ISSN:0975-9646

Survey on Presentation Slides Generation for Academic Papers

Ektaa G Meshram, Mrs. D. A. Phalke
Department of Computer Engg. Savitribai Phule Pune University
DYPCOE, Akurdi, India

Abstract: This paper discusses a method for automatically generating summary slides from a text, studying the automatic generation of presentation slides from a technical paper and also examines the challenging task of continuously creating presentation slides from academic papers. The created slides can be used as a draft to help moderators setup their systematic slides in a quick manner. This paper introduces a novel system called PPSGen to help moderators create such slides. The system uses the backslide procedure to determine the Noteworthiness Score of the sentences in an educational paper and then uses the whole number Integer Linear Programming (ILP) system to create well-organized slides by selecting and adjusting key expressions and sentences. Evaluation results, based on a test game plan of 200 arrangements of papers and slides assembled on the web, display that our proposed PPSGen structure can create slides with better quality. A customer study also exhibits that PPSGen has obvious advantage over baseline methods.

Keywords: Abstracting methods, text mining, Support Vector Regression (SVR), ILP, Classification.

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. The research-presenter has numerous programming tools to assist him in setting up the slides, including Microsoft Power-Point and Open Office. 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, in terms of time and efforts, from the researchers. In this work, a strategy is proposed for making presentation slides for academic papers. Aim is to generate draft slides for the research-presenters so as to reduce their time and efforts in setting up presentation slides. Intuitive papers have a relatively consistent structure and contain a couple of different sections like 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. These sentences may be incorporated in some visual sequences. Our strategy attempts to produce run-of-the-mill draft slides to help individuals in setting-up their final slides. Programming slides for academic 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. Each part addresses a specific point and contains topics, which are vital to one another.

In this study, we propose the PPSGen system to make pre-composed presentation slides for academic papers. In our structure, the importance of each sentence in a paper is figured out by using the Support Vector Regression (SVR) model with different accommodating components. The presentation slides for the paper are then created by using the Integer Linear Programming (ILP) model with complicatedly arranged target limits and objectives to pick and alter key expressions and sentences. It has been examined on a test course of 200 paper-slides sets, which demonstrate that our methodology can create slides with better quality over the standard frameworks. Using the ROUGE tool and the Pyramid Appraisal, the slides made by our method can enhance ROUGE scores and Pyramid scores. Additionally, in light of customer focus, our slides can get higher rating scores by human judges in both substance and structure points.

II. GENERIC FRAME WORK

Unlike reviews, we give a more general overview on the overall process of automatic slide generation, which is outlined in fig1. In this paper, it review recent developments and analyze future open directions in automatic slide generator. The key contribution of this survey is as follows 1) Sentence scoring and slide generation is discussed in a clearly organized, hierarchical manner and the interlink between these components is shown 2) To examine the state of the art, each task involved in slide generation is divided into sub-process and various categories of approaches to the sub-process are discussed. The merits and limitation of the different approaches are summarized.

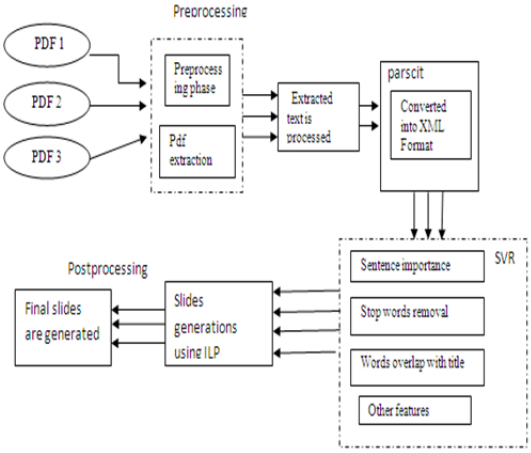

Figure 1: Generic framework for automatic slides generator for academic papers

A. *Sentence selection (scoring)*

Presentation slides are made up of bullet points and corresponding sentences. The selection of these sentences is done using some specific method, so that they can be displayed on the presentation slides. Different methods work differently for selecting the sentences from academic papers. This results in selection of most appropriate and relevant sentences. One more concept i.e. Summarization is used to select specific summary and generate slides using this summary, but with the help of this concept slides are prepared which contains only sentences but not key phrases aligned to the sentences.

