure consisting of a couple of different sections including introduction, system overview, related work, proposed strategy, examinations and conclusions. Different moderators create presentations in different ways; however a moderator generally adjusts slides in the same order as reported in the various areas of the paper. PPSGen maps every area to one or more slides with a typical slide having a title and a few sentences [1]. These sentences may be incorporated in some visual sequences. These methods attempts for producing run-of-the-mill draft slides to help individuals in setting-up their final slides. Programming slides for educational papers proves to be an exceptionally difficult task. Current systems focus on items like sentences from the paper to build the slides. Slides can be isolated into different parts. Every element tackles a precise point and contains topics, which are vital to one another.

In this paper, a system is proposed which automatically generates slides containing graphical elements as well as text data. The proposed system focuses on developing a data-mining technique, which will help in scoring the sentences as well as in generating slides with graphical elements. This system is designed by applying Natural Language Processing

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

(NLP) for scoring the sentences. While obtaining the expected importance score for every sentence from the given input paper, the Integer Linear Programming method (ILP) is utilized to generate well-structured slides. This is done by choosing and bringing into line key phrases and sentences along with diagrams. Different from the existing schemes that produce slides by purely selecting significant sentences and placing sentences on the slides with only text element, this proposed system selects both key phrases and sentences for constructing well structured slides along with graphical elements. This system uses key phrases as bullet points, and sentences applicable to the phrases are positioned below these bullet points. In order to extract key phrases, chunking implemented by the Open NLP library is applied to the sentences and noun phrases are extracted as the candidate key phrases. Images are extracted from the whole document by using the PDF box and PDF documents. Finally the slides are generated and images are added to the slide.

Further part of the paper is structured as follows: Section II depicts information about literature work including different approaches used for slide generation. Section III contains implementation details that include system architecture, algorithm and mathematical model for the system. Section IV describes the dataset used for experimental purpose and result of the project done so far. Section V contains the conclusion of research work done and the future work.

II. RELATED WORK

Slide generation system become increasingly intelligent, since effective presentation slides can be constructed on the basis of SVR, NLP, ILP by means of machine learning and data mining methods. In this section we define the basic theories of above mentioned methods and some new methods in recent years.

Yue Hu and Xiaojun Wan [1] examined an extremely difficult assignment which automatically creates presentation slides for paper. For preparing the slides rapidly the system generated the presentation which can be utilized as a draft. A novel framework termed as PPSGen is designed to tackle this task. Initially it utilizes the regression system to take in the significance score of the sentence in educational paper, and after that endeavors the whole number linear programming system to produce slides.

M. Utiyama and K. Hasida [2] discusses about the distinct approaches for automatically slide generation. This framework inputs a record annotated with the GDA tag-set, a XML tag-set which permits the machine to automatically deduce the semantic structures underlying the rough record. The framework gets imperative points on the premise of semantic dependencies also, gatherings distinguished from the tags. This topic selection additionally relies on association with the audience and further prompts a dynamic adaption of the presentation. Sentences relevant to the selected subject are then extracted and summarized to shape an organized synopsis for the slide.

In [3] presented a sustain method for constructing a presentation slides from a latex report. This framework gives works that assign slides to every section and put objects on a slide. Input to this framework is a specialized paper as a latex record, the number of slides client need to make and keywords of the paper. Initially the framework converts the paper from a text record into a XML report. The XML report can include data of a paper, for example, ID number and term weights. Next the framework evaluate weights of term in the document by the TF*IDF technique. Based on the term weights objects in the report for example, sentences, figures and tables are weighted. Utilizing the weights of the object and slide composition layout, the framework decides how many slides are assigned to each section.

In [4] introduced a methodology for attaining a set of policies for creating presentation sheets by pertaining machine learning methods to various sets of specialized papers also, these presentation sheets gathered from internet. As an initial phase in this paper, a strategy is proposed for aligning specialized papers and presentation sheets and the technique isbased on Jing's system which utilizes a Hidden Markov Model.

T. Shibata and S. Kurohashi [5], proposed a methodology in which slides are produced automatically from raw text, before generation of slides the text is recovered which is most similar to the user query. The method along with

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

generating slides also creates a full automatic presentation. That is to say, input texts are spoken via text-to-speech engine while presenting automatically generated summary slides. Hence, in this system printed texts are mechanically transformed into verbal texts based on paraphrasing strategy and after that are imputed into speech synthesis, but the disadvantage of this slide generation technique is that it contains non-topic parts alongside theme parts.