Y. Yasumura et al [2] introduced a support for making slides from technical papers. The inputs of the system are academic papers in Latex format. The system calculates the weight of the terms in the paper using TF*IDF scores. Using the term weights, objects in the paper like sentences, tables etc. are also weighted. Based on the weights of the objects, sentences can be extracted for each section in the paper and then generate the slides using a slide composition template, which can be edited by the users.

Luhn et al [10] has outlined a technique for automatic creation of abstract using modern electronic data processing devices i.e. Summarization. Experts of technical papers and magazines that serve the purpose of conventional abstracts have been created entirely by automatic means. In the exploratory research described, the complete text of an article in machine readable form is scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the “auto-abstract”.

In [11], another method is proposed for creation of an automatic index from technical documents. The author says that if the sentence falls with relatively high occurrence at some fixed position within the paragraph, it is an easy matter to have the machine select the sentence and record it for compiling an abstract or for extracting the vocabulary to form an index. Examination of 200 paragraphs corroborated the fact that in 85% of the paragraphs the topic sentence was the first one and in 7% of the time it was the last sentence. A simple way to select the topic sentence would be a program the machine to select the first and the last sentence of the paragraph.

B. Slide Generation

After the task of scoring the sentences, the most important sentences are selected using some specific method. Integer Linear Programming method is used to generate well-structured slides by selecting and aligning key phrases and sentences. Key phrases are used as the bullet points and sentences relevant to the phrases are placed below the bullet point. In order to extract the key phrases, chunking implemented by the Open NLP library is applied to the sentences and noun phrases are extracted as the candidate key phrases. Two kinds of phrases are defined: global phrases and local phrases. Any unique phrase in an article is a global phrase, and a local phrase means a global phrase in a particular section. For example, “APPLE” is a global phrase of the paper, while its appearances in different sections are considered different local phrases. “APPLE {Introduction}” and “APPLE {Our Proposed Method}” denote different local phrases, and they represent the appearances of “APPLE” in Sections 1 and 4 of the paper, respectively. So a global phrase that appears in different sections can correspond to a few local phrases. Since an important phrase is always used in many different sections, a global phrase that corresponds to more local phrases should be regarded to be more important and more likely to be selected. Thus, the local phrases are used to generate the bullet points directly for different sections and use the global phrases to address the importance differences between different unique phrases. All the phrases are stemmed and stop words are removed. Moreover, the noun phrases which appear only once in the paper are discarded.

GDA tag set is used in [1] for the automatic slide presentation from semantically annotated documents. Attempted to automatically generate slides from input documents annotated with the GDA tagset.1 GDA tagging can be used to encode semantic structure. The semantic relations include grammatical relations such as subject, thematic relations such as agent, patient, and rhetorical relations such as cause and elaboration. They first detect topics in the input documents and then extract important sentences relevant to the topics to generate slides.

A TF*IDF method is presented in [2] for making presentation slides. A support system is used for making presentation slides from a technical paper. This system provides functions that assign slides to each section and puts objects on a slide. Input to this system is a technical paper as a latex document; the number of slides user wants to make and keywords of the paper. First the system converts a paper from a text document into an XML document. The XML document can include information of a paper such as ID number and term weights. Next, the system calculates weights of term in the document by the TF*IDF method. Based on the term weights, objects in the document such as sentences, figures and tables are weighted. Using the weight of the objects and slide composition template, the system decides how many slides are assigned to each section.

Shibata and Kurohashi [3] exploit a method to automatically generate slides from raw texts. Clauses and sentences are considered as discourse units and coherence relations between the units such as list, contrast, topic chaining and cause are identified. Some of clauses are detected as topic parts and others are regarded as non-topic parts. These different parts are used to generate the final slides based on the detected discourse structure and some heuristic rules.

III. LITERATURE SURVEY

In paper [1] the authors discussed the various methodologies to automatically generate slides. The reported presentation system inputs a document annotated with the GDA tag-set, an XML tag-set which allows the machine to automatically infer the semantic structures underlying the raw document. The system picks up important topics on the basis of semantic dependencies and conferences identified from the tags. This topic selection also depends on interaction with the audience and further leads to a dynamic adaption of the presentation. Sentences relevant to the selected topic are then extracted and paraphrased to form an itemised summary for the slide. Same heuristics are applied for paraphrasing and layout. Since the GDA tag-set is independent of the domain and style of document and applicable for a variety of natural languages, the reported system is also domain independent and easy to adapt to different languages.

In paper [2] Author introduces a support system for making presentation slides from a latex document. This system provides functions that assign slides to each section and put objects on a slide. Input to this system is a technical paper as a latex document, the number of slides user want to make and keywords of the paper. First the system converts the paper from a tex document into a XML document. The XML document can include information of a paper such as ID number and term weights. Next the system calculate weights of term in the document by the TF*IDF method. Based on the term weights objects in the document such as sentences, figures and tables are weighted. Using the weights of the objects and slide composition template, the system decides how many slides are assigned to each section.

In paper [4] author the approach of obtaining a set of rules for generating presentation sheets by applying machine learning techniques to many pairs of technical papers and their presentation sheets collected from world wide web. As a first step in this paper, a method is proposed for aligning technical papers and presentation sheets and the method is based on Jing’s method which uses a Hidden Markov Model.

In paper [3] the slides are generated automatically, but before generation of slides the text are retrieved which is most similar to the users query. Then the system converts the written text into spoken languages and feed them to a speech synthesis engine as written text are not appropriate, because unnatural speech might be produced due to difficult words or long compound nouns, which are unsuitable for speech synthesis. Therefore, written texts are automatically converted into spoken texts based on paraphrasing technique and then are imputed into speech synthesis, but the drawback of this slide generation is that it contains non-topic parts along with topic parts.

In paper [5] A Digital Library consists of only published documents by the researchers. The research works of the researchers are transmitted into written document and slide presentation. As these research documents are very unique and contain very useful information, so instead of referring these two documents separately, it is good to align and present such presentation document pairs together. So such alignment between document and related slides are done in digital library. The three major system components of the Slide-Seer DL: 1) the resource discovery, 2) the fine-grained alignment and 3) the user interface.

In paper [8] Due to availability and accessibility of large Internet-based resources and robust nature of Web pages, the task of information retrieval is becoming more challenging and complex. Agent based autonomous system, automatic report to presentation (ARP), with the notion of autonomous information service emerging as the result of integration among natural language processing, Web intelligence, and character-based agent interaction are the key areas focussed in this paper. The system, ARP, fetches a set of Web-pages; and then parses, summarizes affect-senses and correlates information extracted from those and finally automatically builds a report on a topic and search phrase given by a user. A concise presentation is automatically created by the system and, a group of character based software-agents autonomously present the topic in a story-telling manner. It also employs text- to-speech engine with accompanied content-rich slides, different gestures and effects.

In paper [9] citation-based summarization, text written by several researchers is made use of to identify the important aspects of a target paper. Previously, Extraction (i.e. selecting a representative set of citation sentences that highlight the contribution of the target paper) was the main aspect to work on this problem. Meanwhile, the fluency of the produced summaries has not been given that much importance. For example, the summary which includes diversity, readability, cohesion, and ordering of the sentences have not been thoroughly considered. This led to noisy and confusing summaries. In this work, they present an approach for producing readable and cohesive citation-based summaries. The experiments show that the proposed approach outperforms several baselines in terms of both extraction quality and fluency.