M.Y. Kan [6] displayed Digital Library framework which consists of just distributed records by the specialists. The research works of the researchers are transmitted into composed document and, slide presentation. As these research reports are extremely unique and contain exceptionally valuable data, so instead of referring these two reports independently, it regards adjust furthermore to present such presentation record matches together. So such alignment among document and, related slides are done in digital library. The three noteworthy framework components of the Slide-Seer digital library: the resource detection, the fine grained alignment and the user interface.

D. Galanis, G. Lampouras, and I. Androutsopoulos [8] have introduced reference based summarization; content composed by a few specialists is utilized to distinguish the critical parts of an objective paper. Previously Extraction was the fundamental aspect to take a work at this issue. Meanwhile, the familiarity of the created summaries has not been given that much significance. For instance, the outline which incorporates differing qualities, readability, cohesion, and requesting of the sentences has not been altogether considered. This prompted to noisy and confusing outlines. In this work, they exhibit an approach for creating intelligible and cohesive reference based summaries.

In this paper M. Sravanthi et al. [9] [10] presented the system called QueSTS, which does the above task by filtering and aggregating important query relevant sentences distributed across a set of documents. Our approach captures the contextual relationships among sentences of all input documents and represents them as an “integrated graph”. These relationships are exploited and several sub-graphs of integrated graph which consist of sentences that are highly relevant to the query and that are highly related to each other are constructed. These sub-graphs are ranked by our scoring model. The highest ranked sub-graph which is rich in query relevant information and also has sentences that are highly coherent is returned as a query specific summary.

III. IMPLEMENTATION DETAILS

The main purpose of the proposed system is to improve the efficiency and reduce the time required for automatic slide generation while maintaining the graphical elements with the text elements.

System Overview

The following figure 1 shows the architectural view of the proposed system. The description of the system is as follows:

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

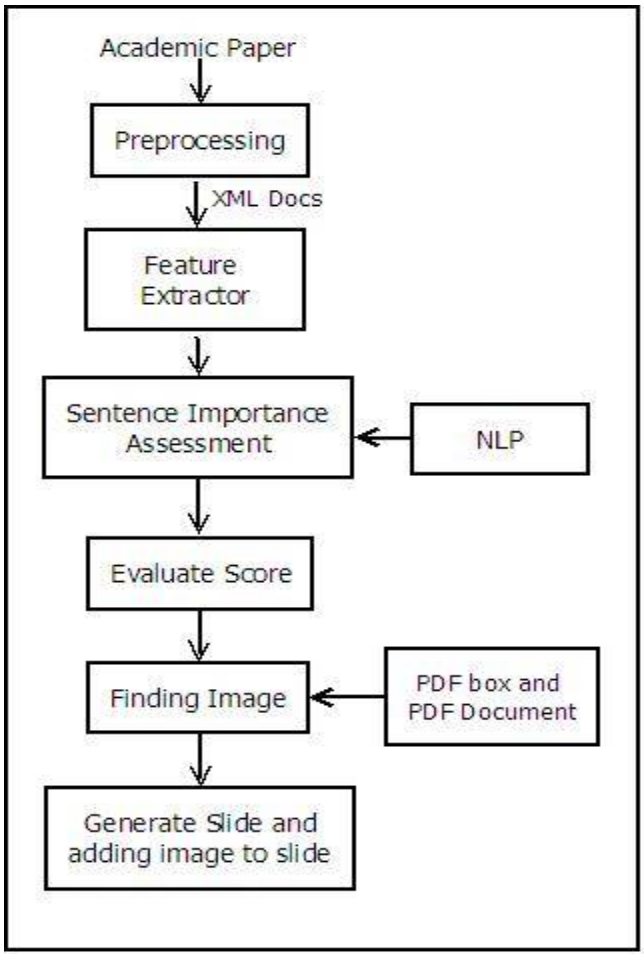

Fig. 3 System Architecture

Proposed system primarily uploads the input file and preprocess the input file in which the process of tokenization is done wherein sentences are splits up into tokens. Stemming is done in which the sentence is converted into their first normal form. Stop words such as like but, also, etc are removed from the data. Postagging of procedure is applied and score is evaluated by using the NLP procedure. The most important objective of NLP is to intend and form such a computer scheme which will observe, recognize and construct NLP. Image is extracted by using PDFbox and PDF document from the input file. Finally, the slides with graphical images are generated.

IV. ALGORITHM

The proposed system implements following algorithm for automatic slide generation along with graphical elements.

Algorithm 1: Proposed System Algorithm

Input: PDF file

Output: Slide Generated with graphical element and text.

Process:

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

1: Read PDF File.
2. Apply Parscit to detect their physical structures of paragraphs, sections and sections
3: Token generation and apply stemmer.
4: Remove Stop words and apply post tagging.
5: Calculate sentence scoring using NLP.
6: Add Image into Slide (See detail in Algorithm 2)
7. Add scoring sentences and image using ILP method
8: Generate the slide.

Algorithm 2: Algorithm for including graphical data into slide

Input: PDF file

Output: Adds Graphical element in slide

Process:
1: Read PDF File.
2: Load PDF file into PDF document.
3: List of all pages in the document.
4: Find image of each single page and store into map data.
5: Repeat Step 4 for all pages.
6: Check label of image and add image according to label.

V. MATHEMATICAL MODEL

Mathematical model of proposed system is stated below. The system S is represented as:

S = {U, P,T, R, F, PP, SP}
Input
Browse Dataset
U = {u1, u2, u3,…., un}
where
U is a set of number of related papers.
u1; u2; u3,……, un are the number of papers.

Process
Parscit Process
P = {p1, p2, p3,….,pn}
Where,
P is represented as a set of Parscit Process
To detect their physical structures of paragraphs, sections and sections by using Parscit.
p1, p2, p3,…..,pn are the number of parscit process.

Preprocessing Method
T={t1,t2,t3,……tn}
Apply token generation, stemming, removing stop words and post tagging.
T1,t2,t3…tn are the number of preprocessing process.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

NLP Model
R = {r1, r2, r3,…., rn}
Where,
R is represented as a set of NLP model
In NLP method, calculate score of the sentence.
r1, r2, r3,…..,rn are the number of NLP model process.

Finding Image
F = {f1, f2, f3,…..,fn}
where,
F is represent as a set of Finding Image
Finding image using PDF document and add all image into data Map.
f1, f2, f3,….., fn are number of Finding Image process.

ILP Method
PP = {pp1, pp2, pp3,…..,ppn}
where,
PP is represent as a set of ILP processing
In ILP method, important sentences are added to the slide along with the images.
pp1, pp2, pp3,…..,ppn are number of ILP processing process.

Output
Slide Transitions
SP = {sp1, sp2, sp3,…..,spn}
where,
SP is represent as a set of slide generation
sp1, sp2, sp3,…..,spn are number of final output.

VI. RESULT S

DataSet

This system uses the academic paper for generating the slides. The input papers must be in PDF format (Latex) which is used as an input and then remove their texts by using PDFlib3 and notice their objective structures of paragraphs, sections and sections by using ParsCit.
Results
This system automatically generates the slides from the research paper. This section shows the result obtained by the proposed system.

The table 1 shows the Average Pyramid Scores of TF-IDF, for the generated presentation slides. Experiment on the paper indicates our method can generate slides with better quality than the baseline methods. Using the ROUGE toolkit and the Pyramid evaluation, the slides generated by our method gets better ROUGE scores and Pyramid scores. Therefore, our slides are considered a better basis for preparing the final
slides.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Table 1: Average Pyramid Scores Comparison

Methods	Average Pyramid Scores
TF-IDF	0.819
SVR-ILP	0.792
NLP-ILP	0.849

The graph shows that the slides generated by our proposed system shows higher pyramid score than that of SVR-ILP and TF-IDP because the proposed system not only contains the text elements but also graphical element like figures from the given input paper.

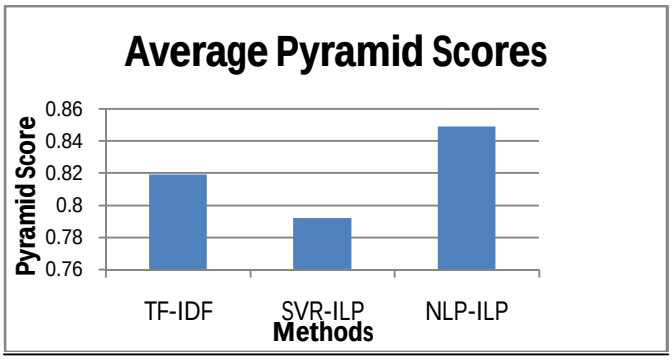

Fig. 2 Average Pyramid Scores Graph

The comparison results over ROUGE metrics are presented in Tables 2, 3. Table 2 shows that our proposed method gets better ROUGE scores, i.e., better content quality when evaluating the whole slides. It means that our generated slides are richer in content and much more similar to the human written slides than those of the baselines as a whole.