IV. Comparison Table

TABLE I SURVEY TABLE

Paper Title/ Year	Approaches Used	Advantages	Future Scope
Automatic slide presentation from semantically annotated documents[1]	Discusses automatic generation of presentation slides from semantically annotated documents.	The reported system is also domain/ style independent and easy to adapt to different languages.	Required more synthetic evaluation.
Alignment between a Technical Paper and Presentation Sheets Using a Hidden Marko Model[4]	Approach consists of obtaining a set of rules for generating presentation sheets by applying machine learning technique to many pairs of technical papers and presentation sheets collected from the WWW.	Use of Alignment method to combine the uses of different features. .	Applying HMM to numerous pairs of papers and presentation sheets, and obtain rules for generating presentation sheets using machine learning techniques.
Auto-presentation: A multi-agent system for building automatic multi-modal presentation of a topic from world wide web information[7]	The system, 'Auto-Presentation', builds a presentation automatically by parsing, summarizing and correlating information collected from the Internet based on knowledge sources after receiving the presentation topic from the user.	Make the presentation more live and quick. Concept building approach around the topic has been implemented by considering presentation template to be filled out by the information miner.	Algorithms are necessary to improve information retrieval.
Identifying Non-explicit Citing Sentences for Citation-based Summarization[12]	Proposed a general framework based on probabilistic inference to extract such context information from scientific papers.	Greater pyramid scores for surveys generated using such context information rather than citrating sentences alone.	Need to combine summarization and bibliometric techniques towards building automatic surveys that employ context information as an important part of the generated surveys.
Automatic slide generation based on discourse Structure Analysis[3]	a method of automatically generating summary slides from a text.	Generated slides are far easier to read.	Trying to reduce non-topic parts in the slides, to obtain greater accuracy

V. Conclusion

This paper proposes a novel system called PPSGen to generate presentation slides from academic papers. Sentence scoring model is trained based on SVR and use the ILP method to align and extract key phrases and sentences for generating the slides. Experimental results show that our strategy can create vastly improved slides than customary routines.

Presently, our system generates slides based on only one given paper, but in future additional information such as other relevant papers and the citation information can be used to improve the generated slides.

Acknowledgment

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

References

[1] M. Utiyama and K. Hasida, "*Automatic slide presentation from semantically annotated documents*," in Proc. ACL Workshop Conf. Its Appl., 1999, pp. 25–30.

[2] Y. Yasumura, M. Takeichi, and K. Nitta, "*A support system for making presentation slides*," Trans. Japanese Soc. Artif. Intell., vol. 18, pp. 212–220, 2003.

[3] T. Shibata and S. Kurohashi, "*Automatic slide generation based on discourse structure analysis*," in Proc. Int. Joint Conf. Natural Lang. Process., 2005, pp. 754–766.

[4] T. Hayama, H. Nanba, and S. Kunifuji, "*Alignment between a technical paper and presentation sheets using hidden Markov model*," in Proc. Int. Conf. Active Media Technol., 2005, pp. 102–106.

[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. 81–90

[6] B. Beamer and R. Girju, "Investigating automatic alignment methods for slide generation from academic papers," in Proc. 13th Conf. Comput. Natural Lang. Learn., Jun. 2009, pp. 111–119.

[7] S. M. A. Masum, M. Ishizuka, and M. T. Islam, "Auto-presentation: A multi-agent system for building automatic multi-modal presentation of a topic from World Wide Web information," in Proc. IEEE/ WIC/ACMInt. Conf. Intell. Agent Technol., 2005, pp. 246–249.

[8] 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.

[9] 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.

[10] H. P. Luhn, "The automatic creation of literature abstracts," IBM J. Res. Develop., vol. 2, pp. 159–165, 1958.

[11] P. B. Baxendale, "Machine-made index for technical literature: an experiment," IBM J. Res. Develop., vol. 2, no. 4, pp. 354–361,1958.

[12] V. Qazvinian and D. R. Radev, "Identifying non-explicit citing sentences for citation-based summarization," in Proc. 48th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2010, pp. 555–564. 1096 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015.