Table 2: ROUGE F-Measure Scores when Evaluating the Whole Slides

Method	Rouge-1	Rouge-2
TF-IDF	0.38859	0.11624
Random Walk	0.39421	0.11555
Mead	0.38778	0.11825
C-Lexrank	0.38712	0.11223
Our Method	0.42185	0.13582

Tables 3 show that our method gets higher ROUGE scores on average when evaluating different parts of the slides. It means that the content texts of the slides generated by our method are more appropriately distributed over different sections. It demonstrates that the slides generated by our method have better structure than that generated by baseline methods.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Table 3: ROUGE-1 F-Measure Scores when Evaluating the Segmented Slides

Method	First 30%	Mid 40%	Last 30%	Avg
TF-IDF	0.28220	0.30732	0.28411	0.29156
Random Walk	0.29241	0.30665	0.28443	0.29456
Mead	0.31132	0.28156	0.25462	0.28225
C-Lexrank	0.28835	0.29845	0.27564	0.28761
Our Method	0.30845	0.35894	0.29782	0.32173

VII. CONCLUSION AND FUTURE SCOPE

The paper proposed the system for automatically generating the slides from the academic paper. This system generates the slides which includes text and graphical element. The graphical elements along with the text data make the generated slides look more comprehensible and vivid. System initially finds the graphical elements of each page from the paper and after that system stores the image to map data. System performs the same operation for each page and finally checks the label of each image and adds that image to the slide according to the label of each image. NLP method is used for sentence scoring and the ILP method is used for slide generation which contain key phrases and the relevant sentences. Presently, the system generate slides on the basis of only one given paper, but in future extra information like additional relevant papers and information of citation can be utilized for improving the slides generation.

REFERENCES

[1] Yue Hu and Xiaojun Wan, ``PPSGen: Learning-Based Presentation Slides Generation for Academic Papers", in IEEE transactions on knowledge and data engineering, vol. 27, no. 4, april 2015.
[2] U. Masao and H. K^oiti, "Automatic slide presentation from semantically annotated documents," in Proceedings of the Workshop on Coreference and its Applications, pp. 25–30, Association for Computational Linguistics, 1999.
[3] Y. Yasumura, M. Takeichi, and K. Nitta, "A support system for making presentation slides," Transactions of the Japanese Society for Artificial Intelligence, vol. 18, pp. 212–220, 2003.
[4] T. Hayama, H. Nanba, and S. Kunifuji, "Alignment between a technical paper and presentation sheets using a hidden markov model," in Active Media Technology, 2005.(AMT 2005). Proceedings of the 2005 International Conference on, pp. 102–106, IEEE, 2005.
[5] T. Shibata and S. Kurohashi, "Automatic slide generation based on discourse structure analysis," in Natural Language Processing–IJCNLP 2005, pp. 754–766, Springer, 2005.
[6] M.-Y. Kan, "Slideseer: A digital library of aligned document and presentation pairs," in Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 81 90, ACM, 2007.
[7] S. M. A. Masum and M. Ishizuka, "Making topic specific report and multimodal presentation automatically by mining the web resources," in Proc. IEEE/WIC/ACM Int. Conf. Web Intell., 2006, pp. 240–246.
[8] D. Galanis, G. Lampouras, and I. Androutsopoulos, "Extractivemulti-document summarization with integer linear programmingand support vector regression," in Proc. COLING, 2012,pp. 911–926.
[9] M. Sravanthi, C. R. Chowdary, and P. S. Kumar, "QueSTS: Aquery specific text summarization approach, in Proc. 21st Int.FLAIRS Conf., 2008, pp. 219–224.
[10] A. Abu-Jbara and D. Radev, "Coherent citation-based summarization of scientific papers," in Proc. 49th Annu.Meeting Assoc. Comput. Linguistics: Human Lang. Technol.-Volume 1, 2011,pp. 500–509.